华北电力大学（北京）

硕士学位论文

正弦加噪的语音信号模型的研究

姓名：李文华

申请学位级别：硕士

专业：通信与信息系统

指导教师：许刚

20050601

摘　　要

因为语音信号是一种非平稳信号，所以建立语音信号谱模型的目的是将一个语音信号变换成一种容易应用的形式，从而去掉和语音感知不相关的信息。正弦加噪模型是一种谱模型，将声音周期分量以正弦的时变频率、幅度和相位形式来表示，剩余的非周期分量以已滤波的噪声形式呈现。因为周期分量通常不稳定，在多重的语音信号中，估计正弦模型的参数是一项十分困难的任务，同时也很难达到较高的时间和频率分辨率。本文主要内容是讨论几种经典算法对周期信号的检测和参数估计。在研究已经存在算法的基础上，本文提出一种新的基于紧密间隔的正弦信号融合的迭代算法，同时也讨论了一种新的基于多音高估计的分离方法。

关键词：谱模型，中层表示，正弦模型，声音源分离

ABSTRACT

Because audio signal is a nonstationary signals, the aim of build audio signal spectrum modeling is to transform a signal to a more easily applicable form, removing the information that is irrelevant in signal perception. Sinusoids plus noise model is a spectral model, in which the periodic components of the sound are represented with sinusoids with time-varying frequencies, amplitudes and phases. The remaining non-periodic components are represented with a filtered noise. In the case of polyphonic music signals, the estimation of the parameters of sinusoids is a difficult task, since the periodic components are usually not stable. A sufficient time and frequency resolution is also difficult to achieve at the same time. A big part of this thesis discusses the detection and parameter estimation of periodic components with several algorithms. In addition to already existing algorithms, a new iterative algorithm is presented, which is based on the fusion of closely spaced sinusoids. Also a new separation method based on the multipitch estimation is explained.

Li Wenhua(Communication and Information System)
Directed by prof. Xu Gang

KEY WORDS: Spectrum model, mid-level representation, sinusoidal modeling, sound source separation

摘　　要

因为语音信号是一种非平稳信号，所以建立语音信号谱模型的目的是将一个语音信号变换成一种容易应用的形式，从而去掉和语音感知不相关的信息。正弦加噪模型是一种谱模型，将声音周期分量以正弦的时变频率、幅度和相位形式来表示，剩余的非周期分量以已滤波的噪声形式呈现。因为周期分量通常不稳定，在多重的语音信号中，估计正弦模型的参数是一项十分困难的任务，同时也很难达到较高的时间和频率分辨率。本文主要内容是讨论几种经典算法对周期信号的检测和参数估计。在研究已经存在算法的基础上，本文提出一种新的基于紧密间隔的正弦信号融合的迭代算法，同时也讨论了一种新的基于多音高估计的分离方法。

关键词：谱模型，中层表示，正弦模型，声音源分离

ABSTRACT

Because audio signal is a nonstationary signals, the aim of build audio signal spectrum modeling is to transform a signal to a more easily applicable form, removing the information that is irrelevant in signal perception. Sinusoids plus noise model is a spectral model, in which the periodic components of the sound are represented with sinusoids with time-varying frequencies, amplitudes and phases. The remaining non-periodic components are represented with a filtered noise. In the case of polyphonic music signals, the estimation of the parameters of sinusoids is a difficult task, since the periodic components are usually not stable. A sufficient time and frequency resolution is also difficult to achieve at the same time. A big part of this thesis discusses the detection and parameter estimation of periodic components with several algorithms. In addition to already existing algorithms, a new iterative algorithm is presented, which is based on the fusion of closely spaced sinusoids. Also a new separation method based on the multipitch estimation is explained.

Li Wenhua(Communication and Information System)
Directed by prof. Xu Gang

KEY WORDS: Spectrum model, mid-level representation, sinusoidal modeling, sound source separation

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

特此申明。

签　名：李文华　日　期：2005.6

关于学位论文使用授权的说明

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

(涉密的学位论文在解密后遵守此规定)

作者签名：李文华　导师签名：许明

日　期：2005.6　日　期：2005.6

第一章 引言

本文主要讨论机器听觉和音乐信号分解的正弦加噪信号模型。在20世纪早期，机器听觉方面的大部分研究工作主要是在语音识别领域，其后学者们研究的兴趣逐步转到一般的计算机声音场景分解。最近在这方面的研究表明，只有在严格限制的条件下的机器听觉才能达到和人耳听觉相近的结果。一般来说，人类的听觉系统比计算机听觉要高级的多，因此，我们自然想要去设计一个新的系统能够接近人类听觉系统所能达到的效果。

一般情况下，标准的脉冲编码调制信号（PCM）主要用来描述人耳的声压水平，但并不是对声音分解最好的表示。通常使用的方法是谱模型，或者是用合适的中层表示去变换语音信号，使其成为一种能够很容易从PCM信号中生成的形式，同时高层的信息也能够很容易获得。正弦加噪模型就是这种谱模型的一种：正弦部分利用一般共振系统的物理属性表示共振成份；噪声模型利用人类对随机信号的精确谱形状或相位无法精确感知的特性表示正弦曲线的残余成分。

在机器听觉方面自动翻译是一个具有很大发展前途的应用领域，其中处理语音的各种各样的应用工具、语音的音高范围的测试、语音变换的频谱和其它特性等问题的研究具有很大的挑战性。本文的正弦加噪模型主要针对的是音乐信号，该模型是建立在几个其它的正弦加噪模型和一些本文提到的原始算法之上，设计的目的是为了能够在处理的每一个阶段测试不同的算法，算法在 Windows 环境下用 Matlab 实现。

正弦加噪有能力去除与信号不相关数据并用较低的比特率对信号进行编码，所以能够用在数字语音和语言编码上。评价一个系统的标准是对合成声音感知质量的评价，但本文主要关心的重点不是系统质量的好坏而是能够有实际应用，如用在语音分离和信号分解等领域。

1．1 正弦加噪模型

由乐器和其它物理系统产生的声音分为确定的部分和随机的部分，或称为正弦部分和噪声残余部分[1]。正弦部分由振动系统产生，通常是谐音；噪声残余信号包含由激励器械和其它不是周期振动产生的能量。

在标准正弦模型中，信号$x(t)$的确定部分用时间函数表示一组正弦轨迹（正弦轨迹的解释见表1－1）：

$$x(t)=\sum_{i=1}^{N}a_i(t)\cos\left(\theta_i(t)\right)+r(t), \qquad (1-1)$$

第一章 引言

本文主要讨论机器听觉和音乐信号分解的正弦加噪信号模型。在20世纪早期，机器听觉方面的大部分研究工作主要是在语音识别领域，其后学者们研究的兴趣逐步转到一般的计算机声音场景分解。最近在这方面的研究表明，只有在严格限制的条件下的机器听觉才能达到和人耳听觉相近的结果。一般来说，人类的听觉系统比计算机听觉要高级的多，因此，我们自然想要去设计一个新的系统能够接近人类听觉系统所能达到的效果。

一般情况下，标准的脉冲编码调制信号（PCM）主要用来描述人耳的声压水平，但并不是对声音分解最好的表示。通常使用的方法是谱模型，或者是用合适的中层表示去变换语音信号，使其成为一种能够很容易从PCM信号中生成的形式，同时高层的信息也能够很容易获得。正弦加噪模型就是这种谱模型的一种：正弦部分利用一般共振系统的物理属性表示共振成份；噪声模型利用人类对随机信号的精确谱形状或相位无法精确感知的特性表示正弦曲线的残余成分。

在机器听觉方面自动翻译是一个具有很大发展前途的应用领域，其中处理语音的各种各样的应用工具、语音的音高范围的测试、语音变换的频谱和其它特性等问题的研究具有很大的挑战性。本文的正弦加噪模型主要针对的是音乐信号，该模型是建立在几个其它的正弦加噪模型和一些本文提到的原始算法之上，设计的目的是为了能够在处理的每一个阶段测试不同的算法，算法在 Windows 环境下用 Matlab 实现。

正弦加噪有能力去除与信号不相关数据并用较低的比特率对信号进行编码，所以能够用在数字语音和语言编码上。评价一个系统的标准是对合成声音感知质量的评价，但本文主要关心的重点不是系统质量的好坏而是能够有实际应用，如用在语音分离和信号分解等领域。

1．1 正弦加噪模型

由乐器和其它物理系统产生的声音分为确定的部分和随机的部分，或称为正弦部分和噪声残余部分[1]。正弦部分由振动系统产生，通常是谐音；噪声残余信号包含由激励器械和其它不是周期振动产生的能量。

在标准正弦模型中，信号$x(t)$的确定部分用时间函数表示一组正弦轨迹（正弦轨迹的解释见表1－1）：

$$x(t)=\sum_{i=1}^{N}a_i(t)\cos\left(\theta_i(t)\right)+r(t), \qquad (1-1)$$

其中$a_i(t)$和$\theta_i(t)$是正弦信号i在t时刻的幅度和相位，$r(t)$是噪声残余信号由随机模型表示。我们假定正弦是局部稳定的，就是说幅度不会任意改变，相位局部线性，整个信号是用正弦信号和随机信号模型建模，因此，残余信号$r(t)$包含所有信号$x(t)$中不在正弦中包含的部分，还包括没有检测出来的正弦部分。

人类的感知对声音非周期信号谱的形状和相位不敏感，我们假定残余信号只有随机分量，因此就能用已滤波的白噪声来表示。残余信号既没有瞬时幅度也没有残余的相位，而是用时变频率整形滤波器或用短时能量在特定频率段如在 Bark 段上建模。考虑到这些因素，正弦加噪模型可以考虑成为一个物理和生理学属性上的模型。

表 1—1:术语定义

术语	定义
轨道、轨迹	包含时变频率、幅度和相位的正弦分量，在时频谱图中以轨迹表示
谐音	振动系统模式，频率是基频的整数倍
残余	整个信号中去掉确定的信号部分后剩下的信号
声音分离	混合在一个信号里的两个以上声音信号从信号中分离的过程

1．2 正弦加噪分解合成系统的的一般结构

目前有很多标准的正弦加噪模型，并有各自的发展。在这里我们提出标准模型的实现，改进部分在第三、四章提出。正弦加噪系统的框图如图 1—1 所示。首先，

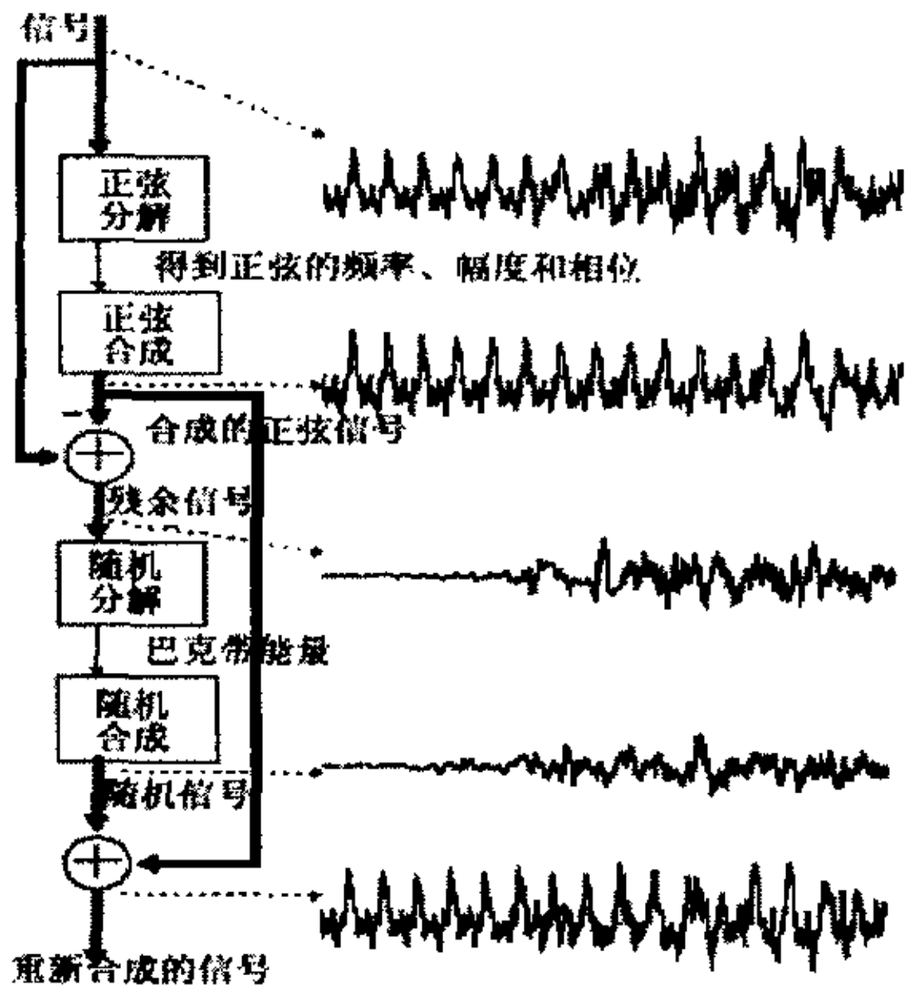

图 1—1 标准正弦加噪模型

其中$a_i(t)$和$\theta_i(t)$是正弦信号i在t时刻的幅度和相位，$r(t)$是噪声残余信号由随机模型表示。我们假定正弦是局部稳定的，就是说幅度不会任意改变，相位局部线性，整个信号是用正弦信号和随机信号模型建模，因此，残余信号$r(t)$包含所有信号$x(t)$中不在正弦中包含的部分，还包括没有检测出来的正弦部分。

人类的感知对声音非周期信号谱的形状和相位不敏感，我们假定残余信号只有随机分量，因此就能用已滤波的白噪声来表示。残余信号既没有瞬时幅度也没有残余的相位，而是用时变频率整形滤波器或用短时能量在特定频率段如在 Bark 段上建模。考虑到这些因素，正弦加噪模型可以考虑成为一个物理和生理学属性上的模型。

表 1—1:术语定义

术语	定义
轨道、轨迹	包含时变频率、幅度和相位的正弦分量，在时频谱图中以轨迹表示
谐音	振动系统模式，频率是基频的整数倍
残余	整个信号中去掉确定的信号部分后剩下的信号
声音分离	混合在一个信号里的两个以上声音信号从信号中分离的过程

1．2 正弦加噪分解合成系统的的一般结构

目前有很多标准的正弦加噪模型，并有各自的发展。在这里我们提出标准模型的实现，改进部分在第三、四章提出。正弦加噪系统的框图如图 1—1 所示。首先，

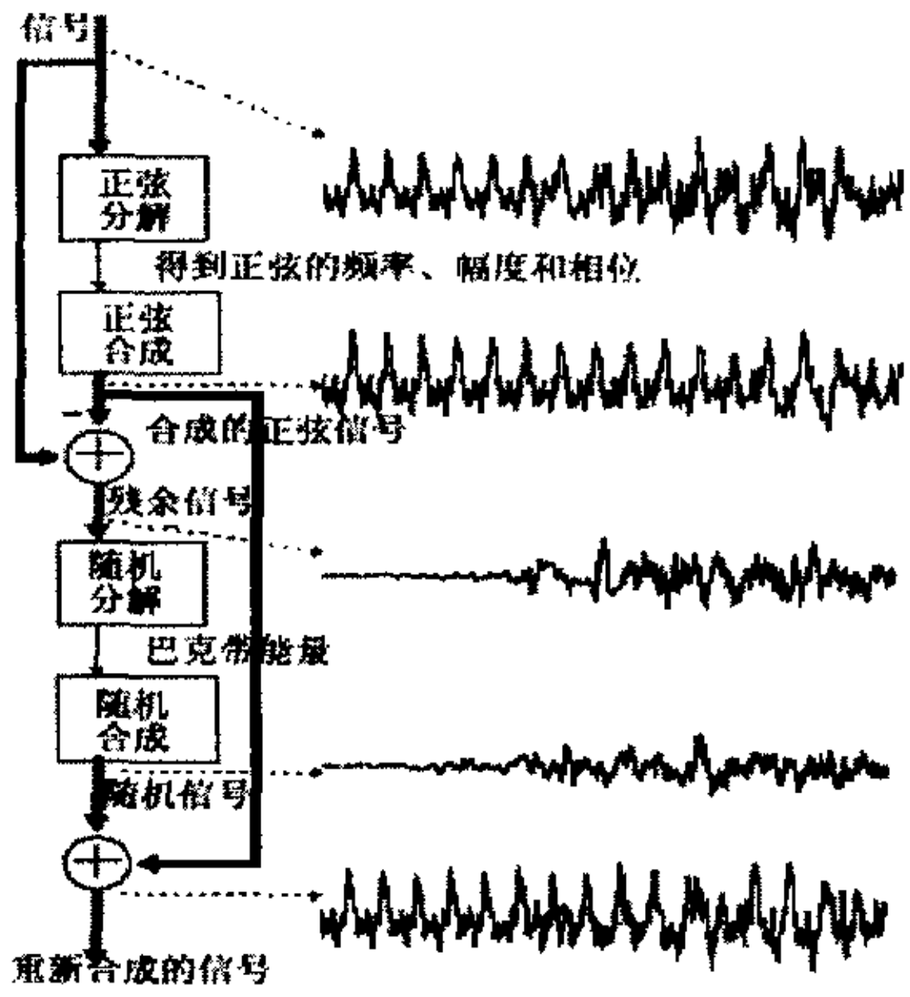

图 1—1 标准正弦加噪模型

对输入的信号进行分解获得正弦信号的时变幅度、频率和相位，然后从原始信号中去掉合成的正弦信号获得噪声残余信号。应用随机分解获得短时 BARK 频带的能量，将随机信号重新合成到合成信号中获得整个重新合成的新信号。

在参数域，我们可以通过修改参数来产生如音高漂移或时间延伸这样的效果，同时进一步分解合成信号或残余信号，或在参数数据中直接完成分解。例如，用短时 BARK 带能量，我们就可以辨别类似鼓声的声学噪音[2]。

正弦信号分解是系统中最复杂的部分，首先，将输入信号分成部分重叠并加窗的帧；第二，通过 DFT 获得每帧的短时谱；第三，分解谱、检测主谱峰并且估计它们的参数，包括幅度、频率和相位。峰值检测法和参数估计在第三章详细讨论。

当被检测的正弦峰的幅度、频率和相位估计出来以后，它们连接起来形成帧间轨迹。峰值延续算法可以从下一帧峰中发现现有轨迹合适的延续，要重新合成正弦必需要获得正弦轨迹所包含的所有信息。在时域里，正弦的合成可以用插值的轨迹参数和结果波形的总和来完成。峰值延续算法和正弦合成在第四章讨论。

在时域里信号的残余部分是原始信号减去合成的正弦部分得到的，这个残余信号是以经过滤波的噪音形式出现的。因为人类听觉感知不能辨别在特定频率带内的能量变化，如 BARK 带内类似噪声的稳定信号，所以不要求精确的谱形状。对于随机信号处理过程，相位也因为在感知上的无关紧要而忽略不计。因此，在每个 BARK 带内对于类似噪音信号唯一需要的信息就是短时能量。在随机分解中，只要计算出复杂的残余信号谱，每个 BARK 带内的短时能量就能估计出来。我们合成复杂谱需要产生一个幅度的随机相位，这个幅度是从 BARK 带中获得，用相邻帧用叠加合成获得。随机模型在第 5 章讨论。

第二章 概论

2.1 信号的分层表示

人类对声音信号的感知在 D.Ellis 和 D.Rosenthal 1995 年的文章中从低到高有一系列的记载[3]：低层表示相当于进入内耳之前的信号；高层表示是我们已有的认知方法，就像电话铃声响起的时候有人在玩低音吉他的效果；在这两层之间我们表示的叫中层表示，正弦加噪模型对实现中层表示功能十分有效。谱模型的思想是丢弃人类听觉感知中所有无用的信息，如非立体声的声音感知中不需要的相位等。

目前人类听觉系统的中层表示方面的知识在一定程度上还十分有限，我们试着在计算听觉场景分解中建立一个和人类听觉系统相似的模型。D.Ellis 和 D.Rosenthal 列出了下列描述听觉中层表示的属性[3]：

(1) 声源分离性。自然声音彼此叠加，人类的听觉有能力从复杂的环境中把声音从它们的声源中分离出来。

(2) 可逆性。从参数的表示中，我们能重新产生原始信号，但是只是在感知上重构而不是完全的重构。

(3) 分量缩减性。原始信号到达中耳可以看做为气压等级的排列，我们重构它的时候目标总数减少的同时有意义的目标将会增加。

(4) 理论本质的优越性。用相应的物理特征而不是具体的算法重构信号的特性。

(5) 生理学上的真实性。从人类听觉生理学的观点看具有生理学上的真实性。

因为我们要从获得的参数中分解和合成信号，并且参数的数目十分有限，所以正弦加噪模型有不同的标准。正弦模型可以分离声源，具体方法在第 7 章中给出，一般情况下，噪声模型则不能分离声源。正弦模型也有其它标准，正弦的开端和声音的开端相关，正弦的频率和声源的共鸣频率相关。

整个正弦加噪模型的生理学上的研究很少，而正弦模型更多的是偏向物理学上的研究，但这也可以认为是该模型引起众多学者感兴趣的地方。由于正弦模型产生出过于简单化的数据，同时只做了少量的推演，因此如果要得到信号高水平的数据信息，信号需要用较高级的分解。

2.2 正弦模型相关的谱模型

加法合成是一个非常接近正弦模型的传统语音合成方法，它用于电子音乐已经有几十年的历史了[4]。类似正弦模型一样，它用一个正弦曲线的时变的幅度，频率和相位表示原始信号[5]。然而，它没有谐音分量和非谐音分量的差别，为了重构非谐

第二章 概论

2.1 信号的分层表示

人类对声音信号的感知在 D.Ellis 和 D.Rosenthal 1995 年的文章中从低到高有一系列的记载[3]：低层表示相当于进入内耳之前的信号；高层表示是我们已有的认知方法，就像电话铃声响起的时候有人在玩低音吉他的效果；在这两层之间我们表示的叫中层表示，正弦加噪模型对实现中层表示功能十分有效。谱模型的思想是丢弃人类听觉感知中所有无用的信息，如非立体声的声音感知中不需要的相位等。

目前人类听觉系统的中层表示方面的知识在一定程度上还十分有限，我们试着在计算听觉场景分解中建立一个和人类听觉系统相似的模型。D.Ellis 和 D.Rosenthal 列出了下列描述听觉中层表示的属性[3]：

(1) 声源分离性。自然声音彼此叠加，人类的听觉有能力从复杂的环境中把声音从它们的声源中分离出来。

(2) 可逆性。从参数的表示中，我们能重新产生原始信号，但是只是在感知上重构而不是完全的重构。

(3) 分量缩减性。原始信号到达中耳可以看做为气压等级的排列，我们重构它的时候目标总数减少的同时有意义的目标将会增加。

(4) 理论本质的优越性。用相应的物理特征而不是具体的算法重构信号的特性。

(5) 生理学上的真实性。从人类听觉生理学的观点看具有生理学上的真实性。

因为我们要从获得的参数中分解和合成信号，并且参数的数目十分有限，所以正弦加噪模型有不同的标准。正弦模型可以分离声源，具体方法在第 7 章中给出，一般情况下，噪声模型则不能分离声源。正弦模型也有其它标准，正弦的开端和声音的开端相关，正弦的频率和声源的共鸣频率相关。

整个正弦加噪模型的生理学上的研究很少，而正弦模型更多的是偏向物理学上的研究，但这也可以认为是该模型引起众多学者感兴趣的地方。由于正弦模型产生出过于简单化的数据，同时只做了少量的推演，因此如果要得到信号高水平的数据信息，信号需要用较高级的分解。

2.2 正弦模型相关的谱模型

加法合成是一个非常接近正弦模型的传统语音合成方法，它用于电子音乐已经有几十年的历史了[4]。类似正弦模型一样，它用一个正弦曲线的时变的幅度，频率和相位表示原始信号[5]。然而，它没有谐音分量和非谐音分量的差别，为了重构非谐

音分量需要非常大量的正弦曲线，因此它只对谐音输入信号的处理可以得到很好的结果。

声码器是谱模型的另一个分支，它在多重并行的信道上重构输入信号，每个信道代表一个特定频带的信号[6]，声码器简化谱信息并且减少了数据量。相位声码器是一个特别类型的声码器，它使用一个复杂的短时谱来保存信号的相位信息。相位声码器与一组带通滤波器或一个 STFT 一起实现，它同正弦模型一样允许时间和音高等级的修改[7]。

源滤波器是联合使用一个时变的滤波器和一个激励信号，激励信号是一列脉冲或白噪声。需要的信号由滤波器宽频激励获得，此方法也被称为减法合成。这个方法接近人类的语音生成系统，而且它常用于语音编码[5]。滤波器系数由线性预测分解获得，对于有声的语音，一个周期脉冲序列当作一个激励使用，而白噪声当作无声的语音。虽然源滤波器中有声的激励才能作为单音信号，但是用已过滤噪声的方法处理非谐音信号和本文的随机合成方法相当接近。我们系统使用 BARK 带的能量来代替时变滤波器，在一般情况下，这种方法在心理声学上的表现是很好的。

2．3 正弦加噪模型系统

正弦模型是由 R.J.McAulay 和 J.O.Smith 等人提出的[8][9]，R.J.McAulay 和 Quatieri 在 1986 年利用正弦模型的目的是用来进行语音编码，J.O.Smith 和 X.Serra 在 1987 年利用正弦模型的目的是要对音乐信号进行计算机合成。虽然当时这两个系统是独立研究的，但其中算法思想有十分相似之处。两个系统只有一小部分稍有不同，如峰值检测。两个系统有相同的思想，都需要正弦分解和合成。其具体流程是：首先给原始信号加窗，并且通过检测短时谱来获得信号显著的频谱峰值；接下来估计出峰值的频率、振幅和相位，而且峰值连接形成正弦轨迹并追踪；最后的轨迹合成中对幅度是使用线性插值方法，而使用三次多项插值方法处理频率和相位。X.Serra 是第一个将信号分解为确定的和随机的两个部分，并和正弦模型一起使用一个随机模型[10]。由此这种信号分解机理已经在很多领域应用，目前大多数的噪声模型系统使用两种方法：一种是用时变滤波器来描绘谱；另一种是用特定的频带内的短时能量来描绘谱。

2．4 暂态模型

当用正弦模型和噪声模型处理一个变化非常复杂的声音信号时，它们只能非常迅速的改变信号分量，这个时刻就叫做暂态。正弦模型可以处理暂态信号，但是因为暂态信号通常有一个巨大的带宽，需要的正弦曲线的数量非常大。同时因为信号

音分量需要非常大量的正弦曲线，因此它只对谐音输入信号的处理可以得到很好的结果。

声码器是谱模型的另一个分支，它在多重并行的信道上重构输入信号，每个信道代表一个特定频带的信号[6]，声码器简化谱信息并且减少了数据量。相位声码器是一个特别类型的声码器，它使用一个复杂的短时谱来保存信号的相位信息。相位声码器与一组带通滤波器或一个 STFT 一起实现，它同正弦模型一样允许时间和音高等级的修改[7]。

源滤波器是联合使用一个时变的滤波器和一个激励信号，激励信号是一列脉冲或白噪声。需要的信号由滤波器宽频激励获得，此方法也被称为减法合成。这个方法接近人类的语音生成系统，而且它常用于语音编码[5]。滤波器系数由线性预测分解获得，对于有声的语音，一个周期脉冲序列当作一个激励使用，而白噪声当作无声的语音。虽然源滤波器中有声的激励才能作为单音信号，但是用已过滤噪声的方法处理非谐音信号和本文的随机合成方法相当接近。我们系统使用 BARK 带的能量来代替时变滤波器，在一般情况下，这种方法在心理声学上的表现是很好的。

2．3 正弦加噪模型系统

正弦模型是由 R.J.McAulay 和 J.O.Smith 等人提出的[8][9]，R.J.McAulay 和 Quatieri 在 1986 年利用正弦模型的目的是用来进行语音编码，J.O.Smith 和 X.Serra 在 1987 年利用正弦模型的目的是要对音乐信号进行计算机合成。虽然当时这两个系统是独立研究的，但其中算法思想有十分相似之处。两个系统只有一小部分稍有不同，如峰值检测。两个系统有相同的思想，都需要正弦分解和合成。其具体流程是：首先给原始信号加窗，并且通过检测短时谱来获得信号显著的频谱峰值；接下来估计出峰值的频率、振幅和相位，而且峰值连接形成正弦轨迹并追踪；最后的轨迹合成中对幅度是使用线性插值方法，而使用三次多项插值方法处理频率和相位。X.Serra 是第一个将信号分解为确定的和随机的两个部分，并和正弦模型一起使用一个随机模型[10]。由此这种信号分解机理已经在很多领域应用，目前大多数的噪声模型系统使用两种方法：一种是用时变滤波器来描绘谱；另一种是用特定的频带内的短时能量来描绘谱。

2．4 暂态模型

当用正弦模型和噪声模型处理一个变化非常复杂的声音信号时，它们只能非常迅速的改变信号分量，这个时刻就叫做暂态。正弦模型可以处理暂态信号，但是因为暂态信号通常有一个巨大的带宽，需要的正弦曲线的数量非常大。同时因为信号

音分量需要非常大量的正弦曲线，因此它只对谐音输入信号的处理可以得到很好的结果。

声码器是谱模型的另一个分支，它在多重并行的信道上重构输入信号，每个信道代表一个特定频带的信号[6]，声码器简化谱信息并且减少了数据量。相位声码器是一个特别类型的声码器，它使用一个复杂的短时谱来保存信号的相位信息。相位声码器与一组带通滤波器或一个 STFT 一起实现，它同正弦模型一样允许时间和音高等级的修改[7]。

源滤波器是联合使用一个时变的滤波器和一个激励信号，激励信号是一列脉冲或白噪声。需要的信号由滤波器宽频激励获得，此方法也被称为减法合成。这个方法接近人类的语音生成系统，而且它常用于语音编码[5]。滤波器系数由线性预测分解获得，对于有声的语音，一个周期脉冲序列当作一个激励使用，而白噪声当作无声的语音。虽然源滤波器中有声的激励才能作为单音信号，但是用已过滤噪声的方法处理非谐音信号和本文的随机合成方法相当接近。我们系统使用 BARK 带的能量来代替时变滤波器，在一般情况下，这种方法在心理声学上的表现是很好的。

2．3 正弦加噪模型系统

正弦模型是由 R.J.McAulay 和 J.O.Smith 等人提出的[8][9]，R.J.McAulay 和 Quatieri 在 1986 年利用正弦模型的目的是用来进行语音编码，J.O.Smith 和 X.Serra 在 1987 年利用正弦模型的目的是要对音乐信号进行计算机合成。虽然当时这两个系统是独立研究的，但其中算法思想有十分相似之处。两个系统只有一小部分稍有不同，如峰值检测。两个系统有相同的思想，都需要正弦分解和合成。其具体流程是：首先给原始信号加窗，并且通过检测短时谱来获得信号显著的频谱峰值；接下来估计出峰值的频率、振幅和相位，而且峰值连接形成正弦轨迹并追踪；最后的轨迹合成中对幅度是使用线性插值方法，而使用三次多项插值方法处理频率和相位。X.Serra 是第一个将信号分解为确定的和随机的两个部分，并和正弦模型一起使用一个随机模型[10]。由此这种信号分解机理已经在很多领域应用，目前大多数的噪声模型系统使用两种方法：一种是用时变滤波器来描绘谱；另一种是用特定的频带内的短时能量来描绘谱。

2．4 暂态模型

当用正弦模型和噪声模型处理一个变化非常复杂的声音信号时，它们只能非常迅速的改变信号分量，这个时刻就叫做暂态。正弦模型可以处理暂态信号，但是因为暂态信号通常有一个巨大的带宽，需要的正弦曲线的数量非常大。同时因为信号

加窗的长度可能比暂态信号的长度大很多，所以使用正常的时间分辨率来进行正弦合成暂态信号就不是理想的办法。对暂态信号使用长窗在声音的编码中还会造成一个问题：回声效应。要避免这个问题可以使用一个分开的模型，即先用短暂的检波器确定暂态位于的地方，其他的信号部分用加参数的正弦加噪模型表示，发现暂态的地方用没参数的变换码表示[11]。在一个系统中同时使用暂态模型和正弦加噪模型已经在很多文献中提到[11][12][13]，其中也有很多好的应用结果。

在听觉场景信号分解的应用中其暂态模型的引入并不能有重大意义的改善，所以暂态模型不包含在本文的系统之中。如果我们要考虑合成的语音质量，增加暂态模型显然会提高合成的质量，然而我们的主要目的是为了构建一个中等水平的声音内容的分解，而不是一个声音编码器，所以并不考虑信号的暂态模型。

2.5 音高同步分解

一般来说正弦模型的参数估计是一项十分困难的工作，其中大部分的问题都涉及到分解窗的长度。如果输入信号是单音，或有不实时重叠的谐音浊音，分解窗长度和语音的基本频率同步。通常浊音谐音分量的频率是基本频率的整数倍，音高同步分解的优势在频域中表现十分明显：谐音分量的频率完全符合 DFT 系数的频率。因为不需要插值，而且能从复杂的谱中直接获得幅度和相位，所以参数的估计非常容易。同时，音高同步分解可以使用和语音的一个周期那么小的窗，而根据估计方法的不同非同步窗必须是周期的 2-4 倍，这意味着用音高同步的分解方法能得到一个更好的时间分辨率。但是音高同步分解法不能用在不同基本频率的一些语音同时发生的情况下。一般情况下，单音记录只是极少的用在音乐信号上，因此音高同步分解方法是不能通用的。为了简化系统的复杂性，在我们的系统之中不包含音高同步分解方法。目前自适应的窗长度已经成功的用于现代声音编码系统，但是方式不同：在信号的稳定部分使用一个长窗，而且当信号迅速的变化发生的时候，窗长变得较短，这样就能对信号的稳定部分有一个较好的频率分辨率，对信号迅速变化的时候有一个较好的时间分辨率，例如 MP3 就是一个很好的例子。

加窗的长度可能比暂态信号的长度大很多，所以使用正常的时间分辨率来进行正弦合成暂态信号就不是理想的办法。对暂态信号使用长窗在声音的编码中还会造成一个问题：回声效应。要避免这个问题可以使用一个分开的模型，即先用短暂的检波器确定暂态位于的地方，其他的信号部分用加参数的正弦加噪模型表示，发现暂态的地方用没参数的变换码表示[11]。在一个系统中同时使用暂态模型和正弦加噪模型已经在很多文献中提到[11][12][13]，其中也有很多好的应用结果。

在听觉场景信号分解的应用中其暂态模型的引入并不能有重大意义的改善，所以暂态模型不包含在本文的系统之中。如果我们要考虑合成的语音质量，增加暂态模型显然会提高合成的质量，然而我们的主要目的是为了构建一个中等水平的声音内容的分解，而不是一个声音编码器，所以并不考虑信号的暂态模型。

2.5 音高同步分解

一般来说正弦模型的参数估计是一项十分困难的工作，其中大部分的问题都涉及到分解窗的长度。如果输入信号是单音，或有不实时重叠的谐音浊音，分解窗长度和语音的基本频率同步。通常浊音谐音分量的频率是基本频率的整数倍，音高同步分解的优势在频域中表现十分明显：谐音分量的频率完全符合 DFT 系数的频率。因为不需要插值，而且能从复杂的谱中直接获得幅度和相位，所以参数的估计非常容易。同时，音高同步分解可以使用和语音的一个周期那么小的窗，而根据估计方法的不同非同步窗必须是周期的 2-4 倍，这意味着用音高同步的分解方法能得到一个更好的时间分辨率。但是音高同步分解法不能用在不同基本频率的一些语音同时发生的情况下。一般情况下，单音记录只是极少的用在音乐信号上，因此音高同步分解方法是不能通用的。为了简化系统的复杂性，在我们的系统之中不包含音高同步分解方法。目前自适应的窗长度已经成功的用于现代声音编码系统，但是方式不同：在信号的稳定部分使用一个长窗，而且当信号迅速的变化发生的时候，窗长变得较短，这样就能对信号的稳定部分有一个较好的频率分辨率，对信号迅速变化的时候有一个较好的时间分辨率，例如 MP3 就是一个很好的例子。

第三章 峰值检测和参数估计

本文中，正弦分解算法的基本的原理和理论在本章和第 4 章中说明，算法的实际应用的范例在第 6 章中讨论。实验模拟和在算法的应用中所用到一些基础知识，如要在实际的应用中如何选择二个不同的算法等，这些都在第 6.5 章中阐述。如图 1—1 所描述的那样，正弦分解是整个正弦加噪系统的一个重要组成部分。正弦分解模块能将信号更进一步的分解成四个部分，如图 3—1：首先检测输入信号的有意义的峰值；其次，峰值被插值替换以获得较好的频率分辨率；第三，将检测到峰值的幅度和相位估计出来，最后连接峰值形成轨迹。这四个步骤还可以有一些其它的方法进行代替。在本章中，我们提出在我们的正弦分解系统中前三个阶段测试的算法，有两个峰值检测算法，两个峰值插值算法和两个迭代参数估计方法，其延续部分在第 4 章中讨论。峰值检测是正弦模型系统中的一个至关重要的部分，因为正弦合成必需用检测到的峰值来完成。估计有意义的峰值和获得它们的参数存在许多难点，这些难点其大部分涉及到分解窗的长度：短窗能随输入信号迅速改变，但是，如果估计正弦曲线的精确频率或区分频谱上接近正弦曲线时就需要一个长窗。

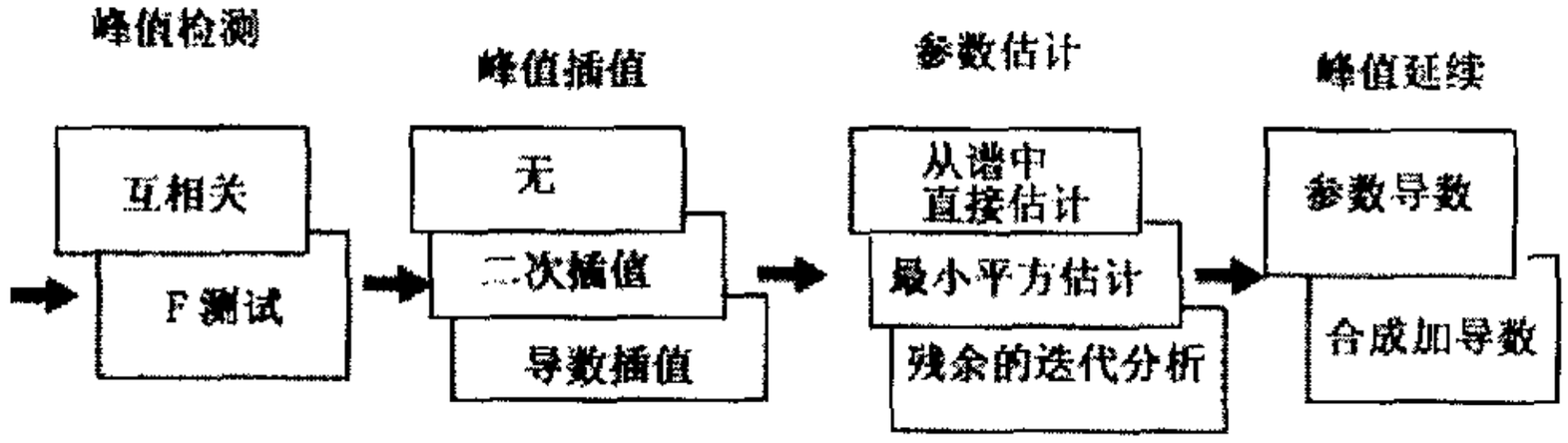

图 3—1 正弦分解过程各阶段和算法

什么是“有意义的峰”是一个比较基本问题，如果允许幅度和频率快速改变，那么，甚至信号的随机部分都能使用大量的正弦曲线来建模，一般情况下这并不是我们在正弦模型中想要的，相反我们只想使用正弦模型来表示周期语音的谐音部分。在几乎所有正弦分解系统中峰值检测和参数估计都是使用 DFT 在频域中完成的，这是因为每个稳定的正弦曲线对应频域中的一个脉冲。因为自然的语音从来都不会是一个长时间稳定的正弦曲线，我们必须使用一个滑动窗和一个短时的 Fourier 变换（STFT）来分解时域信号，通常可以用增加零点可以提高短时谱的频率分辨率[14]。如果 N 是 2 的幂，我们就能使用快速 Fourier 变换（FFT）算法，这样可以提高算法的计算效率，其处理的每个阶段如图 3—2。

一个峰值或者在 STFT 绝对值的局部最大值表明在该频率附近存在正弦曲线，检测信号的正弦曲线的最简单的方法是选择 STFT 绝对值的一个局部最大值。这一个方法的特点是效率高并且快速生产一个固定的位元传输率，因此它经常应用于语

音编码。对于信号分解,使用固定的正弦曲线数目是不实际的：对于非谐音的语音，检测由噪声所引起的峰值，固定正弦曲线的数目会引起接下来的分解问题；对于多音的信号，谐音部分的数量是很大的，而且固定数目的正弦曲线不可能完全满足这些谐音部分。

有一个简单的改进方法就是为峰值检测设一个域值：在域值上的所有 STFT 绝对值的局部极大值都被认为是正弦曲线的峰值。这种方法产生一个变化的峰值，但是

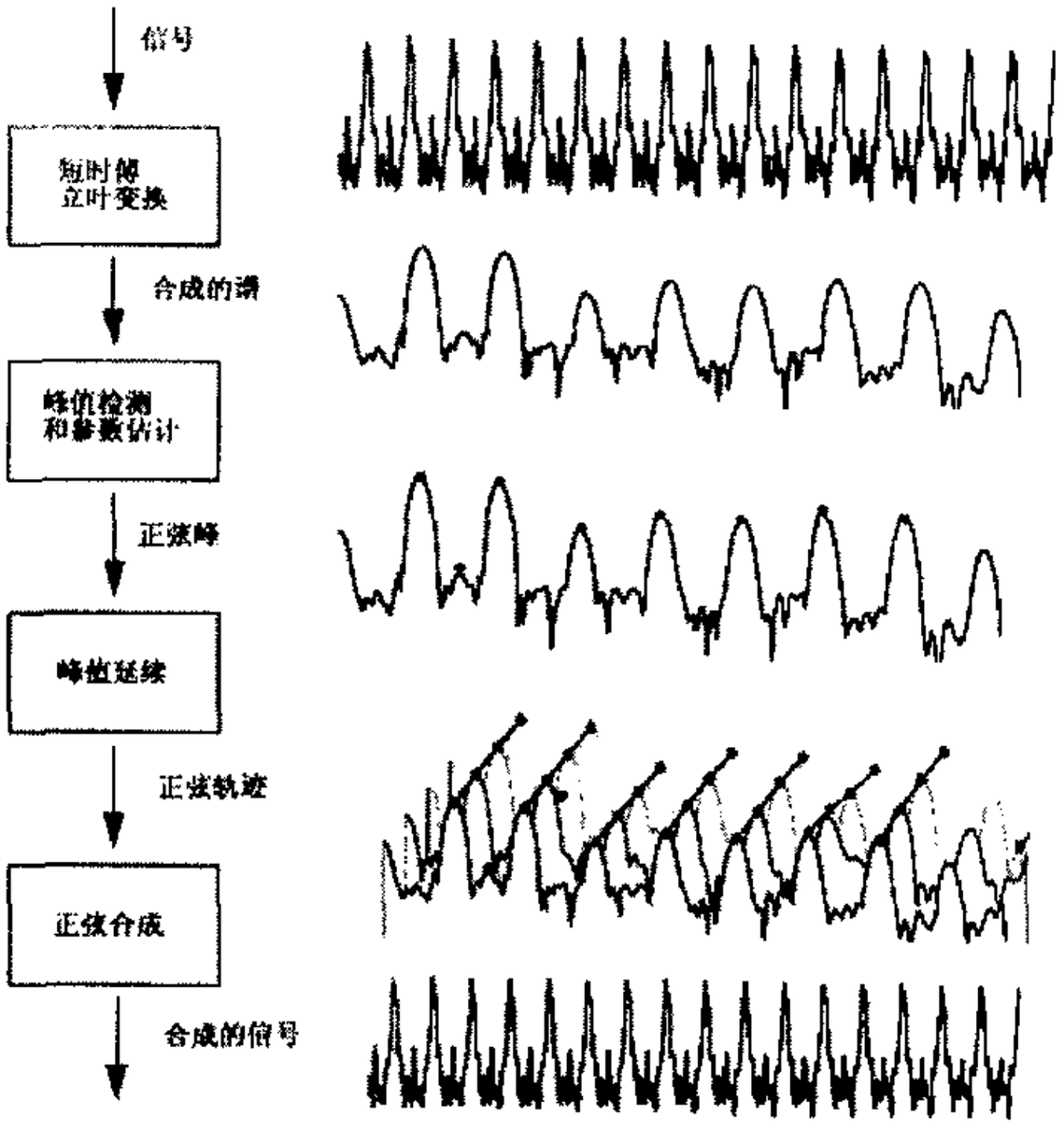

图 3—2 正弦分解系统的实现框图

它不能解决由噪声所引起的峰值,或其它非谐音的声音引起的峰值。当频率函数或自然声音的能量集中在低频时，这种方法还不能估计全部的频谱形状和谐音的幅度。有时较高的谐音部分也会在固定的门限之下,而不能被检测到。出于存在这些问题,我们可以考虑另外两个复杂的峰值检测算法,即互相关法和 F-测试法。

3．1 互相关法

正弦定义为：比相邻的频率包含更多有意义的能量的频率分量。互相关法就是利用的这个思想，首先计算信号的短时谱和从一个理想正弦得到的谱之间的互相关，并且根据整个谱形状依比例进行缩放，这种结果叫做正弦相似测量。

互相关方法成功的应用是在语音编码上[15]，而在音乐信号方面，由于几个正弦分量在频率上彼此接近，所以这种方法很难测量有声分量和无声分量，但是这种方法在很多不同的条件下仍然能够检测出大量的正弦。

音编码。对于信号分解,使用固定的正弦曲线数目是不实际的：对于非谐音的语音，检测由噪声所引起的峰值，固定正弦曲线的数目会引起接下来的分解问题；对于多音的信号，谐音部分的数量是很大的，而且固定数目的正弦曲线不可能完全满足这些谐音部分。

有一个简单的改进方法就是为峰值检测设一个域值：在域值上的所有 STFT 绝对值的局部极大值都被认为是正弦曲线的峰值。这种方法产生一个变化的峰值，但是

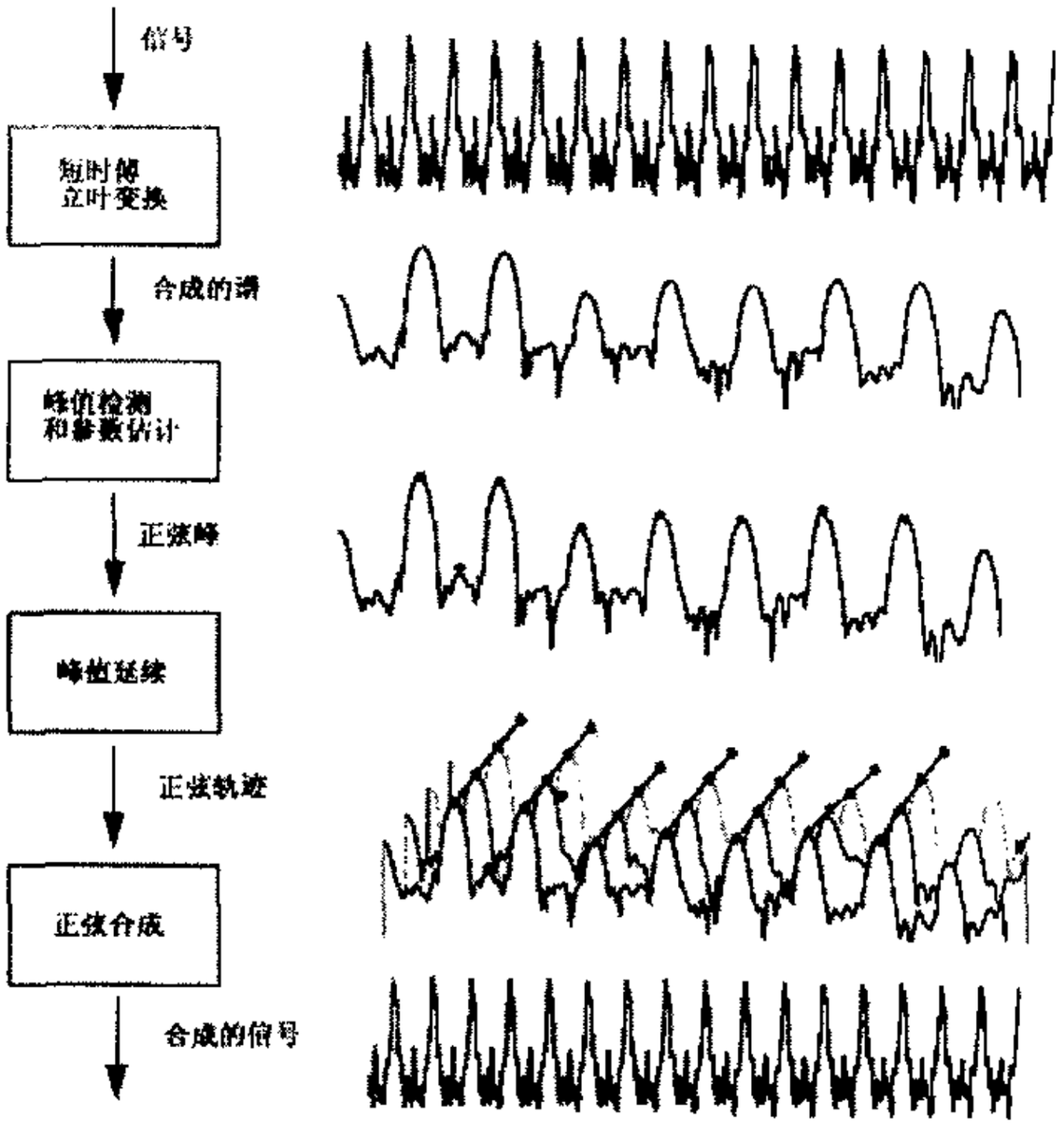

图 3—2 正弦分解系统的实现框图

它不能解决由噪声所引起的峰值,或其它非谐音的声音引起的峰值。当频率函数或自然声音的能量集中在低频时，这种方法还不能估计全部的频谱形状和谐音的幅度。有时较高的谐音部分也会在固定的门限之下,而不能被检测到。出于存在这些问题,我们可以考虑另外两个复杂的峰值检测算法,即互相关法和 F-测试法。

3．1 互相关法

正弦定义为：比相邻的频率包含更多有意义的能量的频率分量。互相关法就是利用的这个思想，首先计算信号的短时谱和从一个理想正弦得到的谱之间的互相关，并且根据整个谱形状依比例进行缩放，这种结果叫做正弦相似测量。

互相关方法成功的应用是在语音编码上[15]，而在音乐信号方面，由于几个正弦分量在频率上彼此接近，所以这种方法很难测量有声分量和无声分量，但是这种方法在很多不同的条件下仍然能够检测出大量的正弦。

单正弦谱$S(\omega_k)$是$H(\omega_k-\Omega)$的缩放和相位偏移，这是在频率Ω上分解窗变换的谱[16]。在有谐音的声音信号中，我们把几个$H(\omega_k)$的和进行变换，缩放和移动相位来得到不同的频率，幅度和相位。自然的就要研究$H(\omega_k)$和$X(\omega_k)$之间的互相关函数，和加窗信号的 STFT。通常在高频时$H(\omega_k)$值很小，所以我们能够用$H(\omega_k)$的一个窄带宽$[-W,W]$来计算互相关：

$$r(\omega)=\sum_{k,|\omega-\omega_k|<W} H(\omega-\omega_k)X(\omega_k) \tag{3-1}$$

如果我们在频率Ω为$H(\omega_k)$和$X(\omega_k)$定义标准：

$$|H|_\Omega^2=\sum_{k,|\omega-\omega_k<W|}|H(\Omega-\omega_k)|^2 \quad 和 \quad |X|_\Omega^2=\sum_{k,|\omega-\omega_k|<W}|X(\Omega-\omega_k)|^2 \tag{3-2}$$

在观察峰值和及从理想正弦中得到的峰值之间得到一个估计值v_Ω

$$v_\Omega=\frac{|r(\Omega)|}{|H|_\Omega|X|_\Omega} \tag{3-3}$$

v_Ω的值在 0 和 1 之间，$v_\Omega=1$是理想的正弦，没有一点噪声。我们也可以得到一个幅度的估计值 A、相位估计值φ和频率估计值Ω：

$$A(\Omega)=\frac{|r(\Omega)|}{|H|_\Omega^2}和 \tag{3-4}$$

$$\varphi(\Omega)=Arg[r(\Omega)] \tag{3-5}$$

我们通过设置固定的限制（限制在 0 和 1 之间）用v_Ω来检测正弦和它的频率，在固定域值之上选择本地最大v_Ω的频率点。

在图 3—3 中给出了加窗的小提琴样本和对同一个样本的正弦相似测量计算的幅度谱。我们从幅度谱中可以看到，整个谱水平在高频部分很低。在正弦相似测量中考虑到：即使高谐音部分的幅度比低谐音部分低 20dB，但在高频部分谐音有很大的正弦相似测量值。正弦相似测量对高谐音部分是低比特率的，有个很好的例子就是小提琴有高频噪音，因此高谐音不是理想的正弦。

互相关是一个卷积，这里其它信号的时间尺度是反相的[17]。频域信号的互相关因此能用时域信号的乘积来实现。对于大的带宽W，能对$x(t)$加长度是分解窗$h(t)$两倍的窗来做离散 Fourier 变换计算$r(\omega)$。因为$|x|_\Omega^2$的计算可以看做一个用 FIR 滤波器的滤波操作，所有系数都是 1，FIR 滤波器也可以用 IIR 滤波器替换，其中只有两个系数不等于零：一个延迟用所有输入信号的累加；另一个延迟是减去窗口结尾的值，但是这样的缺点是计算V_Ω非常困难。

正弦相似测量假定在带宽W内只有一个正弦，在大多数情况下，我们必需用小带宽控制谐音分音密集的部分。另一方面，噪音的存在要求域值尽量小才能检测低幅度的正弦。由此得出结论：在没有正弦曲线的情况下，同时带宽和域值较小的频域里，会有由噪声或其它正弦曲线的旁瓣产生的峰值被误认为是正弦曲线。因此在

一个较小帧的里，我们不能在特定的频率上准确的判断是否存在正弦，为了准确的正弦分解还需要从相邻帧里得到信息。

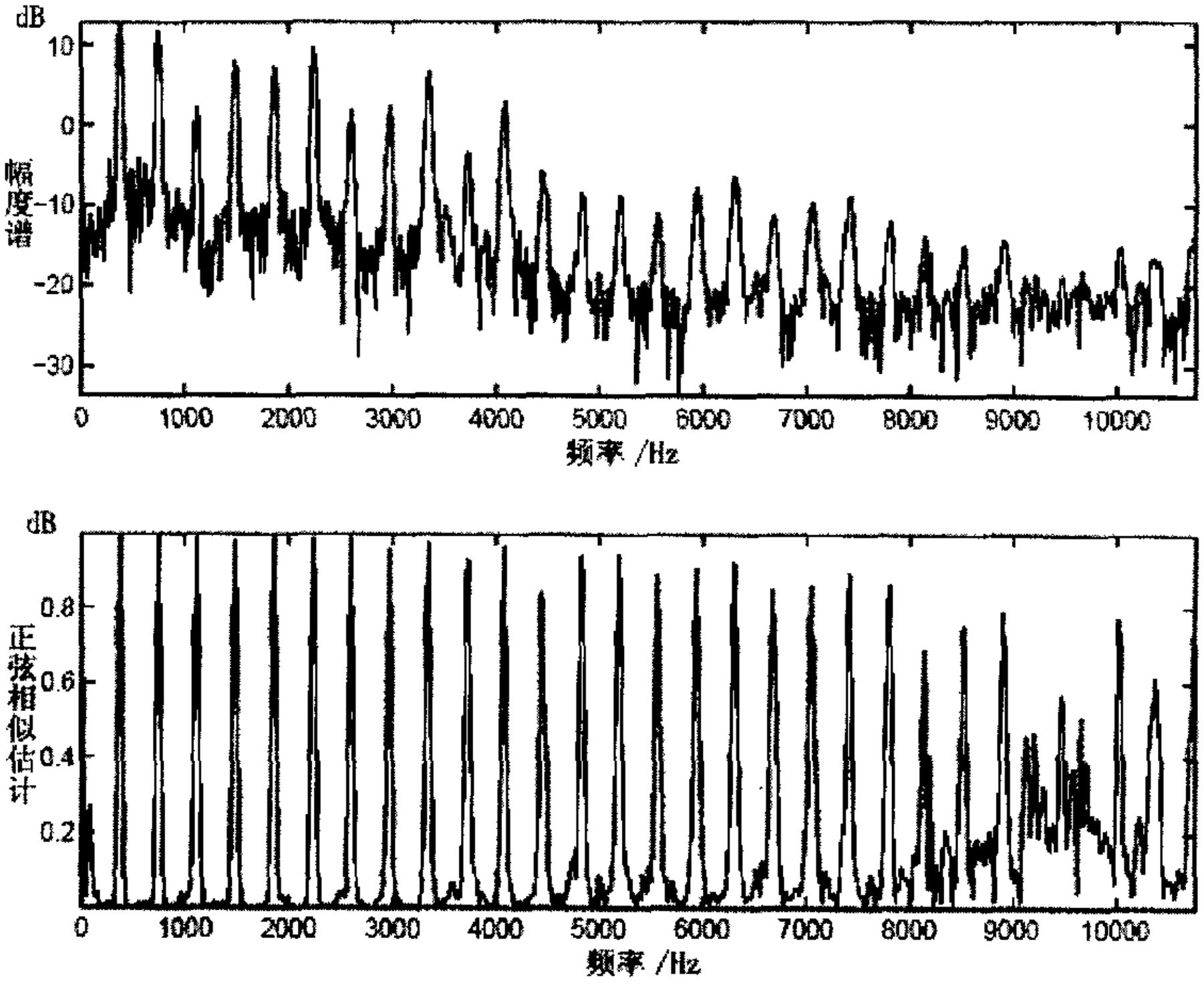

图 3—3 上图是小提琴音的幅度谱；下图是对样本的正弦相似测量

从心理声学的观点上看，用正弦来重构出噪声引起的峰值并不一定是错误的。在有些条件下，人类听觉系统试图对不是周期的信号分量指定音高，例如对延时的宽带噪声或重复噪声脉冲等[18]，但是我们的目标只是用正弦模型重构信号的周期的部分。实践中，我们发现互相关方法对动态参数特别是在时变频率中有很好的稳定性。

3．2 F 测试

在 1982 年由 D. J. Thomson 提出的一种统计学测试最早应用在地球物理学上，但是现在它已经成功的应用在声音的正弦检测上[11]。这种方法使用了一组正交窗，叫做离散扁长椭球序列。对于偏差和光滑的问题，其解决方法是用几个数据窗的加权平均来计算出谱的估计值。

与互相关方法一样，F 测试对每一个频率分量给一个值，这将会决定在这个频率内包含正弦的可能性有多大。在 F 测试中这个值叫做 f 值。我们可以设定一个固定域，f 值的频率系数是局部最大并比该固定域值还要大，所以频率系数用此频率

一个较小帧的里，我们不能在特定的频率上准确的判断是否存在正弦，为了准确的正弦分解还需要从相邻帧里得到信息。

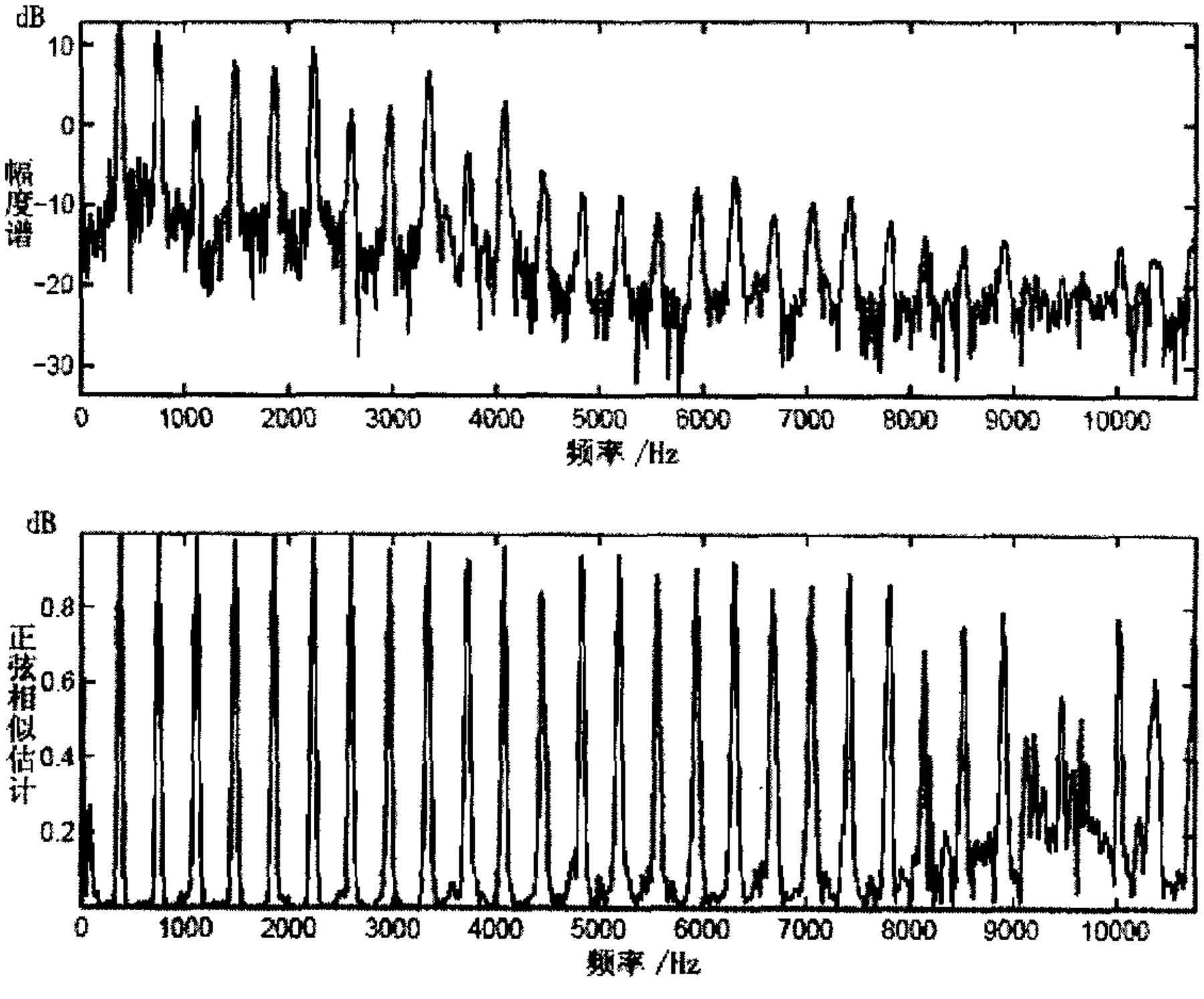

图 3—3 上图是小提琴音的幅度谱；下图是对样本的正弦相似测量

从心理声学的观点上看，用正弦来重构出噪声引起的峰值并不一定是错误的。在有些条件下，人类听觉系统试图对不是周期的信号分量指定音高，例如对延时的宽带噪声或重复噪声脉冲等[18]，但是我们的目标只是用正弦模型重构信号的周期的部分。实践中，我们发现互相关方法对动态参数特别是在时变频率中有很好的稳定性。

3．2 F 测试

在 1982 年由 D. J. Thomson 提出的一种统计学测试最早应用在地球物理学上，但是现在它已经成功的应用在声音的正弦检测上[11]。这种方法使用了一组正交窗，叫做离散扁长椭球序列。对于偏差和光滑的问题，其解决方法是用几个数据窗的加权平均来计算出谱的估计值。

与互相关方法一样，F 测试对每一个频率分量给一个值，这将会决定在这个频率内包含正弦的可能性有多大。在 F 测试中这个值叫做 f 值。我们可以设定一个固定域，f 值的频率系数是局部最大并比该固定域值还要大，所以频率系数用此频率

的正弦来进行插值。与互相关方法一样，假定残余谱光滑，F 测试也能够测量对连续的谐音分量谱与非谐音分量谱的比率。

用离散扁长球体序列做窗函数，在一个有限频率区间内，这些窗的能量就是最集中的[13]。输入信号中对每个序列加窗，并且对每个加窗信号做 FFT 会得到谱的估计值。因为序列是相互正交的，它们彼此并不相关。

估计谱的变化是依靠谱的局部连续部分，并给出背景谱的估计值，通过比较谱的连续部分的特定频率的能量我们得到了 f 值。由于 F 测试要求做好几次 FFT，因此计算起来比互相关方法更加费时。在理想条件下，能非常可靠并且是能在没有挑出噪音峰的条件下检测正弦。在非理想条件下，例如正弦密集的空间或幅度和频率迅速改变的时候，它就没有互相关算法理想。如果正弦的频率彼此接近，它们就可能彼此相互抵消。如果序列的窗长比正弦波长小，F 测试的性能就明显减弱。

小提琴样本的 f 值，幅度谱见图 3—3 和图 3—4，小提琴有些颤音或有频率调制，这些对 f 值影响要比对正弦相似测量的影响大。在一般情况下 F 测试和正弦相似测量的差别不是很大，见图 3—3 和图 3—4。

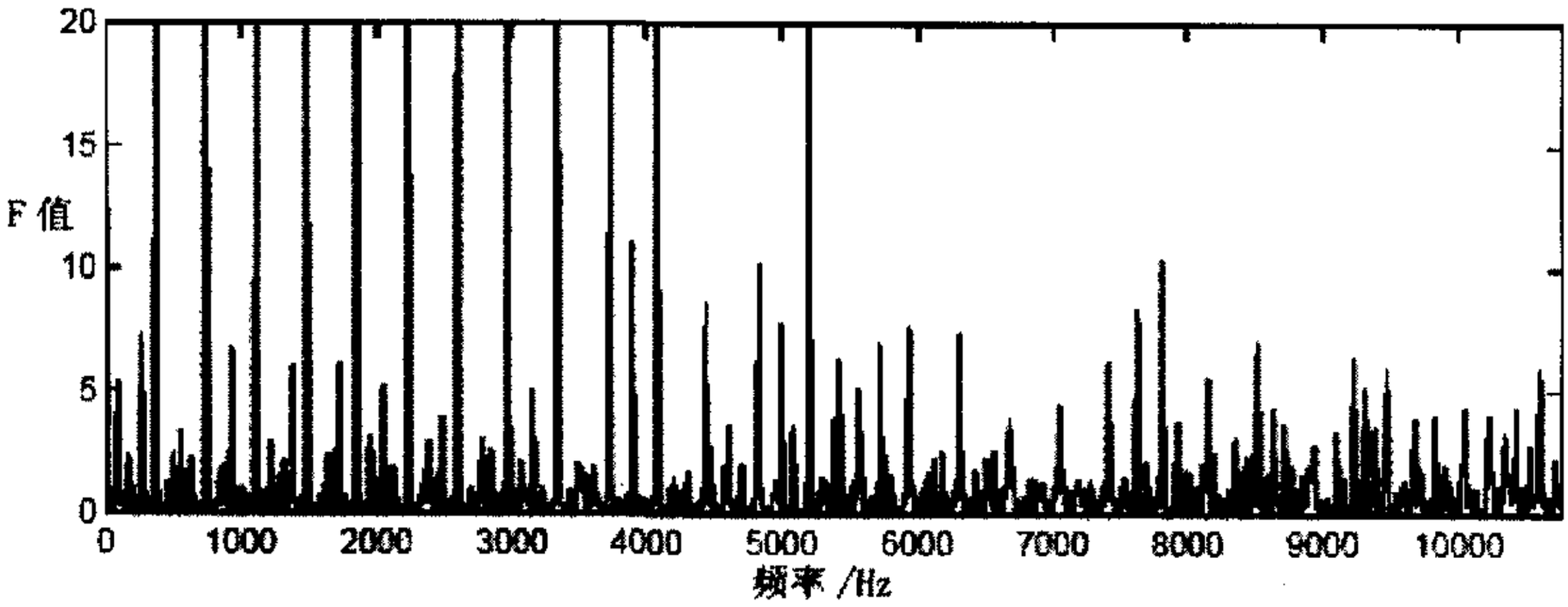

图 3—4 小提琴信号的 F 值

3．3 二次插值

根据 Heisenberg 不确定性原理，频率分辨率限制在一个有限的时间帧里。然而，如果一个正弦只在它周围有有效分量，那么添加零点可以用来获得更好的分辨率，这就使正弦的谱形状和位置更清晰，并且参数估计更精确。

每一个 DFT 参数代表一个 F_s/N 频率区间，这里 F_s 是样本频率，N 是 DFT 的长度。例如，相邻音符的一个半音在西方乐器中低音部是差 1Hz，对于高质量的采样频率，需要 10K 甚至 100K 样本的 DFT 长度，这是不实际的，所以需要不同的方法来获得正弦的精确频率。1987 年由 J.O.Smith 和 X.Serra 第一次提出了用二次方程来获得正弦分量的精确频率[9]。

的正弦来进行插值。与互相关方法一样，假定残余谱光滑，F 测试也能够测量对连续的谐音分量谱与非谐音分量谱的比率。

用离散扁长球体序列做窗函数，在一个有限频率区间内，这些窗的能量就是最集中的[13]。输入信号中对每个序列加窗，并且对每个加窗信号做 FFT 会得到谱的估计值。因为序列是相互正交的，它们彼此并不相关。

估计谱的变化是依靠谱的局部连续部分，并给出背景谱的估计值，通过比较谱的连续部分的特定频率的能量我们得到了 f 值。由于 F 测试要求做好几次 FFT，因此计算起来比互相关方法更加费时。在理想条件下，能非常可靠并且是能在没有挑出噪音峰的条件下检测正弦。在非理想条件下，例如正弦密集的空间或幅度和频率迅速改变的时候，它就没有互相关算法理想。如果正弦的频率彼此接近，它们就可能彼此相互抵消。如果序列的窗长比正弦波长小，F 测试的性能就明显减弱。

小提琴样本的 f 值，幅度谱见图 3—3 和图 3—4，小提琴有些颤音或有频率调制，这些对 f 值影响要比对正弦相似测量的影响大。在一般情况下 F 测试和正弦相似测量的差别不是很大，见图 3—3 和图 3—4。

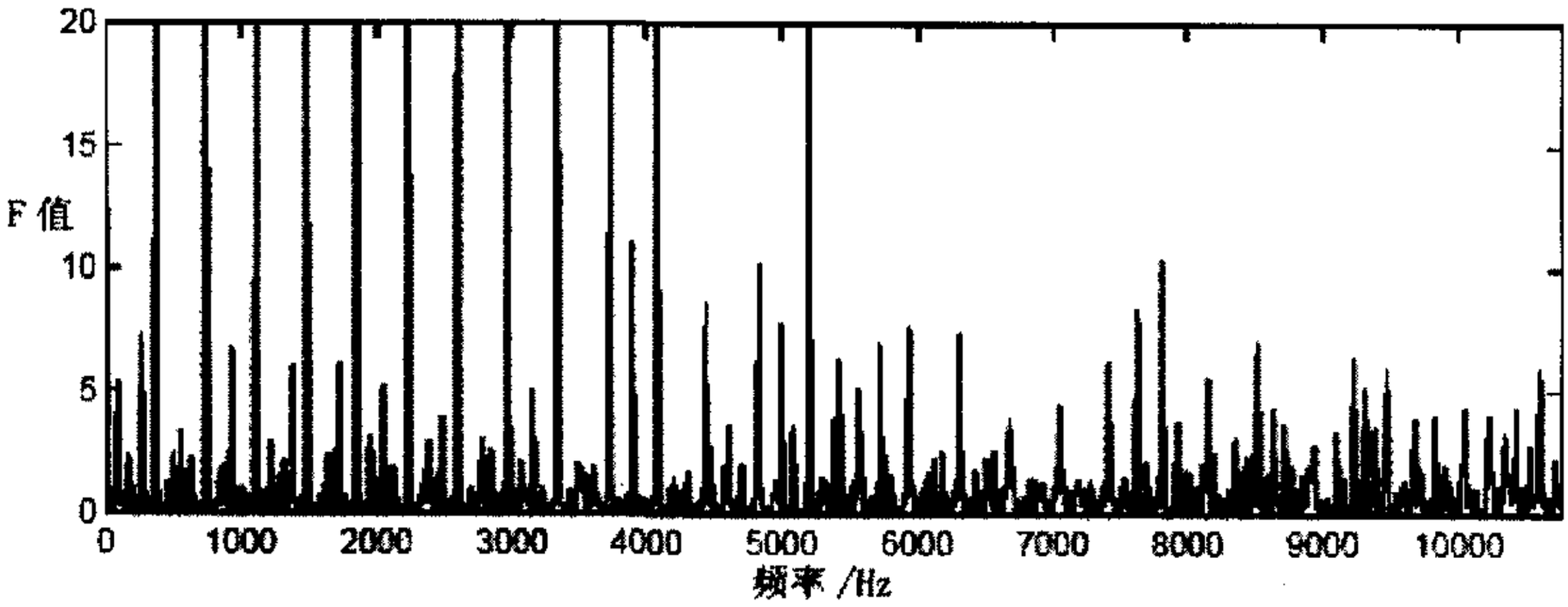

图 3—4 小提琴信号的 F 值

3．3 二次插值

根据 Heisenberg 不确定性原理，频率分辨率限制在一个有限的时间帧里。然而，如果一个正弦只在它周围有有效分量，那么添加零点可以用来获得更好的分辨率，这就使正弦的谱形状和位置更清晰，并且参数估计更精确。

每一个 DFT 参数代表一个 F_s/N 频率区间，这里 F_s 是样本频率，N 是 DFT 的长度。例如，相邻音符的一个半音在西方乐器中低音部是差 1Hz，对于高质量的采样频率，需要 10K 甚至 100K 样本的 DFT 长度，这是不实际的，所以需要不同的方法来获得正弦的精确频率。1987 年由 J.O.Smith 和 X.Serra 第一次提出了用二次方程来获得正弦分量的精确频率[9]。

局部最大值$|X(\omega)|$是窗口信号谱的绝对值，表示频率附近的正弦。加窗正弦的形状是$|H(\omega-\Omega)|$，窗口函数的 DFT 样本形状转化为频率Ω。如果窗口函数$h(t)$是均衡的，中心在Ω的一个二次函数在Ω周围给出$|H(\omega-\Omega)|$的很好的近似值，我们可以仅用 DFT 谱的三个点来估计函数的参数[16]。对于所用的窗口函数，取 H 的对数比只用 H 效果要好的多。假定 H 是零附近的高斯函数，则 H 的对数是二次的。如果$|X(\omega_{\lambda-1})|$、$|X(\omega_{\lambda})|$和$|X(\omega_{\lambda+1})|$是谱绝对值的相邻值，$|X(\omega_{\lambda})|$是最大值，二次方程是：

$$f(\omega)=a\omega^2+b\omega+c=\log|X(\omega)|,\omega\approx\omega_{\lambda} \tag{3-6}$$

其中a、b和c的值由二次方程来获得，设二次方程的导数等于零，就可以得到频率和幅度的估计值：

$$a_{peak}=|X(\omega_{\lambda})|+\frac{1}{8}\frac{\log|X(\omega_{\lambda+1})|-\log|X(\omega_{\lambda-1})|}{\log|X(\omega_{\lambda+1})|+\log|X(\omega_{\lambda-1})|-2\log|X(\omega_{\lambda})|} \tag{3-7}$$

$$\omega_{peak}=\omega_{\lambda}+\frac{\log|X(\omega_{\lambda+1})|-\log|X(\omega_{\lambda-1})|}{\log|X(\omega_{\lambda+1})|+\log|X(\omega_{\lambda-1})|-2\log|X(\omega_{\lambda})|}(\omega_{\lambda+1}-\omega_{\lambda}) \tag{3-8}$$

在得到频率后，相位谱用与精确频率最近的两个 DFT 系数的加权平均值来做插值。单正弦的二次插值见图 3—5。

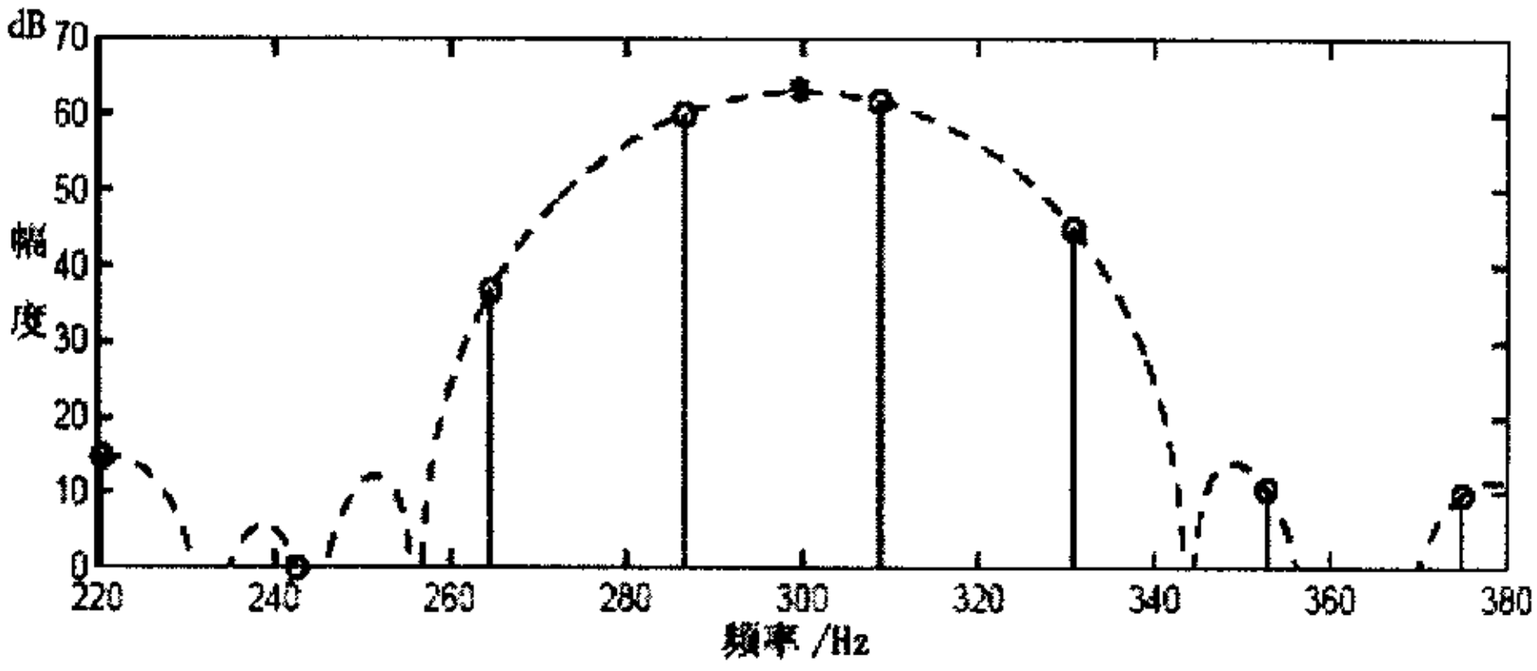

图 3—5 300Hz 信号的二次插值

虽然这种方法是基于理想的正弦信号，或者是没有噪音或没有其他正弦出现的时候，但是在多音信号的时候也不失为一种好的插值替换频率的方法。应该注意到二次插值法是假定中心幅度$|X(\omega_{\lambda})|$是比相邻幅度$|X(\omega_{\lambda-1})|$和$|X(\omega_{\lambda+1})|$都要大。而某些时候，就不能仅仅选择局部最大幅度谱方法，而是要用更加复杂的峰值检测算法，它的中心幅度可能并不是最大的，这时二次插值法就不能使用了。

3．4 信号导数插值法

Desainte—Catherine 和 S.Marchand 证明了信号的 DFT 和它的导数能够用来获得谱分量的精确频率和幅度[20]。对信号取导数并不影响正弦的频率，在理想情

局部最大值$|X(\omega)|$是窗口信号谱的绝对值，表示频率附近的正弦。加窗正弦的形状是$|H(\omega-\Omega)|$，窗口函数的 DFT 样本形状转化为频率Ω。如果窗口函数$h(t)$是均衡的，中心在Ω的一个二次函数在Ω周围给出$|H(\omega-\Omega)|$的很好的近似值，我们可以仅用 DFT 谱的三个点来估计函数的参数[16]。对于所用的窗口函数，取 H 的对数比只用 H 效果要好的多。假定 H 是零附近的高斯函数，则 H 的对数是二次的。如果$|X(\omega_{\lambda-1})|$、$|X(\omega_{\lambda})|$和$|X(\omega_{\lambda+1})|$是谱绝对值的相邻值，$|X(\omega_{\lambda})|$是最大值，二次方程是：

$$f(\omega)=a\omega^2+b\omega+c=\log|X(\omega)|,\omega\approx\omega_{\lambda} \tag{3-6}$$

其中a、b和c的值由二次方程来获得，设二次方程的导数等于零，就可以得到频率和幅度的估计值：

$$a_{peak}=|X(\omega_{\lambda})|+\frac{1}{8}\frac{\log|X(\omega_{\lambda+1})|-\log|X(\omega_{\lambda-1})|}{\log|X(\omega_{\lambda+1})|+\log|X(\omega_{\lambda-1})|-2\log|X(\omega_{\lambda})|} \tag{3-7}$$

$$\omega_{peak}=\omega_{\lambda}+\frac{\log|X(\omega_{\lambda+1})|-\log|X(\omega_{\lambda-1})|}{\log|X(\omega_{\lambda+1})|+\log|X(\omega_{\lambda-1})|-2\log|X(\omega_{\lambda})|}(\omega_{\lambda+1}-\omega_{\lambda}) \tag{3-8}$$

在得到频率后，相位谱用与精确频率最近的两个 DFT 系数的加权平均值来做插值。单正弦的二次插值见图 3—5。

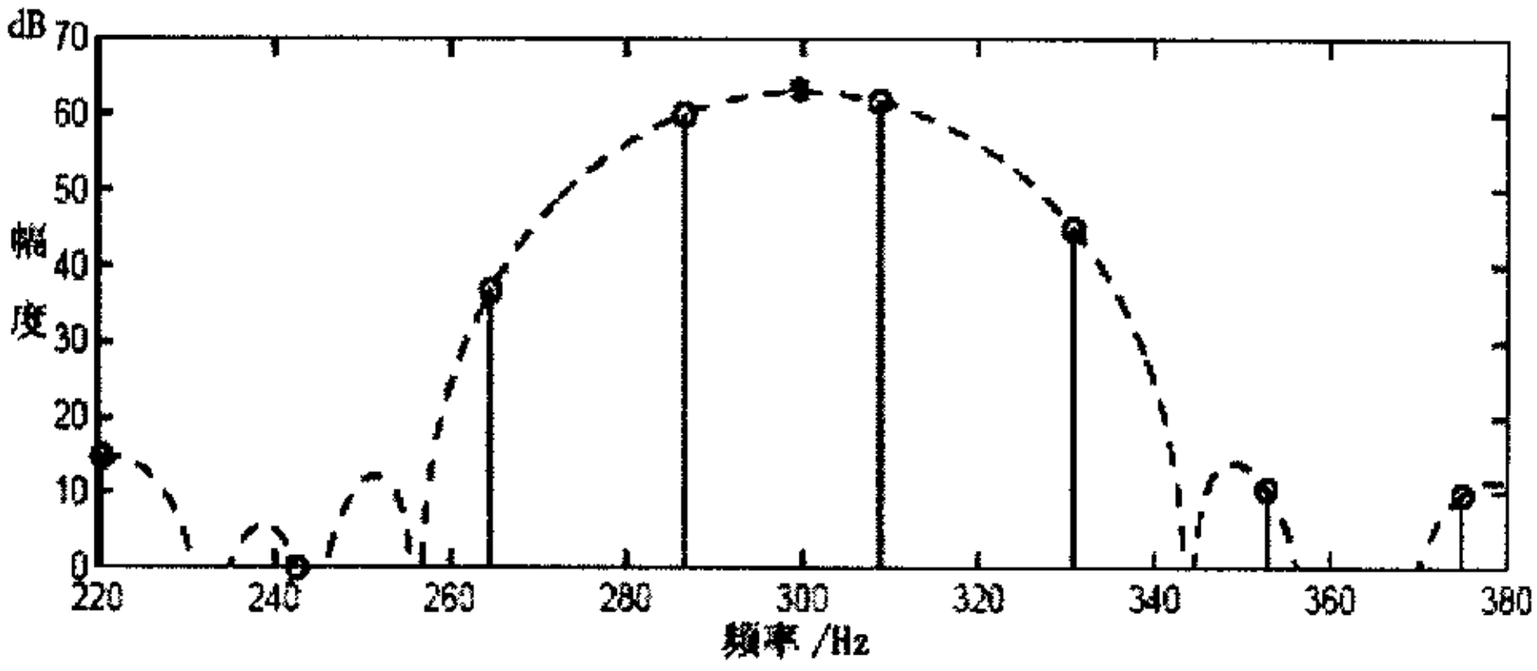

图 3—5 300Hz 信号的二次插值

虽然这种方法是基于理想的正弦信号，或者是没有噪音或没有其他正弦出现的时候，但是在多音信号的时候也不失为一种好的插值替换频率的方法。应该注意到二次插值法是假定中心幅度$|X(\omega_{\lambda})|$是比相邻幅度$|X(\omega_{\lambda-1})|$和$|X(\omega_{\lambda+1})|$都要大。而某些时候，就不能仅仅选择局部最大幅度谱方法，而是要用更加复杂的峰值检测算法，它的中心幅度可能并不是最大的，这时二次插值法就不能使用了。

3．4 信号导数插值法

Desainte—Catherine 和 S.Marchand 证明了信号的 DFT 和它的导数能够用来获得谱分量的精确频率和幅度[20]。对信号取导数并不影响正弦的频率，在理想情

况下，幅度对于频率线性变化。如果 $v(t)$ 是的 $x(t)$ 的导数， $x(t)$ 的 Fourier 变换是 $X(\omega)$，则 $v(t)$ 的 Fourier 变换是 $V(\omega)$，

$$V(\omega) = j\omega X(\omega) \tag{3-9}$$

系数 $j\omega$ 是理论上的并不用在离散时间处理上。信号的导数必需近似等于一阶差分，一阶差分可以看做用一阶线性滤波器进行一次滤波。在近似的信号导数谱和理想值之间的误差可以通过改变因子 F 来纠正[20]：

$$F(\omega) = \frac{\omega}{2\sin(\omega/2)} \tag{3-10}$$

当信号导数的 DFT （DFT^1）通过缩放因子 F 纠正以后，正弦的频率 ω_{peak} 可以用 DFT^1 除原始的信号的 DFT 近似得到：

$$\omega_{peak} = \frac{1}{2\pi}\frac{DFT^1(\omega)}{DFT^0(\omega)} \tag{3-11}$$

两个信号都是在加窗以后做 DFT。

我们最初使用这种方法中并没有得到用二次插值法所得到的精确的频率估计，甚至和没有任何插值的参数估计比较也没有明显的改善。特别是在有噪音的条件下，二次插值所表现出来的结果更好。由于二次插值方法简单，我们将重点对它进行讨论。对信号进行详细测试的结果表明导数插值和二次插值的效果几乎是相同（见 6.2 章）。

3.5 迭代最小平方估计

目前即使用最好的方法，也很难在一帧的时间里对复杂声音只进行一次分解就能够精确的估计出正弦参数。比较理想的方法是先用一种简单的估计方法进行参数估计，然后利用迭代来改进参数组[21][22]。

如果我们假定正弦的幅度 a_k 和频率 ω_k 能在一帧里保持相对稳定，对一帧的正弦模型是：

$$\hat{s}(n) = \sum_{k=1}^{k} a_k \cos(2\pi\omega_k n + \phi_k) \tag{3-12}$$

这里 K 是正弦的总数，ϕ_k 是第 k 个正弦的初始相位。对模型 $\hat{s}(n)$ 的 STFT 估计值是：

$$\hat{s}(\omega) = \sum_{k=1}^{k} \frac{a_k}{2}(e^{j\phi_k}H(\omega-\omega_k) + e^{-j\phi_k}H(\omega+\omega_k)) \tag{3-13}$$

其中 $H(\omega)$ 是分解窗的 Fourier 变换。我们的目标是找到参数 a_k, ω_k 和 ϕ_k，最小化真实的 STFT 和估计的 STFT 之间的最小均方差值 $\|S-\hat{S}\|$。两个 STFT 都是在 N 个等距频率中测量的，其中 $\omega_i = i/N$， $i = 0,...,N-1$。等式（3－13）中 $\hat{S}$ 根据 ω_k 非线性变化，即使 $\hat{S}$ 和 ω_k 是线性的，表达式中 $\hat{S}$ 还包含未知参数的乘积，如正弦的总数 K 是

况下，幅度对于频率线性变化。如果 $v(t)$ 是的 $x(t)$ 的导数， $x(t)$ 的 Fourier 变换是 $X(\omega)$，则 $v(t)$ 的 Fourier 变换是 $V(\omega)$，

$$V(\omega) = j\omega X(\omega) \tag{3-9}$$

系数 $j\omega$ 是理论上的并不用在离散时间处理上。信号的导数必需近似等于一阶差分，一阶差分可以看做用一阶线性滤波器进行一次滤波。在近似的信号导数谱和理想值之间的误差可以通过改变因子 F 来纠正[20]：

$$F(\omega) = \frac{\omega}{2\sin(\omega/2)} \tag{3-10}$$

当信号导数的 DFT （DFT^1）通过缩放因子 F 纠正以后，正弦的频率 ω_{peak} 可以用 DFT^1 除原始的信号的 DFT 近似得到：

$$\omega_{peak} = \frac{1}{2\pi}\frac{DFT^1(\omega)}{DFT^0(\omega)} \tag{3-11}$$

两个信号都是在加窗以后做 DFT。

我们最初使用这种方法中并没有得到用二次插值法所得到的精确的频率估计，甚至和没有任何插值的参数估计比较也没有明显的改善。特别是在有噪音的条件下，二次插值所表现出来的结果更好。由于二次插值方法简单，我们将重点对它进行讨论。对信号进行详细测试的结果表明导数插值和二次插值的效果几乎是相同（见 6.2 章）。

3.5 迭代最小平方估计

目前即使用最好的方法，也很难在一帧的时间里对复杂声音只进行一次分解就能够精确的估计出正弦参数。比较理想的方法是先用一种简单的估计方法进行参数估计，然后利用迭代来改进参数组[21][22]。

如果我们假定正弦的幅度 a_k 和频率 ω_k 能在一帧里保持相对稳定，对一帧的正弦模型是：

$$\hat{s}(n) = \sum_{k=1}^{k} a_k \cos(2\pi\omega_k n + \phi_k) \tag{3-12}$$

这里 K 是正弦的总数，ϕ_k 是第 k 个正弦的初始相位。对模型 $\hat{s}(n)$ 的 STFT 估计值是：

$$\hat{s}(\omega) = \sum_{k=1}^{k} \frac{a_k}{2}(e^{j\phi_k}H(\omega-\omega_k) + e^{-j\phi_k}H(\omega+\omega_k)) \tag{3-13}$$

其中 $H(\omega)$ 是分解窗的 Fourier 变换。我们的目标是找到参数 a_k, ω_k 和 ϕ_k，最小化真实的 STFT 和估计的 STFT 之间的最小均方差值 $\|S-\hat{S}\|$。两个 STFT 都是在 N 个等距频率中测量的，其中 $\omega_i = i/N$， $i = 0,...,N-1$。等式（3－13）中 $\hat{S}$ 根据 ω_k 非线性变化，即使 $\hat{S}$ 和 ω_k 是线性的，表达式中 $\hat{S}$ 还包含未知参数的乘积，如正弦的总数 K 是

未知的，因此无法得到最小均方问题的分解结果。

我们也可以用其它一些方法从 a_k、ω_k 和 ϕ_k 的估计开始，通过迭代方法改善估计值。首先，假定频率是正确的，改善幅度和相位的问题。然后，假定幅度和相位是正确的，频率估计也改善了。这个处理过程重复多次，在每次迭代后得到改善的估计。在迭代过程中，正弦总数能被改变，所以我们能在需要的时候添加或减少正弦。

3．5．1 幅度和相位估计

假定正弦的总数和每个频率是已知的，等式（3－13）的谱估计可以写为：

$$\hat{X}(\omega)=\sum_{k=1}^{2k} p_k R_k(\omega) \tag{3-14}$$

其中参数 p_k 是：

$$\begin{cases} p_k=\dfrac{a_k}{2}\cos\phi_k & \mathrm{k}\in[1,\mathrm{K}] \\ \mathrm{p_{K+k}}=\dfrac{a_k}{2}\sin\phi_k & \mathrm{k}\in[1,\mathrm{K}] \end{cases} \tag{3-15}$$

已知 $2K$ 表达式相关窗口函数的 Fourier 变换：

$$\begin{cases} R_k(\omega)=H(\omega-\omega_k)+H(\omega+\omega_k) \\ R_{K+k}(\omega)=jH(\omega-\omega_k)-H(\omega-\omega_k) \end{cases} \tag{3-16}$$

如果我们定义一个 R 维矩阵 $N\times 2K$，其中 $R_{i,k}=R_k(\omega_i)$，并且未知参数 $[p\ ,\ldots,p_{2k}]'$ 的向量是 $\vec{p}$，谱估计可以写做：

$$\hat{\vec{X}}=R\vec{p} \tag{3-17}$$

最小平方解是[23]

$$\vec{p}=(R^H R)^{-1}R^H\vec{X} \tag{3-18}$$

从中我们可以得到幅度

$$a_k=2\sqrt{p_k^2+p_{K+k}^2} \tag{3-19}$$

和相位

$$\phi_k=\arg(p_k+jp_{K+k}) \tag{3-20}$$

对于已知频率，这种方法会得到一个好的结果，特别是在正弦频率彼此接近的时候。在这种条件下，其他方法通常表现很差。但是，如果频率彼此太接近，或者在相同的 F_i 间隔内，R 变成单一的，这时问题就没办法解决了。

3．5．2 频率估计

如果知道正弦的幅度和相位，我们就可以粗略的估计频率，模型线性依靠频率

未知的，因此无法得到最小均方问题的分解结果。

我们也可以用其它一些方法从 a_k、ω_k 和 ϕ_k 的估计开始，通过迭代方法改善估计值。首先，假定频率是正确的，改善幅度和相位的问题。然后，假定幅度和相位是正确的，频率估计也改善了。这个处理过程重复多次，在每次迭代后得到改善的估计。在迭代过程中，正弦总数能被改变，所以我们能在需要的时候添加或减少正弦。

3．5．1 幅度和相位估计

假定正弦的总数和每个频率是已知的，等式（3－13）的谱估计可以写为：

$$\hat{X}(\omega)=\sum_{k=1}^{2k} p_k R_k(\omega) \tag{3-14}$$

其中参数 p_k 是：

$$\begin{cases} p_k=\dfrac{a_k}{2}\cos\phi_k & \mathrm{k}\in[1,\mathrm{K}] \\ \mathrm{p_{K+k}}=\dfrac{a_k}{2}\sin\phi_k & \mathrm{k}\in[1,\mathrm{K}] \end{cases} \tag{3-15}$$

已知 $2K$ 表达式相关窗口函数的 Fourier 变换：

$$\begin{cases} R_k(\omega)=H(\omega-\omega_k)+H(\omega+\omega_k) \\ R_{K+k}(\omega)=jH(\omega-\omega_k)-H(\omega-\omega_k) \end{cases} \tag{3-16}$$

如果我们定义一个 R 维矩阵 $N\times 2K$，其中 $R_{i,k}=R_k(\omega_i)$，并且未知参数 $[p\ ,\ldots,p_{2k}]'$ 的向量是 $\vec{p}$，谱估计可以写做：

$$\hat{\vec{X}}=R\vec{p} \tag{3-17}$$

最小平方解是[23]

$$\vec{p}=(R^H R)^{-1}R^H\vec{X} \tag{3-18}$$

从中我们可以得到幅度

$$a_k=2\sqrt{p_k^2+p_{K+k}^2} \tag{3-19}$$

和相位

$$\phi_k=\arg(p_k+jp_{K+k}) \tag{3-20}$$

对于已知频率，这种方法会得到一个好的结果，特别是在正弦频率彼此接近的时候。在这种条件下，其他方法通常表现很差。但是，如果频率彼此太接近，或者在相同的 F_i 间隔内，R 变成单一的，这时问题就没办法解决了。

3．5．2 频率估计

如果知道正弦的幅度和相位，我们就可以粗略的估计频率，模型线性依靠频率

未知的，因此无法得到最小均方问题的分解结果。

我们也可以用其它一些方法从 a_k、ω_k 和 ϕ_k 的估计开始，通过迭代方法改善估计值。首先，假定频率是正确的，改善幅度和相位的问题。然后，假定幅度和相位是正确的，频率估计也改善了。这个处理过程重复多次，在每次迭代后得到改善的估计。在迭代过程中，正弦总数能被改变，所以我们能在需要的时候添加或减少正弦。

3．5．1 幅度和相位估计

假定正弦的总数和每个频率是已知的，等式（3－13）的谱估计可以写为：

$$\hat{X}(\omega)=\sum_{k=1}^{2k} p_k R_k(\omega) \tag{3-14}$$

其中参数 p_k 是：

$$\begin{cases} p_k=\dfrac{a_k}{2}\cos\phi_k & \mathrm{k}\in[1,\mathrm{K}] \\ \mathrm{p_{K+k}}=\dfrac{a_k}{2}\sin\phi_k & \mathrm{k}\in[1,\mathrm{K}] \end{cases} \tag{3-15}$$

已知 $2K$ 表达式相关窗口函数的 Fourier 变换：

$$\begin{cases} R_k(\omega)=H(\omega-\omega_k)+H(\omega+\omega_k) \\ R_{K+k}(\omega)=jH(\omega-\omega_k)-H(\omega-\omega_k) \end{cases} \tag{3-16}$$

如果我们定义一个 R 维矩阵 $N\times 2K$，其中 $R_{i,k}=R_k(\omega_i)$，并且未知参数 $[p\ ,\ldots,p_{2k}]'$ 的向量是 $\vec{p}$，谱估计可以写做：

$$\hat{\vec{X}}=R\vec{p} \tag{3-17}$$

最小平方解是[23]

$$\vec{p}=(R^H R)^{-1}R^H\vec{X} \tag{3-18}$$

从中我们可以得到幅度

$$a_k=2\sqrt{p_k^2+p_{K+k}^2} \tag{3-19}$$

和相位

$$\phi_k=\arg(p_k+jp_{K+k}) \tag{3-20}$$

对于已知频率，这种方法会得到一个好的结果，特别是在正弦频率彼此接近的时候。在这种条件下，其他方法通常表现很差。但是，如果频率彼此太接近，或者在相同的 F_i 间隔内，R 变成单一的，这时问题就没办法解决了。

3．5．2 频率估计

如果知道正弦的幅度和相位，我们就可以粗略的估计频率，模型线性依靠频率

变化。我们的目标是估计 $\Delta_k = \omega_k - \hat{\omega}_k$，是估计的频率 $\hat{\omega}_k$ 和正确的频率 ω_k 之间的距离。对每个频率测量点 ω_i，我们使频率线性化，使用 $H(\omega)$ 的级数展开。分解窗的 Fourier 变换能写成：

$$H(\omega \mp \omega_k) = H(\omega \mp \hat{\omega}_k) \mp H'(\omega \mp \hat{\omega}_k)\Delta_k + o(\Delta_k^2) \tag{3-21}$$

其中导数 $H'(\omega)$ 在离散频率点既能用 $H(\omega)$ 的一阶微分也能用 $h(t)t$ 乘积的 DFT 来估计。如果我们定义矩阵 Ω 如下：

$$(\Omega)_{i,k} = \frac{a_k}{2}(-e^{j\phi_k} H'(\omega_i - \hat{\omega}_k) + e^{-j\phi_k} H'(\omega_i + \hat{\omega}_k)) \tag{3-22}$$

我们可以重写谱估计为：

$$\hat{X} = \tilde{X} + \Omega\Delta \tag{3-23}$$

其中 $\tilde{X}$ 是用频率 $\hat{\omega}_k$ 的 STFT 模型估计。对频率的最小平方解是：

$$\omega = \hat{\omega}_k + (\Omega^H\Omega)^{-1}\Omega^H(X - \tilde{X}) \tag{3-24}$$

因为用了 $H(\omega)$ 的一阶展开，这个方法是对分解窗 $h(t)$ 的形状非常敏感。实际上，这意味着分解窗的 Fourier 变换没有旁瓣。Ph.Depalle 和 T.Helie 提出用一个小的带宽和一个小的有效的持续时间设计窗的方法[21]。

在理想的情况或者信号全部由合成的正弦组成时，即使初始值和正确值相差甚远，迭代分解都能得到一个好的参数估计。但是，在更多的有复杂的情况下，例如正弦密集或者含有复杂的多音信号时，算法表现就很差。如果频率估计接近正确值，这种方法能对幅度和相位有很好的估计，但是如果频率估计不正确，算法就不能对复杂信号有很好的估计。这种问题可以通过用分离谱到分开的频率段中，并且在每个段内分别解决分离的参数的办法来解决。

我们发现 LSQ 算法在用非迭代实现时对幅度和相位最有用。当频率用峰值检测算法得到后，幅度和相位能用 LSQ 算法在第一步解决。甚至在紧密相连的正弦中，算法输出正确的参数仍能保证频率是正确的。

3.6 残余信号的迭代分解

目前有一种迭代方法可以完成残余信号的迭代分解[24]。和一个参数融合算法结合起来，这个参数估计程序有两个优势：一是减少正弦分量的数量；二是给单个正弦更加精确的参数。因为迭代分解要经过几个传统分解的步骤，计算起来非常费事。

迭代分解处理过程如下：首先，我们用一些简单的检测方法检测正弦信号；第二步，我们合成信号，然后在时域中从原始信号中减去合成的信号；第三步，在余下的残余信号中检测正弦，再合成并再去掉这些信号。一直重复到一个固定次数的迭代，或者一直到得到了想要的正弦，或者残余信号中没有了有意义的信号分量。用这种算法在感知上得到一个好的结果时，正弦的数目通常已经变得非常大了。每

变化。我们的目标是估计 $\Delta_k = \omega_k - \hat{\omega}_k$，是估计的频率 $\hat{\omega}_k$ 和正确的频率 ω_k 之间的距离。对每个频率测量点 ω_i，我们使频率线性化，使用 $H(\omega)$ 的级数展开。分解窗的 Fourier 变换能写成：

$$H(\omega \mp \omega_k) = H(\omega \mp \hat{\omega}_k) \mp H'(\omega \mp \hat{\omega}_k)\Delta_k + o(\Delta_k^2) \tag{3-21}$$

其中导数 $H'(\omega)$ 在离散频率点既能用 $H(\omega)$ 的一阶微分也能用 $h(t)t$ 乘积的 DFT 来估计。如果我们定义矩阵 Ω 如下：

$$(\Omega)_{i,k} = \frac{a_k}{2}(-e^{j\phi_k} H'(\omega_i - \hat{\omega}_k) + e^{-j\phi_k} H'(\omega_i + \hat{\omega}_k)) \tag{3-22}$$

我们可以重写谱估计为：

$$\hat{X} = \tilde{X} + \Omega\Delta \tag{3-23}$$

其中 $\tilde{X}$ 是用频率 $\hat{\omega}_k$ 的 STFT 模型估计。对频率的最小平方解是：

$$\omega = \hat{\omega}_k + (\Omega^H\Omega)^{-1}\Omega^H(X - \tilde{X}) \tag{3-24}$$

因为用了 $H(\omega)$ 的一阶展开，这个方法是对分解窗 $h(t)$ 的形状非常敏感。实际上，这意味着分解窗的 Fourier 变换没有旁瓣。Ph.Depalle 和 T.Helie 提出用一个小的带宽和一个小的有效的持续时间设计窗的方法[21]。

在理想的情况或者信号全部由合成的正弦组成时，即使初始值和正确值相差甚远，迭代分解都能得到一个好的参数估计。但是，在更多的有复杂的情况下，例如正弦密集或者含有复杂的多音信号时，算法表现就很差。如果频率估计接近正确值，这种方法能对幅度和相位有很好的估计，但是如果频率估计不正确，算法就不能对复杂信号有很好的估计。这种问题可以通过用分离谱到分开的频率段中，并且在每个段内分别解决分离的参数的办法来解决。

我们发现 LSQ 算法在用非迭代实现时对幅度和相位最有用。当频率用峰值检测算法得到后，幅度和相位能用 LSQ 算法在第一步解决。甚至在紧密相连的正弦中，算法输出正确的参数仍能保证频率是正确的。

3.6 残余信号的迭代分解

目前有一种迭代方法可以完成残余信号的迭代分解[24]。和一个参数融合算法结合起来，这个参数估计程序有两个优势：一是减少正弦分量的数量；二是给单个正弦更加精确的参数。因为迭代分解要经过几个传统分解的步骤，计算起来非常费事。

迭代分解处理过程如下：首先，我们用一些简单的检测方法检测正弦信号；第二步，我们合成信号，然后在时域中从原始信号中减去合成的信号；第三步，在余下的残余信号中检测正弦，再合成并再去掉这些信号。一直重复到一个固定次数的迭代，或者一直到得到了想要的正弦，或者残余信号中没有了有意义的信号分量。用这种算法在感知上得到一个好的结果时，正弦的数目通常已经变得非常大了。每

个单独的信号迭代得到的参数通常不是十分精确，这将会导致残余信号估计时出现错误。正弦分量的错误估计也是接近原始信号频率的正弦，并且幅度通常比原始信号的幅度小。在接下来的迭代中我们来检测这些估计错误，因此原始信号的每个谐音分量在重构时不止是用一个正弦，但这并不是我们想要的，每一步迭代以后我们的算法结合正弦见插图 3—6。

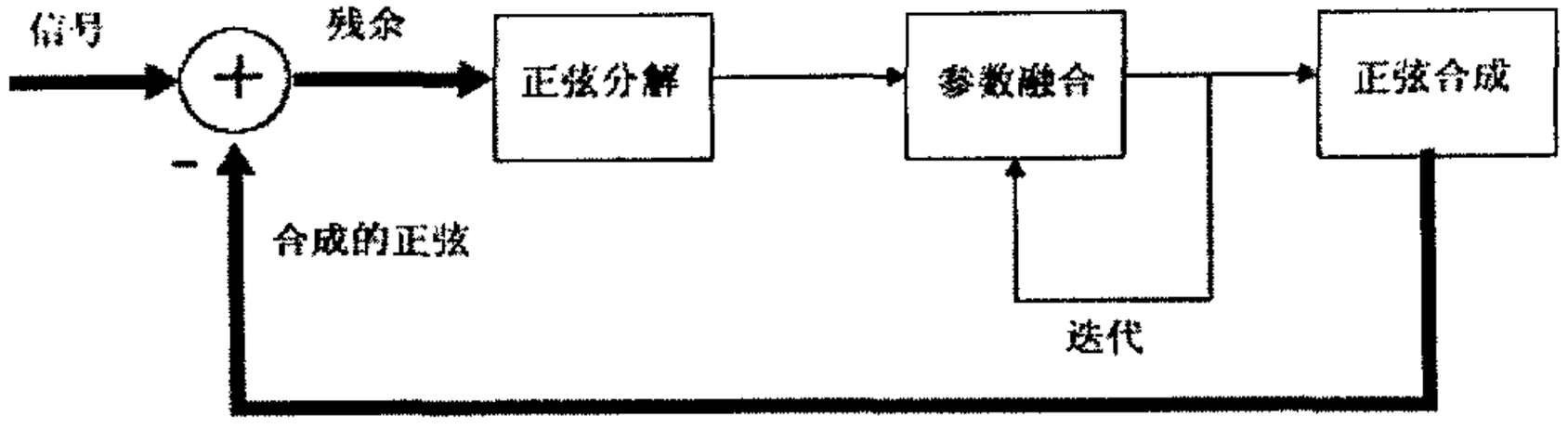

图 3—6 迭代参数估计算法框图

参数合并是基于一个假定条件：两个紧密相邻的正弦已经从原始信号中分离出来。因此我们可以用这样一种方法结合正弦，重构正弦的重要的谐音分量而不是单独处理每一个正弦。新正弦的参数计算出来后，新正弦表示了原始正弦信号的集合。接下来简化公式，我们假定时间 $t=0$，两个原始正弦的幅度，频率和相位是 $a_1, a_2, \omega_1, \omega_2, \varphi_1, \varphi_2$。则在时间 t 正弦总和是：

$$x(t)=a_1\sin(\omega_1 t+\varphi_1)+a_2\sin(\omega_2 t+\varphi_2) \tag{3—25}$$

附录一中解释正弦总和 $x(t)$ 能用正弦来重构，它的幅度和频率是时变的：

$$x(t)=a_3(t)\sin\left(\frac{(\omega_2+\omega_1)t+\varphi_2+\varphi_1}{2}+\int_0^t\omega_3(u)du+\varphi_3(0)\right) \tag{3—26}$$

其中新的幅度 $a_3(t)$，频率 $\omega_3(t)$ 和初始相位 $\varphi_3(0)$ 是

$$a_3(t)=\sqrt{a_1^2+a_2^2+2a_1a_2\cos\left((\omega_2-\omega_1)t+\varphi_2-\varphi_1\right)} \tag{3—27}$$

$$\omega_3(t)=\frac{\left[1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\right]}{1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\left[\frac{(a_2-a_1)}{(a_2+a_1)}\right]^2}\left(\frac{\omega_2-\omega_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)} \tag{3—28}$$

$$\varphi_3(0)=\begin{cases}\operatorname{atan}\left(\tan\left(\frac{\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\pi & \frac{\pi}{2}<\left(\frac{\varphi_2-\varphi_1}{2}\bmod 2\pi\right)<\frac{3\pi}{2}\\ a\tan\left(\tan\left(\frac{\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right) & \text{其他}\end{cases} \tag{3—29}$$

等式（3—26）的时变频率和时变相位（见附录一）都不能直接应用于我们的正弦模型中，因为我们的模型假设幅度和频率是常数，并且在一个帧里相位是线性的。

但是，在一定条件下，我们能够用常数近似表示时变幅度和频率。这些条件是：

1．时间t接近于零。这意味着只有在小的时间帧里估计值有效，帧与帧之间正弦模型的参数不断更新，所以满足这个条件。时间越短，近似效果越好。

2．频率彼此接近。当满足条件 1 和 2 时，项$(\omega_2-\omega_1)$t 在等式（3－27）和（3－28）中可以忽略不计。

3．在时间帧里，两个正弦总和的幅度包络没有局部最大和最小。这和原始信号的相位和频率有关，如果$0\le(\omega_2-\omega_1)S/F_s+(\varphi_2-\varphi_1+\pi/2)\bmod\pi\le\pi$，这种状况就可以实现，其中 S 是样本中帧的长度，F_s是样本频率。

4．有大的幅度率。这种情况发生在在第一步中得到的第一个正弦并且第二个正弦是第一个正弦的误差残余。如果条件满足，等式（3－28）中的项$\left[\frac{(a_2-a_1)}{(a_2+a_1)}\right]^2$接近于 1。

如果这些条件都满足，时变正弦的参数能用常参数的正弦来近似：

$$x(t)=a_n\sin(\omega_n t+\varphi_n) \tag{3-30}$$

其中常数a_n，ω_n和φ_n是新正弦的参数，用来替换旧的参数。近似值是：

$$a_n=\sqrt{a_1^2+a_2^2+2a_1a_2\cos(\varphi_1-\varphi_2)} \tag{3-31}$$

$$\omega_n=\frac{\omega_1a_1+\omega_2a_2}{a_1+a_2} \tag{3-32}$$

$$\varphi_n=\begin{cases}\operatorname{atan}\left(\tan\left(\frac{\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\frac{\varphi_2+\varphi_1}{2}+\pi & \frac{\pi}{2}<\left(\frac{\varphi_2-\varphi_1}{2}\bmod 2\pi\right)<\frac{3\pi}{2}\\ \operatorname{atan}\left(\tan\left(\frac{\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\frac{\varphi_2+\varphi_1}{2} & \text{其他}\end{cases} \tag{3-33}$$

一个近似值的例子见图 3－7。从图中能够清晰的看到零附近的近似值是最好的。

在参数合并中，我们检测满足所有条件的正弦对，新正弦的参数可以通过对每一帧计算得到，得到新正弦的参数后，用新的正弦替换老的。

在合成时，因为正弦参数在帧与帧之间做插值，所以很难测量单时间帧内近似值的有效性。幅度被线性的插值，并且如果在帧之间没有局部最大值和最小值，插值的作用会很明显。

在得到精确幅度的等式是（3－27），近似幅度的等式是（3－31），如果等式（3－27）的幅度包络导数的符号没有改变，我们能大概判定两个正弦的合并值是有效的，见图 3－8。Cos 的导数是负的 Sin，它的符号在$\pm n\pi, n=0,1,2...$变化。因此，近似值的有效性能公式化为：

$$\mathrm{sgn}[\sin((\omega_2-\omega_1)t+\varphi_2-\varphi_1)]=\mathrm{sgn}[\sin(\varphi_2-\varphi_1)],\forall t\left(0\le t\le\frac{S}{F_s}\right) \tag{3-34}$$

在帧的开始和结尾如果自变量 $(\omega_2-\omega_1)t+\varphi_2-\varphi_1$ 在相同区间 $[-\pi/2+n\pi,\pi/2+n\pi]$ 内（n 是任意整数）等式就可以成立。在点 $t=0$ 和 $t=S/F_s$，区间和自变量值都增加

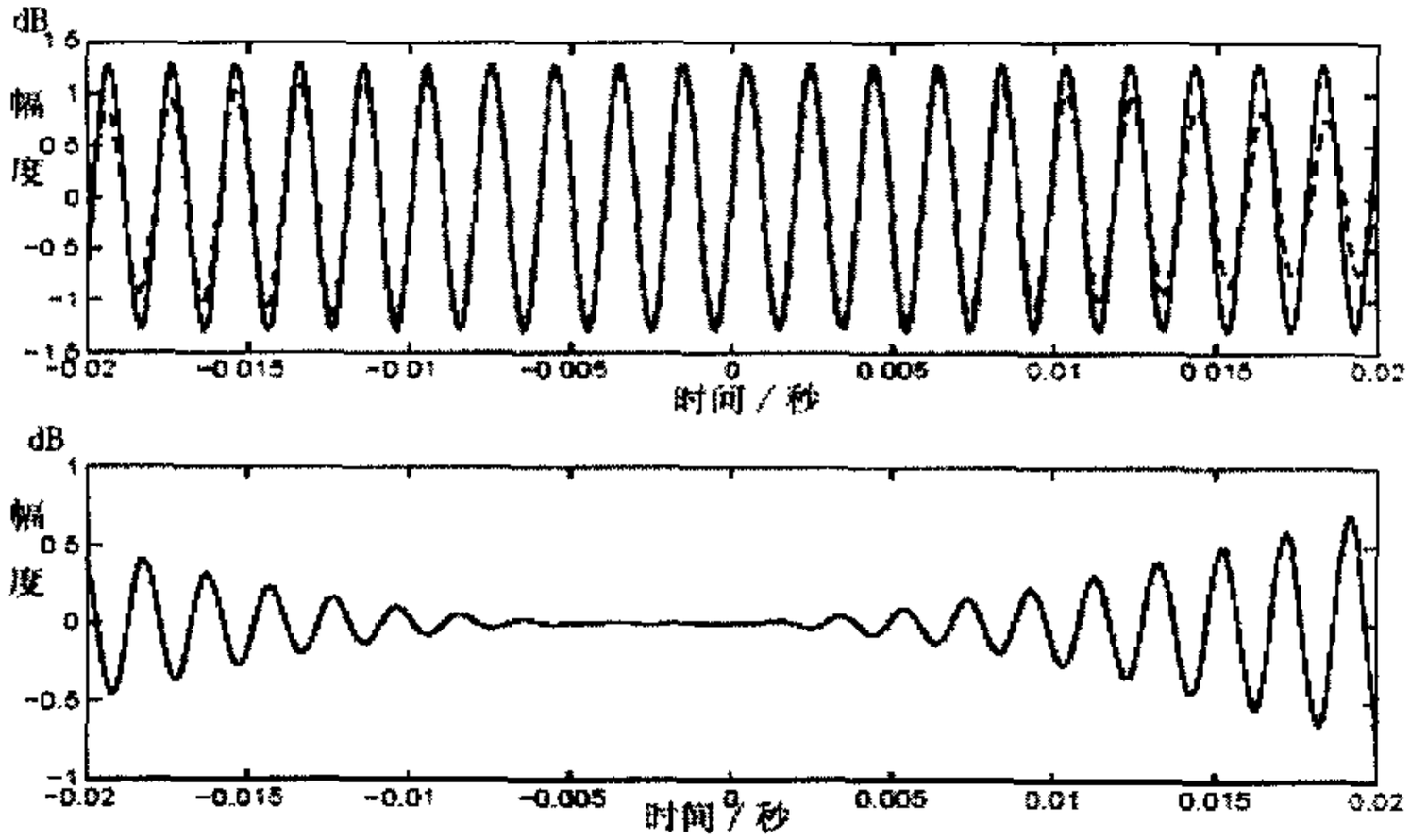

图 3-7 两个正弦融合的例子

$\pi/2$，自变量值在帧的开始和结尾变成 $(\omega_2-\omega_1)S/F_s+\varphi_2-\varphi_1+\pi/2$ 和 $\varphi_2-\varphi_1+\pi/2$，区间变成 $[n\pi,\pi+n\pi]$。$n=\left[\frac{\varphi_2-\varphi_1}{\pi}+\frac{1}{2}\right]$，其中关系：

$$n\pi\le(\omega_2-\omega_1)\frac{S}{F_s}+\varphi_2-\varphi_1+\frac{\pi}{2}\le\pi+n\pi \tag{3-35}$$

减去 $n\pi$ 后得到：

$$0\le(\omega_2-\omega_1)\frac{S}{F_s}+\varphi_2-\varphi_1+\frac{\pi}{2}-\pi\left[\frac{\varphi_2-\varphi_1}{\pi}+\frac{1}{2}\right]\le\pi \tag{3-36}$$

简化为：

$$0\le(\omega_2-\omega_1)\frac{S}{F_s}+\left(\varphi_2-\varphi_1+\frac{\pi}{2}\right)\bmod\pi\le\pi \tag{3-37}$$

对于一大组正弦，对于在频率之间的差别我们可以首先用一个常数限滤掉正弦对：

$$|(\omega_2-\omega_1)|\frac{S}{F_s}\le\pi\Leftrightarrow|(\omega_2-\omega_1)|\le\frac{\pi F_s}{S} \tag{3-38}$$

等式对于频率之间的差别给出一个域值 $\pi F_s/S$，它能够只用频率信息就能滤掉大多数正弦对。因此，如果它们满足其它的条件我们能检测所有的剩余对。

并没有十分有效的方法对未知信号内容的正弦分解做精确的判断，用生成的测试信号比较迭代算法和其他分解方法做了一些测试，结果在第六章给出。

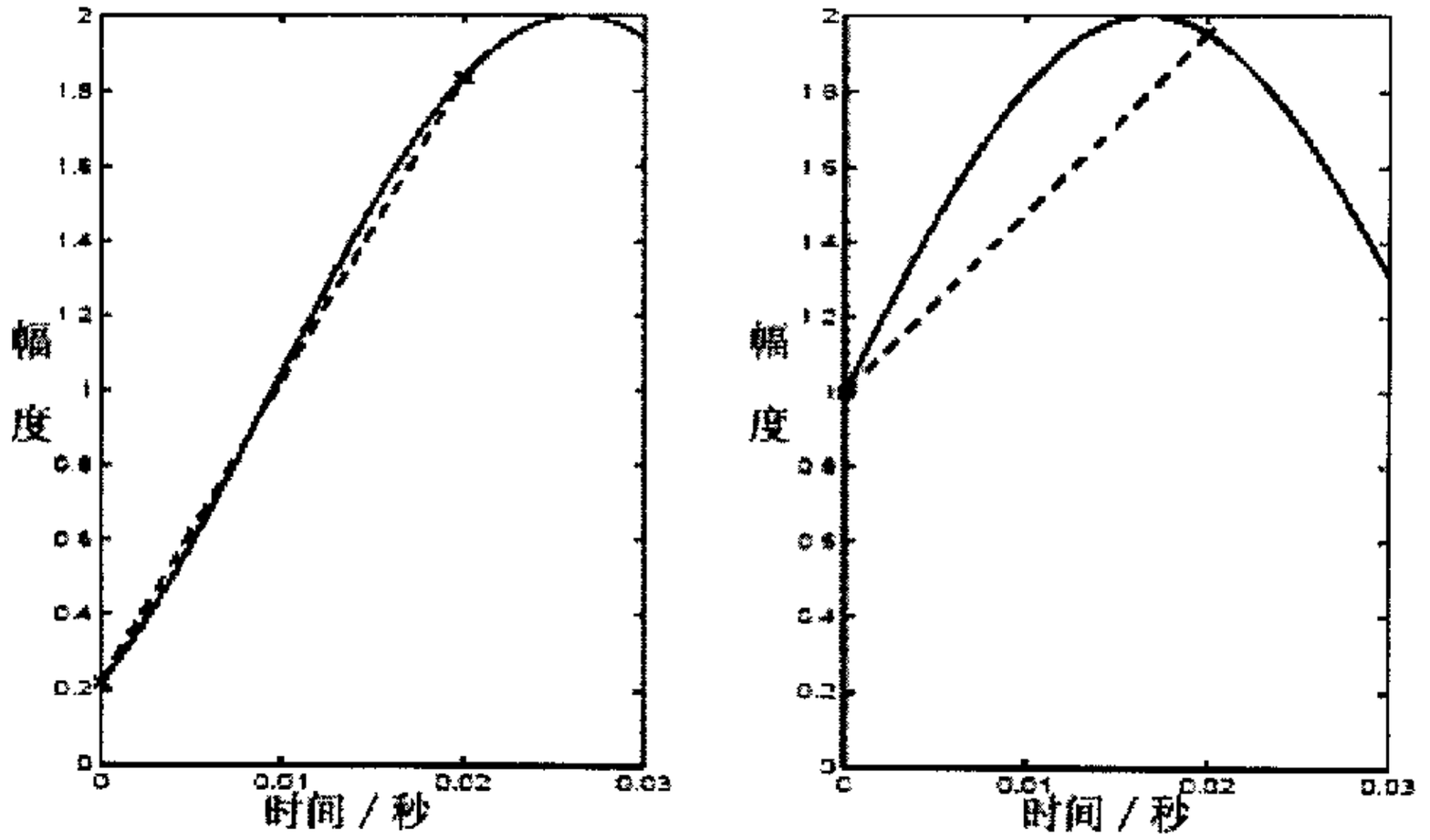

图 3—8 两个合成正弦幅度包络的线性逼近

3．7 多尺度逼近

因为短时谱的频率分辨率与分解窗的长度是线性的，正弦的频率精确的决定窗的长度。同样，需要长分解窗来检测低频正弦，窗长必需是正弦波长的 2—4 倍，具体的值取决于使用不同的分解方法，用长窗的一个缺点是时间分辨率很差。真实的声音经常表现出在频率和幅度上迅速改变，所以假定在一个窗里正弦是稳定的不可能实现。显而易见，在时间和频率分辨率之间要做出选择。

因为在低频时正弦的波长很长，显然这时需要一个长窗。在高频时，波长短并且分量经常迅速改变，因此需要短窗。在中频时窗长在两者之间，同时，音乐的音阶的在频率上有几何上的间隔。考虑到所有这些情况，常 Q 变换（CQT）是个很好的选择。CQT 的频率系数有几何上的间隔，并且窗长度与频率上成反比例[25]，中心频率和频率分辨率的比例是个常数，因此，叫做常 Q 变换。

有限制的 Q 变换（BQT）用对于不同八音度用不同的分辨率和窗，近似一个对数级的频率缩放，所以在每八音度频率系数是个常数。这种方法用 S.Levine 滤波器组来实现[11]。他只用在频率从 0—5KHz 的正弦分解，和三个八音度范围，0—1250Hz，1250—2500Hz 和 2500—5000Hz，并且窗口长度分别是 46ms，23ms 和 11.5ms。

在我们的系统中，正弦分解最高到 10KHz。系统是可变的，分解带不需要是一个八度，甚至可以定义在任何位置。在我们测试的音乐样本中最低的基础频率低音部大约 30Hz。46ms 的分解窗不能可靠的检测低音频率。因为低音部的绝大多数能量在低频区，我们发现对频率从 0 到 200Hz 用 80ms 的分解窗是足够的。46ms 的分解窗用在从 200Hz 到 5KHz。5KHz 以上声音特性是和低音部分有很大不同。46ms 的

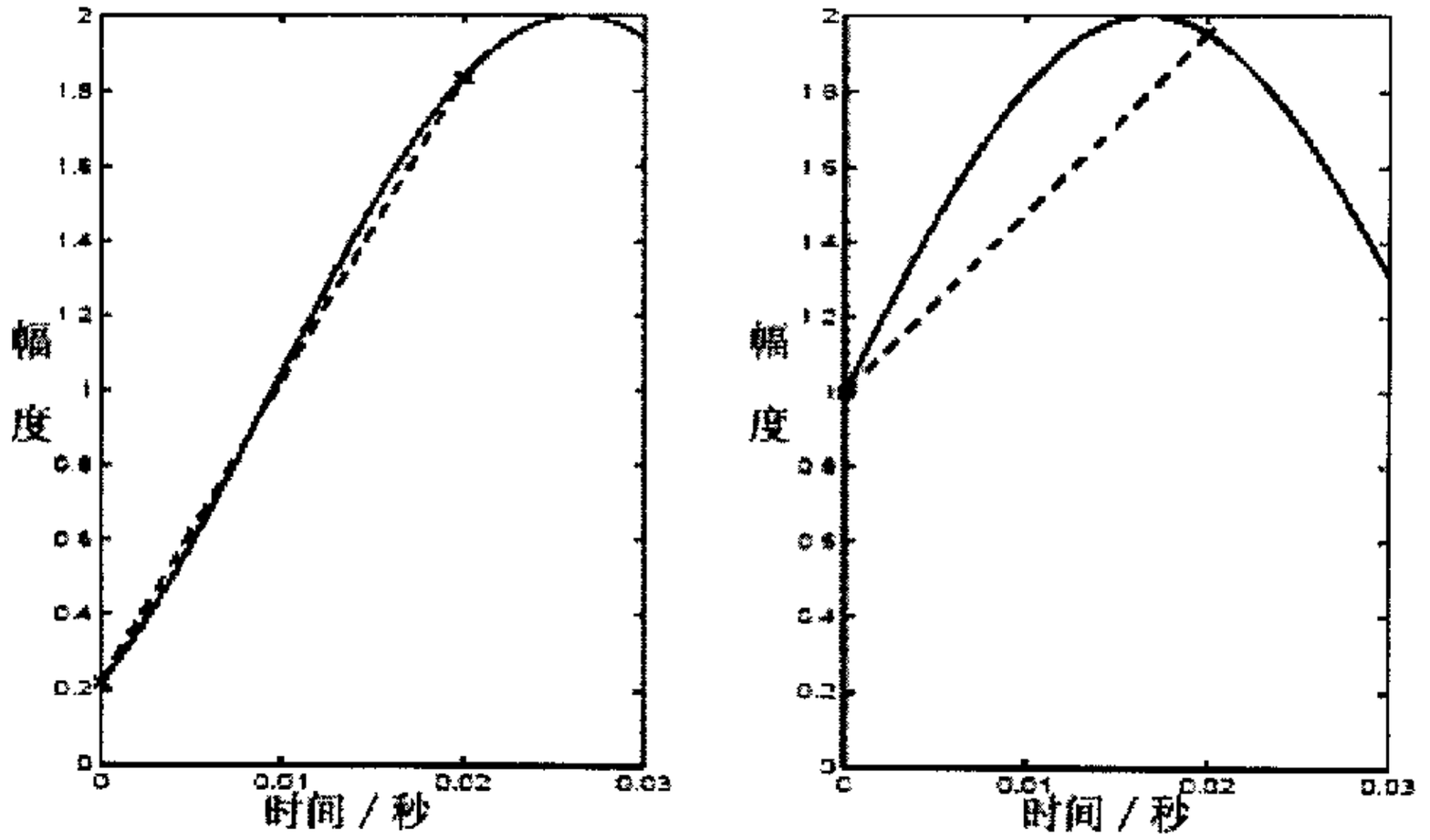

图 3—8 两个合成正弦幅度包络的线性逼近

3．7 多尺度逼近

因为短时谱的频率分辨率与分解窗的长度是线性的，正弦的频率精确的决定窗的长度。同样，需要长分解窗来检测低频正弦，窗长必需是正弦波长的 2—4 倍，具体的值取决于使用不同的分解方法，用长窗的一个缺点是时间分辨率很差。真实的声音经常表现出在频率和幅度上迅速改变，所以假定在一个窗里正弦是稳定的不可能实现。显而易见，在时间和频率分辨率之间要做出选择。

因为在低频时正弦的波长很长，显然这时需要一个长窗。在高频时，波长短并且分量经常迅速改变，因此需要短窗。在中频时窗长在两者之间，同时，音乐的音阶的在频率上有几何上的间隔。考虑到所有这些情况，常 Q 变换（CQT）是个很好的选择。CQT 的频率系数有几何上的间隔，并且窗长度与频率上成反比例[25]，中心频率和频率分辨率的比例是个常数，因此，叫做常 Q 变换。

有限制的 Q 变换（BQT）用对于不同八音度用不同的分辨率和窗，近似一个对数级的频率缩放，所以在每八音度频率系数是个常数。这种方法用 S.Levine 滤波器组来实现[11]。他只用在频率从 0—5KHz 的正弦分解，和三个八音度范围，0—1250Hz，1250—2500Hz 和 2500—5000Hz，并且窗口长度分别是 46ms，23ms 和 11.5ms。

在我们的系统中，正弦分解最高到 10KHz。系统是可变的，分解带不需要是一个八度，甚至可以定义在任何位置。在我们测试的音乐样本中最低的基础频率低音部大约 30Hz。46ms 的分解窗不能可靠的检测低音频率。因为低音部的绝大多数能量在低频区，我们发现对频率从 0 到 200Hz 用 80ms 的分解窗是足够的。46ms 的分解窗用在从 200Hz 到 5KHz。5KHz 以上声音特性是和低音部分有很大不同。46ms 的

分解窗也用在这个频率上，但是分解算法的参数略有不同。为了使下一步的分解更容易，不同频带内的所有窗口定位在同一时间。因为帧率是常数，所以长窗在低频叠加超过短窗在高频叠加。

谐音分音线性间隔的特性不适宜使用频率缩放的对数，基频 50Hz 的声音周期是 20ms，它的第十个谐音分音频率 500Hz 周期 2ms，而相邻谐音间距离还是 50Hz。因为用窗口长度区分两个正弦不仅仅靠正弦的频率，还要看频率之间的差别，我们需要一个长的分解窗处理低音的高谐音分量。在多音信号中，谐音分音的数量比中频时要大，因此即使正弦波长很小也需要长窗。

虽然谐音分音的线性间隔不适宜使用多分辨率分解，我们发现不同频率带用不同长度的窗还是有优势的。因为声音的属性是在不同的频带内是不同的，可以在每个频带内使用最优的检测算法。我们系统的灵活性使对不同频带使用不同的正弦检测算法成为可能，主要使用不同参数的互相关法和 F 测试法。

第四章 正弦延续和合成

如第三章中图 3－1 所示，正弦估计的最后一步是峰值延续分解。本章介绍两个峰值延续算法和正弦合成的原理，算法的性能和实验结果将在第 6 章给出。

当有意义的正弦峰值和它们的参数估计出来后，峰值轨迹连接形成帧间轨迹。在每一帧中，峰值延续算法连接正弦峰值到前面已经存在的帧轨迹里，形成一条光滑的频率和幅度曲线。我们用两个算法测试延续：一个是用正弦参数获得光滑轨道；另一个是在特定的偏差限内合成正弦延续并和原始信号做比较。也有其它系统使用开根的方法，例如隐马尔可夫模型[27]，但是我们这里并没有测试。

4．1 基于导数的延续

我们对频率和幅度求导获得光滑的轨迹，每对峰值的光滑系数由参数的一阶或二阶导数加权和计算得到，算法假定参数是缓慢时变的并且轨道彼此不交叉。

因为人类对音高的感知接近大多数听觉范围的对数，同时大多数乐器产生的基频在对数上分离，所以我们对频率取对数。由于对数相减等于除法的对数，描述频率光滑的因数变成频率比率的对数：

$$\log\left(\omega_{n-1}(i)\right)-\log\left(\omega_n(j)\right)=\log\left(\frac{\omega_{n-1}(i)}{\omega_n(j)}\right)$$

幅度差异的感知同样更具有对数特性，并且也可以做相同的处理。

因为频率是相位的导数，相位的光滑度也依靠频率。我们利用在正弦合成中用的插值系数 α 和 β 来估计相位光滑度（见 4.4 章）。为了避免产生偏差，所以对所有的因子取绝对值。如果只用一阶导数，在帧 $n-1$ 和 n 之间系数 i 和 j 之间的光滑系数是：

$$s_n(i,j)=w_f\left|\log\left(\frac{\omega_{n-1}(i)}{\omega_n(j)}\right)\right|+\omega_a\left|\log\left(\frac{a_{n-1}(i)}{a_n(i)}\right)\right|+\omega_\alpha\left|\alpha_n(i,j)\right|+w_\beta\left|\beta_n(i,j)\right| \qquad (404-1)$$

其中 $\omega_n(i)$ 和 $a_n(i)$ 是在帧 n 的第 i 峰值的频率和幅度，$\alpha_n(i,j)$ 和 $\beta_n(i,j)$ 是峰值间的相位插值系数，并且 w_a, w_f, w_a 和 w_β 是加权的。对频率和幅度差设最大限制是有好处的，这样可以限制可能出现的轨道峰对的数量。

一般情况下，因为组合的数目非常大，虽然设置了参数偏差的最大限制，估计相邻峰值之间所有可能的组合也是不可能的。我们用贪婪算法估计所有信号轨迹峰对的光滑度，然后选择最光滑的延续，例如有最小值 $s_n(i,j)$ 的。于是，去掉最光滑的峰值 i 和 j 最光滑的延续，并且对余下的峰算法不断重复。对某些人工生成的测试

第四章 正弦延续和合成

如第三章中图 3－1 所示，正弦估计的最后一步是峰值延续分解。本章介绍两个峰值延续算法和正弦合成的原理，算法的性能和实验结果将在第 6 章给出。

当有意义的正弦峰值和它们的参数估计出来后，峰值轨迹连接形成帧间轨迹。在每一帧中，峰值延续算法连接正弦峰值到前面已经存在的帧轨迹里，形成一条光滑的频率和幅度曲线。我们用两个算法测试延续：一个是用正弦参数获得光滑轨道；另一个是在特定的偏差限内合成正弦延续并和原始信号做比较。也有其它系统使用开根的方法，例如隐马尔可夫模型[27]，但是我们这里并没有测试。

4．1 基于导数的延续

我们对频率和幅度求导获得光滑的轨迹，每对峰值的光滑系数由参数的一阶或二阶导数加权和计算得到，算法假定参数是缓慢时变的并且轨道彼此不交叉。

因为人类对音高的感知接近大多数听觉范围的对数，同时大多数乐器产生的基频在对数上分离，所以我们对频率取对数。由于对数相减等于除法的对数，描述频率光滑的因数变成频率比率的对数：

$$\log\left(\omega_{n-1}(i)\right)-\log\left(\omega_n(j)\right)=\log\left(\frac{\omega_{n-1}(i)}{\omega_n(j)}\right)$$

幅度差异的感知同样更具有对数特性，并且也可以做相同的处理。

因为频率是相位的导数，相位的光滑度也依靠频率。我们利用在正弦合成中用的插值系数 α 和 β 来估计相位光滑度（见 4.4 章）。为了避免产生偏差，所以对所有的因子取绝对值。如果只用一阶导数，在帧 $n-1$ 和 n 之间系数 i 和 j 之间的光滑系数是：

$$s_n(i,j)=w_f\left|\log\left(\frac{\omega_{n-1}(i)}{\omega_n(j)}\right)\right|+\omega_a\left|\log\left(\frac{a_{n-1}(i)}{a_n(i)}\right)\right|+\omega_\alpha\left|\alpha_n(i,j)\right|+w_\beta\left|\beta_n(i,j)\right| \qquad (404-1)$$

其中 $\omega_n(i)$ 和 $a_n(i)$ 是在帧 n 的第 i 峰值的频率和幅度，$\alpha_n(i,j)$ 和 $\beta_n(i,j)$ 是峰值间的相位插值系数，并且 w_a, w_f, w_a 和 w_β 是加权的。对频率和幅度差设最大限制是有好处的，这样可以限制可能出现的轨道峰对的数量。

一般情况下，因为组合的数目非常大，虽然设置了参数偏差的最大限制，估计相邻峰值之间所有可能的组合也是不可能的。我们用贪婪算法估计所有信号轨迹峰对的光滑度，然后选择最光滑的延续，例如有最小值 $s_n(i,j)$ 的。于是，去掉最光滑的峰值 i 和 j 最光滑的延续，并且对余下的峰算法不断重复。对某些人工生成的测试

信号这些算法可能产生错误的结果，但是对于自然的语音信号工作的确很好。

如果不能发现峰值的合适的延续，那就意味着产生那个频率分量的声音已经减弱并且轨迹逐渐消失。对任意已经存在的轨迹的延续如果当前帧的波峰没有重构，这就意味着产生了一个新的分量开端和一个新的轨迹，如图 4—1 所示[28]。

计算峰值延续以后，我们得到一组用时变幅度、频率和相位表示的正弦轨迹。每个轨迹有开始时间和偏移时间，时间范围在轨迹存在的范围内。要获得一个从光滑过渡到零水平的曲线，加入一个额外的波峰到轨道的开始和结尾处。这个额外的波峰有和前后轨迹波峰相同的频率，但是幅度是零。这样就确保开端或者偏移不会产生模型假象。

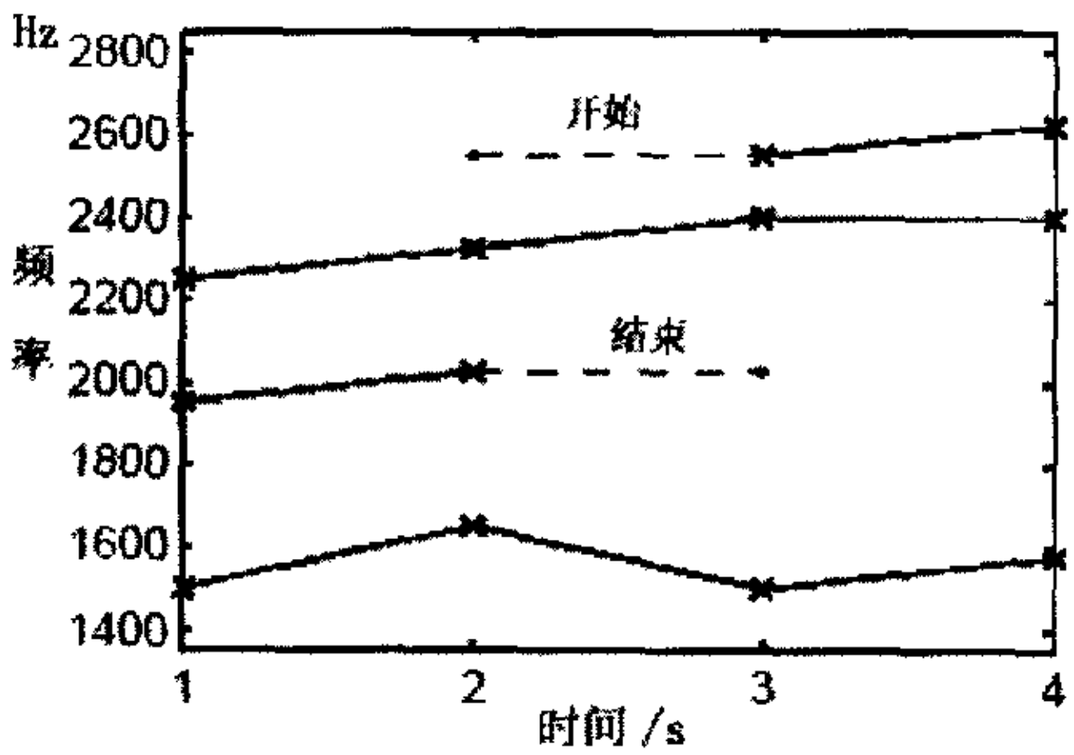

图 4—1 正弦轨迹延续

4．2 基于合成的延续

仅仅基于幅度和频率的偏差及相位插值的延续不是非常稳定的方法。这是因为以下原因：在噪音中检测低幅度谐音分量，峰值检测域必需设的非常低。自然的，这意味着我们也得到很多由噪音引起的波峰。虽然我们对幅度和频率的偏差做了严格的限制，但是由噪声引起的峰值彼此很接近，它们也会进入正弦轨道。

我们解决这个问题的办法是合成所有的在偏差限制内的延续，并和原始信号做比较，正弦合成在 4.4 章介绍。如果一个合成的正弦捕捉到了足够的原始信号能量，我们设想正弦符合原始信号中真实存在的分量，我们用贪婪算法一直捕捉延续使剩余的能量最小化，然后，从原始信号中去掉合成的正弦，并把残余信号与剩下信号的合成延续的可能结果做比较。重复这种方法，直到没有足够的残余合成延续可以减少了，整个处理过程在时域中完成。

这个算法比仅仅基于幅度和相位偏差的延续更加具有可靠性。当然，如果它的延续碰巧和正弦信号匹配的很好，还会有一些噪音波峰出现，但是噪音延续的数量比简单算法要少得多。合成延续的缺点是计算复杂度，比起只用参数偏差算法计算

信号这些算法可能产生错误的结果，但是对于自然的语音信号工作的确很好。

如果不能发现峰值的合适的延续，那就意味着产生那个频率分量的声音已经减弱并且轨迹逐渐消失。对任意已经存在的轨迹的延续如果当前帧的波峰没有重构，这就意味着产生了一个新的分量开端和一个新的轨迹，如图 4—1 所示[28]。

计算峰值延续以后，我们得到一组用时变幅度、频率和相位表示的正弦轨迹。每个轨迹有开始时间和偏移时间，时间范围在轨迹存在的范围内。要获得一个从光滑过渡到零水平的曲线，加入一个额外的波峰到轨道的开始和结尾处。这个额外的波峰有和前后轨迹波峰相同的频率，但是幅度是零。这样就确保开端或者偏移不会产生模型假象。

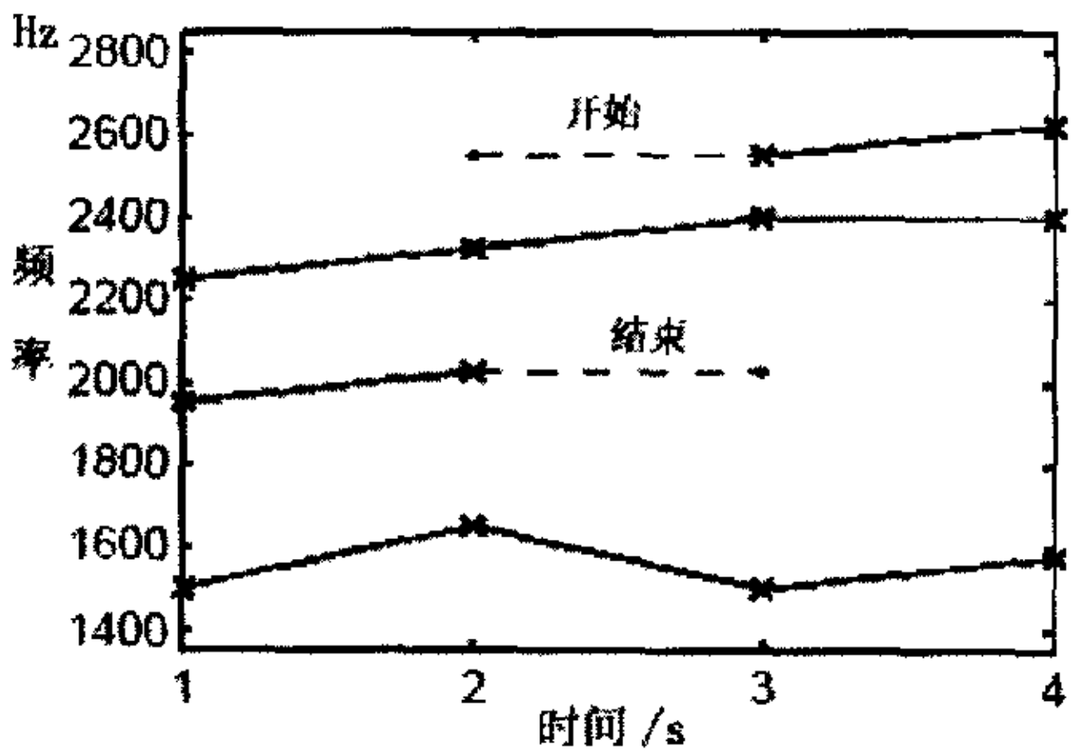

图 4—1 正弦轨迹延续

4．2 基于合成的延续

仅仅基于幅度和频率的偏差及相位插值的延续不是非常稳定的方法。这是因为以下原因：在噪音中检测低幅度谐音分量，峰值检测域必需设的非常低。自然的，这意味着我们也得到很多由噪音引起的波峰。虽然我们对幅度和频率的偏差做了严格的限制，但是由噪声引起的峰值彼此很接近，它们也会进入正弦轨道。

我们解决这个问题的办法是合成所有的在偏差限制内的延续，并和原始信号做比较，正弦合成在 4.4 章介绍。如果一个合成的正弦捕捉到了足够的原始信号能量，我们设想正弦符合原始信号中真实存在的分量，我们用贪婪算法一直捕捉延续使剩余的能量最小化，然后，从原始信号中去掉合成的正弦，并把残余信号与剩下信号的合成延续的可能结果做比较。重复这种方法，直到没有足够的残余合成延续可以减少了，整个处理过程在时域中完成。

这个算法比仅仅基于幅度和相位偏差的延续更加具有可靠性。当然，如果它的延续碰巧和正弦信号匹配的很好，还会有一些噪音波峰出现，但是噪音延续的数量比简单算法要少得多。合成延续的缺点是计算复杂度，比起只用参数偏差算法计算

量是巨大的。我们用三次多项式来插值相位和线性插值幅度，所以合成所有的时域正弦是计算量非常大的。一个合成正弦的 DFT 可以近似为用一系列推导的样本，但是在频域中做整个的处理并没有太大的帮助，因为我们必需计算每个残余信号的剩余能量，不管在时域还是频域，在正弦数量很大时这将需要大量的计算。

4.3 轨迹滤波器

我们从已有先验知识得知，人类的听觉系统有一个特性即在一个大的声音出现的时候，弱的声音会被掩盖而听不见，这个效果叫做掩蔽[30]，一个声音被另一个叫掩蔽音的声音掩蔽。换句话说就是掩蔽可以定义为提高听觉域的过程，如果一个声音在域值以下就听不到了。掩蔽在时域和频域都有发生，分别叫做同步掩蔽和非同步掩蔽。在频域掩蔽，频率上彼此越接近的同步声音，掩蔽效应越强。在时域，大的声音能够掩蔽小的声音，小的声音出现在掩蔽音之后，甚至之前都能被掩蔽。

掩蔽效应能去掉被掩蔽的分量，它可利用在语音编码上，也可以用在正弦模型上，另外听觉域可以用来判断哪个正弦峰或轨迹是由噪声所引起。我们在前面提到，并不是所有的正弦峰都是稳定正弦的产物，因为正弦峰的数量在多音信号中非常大，很有可能存在一些和原始信号匹配很好的假峰，并进一步变成延续进入轨道。这些假轨迹通常很短，只有两个帧长。如果正弦是很短，人类的听觉系统可能不能迅速测定正弦音高，特别是在正弦幅度是很小并且伴有其他信号分量出现时。因此，这里不需要用正弦模型，但是可以用随机模型来模拟残余信号。如果正弦峰明显在听觉域或掩蔽域以下，则这些分量可能不会产生正弦。

S.Levine 用一种方法，就是用平均距离测量隐蔽域和用每个正弦的长度来决定正弦是保留还是滤掉[11]。平均信号掩蔽率是使用正弦的幅度和每一帧计算出来的掩蔽域比较计算出来的。如果 $SMR(i) < 6 - 96 \cdot len(i)$，正弦 i 可以去掉，其中 $SMR(i)$ 是正弦 i 在 0dB 时平均信号掩蔽率，$len(i)$ 是正弦的毫秒级的长度。这意味着一个短的正弦要求一个大的 SMR 不被滤掉，越长的正弦要求越小的 SMR，并且一直保留正弦。

我们系统计算掩蔽域的方法就像 MPEG－2[30]，对每个正弦，在频域计算一个激励模式，它的分辨率是1/25BARK，在 0 到 22.5kHz 之间分了 620 个频带。因此，正弦的能量用一个函数展开分布在 620 频带，在 BARK 域是三角形的，但在频域不是均衡的。所有正弦的激励样本是指数法则的组合：

$$e = \left(\sum_i e(i)^{\alpha} \right)^{1/\alpha},$$

这里 $e(i)$ 是第 i 分量的激励，α 在 1 和 2 之间，我们取 1.5。我们的系统不用异步掩蔽。

量是巨大的。我们用三次多项式来插值相位和线性插值幅度，所以合成所有的时域正弦是计算量非常大的。一个合成正弦的 DFT 可以近似为用一系列推导的样本，但是在频域中做整个的处理并没有太大的帮助，因为我们必需计算每个残余信号的剩余能量，不管在时域还是频域，在正弦数量很大时这将需要大量的计算。

4.3 轨迹滤波器

我们从已有先验知识得知，人类的听觉系统有一个特性即在一个大的声音出现的时候，弱的声音会被掩盖而听不见，这个效果叫做掩蔽[30]，一个声音被另一个叫掩蔽音的声音掩蔽。换句话说就是掩蔽可以定义为提高听觉域的过程，如果一个声音在域值以下就听不到了。掩蔽在时域和频域都有发生，分别叫做同步掩蔽和非同步掩蔽。在频域掩蔽，频率上彼此越接近的同步声音，掩蔽效应越强。在时域，大的声音能够掩蔽小的声音，小的声音出现在掩蔽音之后，甚至之前都能被掩蔽。

掩蔽效应能去掉被掩蔽的分量，它可利用在语音编码上，也可以用在正弦模型上，另外听觉域可以用来判断哪个正弦峰或轨迹是由噪声所引起。我们在前面提到，并不是所有的正弦峰都是稳定正弦的产物，因为正弦峰的数量在多音信号中非常大，很有可能存在一些和原始信号匹配很好的假峰，并进一步变成延续进入轨道。这些假轨迹通常很短，只有两个帧长。如果正弦是很短，人类的听觉系统可能不能迅速测定正弦音高，特别是在正弦幅度是很小并且伴有其他信号分量出现时。因此，这里不需要用正弦模型，但是可以用随机模型来模拟残余信号。如果正弦峰明显在听觉域或掩蔽域以下，则这些分量可能不会产生正弦。

S.Levine 用一种方法，就是用平均距离测量隐蔽域和用每个正弦的长度来决定正弦是保留还是滤掉[11]。平均信号掩蔽率是使用正弦的幅度和每一帧计算出来的掩蔽域比较计算出来的。如果 $SMR(i) < 6 - 96 \cdot len(i)$，正弦 i 可以去掉，其中 $SMR(i)$ 是正弦 i 在 0dB 时平均信号掩蔽率，$len(i)$ 是正弦的毫秒级的长度。这意味着一个短的正弦要求一个大的 SMR 不被滤掉，越长的正弦要求越小的 SMR，并且一直保留正弦。

我们系统计算掩蔽域的方法就像 MPEG－2[30]，对每个正弦，在频域计算一个激励模式，它的分辨率是1/25BARK，在 0 到 22.5kHz 之间分了 620 个频带。因此，正弦的能量用一个函数展开分布在 620 频带，在 BARK 域是三角形的，但在频域不是均衡的。所有正弦的激励样本是指数法则的组合：

$$e = \left(\sum_i e(i)^{\alpha} \right)^{1/\alpha},$$

这里 $e(i)$ 是第 i 分量的激励，α 在 1 和 2 之间，我们取 1.5。我们的系统不用异步掩蔽。

4.4 合成

正弦轨迹包含所有的输入信号谐音部分需要重构的信息：在每帧中每个轨迹的幅度、频率和相位。为避免帧边界的中断，幅度、频率和相位在帧与帧之间用内插值替换，幅度用线性插值，所以在帧n中的轨道p_l的瞬时幅度是：

$$a_{l,n}(m)=a_{l,n}+\left(a_{l,n+1}-a_{l,n}\right)\frac{m}{S}\quad m=0,1,...,S-1 \tag{4-2}$$

其中 S 是样本中帧的长度。

相位的插值更加复杂，因为瞬时频率是相位的导数并且必需要计算 4 个参数（两个相邻帧的频率和相位）。光滑的相位作为一个时间函数用三次多项插值方程计算得到[8]：

$$\theta(t)=\xi+\gamma t+\alpha t^2+\beta t^3 \tag{4-3}$$

其中$\theta(t)$是瞬时相位在时间t的插值，ξ、γ和β是插值系数。设瞬时相位和频率在点$t=0$和$t=S$等于已知频率和相位$\omega_n,\omega_{n+1},\theta_n$和$\theta_{n+1}$，我们得到三次多项式的解是：

$\xi=\theta_n$，

$\gamma=\omega_n$，和

$$\begin{bmatrix}\alpha(M)\\ \beta(M)\end{bmatrix}=\begin{bmatrix}\frac{3}{S^2} & -\frac{1}{S}\\ -\frac{2}{S^3} & \frac{1}{S^2}\end{bmatrix}\begin{bmatrix}\theta_{n+1}-\theta_n-\omega_n S+2\pi M\\ \theta_{n+1}\theta_n\end{bmatrix} \tag{4-4}$$

对于任意整数M，最大的光滑相位或者相位的二阶导数最小，用以下公式获得：

$$M=round\left(\frac{1}{2}\left[\left(\theta_n+\omega_n S-\theta_{n+1}\right)+\left(\omega_{n+1}-\omega_n\right)\frac{S}{2}\right]\right) \tag{4-5}$$

当一帧中每一时间的所有正弦轨迹的瞬时幅度和相位计算出来以后，求所有轨迹的和得到重构的正弦：

$$s(t)=\sum_{i=1}^{N}a_i(t)\cos(\theta_i) \tag{4-6}$$

如果谐音分量变化缓慢，甚至在单帧内几乎是稳定的，使用参数的插值能够改善效果。尽管频率非常小，二次迭代的效果还是很好，在有尖声的时候，幅度值的线性插值不适合谐音分量的实际幅度。这是因为要区分相邻频率相对需要一个长窗并且只有一个值能表示窗内的正弦幅度，所以我们不能得到精确的幅度。有些方法能在一个单窗内提取参数变化的更多的信息[16]，例如幅度和频率调制，但是对于真实的音乐信号，这种方法在有干扰声音出现的时候不够稳定。

相位在感知上不是非常重要，所以在声音编码的应用中我们不需要变换它。在

解码器中每个正弦轨迹能产生一个随机的初始相位，其它的相位是频率的积分：

$$\theta_{n+1} = \theta_n + \omega_{n+1} S \tag{4-7}$$

如果用无相重构，合成信号不再是原始信号的相位排列。如果我们想得到残余信号，在从原始信号中去掉正弦信号部分时必需使用相位信息，之后，相位信息可以忽略。

第五章 随机模型

在时域中，当合成的正弦从原始信号中移除后，我们得到残余信号，理想情况下只包含非谐音分量。与处理确定的信号相比，分解和合成随机信号分量明显要容易，因为人类单耳的听觉感知对相位不敏感，时变的谱形状是唯一需要重构的残余的信息。在生理声学的实验中，发现人耳对固定的 BARK 带内的能量变化不敏感，如类似噪声信号。在 0 道 20kHz 之内有 25 个 BARK 带，或者临界带，并不是线性间隔的，假定残余信号类似噪声，可以通过在每个 BARK 带内计算短时能量来建模。

5.1 分解

随机分解过程如图 5－1 所示。首先将残余分割到帧里，并对每帧做 STFT，能量谱是 STFT 的绝对值的平方。然后，通过对 BARK 带内的能量谱积分计算在每个 BARK 带内的能量。

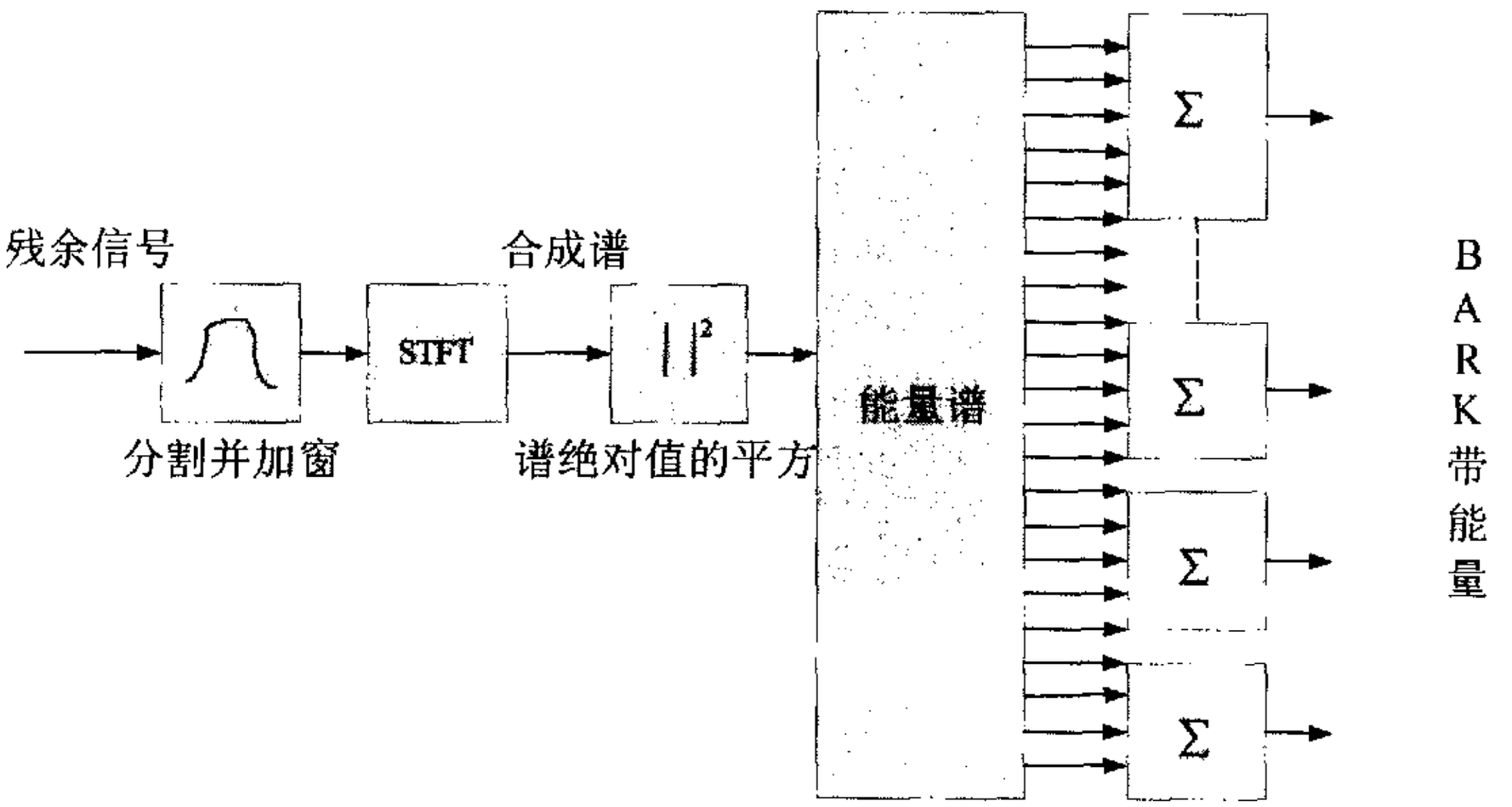

图 5－1 随机分解框图

用 $r(n)$ 和它在频率 ω 和时间 t 内的 STFT $R(\omega,t)$ 表示残余信号，$r(n)$ 短时能量谱是 $|R(\omega,t)|^2$，BARK 带 z 对应频率 f 近似为[31]：

$$z(f)=13a\tan(0.00076f)+3.5a\tan\left[\left(\frac{f}{7500}\right)^2\right] \tag{475－1}$$

角频率和频率之间的关系如下：

$$\omega=2\pi\frac{f}{F_s} \tag{5－2}$$

对每个 BARK 带，分别计算带内短时能量。对带 b 的短时能量是：

第五章 随机模型

在时域中，当合成的正弦从原始信号中移除后，我们得到残余信号，理想情况下只包含非谐音分量。与处理确定的信号相比，分解和合成随机信号分量明显要容易，因为人类单耳的听觉感知对相位不敏感，时变的谱形状是唯一需要重构的残余的信息。在生理声学的实验中，发现人耳对固定的 BARK 带内的能量变化不敏感，如类似噪声信号。在 0 道 20kHz 之内有 25 个 BARK 带，或者临界带，并不是线性间隔的，假定残余信号类似噪声，可以通过在每个 BARK 带内计算短时能量来建模。

5.1 分解

随机分解过程如图 5－1 所示。首先将残余分割到帧里，并对每帧做 STFT，能量谱是 STFT 的绝对值的平方。然后，通过对 BARK 带内的能量谱积分计算在每个 BARK 带内的能量。

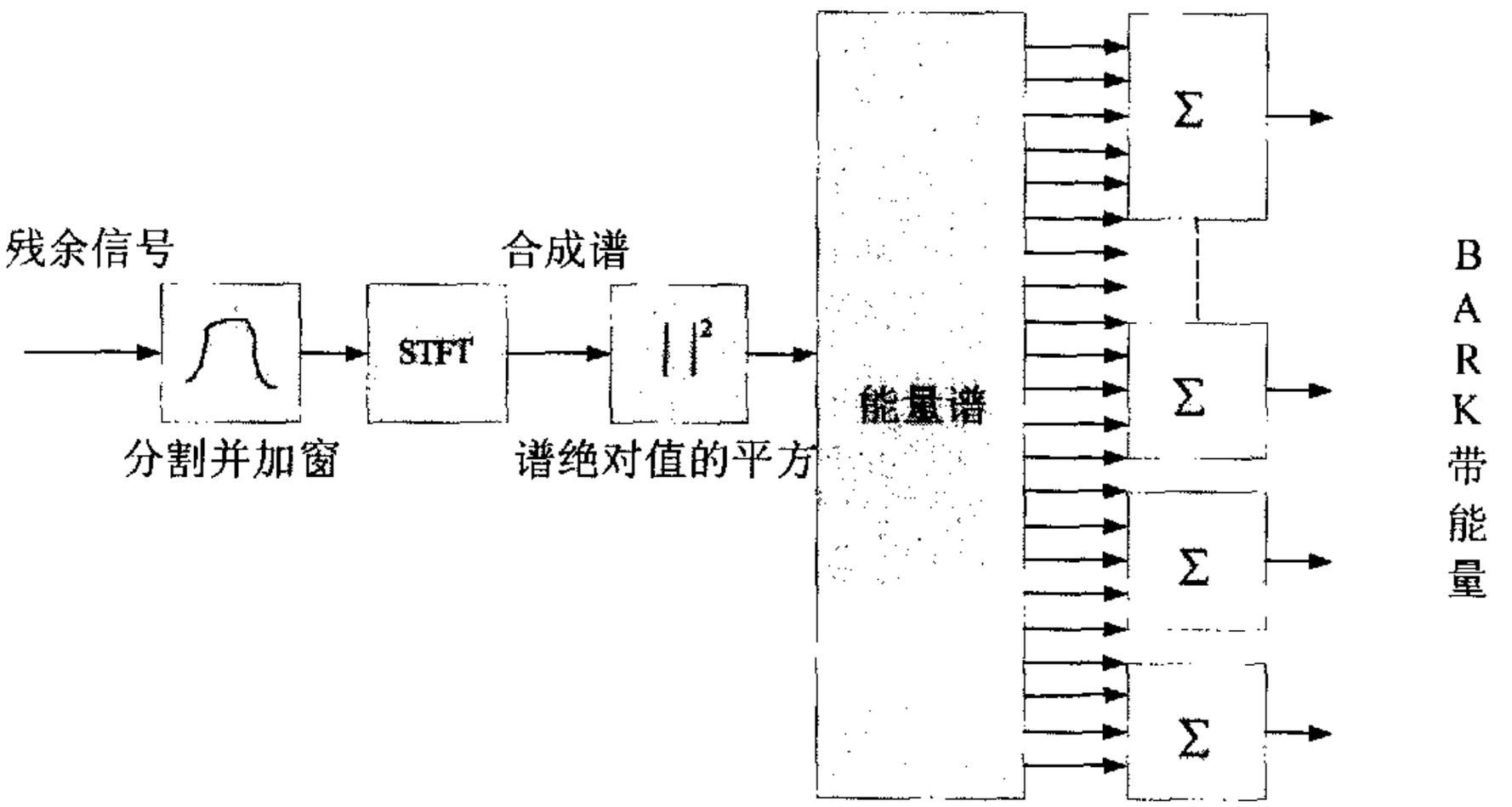

图 5－1 随机分解框图

用 $r(n)$ 和它在频率 ω 和时间 t 内的 STFT $R(\omega,t)$ 表示残余信号，$r(n)$ 短时能量谱是 $|R(\omega,t)|^2$，BARK 带 z 对应频率 f 近似为[31]：

$$z(f)=13a\tan(0.00076f)+3.5a\tan\left[\left(\frac{f}{7500}\right)^2\right] \tag{475－1}$$

角频率和频率之间的关系如下：

$$\omega=2\pi\frac{f}{F_s} \tag{5－2}$$

对每个 BARK 带，分别计算带内短时能量。对带 b 的短时能量是：

$$E(b,t)=\frac{1}{M}\sum_{z(\omega)\approx b}\left|R(\omega,t)\right|^2 \tag{5-3}$$

其中 M 是 STFT 的长度，能够重构残余信号的信息是短时能量和帧率。

5.2 合成

在随机合成里，用分段的 BARK 带的绝对值和随机相位来构建复杂的短时谱。合成处理过程见图 5－2。谱的绝对值等于用每个 BARK 带的能量除相应的带宽取平方根：

$$\left|S(\omega,t)\right|=\sqrt{\frac{E(b,t)}{\beta_b}} \tag{5-4}$$

这里 β_b 是在合成谱 $S(\omega,t)$ 的样本中的带 b 的带宽，在分解阶段也可以除 β_b，于是我

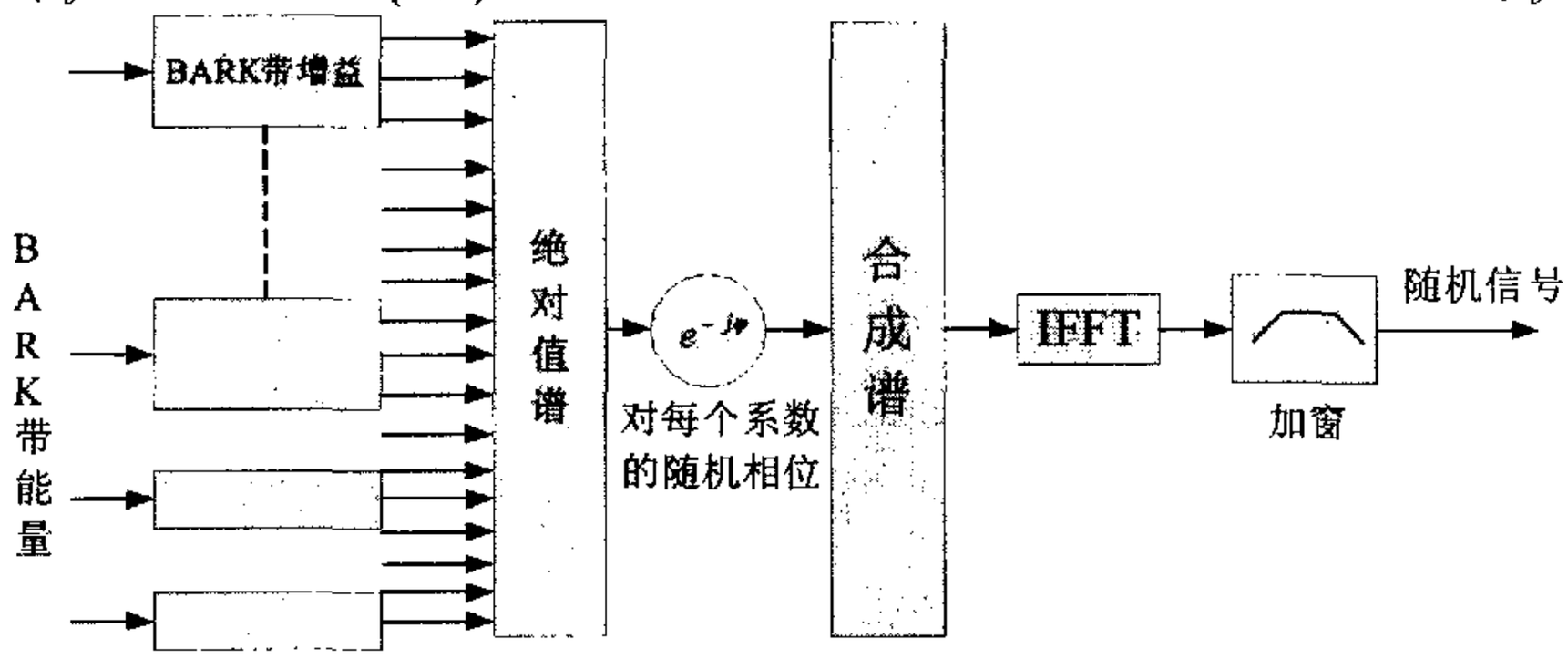

图 5－2 随机合成框图

们不计算在每个带内的能量而是计算每个带内的平均能量谱系数。为了消除带边界的尖峰，使谱能光滑一点，但通常这不是必需的，因为时域窗叠加相位可以引起频域的光滑。

对相位创造一个随机向量产生随机谱，随机相位向量 $\varphi(\omega)$ 一律是分布在区间 $[-\pi,\pi]$ 内，复杂谱是谱的绝对值和随机相位的乘积：

$$S(\omega,t)=\left|S(\omega,t)\right|e^{i\varphi(\omega)} \tag{5-5}$$

通过对每个短时复杂谱取反转的 STFT 来获得随机信号。为防止在帧的边界部分出现咯嚓声，可以使用加窗和叠加的办法，在考虑叠加相邻帧时窗口函数要选能概括总体的。特殊音乐样本的残余信号的 BARK 带能量见图 5－3，鼓声占据了绝大部分残余信号：一般的低音和响鼓能从能量中重构。

与正弦的分解和合成比较起来，随机信号部分的处理明显简单很多，基本上能调整的随机分解的参数只有窗长度和帧率。显然，多分辨率分解需要使用几个不同的窗。我们的一些经验是用谱的绝对值的非线性滤波去掉残余信号可能保留的谐音

$$E(b,t)=\frac{1}{M}\sum_{z(\omega)\approx b}\left|R(\omega,t)\right|^2 \tag{5-3}$$

其中 M 是 STFT 的长度，能够重构残余信号的信息是短时能量和帧率。

5.2 合成

在随机合成里，用分段的 BARK 带的绝对值和随机相位来构建复杂的短时谱。合成处理过程见图 5－2。谱的绝对值等于用每个 BARK 带的能量除相应的带宽取平方根：

$$\left|S(\omega,t)\right|=\sqrt{\frac{E(b,t)}{\beta_b}} \tag{5-4}$$

这里 β_b 是在合成谱 $S(\omega,t)$ 的样本中的带 b 的带宽，在分解阶段也可以除 β_b，于是我

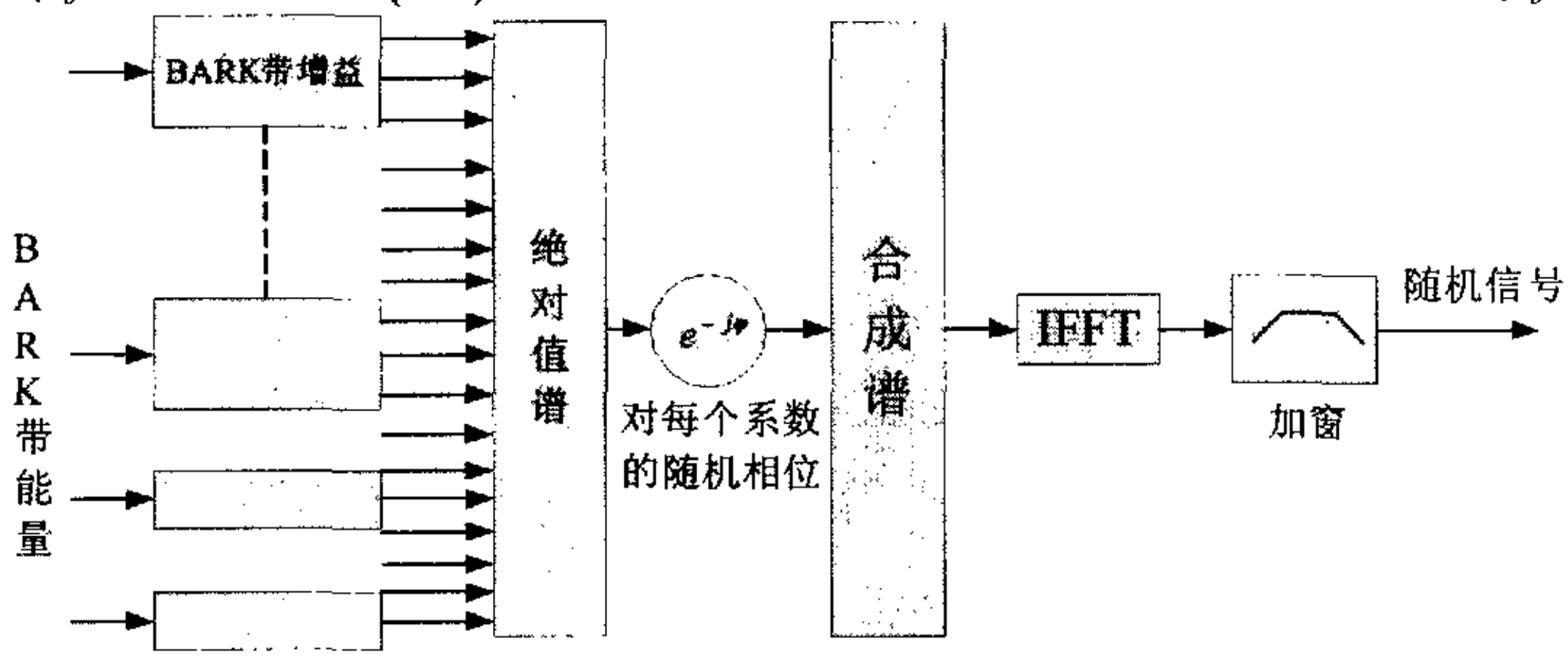

图 5－2 随机合成框图

们不计算在每个带内的能量而是计算每个带内的平均能量谱系数。为了消除带边界的尖峰，使谱能光滑一点，但通常这不是必需的，因为时域窗叠加相位可以引起频域的光滑。

对相位创造一个随机向量产生随机谱，随机相位向量 $\varphi(\omega)$ 一律是分布在区间 $[-\pi,\pi]$ 内，复杂谱是谱的绝对值和随机相位的乘积：

$$S(\omega,t)=\left|S(\omega,t)\right|e^{i\varphi(\omega)} \tag{5-5}$$

通过对每个短时复杂谱取反转的 STFT 来获得随机信号。为防止在帧的边界部分出现咯嚓声，可以使用加窗和叠加的办法，在考虑叠加相邻帧时窗口函数要选能概括总体的。特殊音乐样本的残余信号的 BARK 带能量见图 5－3，鼓声占据了绝大部分残余信号：一般的低音和响鼓能从能量中重构。

与正弦的分解和合成比较起来，随机信号部分的处理明显简单很多，基本上能调整的随机分解的参数只有窗长度和帧率。显然，多分辨率分解需要使用几个不同的窗。我们的一些经验是用谱的绝对值的非线性滤波去掉残余信号可能保留的谐音

分量，但看起来这种方法并不能得到很好的效果。

在正弦部分和随机部分合成以后，能在时域中线性叠加得到完整的重构信号。有的系统中两部分信号都是在时域中合成：用 BARK 带能量产生随机谱，并且将正弦增加到谱中。这种方法不能使用相位的二次插值。

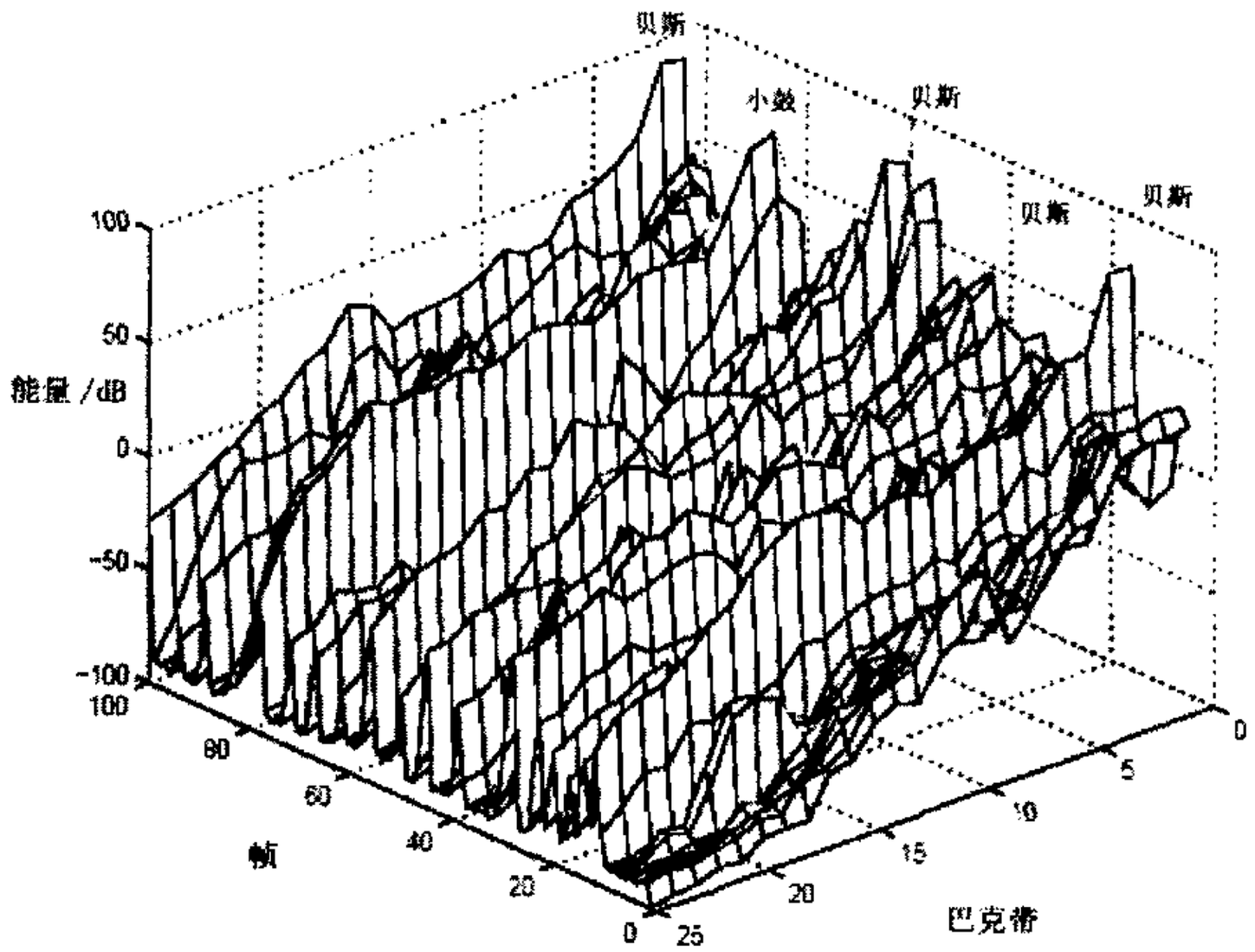

图 5—3 样本音乐信号的 BARK 带能量

第六章 实验结果

在复杂的真实信号中，正弦分量的密度非常高，目前还没有十分精确的方法能测量正弦加噪系统的性能。在系统的实现过程中，使用视觉和声觉评定法描绘正弦峰以及它们的参数，同时通过听正弦信号和残余信号获取每个算法的轨迹。这个信息用数字上的或是用口头上的方法来评估是非常困难的，因为分解算法之间的差异几乎听不出来，因此我们通过计算从一组音乐样本和从人工合成的测试信号中得到的分解和合成结果的统计量来研究分解算法的性能。

因为峰值检测是分解系统十分重要的部分，所以大部分的测试就是比较峰值检测算法。系统中另外一个重要的参数是窗的长度，它是时间和频率分辨率的折中。不同窗长在实现中的影响研究并不在我们的研究范围，但是为了得到最佳的比较效果，所有的算法都使用相同长度的窗。

6.1 音乐信号的峰值检测算法的比较

通常在一个单时间帧内评估峰值检测算法的性能是非常困难的，因此需要用一个延续算法连接峰值到正弦轨迹，而且技术性能分析基于轨迹数据。估计算法产生的大多数假峰在延续阶段去掉了，峰值检测以后，更多的注意未检测的谐音分量，也是因为正弦分解的较后阶段检测丢失的分量十分困难。

对音乐信号测试两个最好的峰值检测方是 F 测试法和互相关法，在表 6—1 中列出了五段音乐的 10 到 20 秒的片断。使用 3 个参数组比较这两个方法：第一组是在实现和算法测试过程中对音乐信号用手工调整使参数达到最好，第二组是比第一组挑出更多的峰，第三组是挑出比第一组较少的峰。用得到的最好方法计算正弦参数，频率使用二次插值法，幅度和相位用最小平方法。延续是基于合成正弦的比较，用互相关算法使用正常参数从“十年”中得到正弦轨迹的频率，见图 6—1。当正弦轨迹用两种算法和参数组得到后，可以合成正弦部分，并且由原始信号减去合成的正弦也得到了残余信号。信号残余率（SSR）是信号能量的比率，SSR 能测量有多少正弦从信号中移掉，还能测量信号全面的特征：如果在信号里有大量的非谐音分量如鼓声，即使正弦分解很好 SSR 还是会很低。总体上，不管正弦是否是正确的还是有不正确的模型噪音，只要检测到的正弦越多，SRR 就会更好。因此，SRR 不能单独测量正弦分解的质量。

检验分解的质量可以用原始信号和合成的残余信号相比的方法，在每个 BARK 带内计算短时能量合成残余信号，然后用通常的方法处理随机合成。如果从残余信号中去掉所有的正弦信号，它的幅度谱应该是光滑的，因此原始的残余信号和合成

第六章 实验结果

在复杂的真实信号中，正弦分量的密度非常高，目前还没有十分精确的方法能测量正弦加噪系统的性能。在系统的实现过程中，使用视觉和声觉评定法描绘正弦峰以及它们的参数，同时通过听正弦信号和残余信号获取每个算法的轨迹。这个信息用数字上的或是用口头上的方法来评估是非常困难的，因为分解算法之间的差异几乎听不出来，因此我们通过计算从一组音乐样本和从人工合成的测试信号中得到的分解和合成结果的统计量来研究分解算法的性能。

因为峰值检测是分解系统十分重要的部分，所以大部分的测试就是比较峰值检测算法。系统中另外一个重要的参数是窗的长度，它是时间和频率分辨率的折中。不同窗长在实现中的影响研究并不在我们的研究范围，但是为了得到最佳的比较效果，所有的算法都使用相同长度的窗。

6.1 音乐信号的峰值检测算法的比较

通常在一个单时间帧内评估峰值检测算法的性能是非常困难的，因此需要用一个延续算法连接峰值到正弦轨迹，而且技术性能分析基于轨迹数据。估计算法产生的大多数假峰在延续阶段去掉了，峰值检测以后，更多的注意未检测的谐音分量，也是因为正弦分解的较后阶段检测丢失的分量十分困难。

对音乐信号测试两个最好的峰值检测方是 F 测试法和互相关法，在表 6—1 中列出了五段音乐的 10 到 20 秒的片断。使用 3 个参数组比较这两个方法：第一组是在实现和算法测试过程中对音乐信号用手工调整使参数达到最好，第二组是比第一组挑出更多的峰，第三组是挑出比第一组较少的峰。用得到的最好方法计算正弦参数，频率使用二次插值法，幅度和相位用最小平方法。延续是基于合成正弦的比较，用互相关算法使用正常参数从“十年”中得到正弦轨迹的频率，见图 6—1。当正弦轨迹用两种算法和参数组得到后，可以合成正弦部分，并且由原始信号减去合成的正弦也得到了残余信号。信号残余率（SSR）是信号能量的比率，SSR 能测量有多少正弦从信号中移掉，还能测量信号全面的特征：如果在信号里有大量的非谐音分量如鼓声，即使正弦分解很好 SSR 还是会很低。总体上，不管正弦是否是正确的还是有不正确的模型噪音，只要检测到的正弦越多，SRR 就会更好。因此，SRR 不能单独测量正弦分解的质量。

检验分解的质量可以用原始信号和合成的残余信号相比的方法，在每个 BARK 带内计算短时能量合成残余信号，然后用通常的方法处理随机合成。如果从残余信号中去掉所有的正弦信号，它的幅度谱应该是光滑的，因此原始的残余信号和合成

残余信号的幅度谱彼此接近。原始残余信号和合成残余信号的短时幅度谱之间的均方差由在每帧中每个 BARK 带内计算得到，并遍及所有时间和所有 25 个 BARK 带。合成残余信号的谱是光滑的，所以产生的误差测量幅度谱的不规则性。因此，在合成的残余信号和原始残余信号之间的误差是留在残余信号中谐音分量数量的估计，并试图去掉所有类似噪音的分量，得到的不规则的谱和 SRR 见图 6－2。从图中看到，不同的音乐样本之间的差距大于由使用不同分解算法和使用不同参数组引起的差

表 6－1：音乐测试信号

音乐名	类型	乐器
十年	流行	男声、声学吉他、钢琴
愉快	拉丁	女声、贝斯、手风琴
大地	摇滚	男声、电吉他
春	经典	管弦乐
炫光	爵士	钢琴、电贝斯、鼓、电吉他、键盘

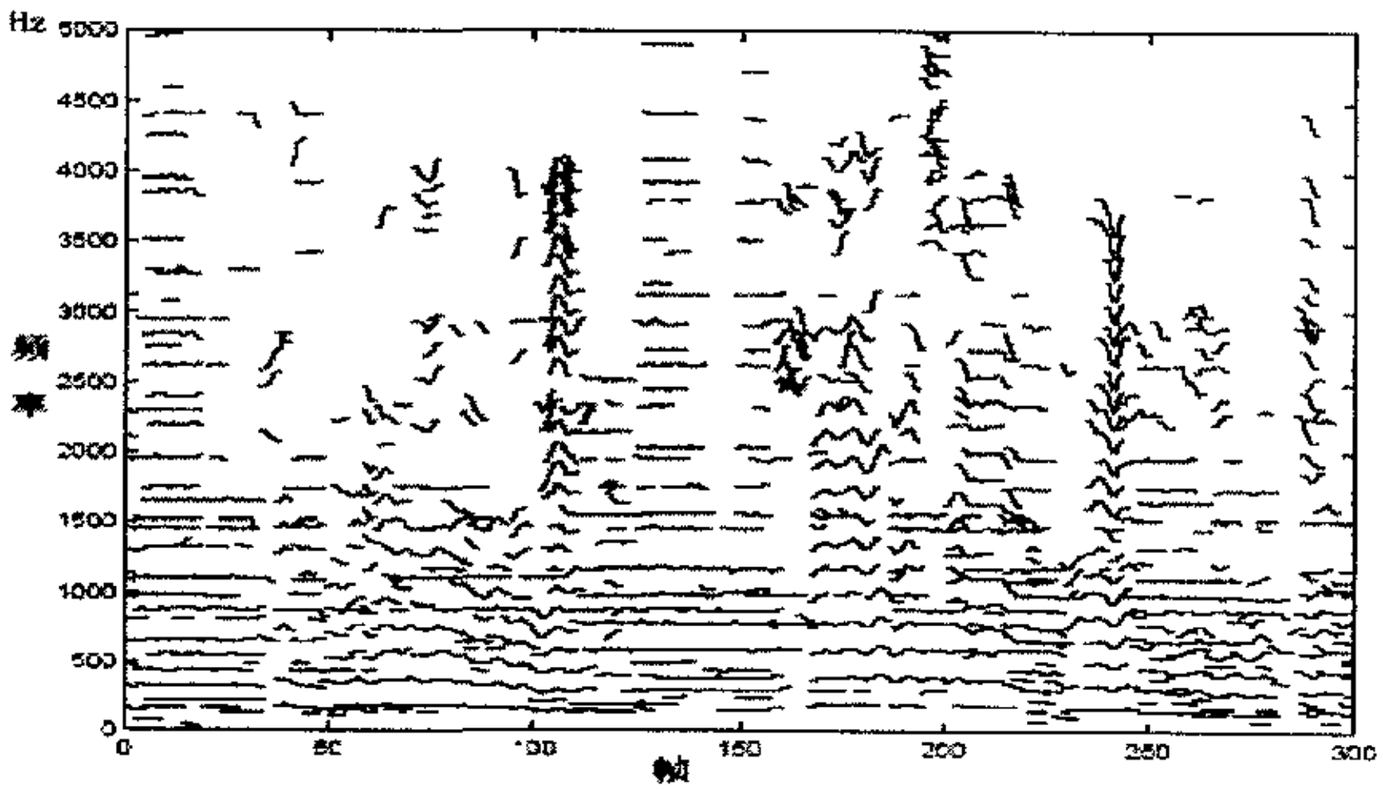

图 6－1 从样本中得到的正弦频率

距。我们试图用去掉每个信号平均值的方法比较不同算法和参数，结果显然是提取最大正弦数量的参数组 SRR 效果会更好，谱的不规则性变低。不同算法之间的差异依然很小，F 测试比互相关算法产生更多变的结果。

通过听合成信号和残余信号，并扫描得到的正弦轨迹的频率和幅度曲线的方法来研究结果时，算法之间的差异几乎听不出来，甚至听起来互相关方法比 F 测试法更好，特别是在有快速频率变化或有颤音的时候尤为明显，这已经用检验频率曲线的方法证实了。

6．2 用产生的测试信号比较正弦分解算法

正如本章开始所提到的，对于复杂真实信号，目前还没有一个好的数值标准来测量正弦分解算法的好坏，我们生成一个只包含正弦的测试信号试图来解决这个问题。测试信号会引入在音乐信号中经常遇到的现象：不同的幅度和频率的变化，正弦组成的谐音声音彼此叠加、碰撞等。生成的测试信号被分成 10 段，详见表 6－2。生成的测试信号在三个不同噪音条件下分解：一是没有附加白噪声，二是附加白噪声－14dB，三是附加白噪声＋6dB。参考水平 0dB 是一个单一幅度的单正弦，噪声水平覆盖 0-22kHz 整个频率范围。

表 6－2：生成的测试信号描述

段	信号描述。幅度为 0dB（除特殊说明外）
1	不同频率的稳定的正弦，在一个时刻的一个正弦
2	扫描正弦的频率从 20Hz 到 10kHz，速度是频率的指数
3	单正弦幅度的减少从 0dB 到－40dB
4	用不同幅度和频率调制的混和正弦。调制频率从 0 到 20Hz 变化，幅度从 0 到 1 变化，频差从 0 到 1.5 个半音变化
5	几个不同频率之间穿过两个正弦的频率
6	不同基频的稳定的合成声音。所有的声音有 10 个一次谐音分音，幅度一致
7	扫描一个谐音的频率，10 个谐音分音
8	谐音颤音，调制频率和颤音的深度同 4 段变化一致
9	不同种的干扰音，谐音在 100、200、400，...，3200，6400Hz
10	扫描一个谐音的频率，用一个常谐音混和

对于正弦分解的每一步有 2 到 3 种算法，并且每一步的表现都要受到前一步的影响。例如：如果检测到的峰值频率是错误的，就不可能包含正确的幅度和相位；有错误参数的峰值很容易持续出错，甚至参数估计和正确峰的延续都要受到假峰的影响。因此，组合前面所有提到的算法来测试每种算法是最理想的。但是，所有组合的数量是 48 个，因此只能选用其中几种算法。

我们选择了 8 种不同的正弦分解系统，在这些系统中，分别比较每种算法和其它每个阶段可能的算法。算法组在表 6－3 中描述。

算法组 2 对应标准的 McAulay－Quatieri 算法，它直接从幅度谱中挑选峰值。由于这种方法没有考虑谱的整体水平，因此用户必须自己定义检测域值。因为必须随测试信号调整域值，所以算法组 1 和 2 与其它信号独立峰值检测算法相比略有优势。其它算法的参数要先对音乐信号调整，然后固定。

本文的系统中所有的算法对所有的频率都使用相同的 46 毫秒的分解窗。测试

信号的波长大约等于分解窗的长度，为检测正弦分量，用正弦分解法时窗长应设为

表 6－3：分解算法组

组	峰值检测	峰值插值	参数估计	峰值延续
1	固定	无	STFT	导数
2	固定	二次插值	STFT	导数
3	互相关	二次插值	STFT	导数
4	互相关	二次插值	LSQ	导数
5	互相关	二次插值	LSQ	合成
6	互相关	信号导数	LSQ	合成
7	F 测试	二次插值	LSQ	合成
8	同第 5 组的算法相同，用迭代分解			

正弦周期的 2 到 4 倍，因此，用多分辨率分解方法特别是在低频时能得到较好的结果。但是对所有频率使用相同的分解窗却是最简单的。

因为使用正弦产生测试信号，所以每个正弦的准确的频率、幅度和相位是已知的。比较分解中得到的正弦的参数和准确正弦的参数，分解完后计算几个统计量，这些统计量包括没有发现的正弦的百分比、发现的额外峰值、正弦中的断点、错误的延续和平均频率、幅度和相位误差。详细的统计值共计有 4 个表，见表 6－4 到表 6－7。

在表 6－4 中给出每个在原始信号中的正弦信号的丢失峰加额外峰的百分比。算法组 7 使用 F 测试进行峰值检测，与其它算法比较明显地错误较多，大多数错误是由 F 测试在窗长很小且频率很低时检测正弦不稳定造成的。F 测试在有颤音、震音和碰撞正弦时也有明显的错误见第 4 和第 5 段，在有谐音音的时候比互相关算法要好，见第 6 和第 10 段。幅度谱域值用在算法组 1 和算法组 2 中特别好，只比第 3

表 6－4：峰值检测：丢失峰和额外峰的百分比

	信号段									
算法组	1	2	3	4	5	6	7	8	9	10
1	0	6	77	4	18	23	10	0	40	17
2	0	6	77	4	18	23	10	0	40	17
3	0	7	28	0	27	67	78	68	33	89
4	0	7	28	0	27	67	78	68	33	89
5	0	4	28	1	23	64	64	21	31	74
6	0	2	28	1	23	64	64	21	31	74
7	25	49	25	22	53	47	82	32	33	63
8	0	5	28	1	18	62	57	10	31	67

表 6—5：峰值插值：平均频率误差（Hz）

	信号段									
算法组	1	2	3	4	5	6	7	8	9	10
1	1.4	3.0	2.4	4.5	2.1	4.8	3.6	6.4	1.7	3.3
2	0.3	1.9	0.3	3.9	1.2	3.9	2.8	6.2	0.5	2.7
3	0.5	1.1	0.8	3.7	1.6	0.3	2.2	4.1	0.7	1.7
4	0.5	1.1	0.8	3.7	1.5	0.3	2.3	4.1	0.7	1.7
5	0.5	1.3	0.8	3.6	1.5	0.3	2.4	5.1	0.7	1.8
6	0.4	1.4	0.7	3.8	1.2	0.3	2.5	5.2	0.9	1.8
7	0.5	0.4	0.8	1.2	0.4	0.6	0.7	5.0	0.6	0.9
8	0.5	1.3	0.8	3.6	1.6	0.4	2.3	5.4	0.7	1.8

表 6—6：参数估计：平均幅度差和相位差

	信号段									
算法组	1	2	3	4	5	6	7	8	9	10
1	0.2	1.6	0.2	0.9	1.2	1.0	2.0	0.8	0.6	1.6
2	0.2	1.6	0.2	0.9	1.2	1.0	2.0	0.8	0.5	1.6
3	0.2	1.1	0.2	0.7	1.1	0.2	1.9	0.3	0.4	1.7
4	0.7	0.6	0.2	0.4	1.3	0.2	0.7	0.2	0.3	0.9
5	0.7	0.7	0.2	0.4	1.3	0.2	0.8	0.2	0.3	1.0
6	0.7	0.7	0.2	0.4	1.3	0.2	0.8	0.2	0.3	1.0
7	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.7
8	0.7	0.7	0.2	0.4	1.7	0.3	0.8	0.3	0.3	1.0

表 6—7：峰值延续：延续失败和轨迹中断的百分比

	信号段									
算法组	1	2	3	4	5	6	7	8	9	10
1	0	6	0	8	11	8	11	3	30	18
2	0	4	0	8	7	9	10	3	30	17
3	0	2	2	1	11	0	3	4	30	2
4	0	2	2	1	10	0	2	4	30	2
5	0	2	2	2	11	0	4	8	30	4
6	0	0	2	2	11	0	4	8	30	4
7	0	1	2	1	2	13	1	6	30	6
8	0	1	2	2	9	0	5	8	30	7

段和第 9 段的平均水平差点，这也是正常的，因为这两段包含的声音幅度和整体水

平是不同的。

平均频率错误见表 6—5。除了使用二次插值，算法组 2 其它地方和算法组 1 相似，平均频率错误清楚地显示二次插值改进了分解过程。二次插值和导数插值比起来（见算法组 5 和算法组 6）性能十分相似，所以不能说哪种插值方法更好。

通过在假想空间中，计算到正确点的距离的方法来测量幅度和相位估计的错误，到正确点的平均距离见表 6—6。通过检查算法组 3 和算法组 4 的性能，看到 LSQ 估计和简单方法的差异。在第一段中使用谱系数的方法明显的好，因为 LSQ 方法在低频时出现很多的错误。谱系数方法在第 5 段也很好，这很奇怪，因为 LSQ 应该在紧密接近的正弦中表现很好才对。但总体上，LSQ 的表现还是要比谱系数方法好。

假延续的百分比或正弦轨迹的断点在表 6—7 中列出。算法组 5—7 用合成延续，其它的用导数延续。使用相同延续算法但用不同的峰值检测和参数估计明显有很大的性能变化，特别是在延续阶段用简单参数估计的幅度谱检测会引起明显的错误。在使不使用合成的延续上差别很小，不能说哪种方法更好，但是在有交叉分音时，基于参数导数的延续经常有错误发生，合成的延续更能得到正确的延续。

算法组 8 和其它的有很大不同，因为它对残余信号使用了一次迭代分解，在峰值检测时，它表现的比没有迭代时要稍好一点，当它达到平均频率、幅度和相位差时，表现几乎相同。同样，延续错误与非迭代算法组也是相差无几，因为大多数的延续是在第一次迭代时做的，在第二次迭代只提取了没有被发现的峰，或者改进参数，不想改变已经做的延续。这些统计中，迭代分解不比非迭代算法好，但是如果附加分量的数量增加了，迭代算法能比其它的算法产生更好的结果[24]。

6.3 计算效率的考虑

在整个正弦加噪模型的分解和合成过程中，正弦分解明显是最费时的部分，占总时间 50%以上。在图 6—1 中表示的是正弦加噪分解和合成过程中时间消耗百分比和使用算法组 5 的分解时间百分比，我们注意到，正弦分解和合成的时间与信号紧密相关：如果没发现正弦，分解和合成就非常快，然而在有丰富谐音声音的时候就要花费很长的时间。随机信号的分解和合成的复杂度与信号无关，因为它仅仅依靠特定频带的能量的计算。

显然，正弦分解时间由所用的算法决定，如图 6—2 所示。当用 matlab 实现时，三个算法都使用第一算法组，分解时间只是真实时间的 4 倍。这些算法组使用最简单的峰值检测算法，幅度谱域值和互相关算法，并且直接从谱中得到参数，因此这些算法组要求在每个分解帧中有效地实现一个 FFT。用这些算法组正弦分解时间与正弦合成、随机分解和随机合成的时间基本相同。

平是不同的。

平均频率错误见表 6－5。除了使用二次插值，算法组 2 其它地方和算法组 1 相似，平均频率错误清楚地显示二次插值改进了分解过程。二次插值和导数插值比起来（见算法组 5 和算法组 6）性能十分相似，所以不能说哪种插值方法更好。

通过在假想空间中，计算到正确点的距离的方法来测量幅度和相位估计的错误，到正确点的平均距离见表 6－6。通过检查算法组 3 和算法组 4 的性能，看到 LSQ 估计和简单方法的差异。在第一段中使用谱系数的方法明显的好，因为 LSQ 方法在低频时出现很多的错误。谱系数方法在第 5 段也很好，这很奇怪，因为 LSQ 应该在紧密接近的正弦中表现很好才对。但总体上，LSQ 的表现还是要比谱系数方法好。

假延续的百分比或正弦轨迹的断点在表 6－7 中列出。算法组 5－7 用合成延续，其它的用导数延续。使用相同延续算法但用不同的峰值检测和参数估计明显有很大的性能变化，特别是在延续阶段用简单参数估计的幅度谱检测会引起明显的错误。在使不使用合成的延续上差别很小，不能说哪种方法更好，但是在有交叉分音时，基于参数导数的延续经常有错误发生，合成的延续更能得到正确的延续。

算法组 8 和其它的有很大不同，因为它对残余信号使用了一次迭代分解，在峰值检测时，它表现的比没有迭代时要稍好一点，当它达到平均频率、幅度和相位差时，表现几乎相同。同样，延续错误与非迭代算法组也是相差无几，因为大多数的延续是在第一次迭代时做的，在第二次迭代只提取了没有被发现的峰，或者改进参数，不想改变已经做的延续。这些统计中，迭代分解不比非迭代算法好，但是如果附加分量的数量增加了，迭代算法能比其它的算法产生更好的结果[24]。

6．3 计算效率的考虑

在整个正弦加噪模型的分解和合成过程中，正弦分解明显是最费时的部分，占总时间 50%以上。在图 6－1 中表示的是正弦加噪分解和合成过程中时间消耗百分比和使用算法组 5 的分解时间百分比，我们注意到，正弦分解和合成的时间与信号紧密相关：如果没发现正弦，分解和合成就非常快，然而在有丰富谐音声音的时候就要花费很长的时间。随机信号的分解和合成的复杂度与信号无关，因为它仅仅依靠特定频带的能量的计算。

显然，正弦分解时间由所用的算法决定，如图 6－2 所示。当用 matlab 实现时，三个算法都使用第一算法组，分解时间只是真实时间的 4 倍。这些算法组使用最简单的峰值检测算法，幅度谱域值和互相关算法，并且直接从谱中得到参数，因此这些算法组要求在每个分解帧中有效地实现一个 FFT。用这些算法组正弦分解时间与正弦合成、随机分解和随机合成的时间基本相同。

当转用更加复杂的 LSQ 参数估计（第 4 组）时，会增加大量分解时间，这很好解释，因为每一时间帧 LSQ 需要计算大矩阵的转置。当峰值检测（第 5 组）使用合成的方法时，分解时间变得很长。导数差值用在第 6 组比二次差值用在第 2 到第 5 组用了较长的时间。用在第 7 组的 F 测试需要几个 FFT 和几个其它的计算复杂的操作，并因此明显的比在第 2 到第 6 组中的互相关方法要慢。迭代算子组 8 分解信号两次，因此需要大约两倍分解时间。它的分解时间包括合成和在一次分解后减去正弦的时间。

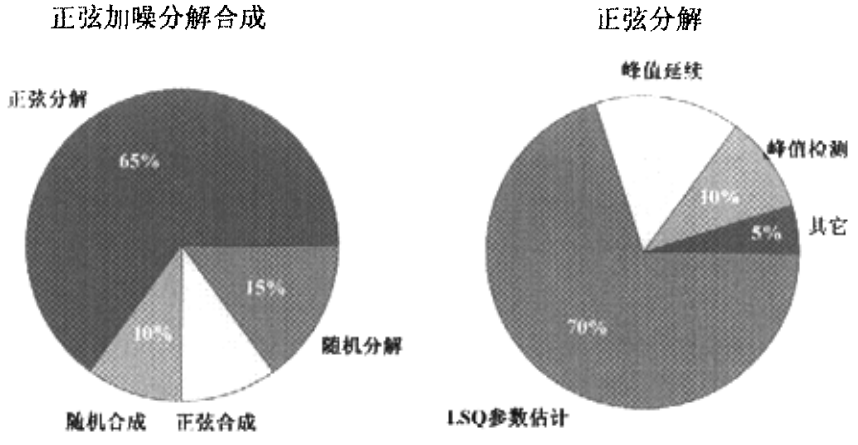

图 6－1 正弦加噪分解和合成过程中时间消耗百分比
和算法组 5 的分解时间百分比

在 Matlab 编程环境下实现这些算法，尽可能使用快速矩阵操作。通常，大多数的计算时间时花费在几个运算复杂的 FFT、均方差或者矩阵转置上。因为在 Matlab 中循环操作非常慢，至少还可以用其它一些编程语言来加速贪婪延续算法，这里就不在做介绍了。

6．4 不同正弦加噪系统的比较

虽然有很多正弦加噪系统，但是能够免费得到的却不多，本文的系统通过人工听合成的信号的方法来和另外两个系统来做比较。系统用了几个和 S.Levine 的系统相似的算法，因此，先要和他的系统做比较一下。有一个软件叫 SNDAN 包括了标准 McAulay－Quatieri 算法并且能够免费得到，另外一个系统就是选用了这种算法。

S.Levine 的系统包括了暂态模型，而本文的系统中没有考虑，因此声音的质量不能直接比较。因为我们没权使用 S.Levine 的系统，只能用他的网页上的两个音乐信号来测试本文的系统。

考虑到本系统不包括暂态模型，系统之间感知上的差距很小，但声音上的区别

是能听的出来的，只是并不是像想象中的那么明显。在两个系统中，合成的噪音基本上相同，这很自然，因为两个系统在随机模型中都使用了相同的BARK带。S. Levine的系统整体感知质量较好，但是很难说是不是暂态模型的结果。

SNDAN比本文的系统多用了很多的帧率，它的峰值检测域值也是手工设的，如果域值很低，系统会检测出巨大的正弦并产生一些像相位合成机式的声音。SNDAN用正弦重构了输入信号的所有分量，如果域值较高，只有谐音分量能用正弦重构，这

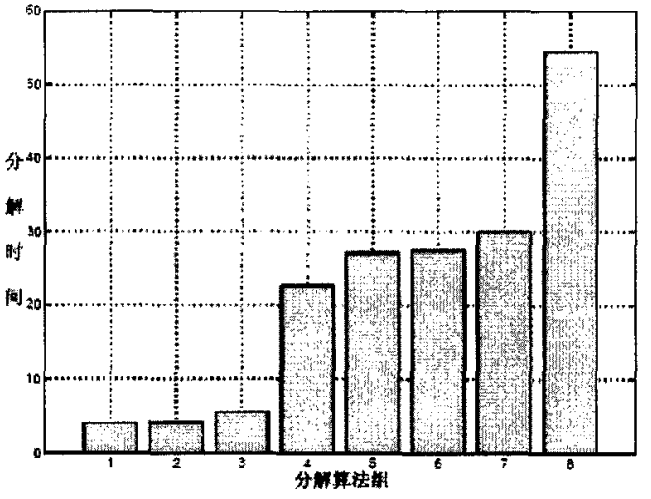

图 6－2 正弦分解时间用不同方法的比较

是最理想的状态。声音的质量靠的是信号的特征，如果信号中有大量的动态变化，固定域值就不能很好的工作了：如在安静的部分，所有在域值以下的分量和没有峰值的分量都会被检测到。

有些时候合成信号的质量与本文的系统相当，但在大多数情况下它的质量比较差。有时，快速的帧率会产生恼人的声音效果。在SNDAN中，正弦轨迹不能彼此交叉，即使McAulay－Quatieri算法的一些实现方法也要求如此。

6．5 算法选择对效率和质量的影响

本文的正弦分解系统是在其它几个正弦分解系统的基础上，结合自己的一些想法所组成，重点是从计算机声学场景分解的角度出发，研究语音合成正弦的质量。在其中的一些算法中，为了减少计算复杂度必需做少量的折衷。

在实验和算法测试中明显的看到，在分解过程中没有哪种算法的结合是最好的，也没有对所有信号都适用的算法，这在第6章的结果中可以明显看到，因此，对于不同的应用系统要建立和使用不同的算法。用户不想自己指定特定的算法时用两个缺省的算法组：一个是速度优先，它要求速度快但还要产生可应用的结果；另

是能听的出来的，只是并不是像想象中的那么明显。在两个系统中，合成的噪音基本上相同，这很自然，因为两个系统在随机模型中都使用了相同的BARK带。S. Levine的系统整体感知质量较好，但是很难说是不是暂态模型的结果。

SNDAN 比本文的系统多用了很多的帧率，它的峰值检测域值也是手工设的，如果域值很低，系统会检测出巨大的正弦并产生一些像相位合成机式的声音。SNDAN用正弦重构了输入信号的所有分量，如果域值较高，只有谐音分量能用正弦重构，这

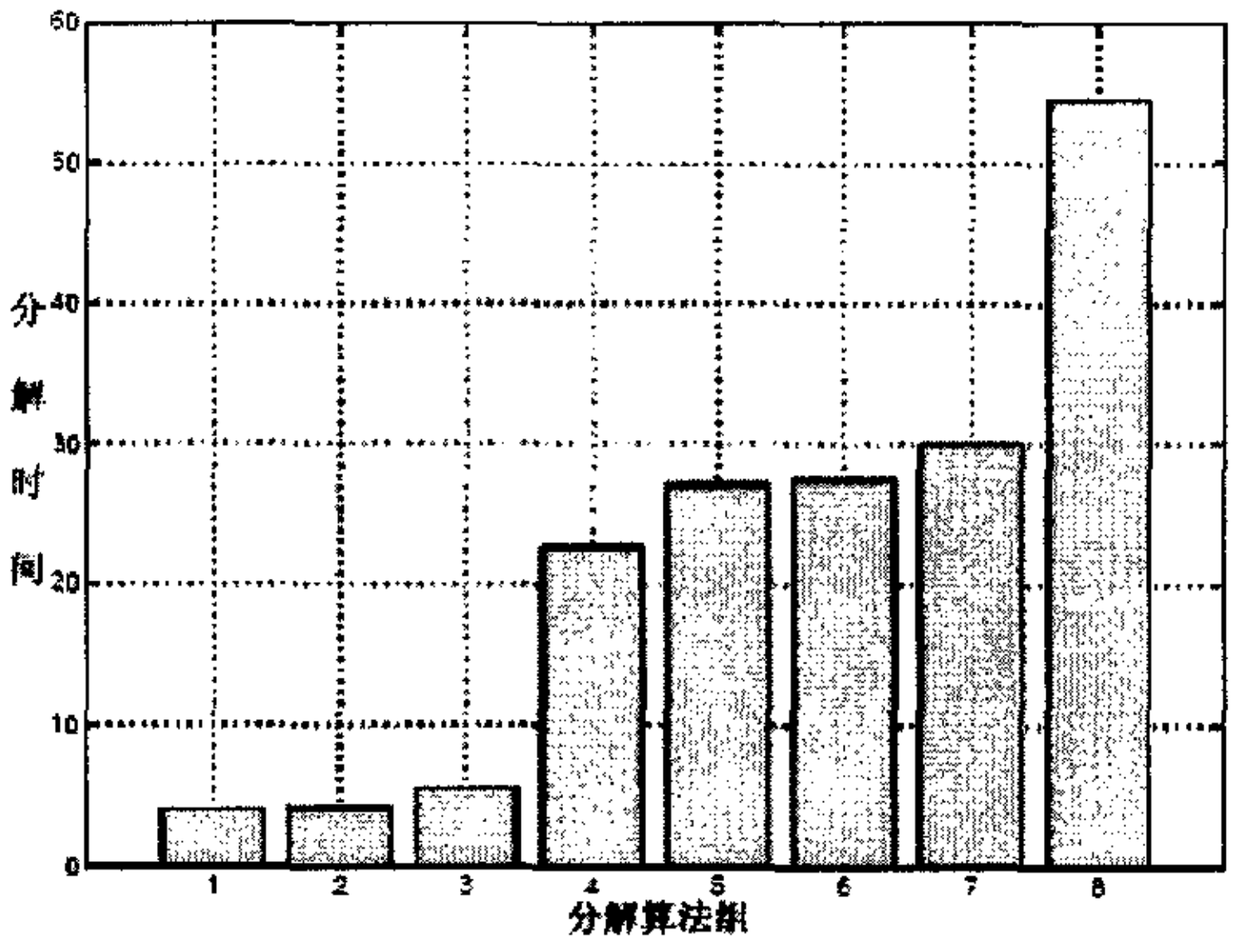

图 6—2 正弦分解时间用不同方法的比较

是最理想的状态。声音的质量靠的是信号的特征，如果信号中有大量的动态变化，固定域值就不能很好的工作了：如在安静的部分，所有在域值以下的分量和没有峰值的分量都会被检测到。

有些时候合成信号的质量与本文的系统相当，但在大多数情况下它的质量比较差。有时，快速的帧率会产生恼人的声音效果。在 SNDAN 中，正弦轨迹不能彼此交叉，即使 McAulay—Quatieri 算法的一些实现方法也要求如此。

6. 5 算法选择对效率和质量的影响

本文的正弦分解系统是在其它几个正弦分解系统的基础上，结合自己的一些想法所组成，重点是从计算机声学场景分解的角度出发，研究语音合成正弦的质量。在其中的一些算法中，为了减少计算复杂度必需做少量的折衷。

在实验和算法测试中明显的看到，在分解过程中没有哪种算法的结合是最好的，也没有对所有信号都适用的算法，这在第 6 章的结果中可以明显看到，因此，对于不同的应用系统要建立和使用不同的算法。用户不想自己指定特定的算法时用两个缺省的算法组：一个是速度优先，它要求速度快但还要产生可应用的结果；另

一个是质量优先，分解的质量是需要优先考虑的，但需要牺牲分解时间为代价，同时分解时间是可以容忍的。算法组选择的结合见表 6－3 中算法组 3 和算法组 5。

两个系统都是用互相关法做峰值检测，用二次插值做峰值插值。在速度优先中，对所有频率都使用单一的 46ms 的窗。在质量优先中，多分辨率分解用在三个频带 20－200Hz，200－5000Hz 和 5－10kHz，窗长分别是 86ms、46ms 和 46ms。两个最高频带用相同的窗长，但是参数稍有不同。声音的特征在最高频带不同，不同的参数会得到不同的效果。

在速度优先中，参数直接从插值谱中获得，用参数导数得到的峰是连续的。在质量优先中，对幅度和相位的估计参数用 LSQ 方法，合成的延续用来延续峰。在速度优先和质量优先中，用掩蔽域计算出掩蔽曲线，并且滤掉错误轨迹。

速度优先和质量优先是有关系的，主要由长分解窗和轨迹滤波器引起算法的延迟。实际中，最长 65ms 的轨迹能被滤掉。最长的分解窗用 86ms，所以算法延时小于 100ms。因此，如果计算源允许的话速度优先和质量优先都能够实时的实现。

如果要得到特别好的质量，输入信号的数量和长度较小，使用插值分解可以获得较好的结果。其次，算法的使用要看具体应用：如果重点在噪音部分并且想去掉所有的谐音分量，就不需要正弦的参数，因此就不需要参数融合。

第七章 语音分离和处理的应用

本章介绍在语音分离的中层再现中使用正弦加噪模型，主要讨论使用轨迹间感知距离进行语音分离[32]，还简单介绍了一个更可靠的语音分离方法，它就是基于多音高的估计模型[33]。

混合语音的分离在语音信号的分解、编辑和处理中有很多的应用，其中包括：语音编码、音乐自动录制、语音增强和计算机听觉场景分析等。目前，声音分离的大部分研究都是在计算机语言学听觉分析领域展开的。

正弦加噪模型允许在参数域进行分离声音的处理，可以在不改变合成声音的质量的情况下，修改信号的音高和时间尺度，这个方法将在本节最后有简单的描述。

7．1 语音分离

当两个声音在时间和频率上叠加的时候，分离是很困难的，没有一个常规的方法来分解声音。然而，如果能够对混合的声音做一些假设的话，就可以合成出感知上近似原声的声音。假定主要的声音是谐音，而且它们有不同的基本频率。使用正弦模型可以将输入信号分解成频谱分量，使用一系列的感知相关提示将它们分配到声源，然后分别的合成这些声音。计算过程如下：首先，系统使用正弦模型重构用正弦轨迹表示的信号。第二，在输出正弦曲线中，通过插值轨迹去除由调幅、暂态和噪声产生的中断。第三，系统通过计算缩放的频率和幅度的差异和轨迹的谐音一致性，估计感知上轨迹的紧密度；然后将这些轨迹分类归入声源。系统断定哪些轨迹是碰撞谐音的结果，然后把这些轨迹分开为两部分；最后，当轨迹都被分类、分开后，系统就能够分别的合成原始的两个声音了。

分类部分本身是目前该系统中最落后的部分。仿真时，分类器假定两个声源，使用它们的不同起始时间来对它们初始化。这样，要求声音的起始时间差必须大于100ms，如果没有这个限制，可以通过计算所有正弦的感知距离，然后用常规的分类算法把它们分类即可。

方法本身可以应用到更复杂的情况，甚至同时发生的声音。当只有正弦加噪模型时，从大量的混谐音音中要获得很好的结果是很困难的，因为它们中有些声音的谐音分音很难被检测到。

7．2 标准正弦模型的修改

由正弦模型得到的所有轨迹通常不能重构声音的全部分音，最常见的错误是轨

第七章 语音分离和处理的应用

本章介绍在语音分离的中层再现中使用正弦加噪模型，主要讨论使用轨迹间感知距离进行语音分离[32]，还简单介绍了一个更可靠的语音分离方法，它就是基于多音高的估计模型[33]。

混合语音的分离在语音信号的分解、编辑和处理中有很多的应用，其中包括：语音编码、音乐自动录制、语音增强和计算机听觉场景分析等。目前，声音分离的大部分研究都是在计算机语言学听觉分析领域展开的。

正弦加噪模型允许在参数域进行分离声音的处理，可以在不改变合成声音的质量的情况下，修改信号的音高和时间尺度，这个方法将在本节最后有简单的描述。

7．1 语音分离

当两个声音在时间和频率上叠加的时候，分离是很困难的，没有一个常规的方法来分解声音。然而，如果能够对混合的声音做一些假设的话，就可以合成出感知上近似原声的声音。假定主要的声音是谐音，而且它们有不同的基本频率。使用正弦模型可以将输入信号分解成频谱分量，使用一系列的感知相关提示将它们分配到声源，然后分别的合成这些声音。计算过程如下：首先，系统使用正弦模型重构用正弦轨迹表示的信号。第二，在输出正弦曲线中，通过插值轨迹去除由调幅、暂态和噪声产生的中断。第三，系统通过计算缩放的频率和幅度的差异和轨迹的谐音一致性，估计感知上轨迹的紧密度；然后将这些轨迹分类归入声源。系统断定哪些轨迹是碰撞谐音的结果，然后把这些轨迹分开为两部分；最后，当轨迹都被分类、分开后，系统就能够分别的合成原始的两个声音了。

分类部分本身是目前该系统中最落后的部分。仿真时，分类器假定两个声源，使用它们的不同起始时间来对它们初始化。这样，要求声音的起始时间差必须大于100ms，如果没有这个限制，可以通过计算所有正弦的感知距离，然后用常规的分类算法把它们分类即可。

方法本身可以应用到更复杂的情况，甚至同时发生的声音。当只有正弦加噪模型时，从大量的混谐音音中要获得很好的结果是很困难的，因为它们中有些声音的谐音分音很难被检测到。

7．2 标准正弦模型的修改

由正弦模型得到的所有轨迹通常不能重构声音的全部分音，最常见的错误是轨

迹的中断，可能是因为暂态、同时出现的噪声或者是谐音本身振幅太小以至于无法估计正弦轨迹。这些错误通常发生在对信号强幅度调制的情况下，这时的幅度趋向于 0，这种情况的很简单的例子是小提琴的高阶谐音，如图 7—1。

如果两个频率分量的差别很小，频率和幅度彼此很接近，人的听觉系统就把声音关联起来。本系统试图通过关联互相接近的轨迹来模仿它，轨迹间的中断点用内插值替换。内插也增加了系统的鲁棒性，比如一个谐音由一个长的轨迹重构，而不是由好多个短的轨迹重构。对于相关联的轨迹，可以通过比较断点附近的起始时刻、时间偏移量、频率和振幅，找到那些最可能是同一个谐音的轨迹。然后在这些断点的地方通过内插频率和振幅消除断点。在本系统中，可以使用线性内插，因为这就能很好达到分离的目的。因为允许太长的断点或者太大的频差和幅差将导致关联到错误的轨迹，所以不是每个断点都要内插。在本例中，轨迹的数目从 87 减少至 48。

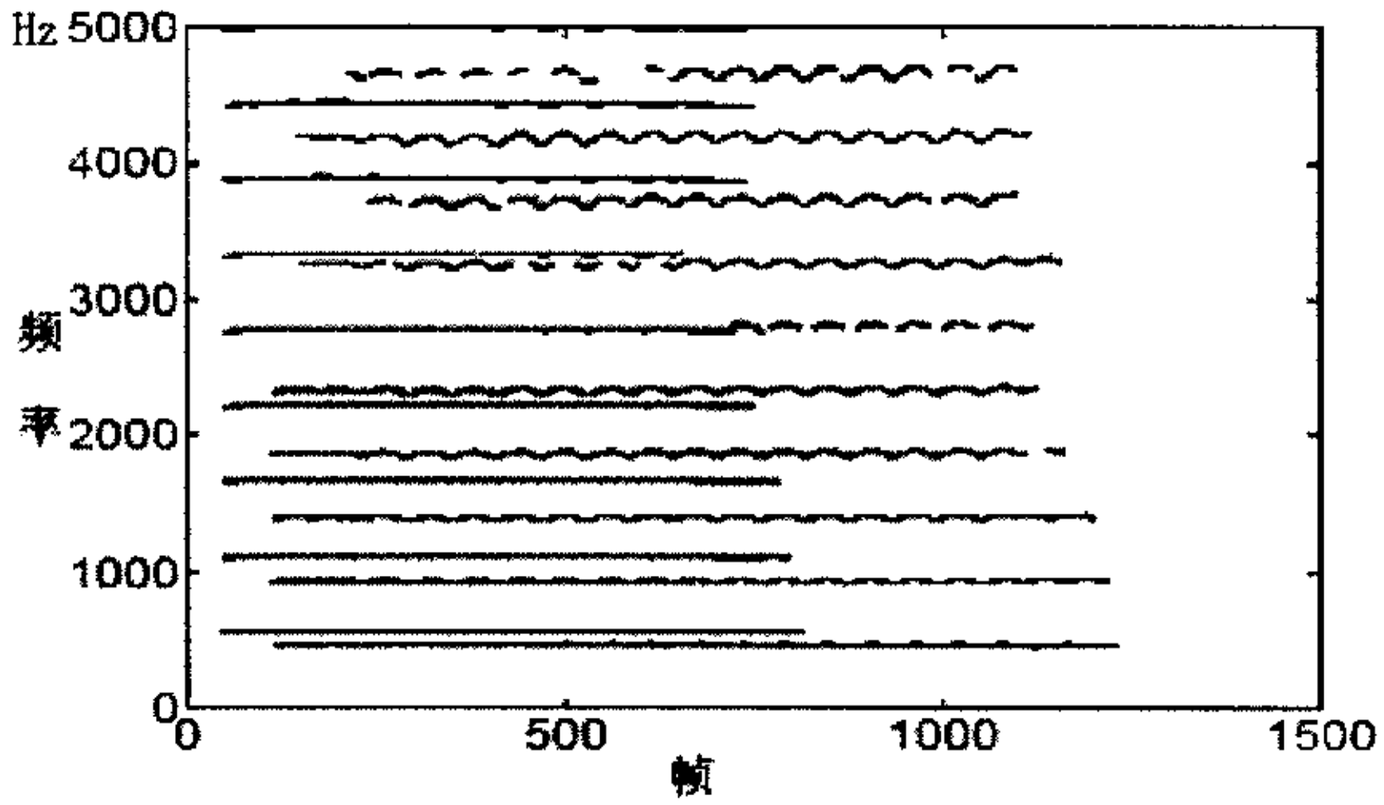

图 7—1 由双簧管和小提琴组成信号的正弦轨迹（小提琴开始较晚）

7．3 感知距离测量

A. S. Bregman 列举了下列人类听觉系统的关联环境[33]：

(1) 频谱 F 接近性（时间和频率接近）

(2) 谐音一致性

(3) 分量的同步变化：a）共同起始，b）共同偏移，c）共同的幅度调制，d）共同的频率调制，e）频谱中同方向移动

(4) 空间接近性

在本文中，我们把重点放在分量的同步变化，在某种程度上还要考虑谐音的一致性。

迹的中断，可能是因为暂态、同时出现的噪声或者是谐音本身振幅太小以至于无法估计正弦轨迹。这些错误通常发生在对信号强幅度调制的情况下，这时的幅度趋向于 0，这种情况的很简单的例子是小提琴的高阶谐音，如图 7—1。

如果两个频率分量的差别很小，频率和幅度彼此很接近，人的听觉系统就把声音关联起来。本系统试图通过关联互相接近的轨迹来模仿它，轨迹间的中断点用内插值替换。内插也增加了系统的鲁棒性，比如一个谐音由一个长的轨迹重构，而不是由好多个短的轨迹重构。对于相关联的轨迹，可以通过比较断点附近的起始时刻、时间偏移量、频率和振幅，找到那些最可能是同一个谐音的轨迹。然后在这些断点的地方通过内插频率和振幅消除断点。在本系统中，可以使用线性内插，因为这就能很好达到分离的目的。因为允许太长的断点或者太大的频差和幅差将导致关联到错误的轨迹，所以不是每个断点都要内插。在本例中，轨迹的数目从 87 减少至 48。

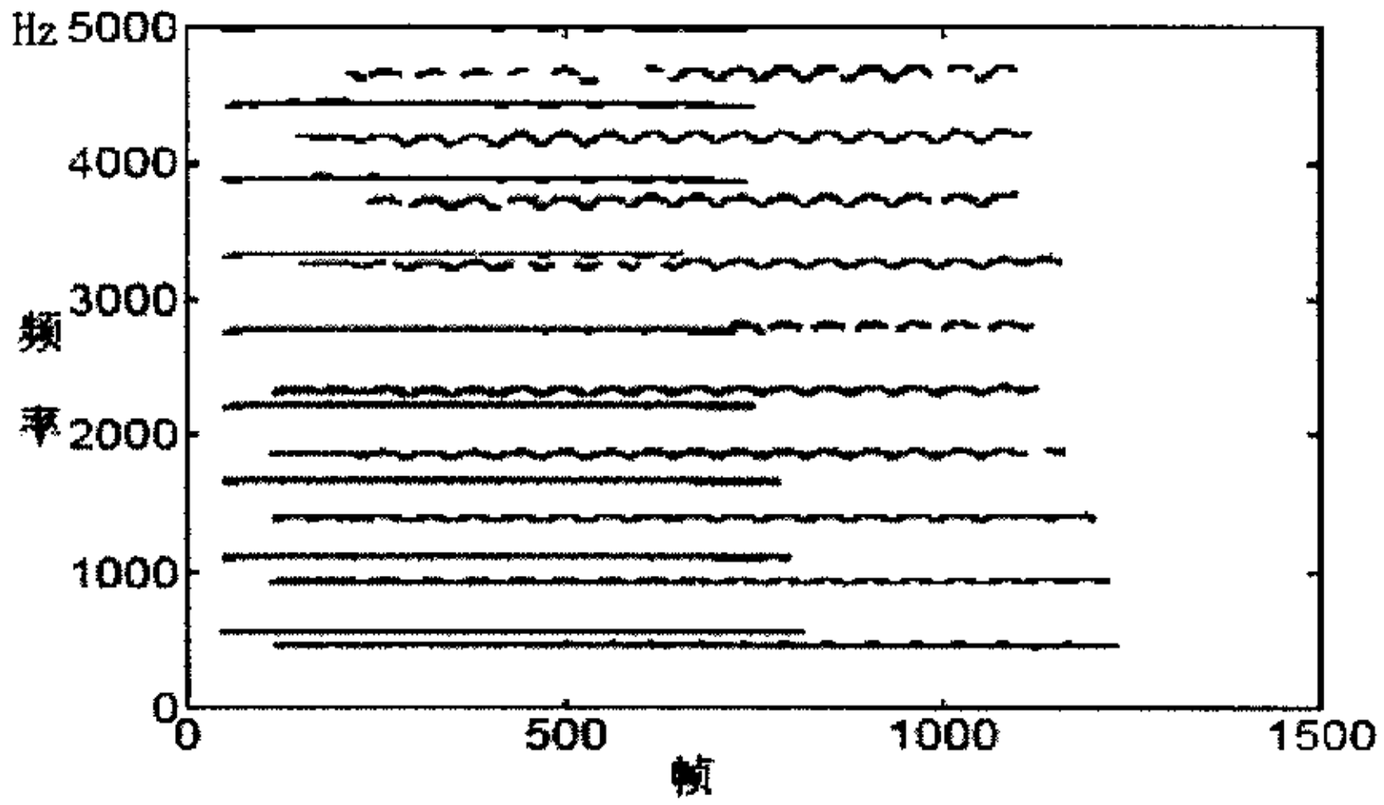

图 7—1 由双簧管和小提琴组成信号的正弦轨迹（小提琴开始较晚）

7．3 感知距离测量

A. S. Bregman 列举了下列人类听觉系统的关联环境[33]：

(1) 频谱 F 接近性（时间和频率接近）

(2) 谐音一致性

(3) 分量的同步变化：a）共同起始，b）共同偏移，c）共同的幅度调制，d）共同的频率调制，e）频谱中同方向移动

(4) 空间接近性

在本文中，我们把重点放在分量的同步变化，在某种程度上还要考虑谐音的一致性。

7．3．1 振幅和频率变化的测量

当在研究测量共同的幅度调制和频率调制的时候，可以发现在某种程度上，调制可以用两个量来表示：调制频率和指数。然而，仅仅用这两个量来表示幅度和频率调制是不够的，因为调制只在时间域上变化，因此需要很多的测量来概括时间的变化。同样，总的长度上的声音强度变化有时使得测量声音的调制特性很困难。

不同的谐音分音具有宽范围的振幅值，有时它们长时间的发展并不相似。然而，通过测量每个分音振幅的平均值，它们的结果曲线却是很接近。在频率的情况下，这个方法甚至更加精确，因为频率不像振幅那样在时间内变化那么大。如图 7－2 所示。这些成比例的频率的均方误差测量出了正弦轨迹的频率差。

$$d_f(i,j)=\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}\left(\frac{f_i(t)}{f_i}-\frac{f_j(t)}{f_j}\right)^2 \qquad (7-1)$$

其中 $f_i(t)$ 是轨迹 p_i 在 t 时刻的频率。选择时间 t_1 和 t_2 使两条曲线 p_i 和 p_j 存在的时间满足 $t_1<t<t_2$。f_i 和 f_j 是曲线 p_i 和 p_j 分别在时间 t_1 和 t_2 上的平均频率。同样的原则用于得到由幅值差 $d_a(i,j)$ 引起的感知距离，幅值差即幅值 $a_i(t)$ 和 $a_j(t)$ 与他们的平均值 a_i 和 a_j 的差值。

7．3．2 谐音一致性的测量

谐音分音频率 f_p 很接近谐音声音的基本频率 f_0 的整数倍。在本文中，不知道语音的基频，尽管可以从原始信号中估计出来，但也不想尝试在那些混合语音中计算出基频[34]。其实，我们研究出一个可以模拟任何两条正弦曲线的谐音一致性测量的方法，如果已知一个谐音源的两条正弦曲线，那么这两条曲线的频率比值即是一个很小的两个正整数之间的比值：

$$\frac{f_i}{f_j}=\frac{a}{b} \qquad (7-2)$$

其中 f_i 和 f_j 是正弦曲线 p_i 和 p_j 的频率，并且语音的谐音是 a^{th} 和 b^{th}。由于不知道哪条曲线属于那个语音或者哪条曲线是哪个谐音，假定基波频率不可能比正弦模型中的最小频率小。这样，可以得到 a 和 b 的上极限：

$$a=1,2,\cdots\cdots,\left|\frac{f_i}{f_{\min}}\right|,\ b=1,2,\cdots\cdots,\left|\frac{fj}{f_{\min}}\right|, \qquad (7-3)$$

其中 $f_{\min}$ 是正弦模型的最小频率。

确定 a 和 b 的极限之后，可以计算出所有可能的 a 和 b 的比率，并且选择出最好

7．3．1 振幅和频率变化的测量

当在研究测量共同的幅度调制和频率调制的时候，可以发现在某种程度上，调制可以用两个量来表示：调制频率和指数。然而，仅仅用这两个量来表示幅度和频率调制是不够的，因为调制只在时间域上变化，因此需要很多的测量来概括时间的变化。同样，总的长度上的声音强度变化有时使得测量声音的调制特性很困难。

不同的谐音分音具有宽范围的振幅值，有时它们长时间的发展并不相似。然而，通过测量每个分音振幅的平均值，它们的结果曲线却是很接近。在频率的情况下，这个方法甚至更加精确，因为频率不像振幅那样在时间内变化那么大。如图 7－2 所示。这些成比例的频率的均方误差测量出了正弦轨迹的频率差。

$$d_f(i,j)=\frac{1}{t_2-t_1+1}\sum_{t=t_1}^{t_2}\left(\frac{f_i(t)}{f_i}-\frac{f_j(t)}{f_j}\right)^2 \qquad (7-1)$$

其中 $f_i(t)$ 是轨迹 p_i 在 t 时刻的频率。选择时间 t_1 和 t_2 使两条曲线 p_i 和 p_j 存在的时间满足 $t_1<t<t_2$。f_i 和 f_j 是曲线 p_i 和 p_j 分别在时间 t_1 和 t_2 上的平均频率。同样的原则用于得到由幅值差 $d_a(i,j)$ 引起的感知距离，幅值差即幅值 $a_i(t)$ 和 $a_j(t)$ 与他们的平均值 a_i 和 a_j 的差值。

7．3．2 谐音一致性的测量

谐音分音频率 f_p 很接近谐音声音的基本频率 f_0 的整数倍。在本文中，不知道语音的基频，尽管可以从原始信号中估计出来，但也不想尝试在那些混合语音中计算出基频[34]。其实，我们研究出一个可以模拟任何两条正弦曲线的谐音一致性测量的方法，如果已知一个谐音源的两条正弦曲线，那么这两条曲线的频率比值即是一个很小的两个正整数之间的比值：

$$\frac{f_i}{f_j}=\frac{a}{b} \qquad (7-2)$$

其中 f_i 和 f_j 是正弦曲线 p_i 和 p_j 的频率，并且语音的谐音是 a^{th} 和 b^{th}。由于不知道哪条曲线属于那个语音或者哪条曲线是哪个谐音，假定基波频率不可能比正弦模型中的最小频率小。这样，可以得到 a 和 b 的上极限：

$$a=1,2,\cdots\cdots,\left|\frac{f_i}{f_{\min}}\right|,\ b=1,2,\cdots\cdots,\left|\frac{fj}{f_{\min}}\right|, \qquad (7-3)$$

其中 $f_{\min}$ 是正弦模型的最小频率。

确定 a 和 b 的极限之后，可以计算出所有可能的 a 和 b 的比率，并且选择出最好

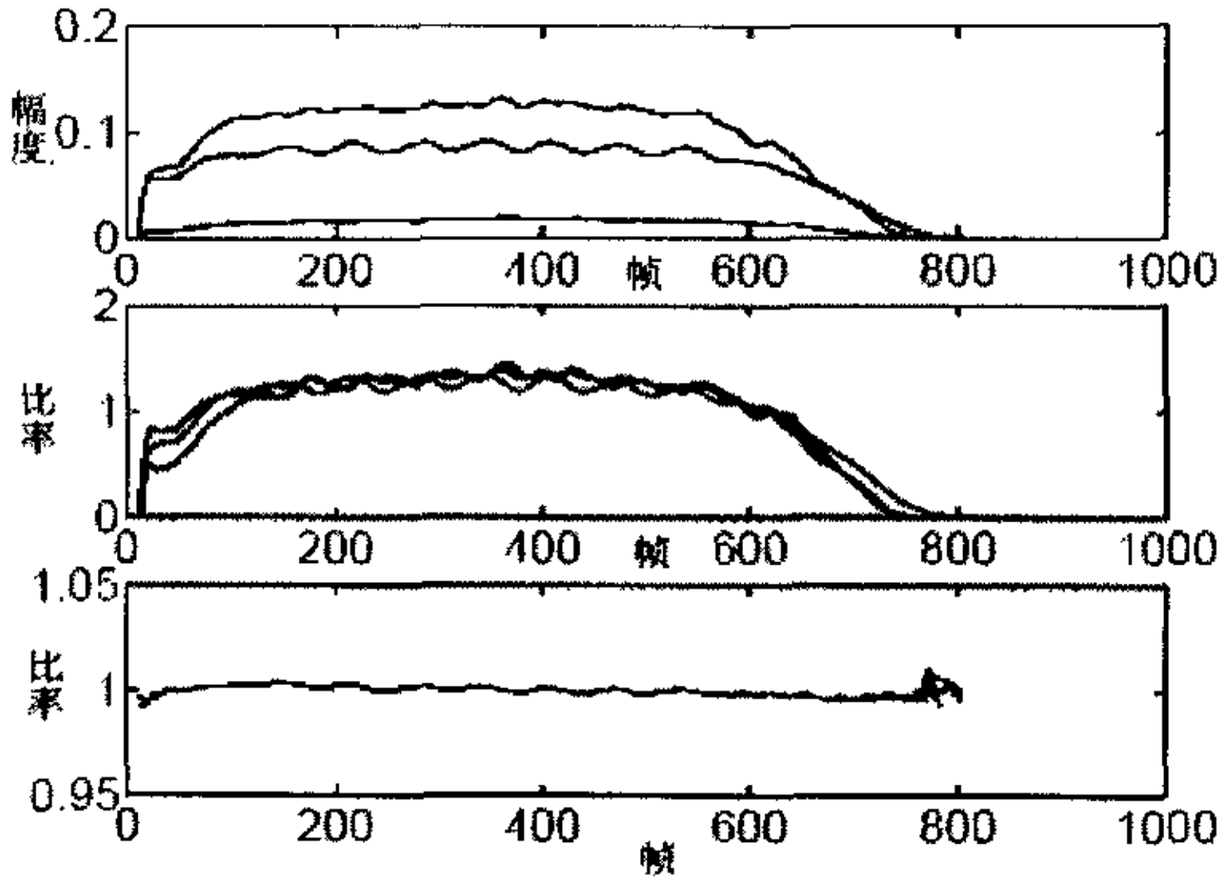

图 7－2 上图是双簧管的三个一次谐音的幅度；

中图是这些曲线缩放后的形状；

下图是同一谐音比例缩放后的曲线

的频率比率。谐音距离的长度是频率比率和a/b的比率之间的误差。归一化处理，取比率的对数的绝对值来测量轨迹之间的谐音误差：

$$d_h(i,j)=\min\left|\log(\frac{f_i/f_j}{a/b})\right| \qquad (7-4)$$

其中a 和b有（7－3）式的限制。

7．3．3 轨迹间总感知距离

任何两条轨迹的总感知距离等于频率、幅度和谐音距离的加权和：

$$d_{all}(i,j)=w_f d_f(i,j)+w_a d_a(i,j)+w_h d_h(i,j) \qquad (7-5)$$

由于语音产生的物理特征，频率常常不随幅度的变化而变化。因此频距必须比幅距的权重。谐音距离可以用另一个方法计算出来，因此他有不同的缩放比例。因为人对同时发生的轨迹的感知很大程度上基于谐音一致性，在一定程度上谐音距离的加权起最大的影响。

当计算总误差时，开始时间并没有直接考虑进去，幅值曲线包含开始时间和偏差时间的信息，在开始前，曲线的幅值总是 0。对于自然语音，幅值曲线常常在开始时上升的很快，然后就慢慢开始衰减。偏差时间过后，幅值再一次变为 0，幅距能够反应所有这些变化。

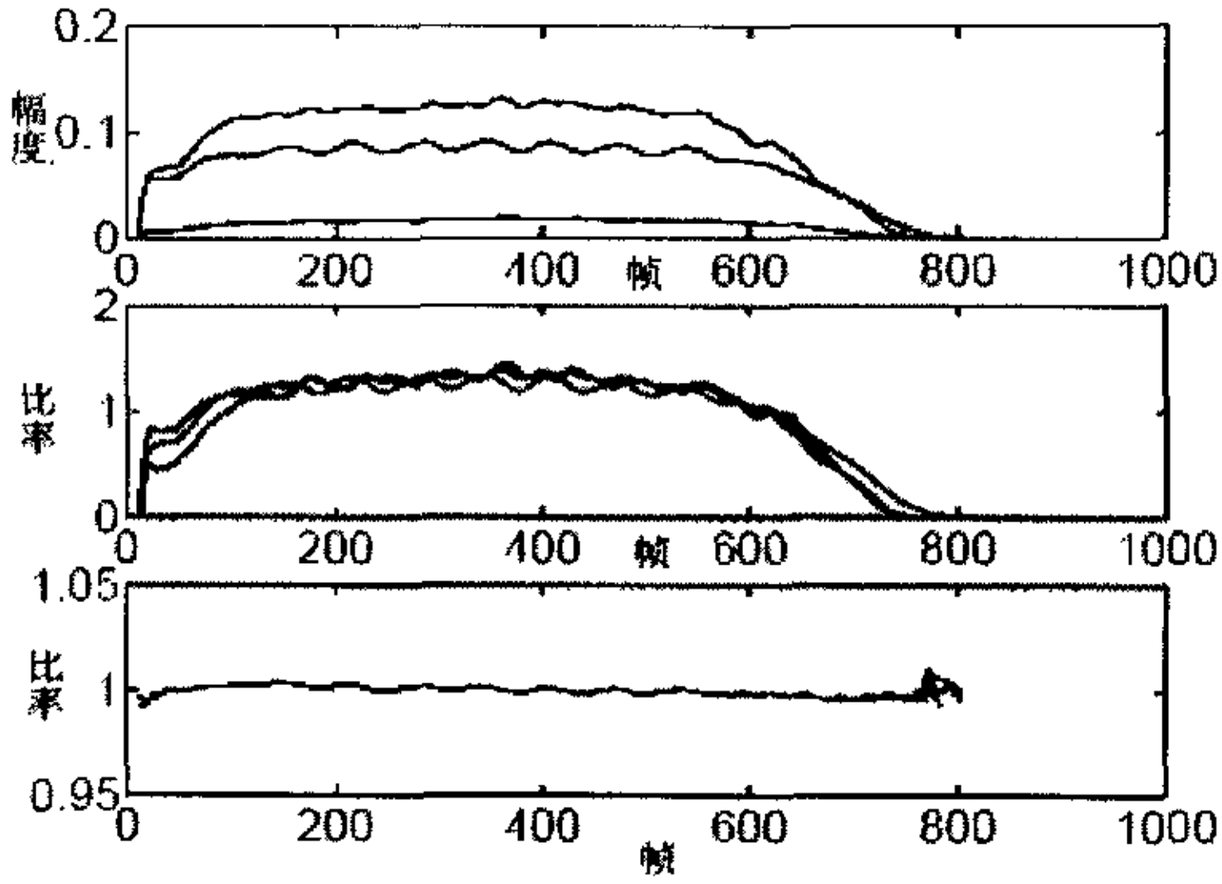

图 7－2 上图是双簧管的三个一次谐音的幅度；

中图是这些曲线缩放后的形状；

下图是同一谐音比例缩放后的曲线

的频率比率。谐音距离的长度是频率比率和a/b的比率之间的误差。归一化处理，取比率的对数的绝对值来测量轨迹之间的谐音误差：

$$d_h(i,j)=\min\left|\log(\frac{f_i/f_j}{a/b})\right| \qquad (7-4)$$

其中a 和b有（7－3）式的限制。

7．3．3 轨迹间总感知距离

任何两条轨迹的总感知距离等于频率、幅度和谐音距离的加权和：

$$d_{all}(i,j)=w_f d_f(i,j)+w_a d_a(i,j)+w_h d_h(i,j) \qquad (7-5)$$

由于语音产生的物理特征，频率常常不随幅度的变化而变化。因此频距必须比幅距的权重。谐音距离可以用另一个方法计算出来，因此他有不同的缩放比例。因为人对同时发生的轨迹的感知很大程度上基于谐音一致性，在一定程度上谐音距离的加权起最大的影响。

当计算总误差时，开始时间并没有直接考虑进去，幅值曲线包含开始时间和偏差时间的信息，在开始前，曲线的幅值总是 0。对于自然语音，幅值曲线常常在开始时上升的很快，然后就慢慢开始衰减。偏差时间过后，幅值再一次变为 0，幅距能够反应所有这些变化。

7．4 轨迹分类

经过估计每一对正弦轨迹的紧密度之后，应该把它们分类成独立的语音源。因为所有的轨迹除了距离之外，就没有任何共同的特征了，只好把一类轨迹中具有最小误差的轨迹分类：

$$\min\left(\frac{1}{|S_1|}\sum_{i,j\in S_1} d_{all}(i,j)+\frac{1}{|S_2|}\sum_{k,l\in S_2} d_{all}(k,l)\right) \tag{7-6}$$

$$S_1 \cup S_2 = S, S_1 \cap S_2 = \phi$$

其中 S_1 和 S_2 是两个语音的轨迹集合，S 是所有曲线的集合，$|S|$ 是集合的基数。

选择分类的理想解决办法是：计算所有的排列，选择最好的那个。但是，由于正弦轨迹的实际数量很多，计算量会很大（轨迹数的平方），所以应该用其他的分类方法。解决这个问题的有效方法是选择每一类轨迹的初始集合，逐个增加与前一轨迹的误差最小的曲线到分类中。最好的轨迹初始集合可以由以下方法得到：选择所有轨迹中开始时间最接近语音的估计开始时间 t_0 的曲线，然后计算所有可能的曲线子集。最小化误差子集一般包含轨迹较好的正弦，不包含估计误差或者碰撞的正弦。为了突出长的稳定的轨迹，轨迹的长度可作为缩放因子，接近开始时间的轨迹数一般很少，所以可以估计出所有的排列：

$$e = \min\left(\sum_{i,j\in S_1}\frac{d_{all}(i,j)}{\sqrt{length(p_i)}}\right) \tag{7-7}$$

$$(i\in S_1; |t_1 - t(i)| < t_{\lim it}), |S_1| = c, S_1 \subset S,$$

其中 t_1 是语音 1 的估计开始时间，$t(i)$ 是曲线 i 的开始时间，$t_{\lim it}$ 是曲线与初始值的最大误差，c 是初始子集的大小。初始时间可以由很多种方法得到，在本系统中采用的是，首先在每条轨迹开始处求幅差的总和，此时轨迹已经用三角窗平滑过，然后选择平滑过的幅值之和与轨迹的最大值之和为开始时间。这种开始时间测试法可以用在声音的开始没有强暂态时发生的情况下，比如小提琴的声音。

当已经估计出每个声音的初始子集，就开始逐个的把轨迹的剩余部分加上去，要一直选择子集和轨迹距离最小的。子集和轨迹间的距离是轨迹和子集中的轨迹的简单平均值。在所有的轨迹分类完之前会一直迭代下去，分类的结果在图 7—3 中表示。有时一个重叠的轨迹属于两个声音，这些碰撞的轨迹的检测在下部分讨论。

7．5 碰撞轨迹

前面已经提到，音乐信号里的两个声音的谐音经常会重叠。谐音重叠产生的正弦轨迹，称作碰撞轨迹。谐音是否重叠取决于声音的间隔。如果具有类似大三度和

7．4 轨迹分类

经过估计每一对正弦轨迹的紧密度之后，应该把它们分类成独立的语音源。因为所有的轨迹除了距离之外，就没有任何共同的特征了，只好把一类轨迹中具有最小误差的轨迹分类：

$$\min\left(\frac{1}{|S_1|}\sum_{i,j\in S_1} d_{all}(i,j)+\frac{1}{|S_2|}\sum_{k,l\in S_2} d_{all}(k,l)\right) \tag{7-6}$$

$$S_1 \cup S_2 = S, S_1 \cap S_2 = \phi$$

其中 S_1 和 S_2 是两个语音的轨迹集合，S 是所有曲线的集合，$|S|$ 是集合的基数。

选择分类的理想解决办法是：计算所有的排列，选择最好的那个。但是，由于正弦轨迹的实际数量很多，计算量会很大（轨迹数的平方），所以应该用其他的分类方法。解决这个问题的有效方法是选择每一类轨迹的初始集合，逐个增加与前一轨迹的误差最小的曲线到分类中。最好的轨迹初始集合可以由以下方法得到：选择所有轨迹中开始时间最接近语音的估计开始时间 t_0 的曲线，然后计算所有可能的曲线子集。最小化误差子集一般包含轨迹较好的正弦，不包含估计误差或者碰撞的正弦。为了突出长的稳定的轨迹，轨迹的长度可作为缩放因子，接近开始时间的轨迹数一般很少，所以可以估计出所有的排列：

$$e = \min\left(\sum_{i,j\in S_1}\frac{d_{all}(i,j)}{\sqrt{length(p_i)}}\right) \tag{7-7}$$

$$(i\in S_1; |t_1 - t(i)| < t_{\lim it}), |S_1| = c, S_1 \subset S,$$

其中 t_1 是语音 1 的估计开始时间，$t(i)$ 是曲线 i 的开始时间，$t_{\lim it}$ 是曲线与初始值的最大误差，c 是初始子集的大小。初始时间可以由很多种方法得到，在本系统中采用的是，首先在每条轨迹开始处求幅差的总和，此时轨迹已经用三角窗平滑过，然后选择平滑过的幅值之和与轨迹的最大值之和为开始时间。这种开始时间测试法可以用在声音的开始没有强暂态时发生的情况下，比如小提琴的声音。

当已经估计出每个声音的初始子集，就开始逐个的把轨迹的剩余部分加上去，要一直选择子集和轨迹距离最小的。子集和轨迹间的距离是轨迹和子集中的轨迹的简单平均值。在所有的轨迹分类完之前会一直迭代下去，分类的结果在图 7—3 中表示。有时一个重叠的轨迹属于两个声音，这些碰撞的轨迹的检测在下部分讨论。

7．5 碰撞轨迹

前面已经提到，音乐信号里的两个声音的谐音经常会重叠。谐音重叠产生的正弦轨迹，称作碰撞轨迹。谐音是否重叠取决于声音的间隔。如果具有类似大三度和

完全五度之间的间隔[34]，多数的低谐音就会重叠，因为它对于基本频率的比率是一个很小的整数的比率。在有不协和的间隔时，低谐音是不重叠的，但是它们还是非常接近，这就可能导致正弦模型中的估计误差。

不难发现，探测重叠正弦的有效途径是找到对两个声音谐音匹配的轨迹，或者两个声音的谐音距离很小的轨迹：

$$\frac{1}{|S_1|}\sum_{j\in S_1} d_h(i,j)+\frac{1}{|S_2|}\sum_{k\in S_2} d_h(i,k)<c_{limit} \tag{7-8}$$

其中 S_1 和 S_2 是属于声音 1 和声音 2，c_{limit} 是常数。如果 p_i 轨迹方程成立，那么 p_i 就可能包含了来自两个声音的谐音分音。

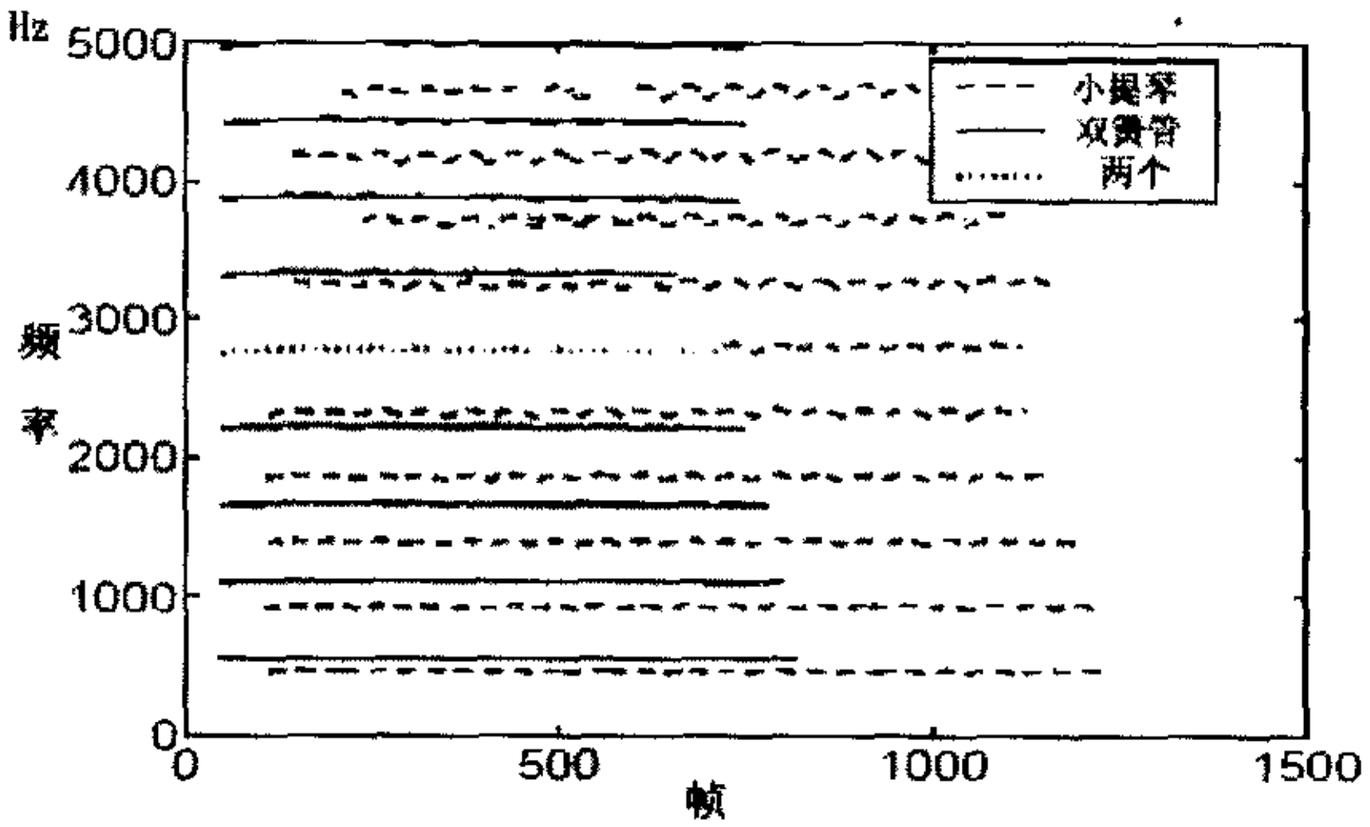

图 7—3 轨迹分类

从两个或两个以上在频率上互相接近的正弦波检测到的振幅，会受正弦波的相位差影响。在实际的声音中，由于频率和幅度的调制，精确的幅度和频率估计是很复杂的，因为经过正弦建模，就可以不再有精确的频谱信息，只有检测到的正弦。

全面的解决波形重叠问题的方法是超出本文的范围的[22]，不过要得到较好的感知质量，可以对重要的谐音做近似估计。本文用其它的非碰撞的正弦波的振幅曲线插入碰撞轨迹的幅度，可以得到完整的的频率。

最后，当检测并分离碰撞的轨迹后，就可以重构分离的信号并合成它们，分离轨迹的介绍见图 7—4。

通过实验和分离声音的感知质量证明，这些方法可以用来产生有用的结果。残留的问题将在以后做解决，包括多混和声音数量的动态检测，碰撞频率分音振幅的较好估计和同时发生的声音的分离等。

7．6 用多音高估计分离

由于使用单正弦模型的信号的分离，在多声音的情况下会变得很困难，所以建

完全五度之间的间隔[34]，多数的低谐音就会重叠，因为它对于基本频率的比率是一个很小的整数的比率。在有不协和的间隔时，低谐音是不重叠的，但是它们还是非常接近，这就可能导致正弦模型中的估计误差。

不难发现，探测重叠正弦的有效途径是找到对两个声音谐音匹配的轨迹，或者两个声音的谐音距离很小的轨迹：

$$\frac{1}{|S_1|}\sum_{j\in S_1} d_h(i,j)+\frac{1}{|S_2|}\sum_{k\in S_2} d_h(i,k)<c_{limit} \tag{7-8}$$

其中 S_1 和 S_2 是属于声音 1 和声音 2，c_{limit} 是常数。如果 p_i 轨迹方程成立，那么 p_i 就可能包含了来自两个声音的谐音分音。

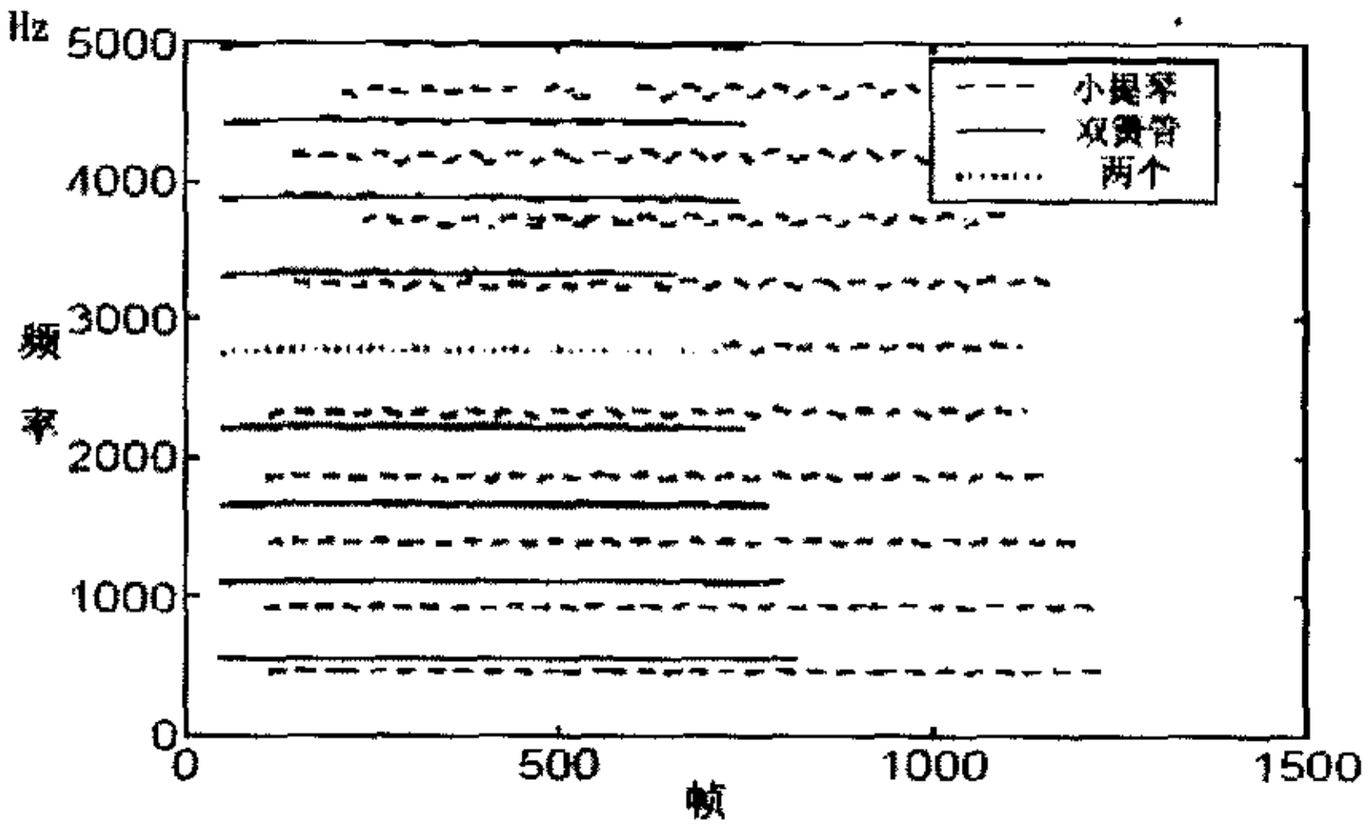

图 7—3 轨迹分类

从两个或两个以上在频率上互相接近的正弦波检测到的振幅，会受正弦波的相位差影响。在实际的声音中，由于频率和幅度的调制，精确的幅度和频率估计是很复杂的，因为经过正弦建模，就可以不再有精确的频谱信息，只有检测到的正弦。

全面的解决波形重叠问题的方法是超出本文的范围的[22]，不过要得到较好的感知质量，可以对重要的谐音做近似估计。本文用其它的非碰撞的正弦波的振幅曲线插入碰撞轨迹的幅度，可以得到完整的的频率。

最后，当检测并分离碰撞的轨迹后，就可以重构分离的信号并合成它们，分离轨迹的介绍见图 7—4。

通过实验和分离声音的感知质量证明，这些方法可以用来产生有用的结果。残留的问题将在以后做解决，包括多混和声音数量的动态检测，碰撞频率分音振幅的较好估计和同时发生的声音的分离等。

7．6 用多音高估计分离

由于使用单正弦模型的信号的分离，在多声音的情况下会变得很困难，所以建

立了同时使用基础频率和它们的谐音分音的方法[32]。实际使用的系统和标准正弦模型有一些不同。谐音分量的频率由多基因估计量产生（MPE），可以推导找出哪个分量是属于哪个声音。这样，就不是必须使用峰值检测了，使用 LSQ 算法可以得到分量的振幅和相位。也就不是必须使用峰值延续，因为已经假定谐音分量的频率包含在 MPE 窗口中，而 MPE 窗口比单独的正弦模型帧要长的多。不过，该方法无法检测到基因频率的过小的改变，例如颤音。

参数估计完以后，由多谐音分音产生的正弦可以从混合声音中得到。如果两个分量的频率不完全相同，全部分量振幅包络总和以分量频率差调制。假定原始振幅包络是缓慢变化的，就可以通过以下的方法来解决：混和的分量通过低通滤波器获

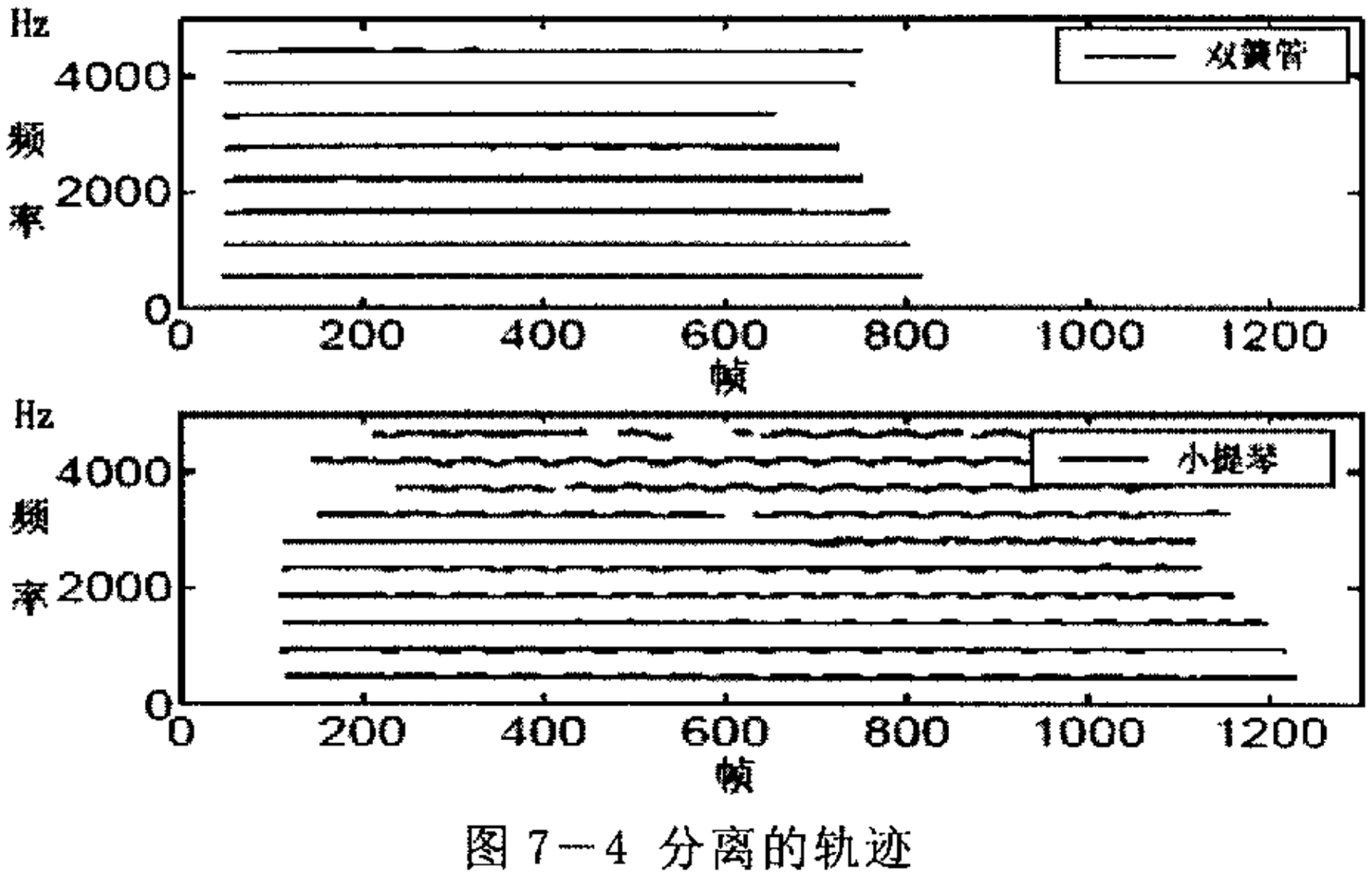

图 7—4 分离的轨迹

得到第一个幅度包络，剩余的是从原始信号减去前边得到的，然后半波矫正，低通滤波滤除差值。接下来通过比较其它的不再重叠的振幅包络，将两个分离的幅度曲线和它们各自产生的声源联系。这种比较可以通过使用前面提到的感知测量法，如果超过两个的谐音分量重叠，它们的振幅只是利用对每个声音处理的分量简单的插值。

7．7 音高和时间尺度修改

正弦模型和随机模型允许改变音高而不影响时间尺度，或者改变时间尺度而不影响音高。这个改变是对参数数据而言，这样可以分解原始声音信号，对必要的参数进行修改然后合成该信号。这种修正信号后合成信号的质量和没有修正的合成信号的质量一样。同时，修正也是非常简单的：不需要 FFT 或者窗口，只是一组乘法和加法。

得到了确定部分的频率 $\omega(t,i)$，振幅 $a(t,i)$ 和相位 $\varphi(t,i)$，不确定部分的 BARK 带能量 $S(t,i)$，对修正来说，还需要跃阶 S。通过因数 ρ_t 拉长时间尺度，这意味着

立了同时使用基础频率和它们的谐音分音的方法[32]。实际使用的系统和标准正弦模型有一些不同。谐音分量的频率由多基因估计量产生（MPE），可以推导找出哪个分量是属于哪个声音。这样，就不是必须使用峰值检测了，使用 LSQ 算法可以得到分量的振幅和相位。也就不是必须使用峰值延续，因为已经假定谐音分量的频率包含在 MPE 窗口中，而 MPE 窗口比单独的正弦模型帧要长的多。不过，该方法无法检测到基因频率的过小的改变，例如颤音。

参数估计完以后，由多谐音分音产生的正弦可以从混合声音中得到。如果两个分量的频率不完全相同，全部分量振幅包络总和以分量频率差调制。假定原始振幅包络是缓慢变化的，就可以通过以下的方法来解决：混和的分量通过低通滤波器获

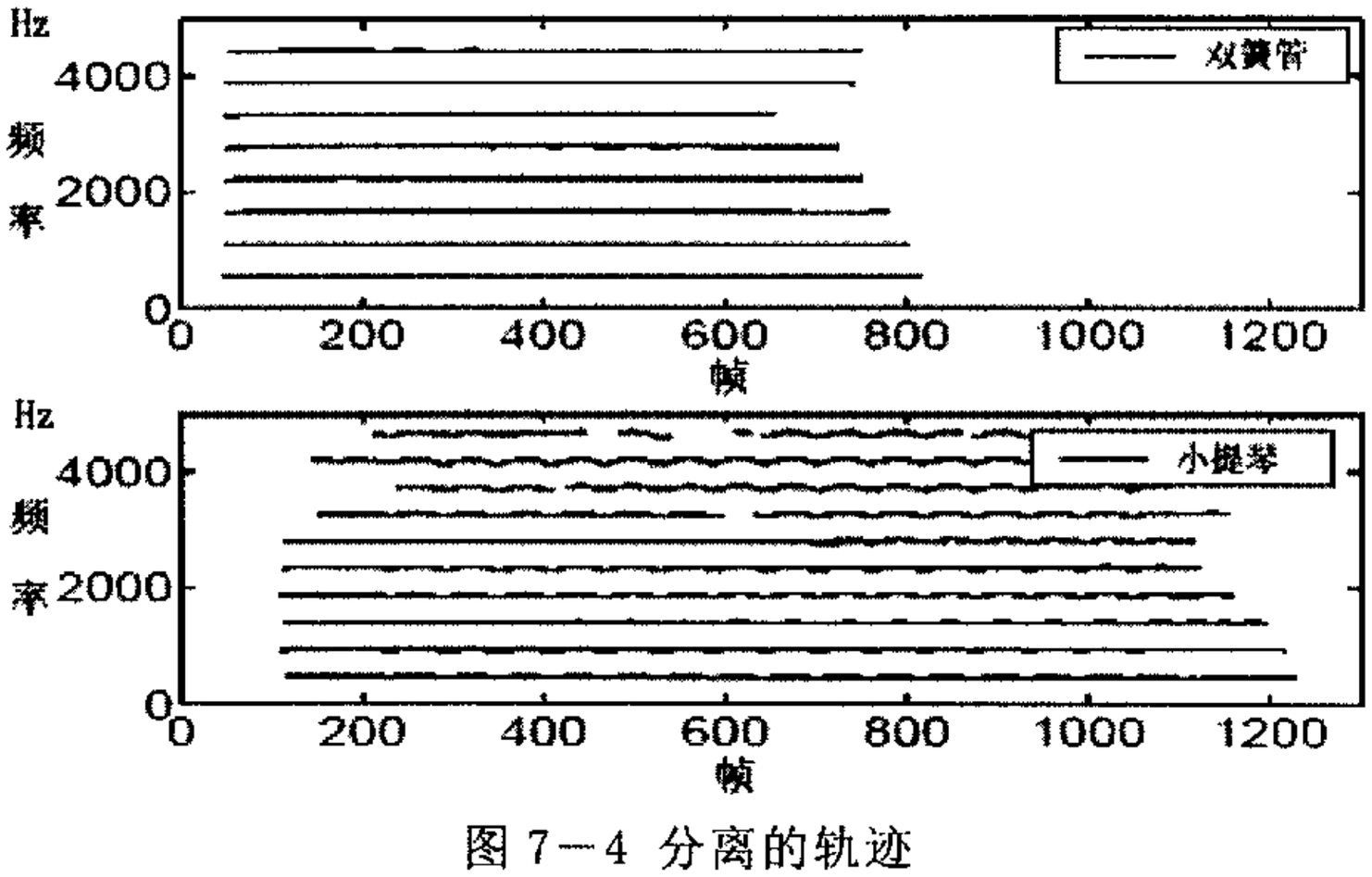

图 7—4 分离的轨迹

得到第一个幅度包络，剩余的是从原始信号减去前边得到的，然后半波矫正，低通滤波滤除差值。接下来通过比较其它的不再重叠的振幅包络，将两个分离的幅度曲线和它们各自产生的声源联系。这种比较可以通过使用前面提到的感知测量法，如果超过两个的谐音分量重叠，它们的振幅只是利用对每个声音处理的分量简单的插值。

7．7 音高和时间尺度修改

正弦模型和随机模型允许改变音高而不影响时间尺度，或者改变时间尺度而不影响音高。这个改变是对参数数据而言，这样可以分解原始声音信号，对必要的参数进行修改然后合成该信号。这种修正信号后合成信号的质量和没有修正的合成信号的质量一样。同时，修正也是非常简单的：不需要 FFT 或者窗口，只是一组乘法和加法。

得到了确定部分的频率 $\omega(t,i)$，振幅 $a(t,i)$ 和相位 $\varphi(t,i)$，不确定部分的 BARK 带能量 $S(t,i)$，对修正来说，还需要跃阶 S。通过因数 ρ_t 拉长时间尺度，这意味着

原始信号的长度T变成$\rho_t T$。另外，通过因数ρ_ω改变信号的音高，或者通过因数ρ_ω乘以基本频率，在音乐中，用$\rho_\omega = 2^{(s/12)}$得到变化的s个半音。

假定信号的非确定部分在音高改变的时候保持不变，这样 BARK 带能量就不需要改变音高。对正弦曲线，新的频率$\omega'(t,i)$只是在以前的频率乘以音高改变因子：

$$\omega'(t,i) = \rho_\omega \omega(t,i) \tag{7-9}$$

修改技术不能保留信号的共振峰结构。正弦和随机分量的合成允许通过把阶跃长度S乘以时间拉伸因子改变时间尺度：

$$S' = \rho_t S \tag{7-10}$$

然而，如果音高或者时间尺度发生改变，实际的波形不能保留，所以必需由修正频率的积分得到的相位：

$$\varphi'(t,i) = S' \sum_{\theta=0}^{i} \omega'(t,i) \tag{7-11}$$

模型的修正功能还需要大量的测试。

第八章 结论

本文研究了正弦加噪模型，目标是应用它来做一个中水平的计算机听觉场景分析。论文的目的就是研究现存的分解合成的算法，试着做一点改善，模型的可用性在声音分离和实验中得到验证。

整个正弦加噪模型采用了人类对声音的感知特性，但是特殊的正弦模型可以考虑为物理的模型而不是心里声学模型。对于复杂的真实信号，在一个单分解帧里检测有意义的峰并且估计它们的参数是很困难的。本文中，使用了几乎所有的用在正弦模型上的方法：首先检测有意义的峰，然后独立的跟踪到轨迹内。在人类对于声音的感知中，这个过程是困难的；声音的幅度和长度，还有其它的干扰音对感知的影响非常大，轨迹滤波器是系统中唯一的试着用来处理这种现象的部分。

即使几个先进的正弦分解算法结合在一起，最终的实验结果表明，没有一个是正弦分解的最好的方法。还有许多基础的问题在参数估计中需要解决，大部分是相关于有限的时间和频率分辨率。

感知上合成声音的质量对于高质量的语音编码还不够好，但是模型想要的只是到达中级水平，重要的是它在重构信号时减少了大量的数据。

实验表明系统是可以用在声音分离上的，单独使用正弦模型分离是有限制的，只有在混和声音的数量很少时能产生好的结果。这时就要使用多音高估计来分离，但是，对于有丰富多音的真实信号要达到高质量声音分离还是有很多工作要做。

我们在分解和合成系统的发展和实现中学到了大量的声音信号处理知识。下一步的研究是这个系统在声音分离和其他计算机听觉场景分析领域发展。

参考文献

[1]X.Serra.Musical Sound Modeling with Sinusoids plus Noise.Roads C. & Pope S. & Picialli G. & De Poli G. (eds).Musical Signal Processing. Swets & Zeitlinger Publishers, 1997

[2]J.Sillanpää,A.Klapuri,J.Seppänen and T.Virtanen,Recognition of Acoustic Noise Mixtures by Combined Bottom-up and Top-down Processing. *European Signal Processing Conference, Tampere, Finland 2000*

[3]D.Ellis and D.Rosenthal,Mid-level representations for Computational Auditory Scene Analysis. International Joint Conference on Artificial Intelligence-Workshop on Computational Auditory Scene Analysis, Montreal, Quebec, August 1995

[4]C.Roads,The Computer Music Tutorial,MIT Press,Cambridge, Massachusetts,USA, 1995

[5] Y.Li,K.Sugahara,T.Osaki and R.Konishi,On the frequency estimation of signal in the noisy circumstance SICE 2001. Proceedings of the 40th SICE Annual Conference. International Session Papers ,2001.7.25-27

[6] 周江扬，柴佩琪. 基于正弦模型的汉语语音时长和音高的修正. 同济大学学报(自然科学版)，2001，29 (3)：312～316

[7]S.Abeysekera,K.P.Padhi,J.Absar and S.George,Investigation of different frequency estimation techniques using the phase vocoder Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on , Volume: 2 ,2001.5.6-9

[8] R.J.McAulay and T.F.Quatieri,Speech Analysis/Synthesis Based on a Sinusoidal Representation. IEEE Transactions on Acoustics,Speech, And Signal Processing, Volume: 34(4),1986.8.744-754

[9]J.O.Smith and X.Serra,PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation, Proceedings of the International Computer Music Conference, 1987

[10] X.Serra.A system for analysis/ transformation /synthesis based on a deterministic plus stochastic decomposition. Ph.D. thesis, Stanford

University, 1989

[11]S.Levine.Audio Representation for Data Compression and Compressed Domain Processing. Ph.D. thesis. Stanford University, 1998

[12]T.Abe and M.Honda.Sinusoidal modeling based on instantaneous frequency attractors. 2003 IEEE International Conference on , Volume: 6,2003.4.6-10

[13]T.S.Verma.A Perceptually Based Audio Signal Model with Application to Scalable Audio Compression.Ph.D.thesis.Stanford University, October 1999

[14]J.Laroche.Time and Pitch Scale Modifications of Audio Signals.Kahrs, M. & Branderburg, K. (eds). Applications of Digital Signal Processing to Audio and Acoustics. Kluwer Academic Publishers, Boston / Dordrecht / London, 1998

[15] D.Griffin and J.Lim.A New Model-Based Speech Analysis/Synthesis System. IEEE International Conference on Acoustics, Speech, and Signal Processing,Tampa, Florida 1985

[16]X.Rodet.Musical Sound Signal Analysis/Synthesis: Sinusoidal+ Residual and Elementary Waveform Models. IEEE Time-Frequency and Time-Scale Workshop 1997, Coventry, Grande Bretagne, 1997

[17]W.M.Hartmann.Signals, Sound, and Sensation. Springler-Verlag New York Inc., 1997

[18]R.Meddis and J.Hewitt.Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. Journal of Acoustic Society of America, June 1991

[19]M.Desainte-Catherine and S.Marchand,High-Presicion Fourier Analysis of Sounds Using Signal Derivatives. Journal of Acoustic Engineering Society, Volume 48(7), 2000.7/8

[20]Ph.Depalle and T.Hélie.Extraction of Spectral Peak Parameters Using a Short-Time Fourier Transform And No Sidelobe Windows. IEEE 1997 Workshop on Applications of Signal Processing to Audio and Acoustics. Mohonk, New York,1997

[21]T.Tolonen.Methods for Separation of Harmonic Sound Sources using Sinusoidal Modeling. AES 106th Convention, Munich, Germany,1999.5

[22]M.Kay.Fundamentals of Statistical Signal Processing: Estimation Theory.PTR Prentice-Hall, Englewood Cliffs, New Jersey 1993

[23]T.Virtanen.Accurate Sinusoidal Model Analysis and Parameter Reduction by Fusion of Components, to be presented in AES 110th convention, Amsterdam, Netherlands,2001.5

[24]J.C.Brown.Calculation of a constant Q Spectral transform. Journal, of Acoustic Society of America, Volume 89(1), 1991,1

[25]J.C.Brown and M.S.Puckette. An efficient algorithm for the calculation of a constant Q transform. Journal of Acoustic Society of America, Volume 92(5), 1992.11

[26]Ph.Depalle,G.Garcia and X.Rodet.Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Minneapolis, Minnesota,USA 1993

[27]B.C.J.Moore.An Introduction to the Psychology of Hearing. Academic Press, 1997

[28]郑新春，柴佩琪. 基于正弦模型的语音频域参数编码. 计算机应用与软件 2002,7：53～56

[29]C.Colomes,M.Lever,J.B.Rault and Y.F.Dehery.A Perceptual Model Applied to Audio Bit-Rate Reduction. Journal of Audio Engineering Society, Volume 43(4), New York, 1995.10

[30]E.Zwicker and H.Fastl.Psychoacoustics: Facts and Models.Springer-Verlag Berlin Heidelberg, 1999

[31]T.Virtanen and A.Klapuri.Separation of Harmonic Sound Sources Using Sinusoidal Modeling. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey 2000

[32]A.Klapuri and T.Virtanen and J.A.Holm.Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Musical Signals. Proceedings of the COST G-6 Conference on Digital Audio Effects, Verona,

Italy, 2000.12

[33]A.S.Bregman.Auditory Scene Analysis. MIT Press, 1990

[34]A.Klapuri.Number Theoretical Means of Resolving a Mixture of Several Harmonic Sounds. Proceedings of the European Signal Processing Conference,1998

致　　谢

硕士研究生的学习和生活就要随着这篇论文的答辩而结束了。有许许多多的舍不得，也有许许多多的感谢要说。

首先要衷心感谢的是我可敬可亲的导师许刚老师！您三年来对我学习和研究的悉心指导和谆谆教诲令我终身受益。在您的指导下，我在各方面的能力都得到了相应的提高。您的睿智、对知识孜孜不倦的追求、对教育科学研究的热爱、严谨的治学态度让我学到了如何做事，您在生活中的幽默、宽容、豁达教会了我如何做人。千言万语在此刻化为了一句谢谢您！。

感谢所有教育过我的老师！你们传授给我的专业知识是我不断成长的源泉，也是完成本论文的基础。

感谢所有在 623 学习和生活过的同门：郑普亮、孟静、彭柏、王珊、何玉斌、穆雨等！特别要感谢的是和我一起学习和生活的杜丹同学，感谢她多年来对我的支持和鼓励。还要特别感谢王晓强——这个和我一起学习和生活了的好朋友、好舍友、好同学，你和我一起经历了许多重要的时刻，你在我论文完成的过程中给了我许多鼓励和帮助。

感谢我的父母！你们给我生活上的关怀和精神上的鼓励是我学习的动力。

再次对所有关心、帮助我的人说一声谢谢。

附录 1：两个正弦的融合：等式推导

从两个正弦的的总和开始，包括幅度、频率和相位a_1、a_2、ω_1、ω_2、φ_1和φ_2，我们用一个正弦重构正弦和，它的幅度和相位是时变的。正弦的和在时间t表示为$x(t)$：

$$x(t)=a_1\sin(\omega_1 t+\varphi_1)+a_2\sin(\omega_2 t+\varphi_2) \tag{1}$$

让$\operatorname{atan}\left(\frac{a_1}{a_2}\right)=\Phi$，$a_1\neq 0$，$a_2\neq 0$。则幅度是

$$a_1=\sqrt{a_1^2+a_2^2\sin(\Phi)}\,,\quad a_2=\sqrt{a_1^2+a_2^2\cos(\Phi)}\,。$$

$x(t)$变成

$$x(t)=\sqrt{a_1^2+a_2^2}\left[\sin(\Phi)\sin(\omega_1 t+\varphi_1)+\cos(\Phi)\sin(\omega_2 t+\varphi_2)\right]$$

通过三角变换，得到

$$x(t)=\frac{\sqrt{a_1^2+a_2^2}}{2}\left[\cos(\Phi-\omega_1 t-\varphi_1)-\cos(\Phi+\omega_1 t+\varphi_1)+\sin(\Phi+\omega_2 t+\varphi_2)-\sin(\Phi-\omega_2 t+\varphi_2)\right]$$

$$=\frac{\sqrt{a_1^2+a_2^2}}{2}$$

$$\left[\sin\left(\Phi-\omega_1 t-\varphi_1+\frac{\pi}{2}\right)-\sin\left(\Phi+\omega_1 t+\varphi_1+\frac{\pi}{2}\right)+\sin(\Phi+\omega_2 t+\varphi_2)-\sin(\Phi-\omega_2 t+\varphi_2)\right]$$

$$=\sqrt{a_1^2+a_2^2}\left[\cos\left(\Phi+\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}+\frac{\pi}{4}\right)\sin\left(\frac{(\omega_1-\omega_2)t+\varphi_1-\varphi_2}{2}\right)-\frac{\pi}{4}\right]$$

$$+\cos\left(\Phi-\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}+\frac{\pi}{4}\right)\sin\left(\frac{(\omega_1-\omega_2)t+\varphi_1-\varphi_2}{2}+\frac{\pi}{4}\right)$$

$$=\frac{\sqrt{a_1^2+a_2^2}}{2}\left[\sin\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}-\frac{\pi}{4}\right)\cos\left(\Phi+\frac{\pi}{4}\right)\right]\cos\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)$$

$$+\cos\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}-\frac{\pi}{4}\right)\sin\left(\Phi+\frac{\pi}{4}\right)\sin\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)$$

因为

$$\cos\left(\Phi+\frac{\pi}{4}\right)=\cos\left(a\tan\frac{a_1}{a_2}+\frac{\pi}{4}\right)=\frac{1}{\sqrt{2}\sqrt{\frac{a_1^2}{a_2^2}+1}}-\frac{\frac{a_1}{a_2}}{\sqrt{2}\sqrt{\frac{a_1^2}{a_2^2}+1}}=\frac{a_2-a_1}{\sqrt{2}\sqrt{a_1^2+a_2^2}}$$

和

$$\sin\left(\Phi+\frac{\pi}{4}\right)=\frac{a_2+a_1}{\sqrt{2}\sqrt{a_1^2+a_2^2}}$$

表达式变为：

$$x(t)=\left[\sin\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2-a_1)\cos\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)\right.$$
$$\left.+\cos\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2+a_1)\sin\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)\right] \quad (2)$$

设$x(t)$等于只有一个正弦的表达式，幅度是a_3相位是$\varphi_3(t)$，变为：

$$x(t)=a_3(t)\sin\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}+\varphi_3(t)\right) \quad (3)$$

$$=a_3(t)\left[\sin\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)\cos(\varphi_3(t))+\cos\left(\frac{(\omega_1+\omega_2)t+\varphi_1+\varphi_2}{2}\right)\sin(\varphi_3(t))\right] \quad (4)$$

从等式（2）和（3）我们得到：

$$a_3(t)\cos\varphi_3(t)=\cos\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2+a_1) \quad (5)$$

$$a_3(t)\sin\varphi_3(t)=\sin\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2-a_1) \quad (6)$$

简化得到a_3为：

$$a_3(t)=\sqrt{\left[\sin\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2-a_1)\right]^2+\left[\cos\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)(a_2+a_1)\right]^2}$$

$$=\sqrt{a_1^2+a_2^2+a_1a_2\left[\cos^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)-\sin\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\right]} \quad (7)$$

$$=\sqrt{a_1^2+a_2^2+2a_1a_2\cos\left((\omega_2-\omega_1)t+\varphi_2-\varphi_1\right)}$$

用（6）除以（5）又有：

$$\tan(\varphi_3(t))=\tan\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)} \quad (8)$$

从中，我们能用－tan 得到相位 φ_3。因为－tan 的区间在 $[-\pi/2,\pi/2]$，负的幅度要加入修整项 Φ：

$$\Phi=\begin{cases}\pi & \frac{\pi}{2}<\frac{\varphi_2-\varphi_1}{2}\bmod 2<\frac{3\pi}{2}\\ 0 & \text{其它}\end{cases}$$

相位的等式变为：

$$\varphi_3(t)=a\tan\left(\tan\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\Phi \tag{9}$$

用时变的相位，我们能用一个时变的幅度和相位表示正弦的总和。设时间 $t=0$ 时初始相位为：

$$\varphi_3(0)=a\tan\left(\tan\left(\frac{\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\Phi \tag{10}$$

因为瞬时频率是相位的导数，由式（9）得到瞬时频率 $\omega_3(t)$：

$$\begin{aligned}\omega_3(t)&=\frac{d}{dt}(\varphi_3(t))=\frac{d}{dt}\left(a\tan\left(\tan\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\right)+\Phi\right)\\&=\frac{\frac{d}{dt}\tan\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}}{1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\left[\frac{(a_2-a_1)}{(a_2+a_1)}\right]^2}\\&=\frac{\left[1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\right]\frac{(a_2-a_1)}{(a_2+a_1)}\frac{d}{dt}\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}}{1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\left[\frac{(a_2-a_1)}{(a_2+a_1)}\right]^2}\\&=\frac{1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)}{1+\tan^2\left(\frac{(\omega_2-\omega_1)t+\varphi_2-\varphi_1}{2}\right)\left[\frac{(a_2-a_1)}{(a_2+a_1)}\right]^2}\left(\frac{\omega_2-\omega_1}{2}\right)\frac{(a_2-a_1)}{(a_2+a_1)}\end{aligned} \tag{11}$$

现在我们可以用一个时变幅度和频率的正弦表示两个正弦的总和：

$$x(t)=a_3(t)\sin\left(\frac{(\omega_2+\omega_1)t+\varphi_2+\varphi_1}{2}+\int_0^t\omega_3(u)\,du+\varphi_3(0)\right) \tag{12}$$

附录 2：主要符号表

序号	符号名	意　义
1	PCM	(pulse code modulated) 脉冲编码调制
2	DFT	(discrete Fourier transform) 离散傅立叶变换
3	STFT	(short-time Fourier transform) 短时离散傅立叶变换
4	FFT	(fast Fourier transform) 快速傅立叶变换
5	FIR	(finite impulse response) 有限脉冲响应
6	IIR	(infinite impulse responed) 无限脉冲响应
7	LSQ	(Least Square) 最小平方
8	MPE	(multipitch estimator) 多音高估计量

在学期间发表的学术论文和参加科研情况

[1]李文华，许刚，郑普亮.低速率终端设备的语音信号数字编码.现代电力，2004，21（4）：66～70

[2]李文华，许刚.利用 ADPCM 进行网络环境下实时多点语音通信.计算机工程与应用，2004