

清 华 大 学

综 合 论 文 训 练

题目：单声道音频中人声与伴奏的分离

系 别：电子工程系

专 业：电子信息科学与技术

姓 名：王 赞

指导教师：欧智坚 副研究员

2010 年 6 月 30 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 王赞 导师签名： 陈智勇 日 期： 2010.6.30

中文摘要

声源分离对语音识别、音频内容分析与检索等有重要作用，人声与伴奏的分离是声源分离的一种重要情况。本文研究的课题是从单声道音频中分离出人声与伴奏。

本文首先综述了几种常见的单声道声源分离方法。由于基音轨迹的提取是声源分离的关键步骤，本文也对一些基音轨迹提取算法进行了评价。本研究发现，声源分离的软蒙版法和提取基音轨迹的 **HMM** 分类法具有比较明显的优势。

本文实现了一个完整的单声道人声伴奏分离系统。它采用 **HMM** 分类方法进行伴奏、清音、浊音的判决和基音轨迹的提取，并对人声建立了一种基于非负矩阵分解的源-滤波器模型，使用软蒙版法分离人声与伴奏。在两个国际公开数据库上的性能测试表明，本分离系统的性能达到了人声伴奏分离的前沿水平，并具有更快的处理速度。

论文最后指出了单声道音频中人声和伴奏分离的进一步研究方向，即更准确地提取复杂音频中的基音轨迹，以及更好地实现清音的分离。

关键词：单声道声源分离；源-滤波器模型；基音轨迹提取；**HMM** 分类；基频显著性函数

ABSTRACT

Sound source separation makes a significant contribution to tasks such as speech recognition, and content-based audio analysis and retrieval. The separation of human voice and accompaniment is an important case of sound source separation, and this paper deals with single-channel audios.

This paper first reviews several common approaches to single-channel source separation. Because of the important role of pitch tracking in source separation, a few pitch tracking methods are evaluated. Our research discovers that the soft-masking approach for source separation and the HMM classification approach for pitch tracking display apparent advantages.

A full single-channel separation system is implemented. This system employs HMM classification to perform accompaniment / unvoiced / voiced decision and pitch tracking, and, after building an NMF-based source-filter model for human voice, separates the voice and the accompaniment by soft-masking. Evaluation on two globally available databases shows that our system reaches state-of-the-art performance, and achieves a faster speed.

At the end of the paper we point out the directions for further research: how to track the pitch contour more accurately in complicated audios, and how to separate the unvoiced parts in the human voice.

Keywords: Single-channel sound source separation; source-filter model; pitch tracking; HMM classification; F0 salience function

目 录

第 1 章 研究任务	1
第 2 章 研究现状	2
2.1 声源分离算法分类概述	2
2.1.1 蒙版分离法	2
2.1.2 几种非蒙版分离法简介	8
2.2 基音轨迹提取方法综述	9
2.2.1 基音提取算法分类	9
2.2.2 基音轨迹提取的难点及解决方案	10
第 3 章 人声伴奏分离系统的实现	13
3.2 利用 MFCC 特征进行 A/U/V 判决	14
3.2.1 MFCC 特征的提取	14
3.2.2 HMM 模型的建立与使用	15
3.2.3 A/U/V 判决的效果	16
3.3 利用 ESI 特征提取基音轨迹	16
3.3.1 用谐波叠加法计算显著性函数	17
3.3.2 ESI 特征的计算和 HMM 模型的建立	19
3.4 利用源-滤波器模型分离人声与伴奏	20
3.4.1 “功率谱”的两种含义	20
3.4.2 源-滤波器模型的建立	21
3.4.3 根据源-滤波器模型估计本质功率谱	21
3.4.4 求软蒙版	23
3.4.5 算法中一些参数的选择	25
第 4 章 性能评价	26
4.1 A/U/V 判决的评价	26

4.2 基音轨迹提取的评价	27
4.3 人声伴奏分离的评价	28
4.3.1 评价标准	28
4.3.2 测试集	30
4.3.3 测试结果	31
4.4 效率评价	35
第 5 章 讨论	36
5.1 A/U/V 判决和基音提取能否同时进行	36
5.2 用 GMM 模型建模 ESI 特征是否合适	37
5.3 关于清音的分离	38
第 6 章 结论	39
插图索引	40
表格索引	41
参考文献	42
致 谢	44
声 明	错误！未定义书签。
附录 A 外文资料的书面翻译	46
附录 B KLGLOTT88 模型	58

第1章 研究任务

声源分离，是指从多个声源发出的声音信号叠加形成的混合信号中分离出各个声音信号的过程。声源分离的目的，往往是对分离出来的各个信号进行后续的处理。由于去除了其它声源的信号的干扰，这种处理会变得简单得多。例如，在有噪声的语音识别系统中，噪声往往会严重降低识别率。如果首先使用声源分离技术去掉噪声，得到干净的语音，识别率就可以大大提高。

声源分离有多种情况。按声源类型来分，可分为 a) 歌声与伴奏的分离、b) 人声与背景音乐的分离、c) 人声与背景噪声的分离、d) 多人说话声的分离等等。分离的目的，可能只是要得到其中一部分声源发出的信号，也可能是要得到全部声源发出的信号。按声道的数目来分，可分为单声道分离、双声道分离、多声道分离等。不同的情况下，声源分离任务的难度可能有很大差别。

本文的任务是 a)，即从单声道音频中，分离人声和伴奏两个声源。并且，限定音频中在任一时刻，只有一个人在发声（即没有和声、合唱等）。分离的目标是同时获得人声、伴奏两个声源的信号。之所以选择单声道音频作为研究对象，是因为在现实应用中并不总是容易获得双声道的声音材料。

本文所进行的研究的用途是多方面的。分离获得的人声可用于语音识别，进而用于歌词与旋律对齐；也可以用作语音合成的材料。分离获得的伴奏可用于卡拉 OK、演出等场合。

本文的结构如下：

第 2 章综述目前单声道人声与伴奏分离的几类主要方法；

第 3 章详细叙述我们的人声伴奏分离系统的实现；

第 4 章测试系统性能，并与基准方法进行比较；

第 5 章讨论我们为提高性能所做的一些尝试和可能的改进方案；

第 6 章总结全文。

第2章 研究现状

2.1 声源分离算法分类概述

2.1.1 蒙版分离法

蒙版分离法 (separation by masking) 是声源分离算法中的一个重要类别。它的基本步骤如下：

- 首先将混合信号分解成一系列时频单元，得到其时频域表示；
- 然后按某种规则将每个时频单元中的信号按一定比例分配给各个声源；
- 最后根据每个声源分得信号的时频域表示重新合成时域信号。

用 x 表示混合信号， \hat{x}_k 表示第 k 个声源分得的信号，而它们的时频域表示记作 \mathbf{X} 和 $\hat{\mathbf{X}}_k$ (均为二维矩阵，行表示频带，列表示帧)，则蒙版分离法的核心可以用以下公式概括：

$$\hat{\mathbf{X}}_k = \mathbf{W}_k \mathbf{X} \quad (2-1)$$

其中矩阵 \mathbf{W}_k 在时频单元 (f, t) 处的元素 W_{kft} 就是此单元信号分给声源 k 的比例。对于每个声源 k ，这些比例系数组成的矩阵 \mathbf{W}_k 可以看成是一个“蒙版”，这也是就是“蒙版分离法”名称的由来。

根据对蒙版元素取值的限制，可以将蒙版分离法分为“硬蒙版法”(hard masking) 和“软蒙版法”(soft masking)。“硬蒙版法”又称为“二进制蒙版法”(binary masking)，它要求在同一时频单元处，各个蒙版中有且仅有一个元素为 1，其它元素均为 0。也就是说，每个时频单元中的信号只能完整地分配给某一个声源。与此相对，“软蒙版法”则允许蒙版取 0 和 1 以外的值。一般地，要求软蒙版中的元素取 $[0, 1]$ 区间内的实数，且同一时频单元处各蒙版值的和为 1。其实，这些要求也可以放松。比如，可以允许在同一时频单元处各蒙版元素之和不为 1，此时分离出的各声源叠加不等于混合信号（称为不满足“保守性”）；在采用语谱图作为时频域表示（见下）时，也可以允许蒙版元素为复数，此时分离出的各声源的相位谱与混合信号不同。

采用蒙版法进行声源分离，需要回答下面两个问题：一是如何把信号分解为时频单元，二是如何生成蒙版。第二个问题的答案自然是随具体方法而千变万化的。而信号的分解则主要有两种方法：一是直接进行短时傅里叶变换 (STFT)，

得到语谱图 (spectrogram)，语谱图中的每一个系数就代表了一个时频单元。另一种是采用一个带通滤波器组对混合信号进行滤波，然后把每个频带的输出分帧、加窗，得到时频单元。滤波器组分解法与语谱图分解法相比，有更大的灵活性，如：

- 语谱图分解法的每个时频单元仅用一个傅里叶系数表示，这就限制了其时域波形必须为（加窗）正弦波，而滤波器组分解法的时频单元波形比较自由；
- 由于语谱图分解法是基于 STFT 的，它的频带必须等距等宽，而滤波器组分解法的频带可以自由选取；
- 语谱图分解法的帧长等于 FFT 的长度，频带数为 FFT 长度的一半加 1，这些关系是固定的。而滤波器组分解法不需要进行 FFT，其帧长与频带数独立；

但是，滤波器组分解法的这些灵活性，是与其极高的时间、空间复杂度为代价的。设样本数为 N ，帧数为 T ，帧长为 L ，频带数为 F ，滤波器阶数为 K ，则语谱图分解法的时间复杂度为 $O(TF \log F) \approx O(N \log F)$ ，而滤波器组分解法的时间复杂度为 $O(NFK)$ ，二者可以差出若干个数量级。在空间复杂度方面，语谱图分解法的每个时频单元仅用一个复数表示，而滤波器组分解法则需要 L 个实数。

下面分别介绍几种具体的硬蒙版法和软蒙版法。

2.1.1.1 硬蒙版法

表面上看来，硬蒙版法似乎没有什么道理，因为混合信号的一个时频单元中，一般会包含来自所有声源的信号，把它只分配给某一个声源是不对的。但事实上，硬蒙版法在声源分离的任务中也取得了相当大的成功。这要归功于语音信号所具有的一种性质——“近似窗分离正交性” [6]。称一组信号是“窗分离正交”

（W-disjoint orthogonal, WDO）的，如果它们在某种时频域表示下，每个时频单元处至多只有一个信号的值非零。完美的窗分离正交性当然是过于理想的，但对于语音来说，由于占其中大部分时间的浊音段是有谐波结构的，不同频率的语音的谐波结构重叠很少，所以语音信号具有近似的窗分离正交性。对于音乐信号来说，由于在同一时刻歌声与音乐的基频常常存在简单的整数比关系，所以其中人声和伴奏的窗分离正交性会稍差一些。尽管混合声音信号中的各个声源并不是完全的窗分离正交，但是，由于人耳的听觉具有掩蔽效应，当一个信号中混入其它

信号的能量不高时，人耳往往感觉不到干扰信号的存在。这就是硬蒙版法可行的原因。

假设我们已经知道各个声源的信号 x_k 及其时频域表示 \mathbf{X}_k ，我们可以用以下方法生成蒙版：

$$W_{kft} = \begin{cases} 1, & \text{if } k = \arg \max_k |X_{kft}| \\ 0, & \text{otherwise} \end{cases} \quad (2-2)$$

其含义是，把每个时频单元完全分配给对它贡献最大的声源。使用这样的蒙版时分配给错误声源的信号能量最小，因此该蒙版称为“理想蒙版”（ideal mask）。实验表明，使用理想蒙版分离各个声源，效果是可以接受的。理想蒙版是硬蒙版法的计算目标，也是硬蒙版法性能的上界。

下面介绍两个硬蒙版法的例子，这两个例子处理的都是两个声源的情况，即一个人声一个伴奏。

芬兰研究者 Virtanen、Ryynänen 等人（下称“芬兰派”）[3]提出了一种最简单的硬蒙版，可以称为“谐波蒙版”。它使用的时频域表示就是 STFT 得到的语谱图。在提取出基音轨迹之后，在每一帧处，把基频及其各个倍频处附近（如 $\pm 20\text{Hz}$ ）之内的时频单元的能量全部分配给人声，而其余能量全部分配给伴奏，就得到了蒙版。

“谐波蒙版”的构造非常简单，但效果却不够理想。由于人声与伴奏的谐波常有重叠，所以分离出的人声中会混有一定量的伴奏；又由于人声的调幅、调频等特性以及加窗产生的频谱泄漏，人声并不仅仅局限在谐波附近的几个时频单元内，所以分离出的伴奏中也混有相当多的人声。如果把谐波蒙版与理想蒙版画出来（图 2.1），也可以看出二者的差别非常大。当然，芬兰派并没有止步于此，他们对这种方法的改进将在 2.1.1.2 节叙述。

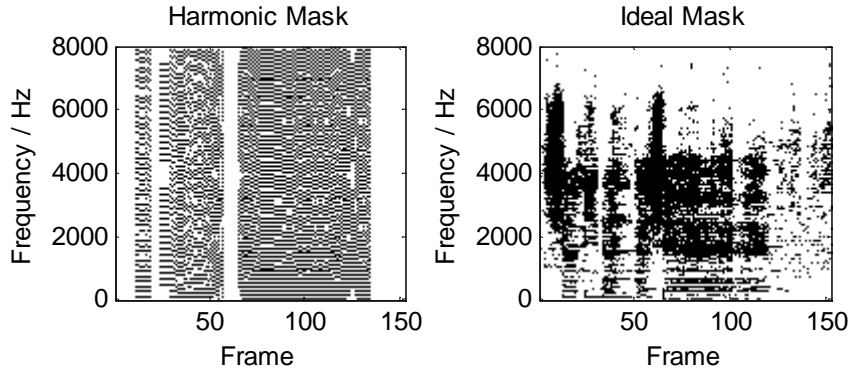


图2.1 谐波蒙版（左）与理想蒙版（右）的比较。

黑色部分代表人声蒙版为1，白色部分代表伴奏蒙版为1。

王德良以及他的学生们的工作[7][8]，可以说是代表了硬蒙版法的最前沿水平。他们采用的方法称为“计算听觉场景分析”（computational auditory scene analysis, CASA），意在尽可能模拟人耳处理声音信号时的行为。他们的信号分解和蒙版构造方法都比较复杂。

CASA 方法使用的信号分解方法为滤波器组方法。首先，混合信号被送入具有 64 或 128 个频带的 Gammatone 滤波器组，以模拟人耳的“临界频带”（critical bands）。这些滤波器组的中心频率是按对数尺度分布的，低频段较密集，高频段较稀疏；各个频带具有相当大的重叠。每个带通滤波器的输出又经过 Meddis 听神经传导模型，转换为听神经纤维的发射速率。将发射速率信号分帧、加窗，就得到时频域表示，称为“耳蜗图”（cochleagram），其中的每个时频单元都是一小段时域信号。

CASA 方法生成蒙版需要经过两个阶段：分块（segmentation）和组合（grouping）。

在分块阶段，首先要挑选出那些与相邻频带单元足够相似的那些时频单元，因为相邻而相似的单元一般会属于同一个声源。这种“相似”性是通过“频带间互相关”——频域上连续的两个时频单元各自的自相关的互相关来刻画的，只有频带间互相关超过某个阈值的单元才被认为是足够相似。在低频段，这个标准可以使用；但在高频段，由于滤波器带宽较大，每个频带内往往包含了多个谐波，这导致时频单元的时域波形具有调幅特性，频带间互相关普遍降低。因此，在高频段，频带间互相关定义为频域上连贯的两个时频单元各自包络的自相关的互相关。被挑选出来的时频单元通过“灌水法”（floodfill）连接成若干个在时、频域上连续的块，将来，同一块中的各个时频单元将被分配给同一个声源。

在组合阶段，首先要判定每一个块应当属于人声还是伴奏。在提取出基音轨迹之后，可以衡量每个时频单元中信号的频率是否与基音一致。如果一个块中有超过半数的时频单元的频率与基音一致，那么这一块就分给人声，否则分给伴奏。然后，再通过一定的规则把分块阶段没有被挑选出的时频单元分配给两个声源。

CASA 方法的亮点在于它的信号分解方法有心理声学理论的支持，但是这种信号分解的时空复杂度都较高。它逼近理想蒙版的效果要远远优于“谐波蒙版”。

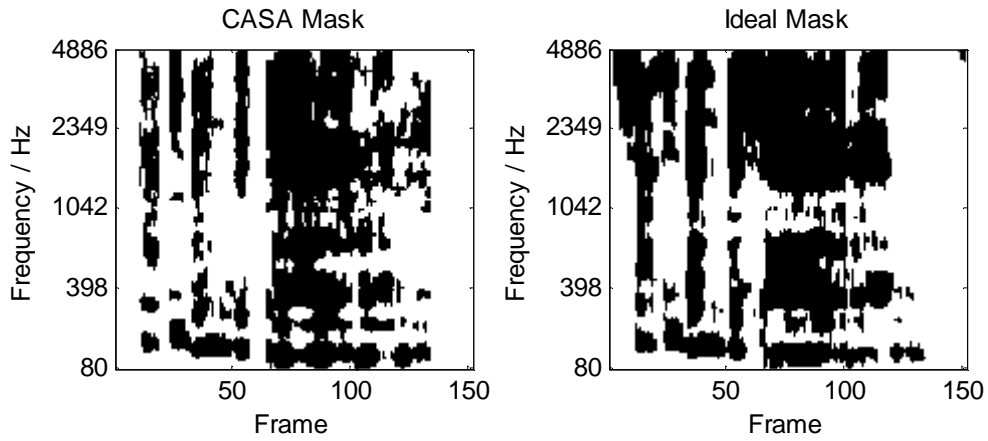


图2.2 CASA蒙版（左）与理想蒙版（右）的比较

尽管理想蒙版的分离效果可以接受，但是从细节上来看，还是有一些不尽如人意的地方。设想这样一种情况：在人声中，有一个谐波稳定地持续较长时间，而在伴奏中，这个频带处的能量忽强忽弱，导致这个频带处的时频单元一会儿被分给人声，一会儿又被分给伴奏。这样，在分离出的人声和伴奏中，都会有一个间断的谐波。尽管这种分配方式能使得分配错误的总能量最小，但从听觉效果上来讲，却远远不是最优的。我们希望能以适当的比例把每个时频单元的能量分配给各个声源，这就需要引入“软蒙版法”。

2.1.1.2 软蒙版法

芬兰派针对上节所述的“谐波蒙版”进行了改进。他们致力于解决人声中混有伴奏，或者说伴奏中位于人声谐波处的时频单元被“挖去”的问题。他们利用伴奏的重复性，对伴奏的幅度语谱进行了建模：

$$\mathbf{M}_{n \times m} \approx \mathbf{B}_{n \times r} \mathbf{A}_{r \times m} = [\mathbf{b}_1 \quad \cdots \quad \mathbf{b}_r] \cdot \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rm} \end{bmatrix} \quad (2-3)$$

这里， \mathbf{M} 代表伴奏的幅度语谱，它是一个二维矩阵，行代表频带，列代表帧。 \mathbf{B} 的每一列代表一个语谱基，而 \mathbf{A} 的矩阵元为这些语谱基的线性组合系数。语谱基的个数 r 比频带数 n 和帧数 m 都小得多，即把每一帧的幅度谱看成是少数几个基的线性组合。这三个矩阵的矩阵元都是非负的，因此这种分解称为“非负矩阵分解”(non-negative matrix factorization, NMF)。 \mathbf{BA} 可以看成是 \mathbf{M} 的一个稀疏表示，它更能体现出伴奏的重复性；其中位于人声谐波处那些时频单元的值，就可以作为伴奏在这些单元处的值的估计。通过某种迭代算法[1]求得一个足够接近 \mathbf{M} 的 \mathbf{BA} 后，将 \mathbf{BA} 在人声谐波处的值从硬蒙版法产生的人声中减去，并加到伴奏中，重新合成后（合成时相位谱使用混合信号的相位谱），就可以得到更纯净的人声了。此时，人声谐波处的时频单元的信号是按一定比例分别分配给人声和伴奏的，因此这种方法属于软蒙版法的范畴。

芬兰派的目标仅仅是获得纯净的人声，而并没有试图获得纯净的伴奏。从他们的网页[26]上下载混合信号和分离后的人声相减，就可以发现按照他们的方法获得的伴奏中还是混有大量的人声的。为了同时获得较为纯净的人声和伴奏，法国的 Durrieu[19]提出了一个更加完整的语谱非负矩阵分解模型。这个模型描述了人发声时声门的激励特性和声道的滤波作用，因此称为“源-滤波器模型”。该模型是建立在功率谱上的，其数学表达如下：

$$\mathbf{D} = (\mathbf{B}_F \mathbf{A}_F) \times (\mathbf{B}_K \mathbf{A}_K) + (\mathbf{B}_M \mathbf{A}_M) \quad (2-4)$$

这里， \mathbf{D} 代表混合信号的功率谱，“ \times ”号表示两个矩阵对应元素相乘。所有的矩阵元也都是非负的。公式右端的三个括号，分别代表了人声的声门、人声的声道和伴奏。 \mathbf{B}_F 的每一列是某一特定基频下的声门激励的功率谱，而 \mathbf{A}_F 中的比例系数则用来选择基频； \mathbf{B}_K 的每一列是某种声道响应的功率谱，可以理解成一种元音，而 \mathbf{A}_K 则用来选择一种元音或几种元音的过渡； \mathbf{B}_M 与 \mathbf{A}_M 的含义与芬兰派方法公式中的 \mathbf{B} 和 \mathbf{A} 相同。与芬兰派方法相比，Durrieu 方法的进步在于针对人声也提出了非负矩阵分解的模型。当然，这种声门—声道模型只适用于语音中的浊音。

在上述模型中， \mathbf{B}_F 是根据 KLGLOTT88 声门模型[21]算出来的，它是固定的；而 \mathbf{A}_F 中非零元的位置是用测出的基音轨迹确定的，它们的值可以变动；其余矩阵则完全自由。通过迭代算法可以求出右端的各个矩阵，也就能得到符合模型的人声功率谱 \mathbf{D}_V 和伴奏功率谱 \mathbf{D}_M ：

$$\mathbf{D}_V = (\mathbf{B}_F \mathbf{A}_F) \times (\mathbf{B}_K \mathbf{A}_K) \quad (2-5)$$

$$\mathbf{D}_M = \mathbf{B}_M \mathbf{A}_M \quad (2-6)$$

Durrieu 并没有把这二者作为最终的人声、伴奏功率谱估计，而是又通过维纳滤波的方法[19]求得幅度语谱上的软蒙版（下标 V 和 M 分别表示人声和伴奏）：

$$W_V = \frac{D_V}{D_V + D_M} \quad (2-7)$$

$$W_M = \frac{D_M}{D_V + D_M} \quad (2-8)$$

这两个蒙版都是实数，即分离出的人声、伴奏的相位语谱与混合信号相同。

用 Durrieu 的方法分离出来的人声和伴奏，既避免了硬蒙版法谐波间断的问题，又解决了芬兰派方法分离出来的伴奏中混有人声较多的问题，效果是相当好的。因此，本文的主体部分将详细叙述这种方法，并提出一些改进。

2.1.2 几种非蒙版分离法简介

2.1.2.1 正弦建模法

正弦建模法[2]也是由芬兰派提出来的。它的第一步也是提取基音轨迹。之后，对于每一帧信号，用基频及其倍频的复指数信号与之求相关，以估计此帧内歌声各次谐波的幅度与相位。对各帧的幅度与相位进行插值以得到连续的幅度与相位曲线，并由此合成人声；从混合信号中减去人声就作为伴奏。

这种方法具有比较明显的局限性。由于人声常常呈现出调幅、调频的特性，即使在一帧的时间范围内，人声的幅度和频率也往往是不稳定的，所以在帧的级别测量人声的频率和幅度是不够精确的。幅度和相位插值在一定程度上解决了这个问题，但完全根据模型合成出的声音，其自然度总是难以保证。另外，由于人声是完全重新合成的，而并不是从混合信号中分离而来，其波形与混合信号中的人声波形就可能有很大的不同。这样，从混合信号中减去人声所得的伴奏中，就会混有大量的人声。由于芬兰派并不试图获得伴奏，所以正弦建模法还可以适用，但在我们的研究中就不适用了。

2.1.2.2 谐波结构分析法

谐波结构分析法是由张长水等人[9][10][11]提出的。这里的“谐波结构”，指的是基波与各次谐波的幅度比例关系。张长水等人认为，多数有音调的乐器，尤其是管乐器，其谐波结构的稳定的，即方差较小；而人声的谐波结构随着所发音素的不同变化很大。在对声音进行多基频提取后，对每帧的每个基频，提取出它的谐波结构，并对所有的谐波结构进行分类。聚集度比较高的那些谐波结构被认

为是属于伴奏的，而分散分布的那些谐波结构则被认为属于人声。使用人声与伴奏两个声源各自包含的那些谐波结构重新合成声音信号，就得到分离结果。

这种方法也具有较多的缺陷。在文献[9]的末尾，作者指出了方法的一些局限性，但我们认为更严重的问题在于：

- 当伴奏中乐器较多时，各乐器的频谱相互重叠会很厉害（即“窗分离正交度”下降），导致提取的谐波结构不准确。事实上，我们并不需要将伴奏中的每种乐器也分离出来。
- 像吉他之类的弦乐器，随着弹奏方法的不同，谐波结构的方差也有可能很大。

此外，作者的实验做得也不够充分。文献[11]中举了四个实验的例子，但其中有三个实验分离的声源根本不是来自同一首歌曲；而且四个实验中，伴奏乐器的种数均不超过 2 种。所以我们认为，这种方法距应用于实际音乐还有一定距离。

2.2 基音轨迹提取方法综述

各种声源分离方法几乎都以基音轨迹提取为前端，基音轨迹提取的准确性在很大程度上决定着分离的效果。鉴于此，我在这里对各种基音轨迹的提取方法作一个比较。

2.2.1 基音提取算法分类

基音提取算法按照在什么域中处理信号，可以分为时域法、频域法、倒频域法等三大类。

时域法的基本思想是把语音信号分帧，对每一帧进行某种“相关”运算。当一帧语音信号具有周期性时，相关函数将具有相同的周期，而一个设计得好的相关函数将在每个周期中显示出一个明显的峰值（或谷值）。因此，相关函数相邻峰值的间隔就等于语音信号的基音周期，其倒数就是基音频率。文献[13][14]是时域法的代表，二者都设计了一种相关函数，前者的优点在于峰比较尖锐，分辨率高；后者的优点在于能够避免各个峰高度的涨落产生的影响。时域法的优点是思路直接，它在单基频提取中发挥了很大的作用；但是在多基频提取的情况下，其理论基础就遭到了破坏。

频域法也需要把信号分帧，但接下来做的事情是求每帧信号的频谱，即从时域变换到频域。如果一帧语音信号具有周期性，那么在频谱上基频整数倍的位置

就会有峰值，相邻两个峰值的间隔就是基频。频域法处理信号频谱，揭示了信号的本质，在单基频、多基频情况下均可使用。但是，直接对一帧信号进行 FFT 得到的频谱的频域分辨率有限，在多基频的情况下可能不够用。

倒频域法又叫语音的同态处理，它建立在语音产生的声管模型上，基本思想是把卷积运算转化为加法，从而通过线性处理把声门激励（包含了基频信息）与声管的传输函数分离开来。文献[12]中提到了线性预测，就是利用了同态处理的原理。倒频域法的优点是可以抛弃与基频无关的声管因素的干扰，但缺点是要求输入的语音是纯净的。这使得它基本不适用于声源分离的场合。

2.2.2 基音轨迹提取的难点及解决方案

(1) 单基频基音提取的困难

在时域法中，基音周期是由“第一个明显的峰”代表的。找“第一个明显的峰”的算法通常有两种：一是按照峰的位置，在允许的基音周期范围内从左到右依次寻找，并判断它是否“明显”；二是按照峰的幅度，由大到小依次寻找，并检验它们是否处在允许的基音周期范围内。在第一种方法中，由于相关函数在 0 时刻处为最大值，在它向右下降的过程中的一点小毛刺也可能被认为是第一个明显的峰，所以提取出的基音周期会偏小；在第二种方法中，由于第二、第三个峰等等也有可能是最高峰，并且处在允许的基音周期范围内，所以提取出的基音周期会偏大。文献[14]精心设计了一种“累积平均归一化幅度差”函数。这个函数在 0 时刻处并不具有峰值；而且它是一种归一化的量度，“明显”这个词变得可以量化，所以按从左到右的寻找方法可以准确地提取出基频。

在频域法中，常常出现的错误是倍频或者半频错误。倍频错误通常发生在二次谐波的幅度显著大于基波时，而半频错误通常发生在存在频率等于基频的一半的干扰时。这些都是由于分析频谱时过于关注细节造成的。Klapuri 提出了一种“谐波叠加法”[4]，它很好地考虑了频谱的整体性：一个频率是基频的可能性，并不是由此频率处频谱的幅度单独决定的，而是由此频率及其各倍频处频谱的幅度加权共同决定的。频率 f 本身和其各倍频处频谱幅度的加权和称为频率 f 的“显著性”。引入“显著性”函数可在一定程度上减少倍频和半频错误。

(2) 多基频基音提取的困难

由于在分离歌声与伴奏的任务中，一帧里往往有多个声源同时发声，所以我们面临的是多基频基音提取的问题。频域法在这种情况下有一定优势，但仍存在如何从多个都很有可能的基频中正确选取的问题。

一种可能的解决方案是，在每一帧的处理中并不急于作出决定，而是给出一些基频的候选值，以及相应的可能性的量度，而最终的决定留给后处理环节去做。上面提到的“显著性”函数就是一个例子。这种思想在单基频基音提取中也可以借鉴。

(3) 基音轨迹的平滑性问题

由于基频是逐帧提取的，而一帧内基频的提取很有可能出现较大的误差，所以对于整段声音信号，提取出来的基音轨迹往往是凹凸不平的，需要进行平滑处理。

比较简单的平滑方法是中值滤波，即把每一帧的基频用它周围若干点的基频值的中位数代替。这种方法具有实现简单的优点，但容易把本来测得比较准的基频用周围并不准确的值代替，即得到的基音轨迹是宏观上平滑、微观上不准确的。

文献[12]中提出了一种动态规划的方法，通过对基频的跳变处以罚分来保证基音轨迹的平滑性。当然，使用动态规划必然要求在每一帧中不做出最终判决。使用动态规划得到的基音轨迹中，每一帧的基频都来自候选值，因此不会出现像中值滤波那样被不准确的值替代的问题。动态规划方法也可以在多基频提取时的多个候选值之间做出选择。但是，动态规划有一个缺点，就是各项罚分的权重分配过于自由，难以找到一组效果又好、又能有理论解释的参数。

台湾的许肇凌等人（下称“台湾派”）[16]和芬兰派[5]则更进一步地采用模式分类的方法，引入了 HMM 模型，以建模帧与帧之间的转移关系。前者以量化的基频值为状态，后者则进一步以每个量化的基频值的开始、保持、结束阶段为状态。对每一个状态，二者均从语谱图中提取特征，并用 GMM 来建模输出概率分布；对于状态之间的转移概率，前者通过对训练集中的状态转移进行统计而获得，后者还引入了各调性歌曲中音符的出现概率和转移概率。文献[16]指出，台湾派方法的性能已经超越了 2005、2006 两年 MIREX 评测（针对基音提取性能的一种国际性评测）的冠军系统。这样的方法的理论基础比较完备，但其性能依赖于特征的选取以及模型是否合适，而且可能面临训练与测试条件不匹配的问题。

(4) 伴奏、清音、浊音（A/U/V）的判决

语音分为浊音和清音，浊音有基频，而清音则没有，其频谱呈现噪声状。另外，在歌曲中，还会有比较长的只有伴奏没有人声的段落。对于清音和伴奏帧，基音提取算法应当给出“没有人声基频”的判断，而不应当对所有帧都给出基频值。

对于朴素的时域法或频域法，可以通过设置阈值过滤掉那些不明显的基频，但阈值的调整总是一个枯燥的过程。对于动态规划算法来说，判断基频的有无稍微容易一些，只需要把“无基频”作为基频的候选值之一，并根据其它候选值的可能性，适当给“无基频”赋予可能性即可。在 Klapuri[5]的模式分类方法中，“无基频”也被作为一个状态。而台湾派[15]则在基音提取之前增加了一个步骤，利用 HMM 对每帧进行 A/U/V 三类的分类，之后只对 V 类进行基音提取。

第3章 人声伴奏分离系统的实现

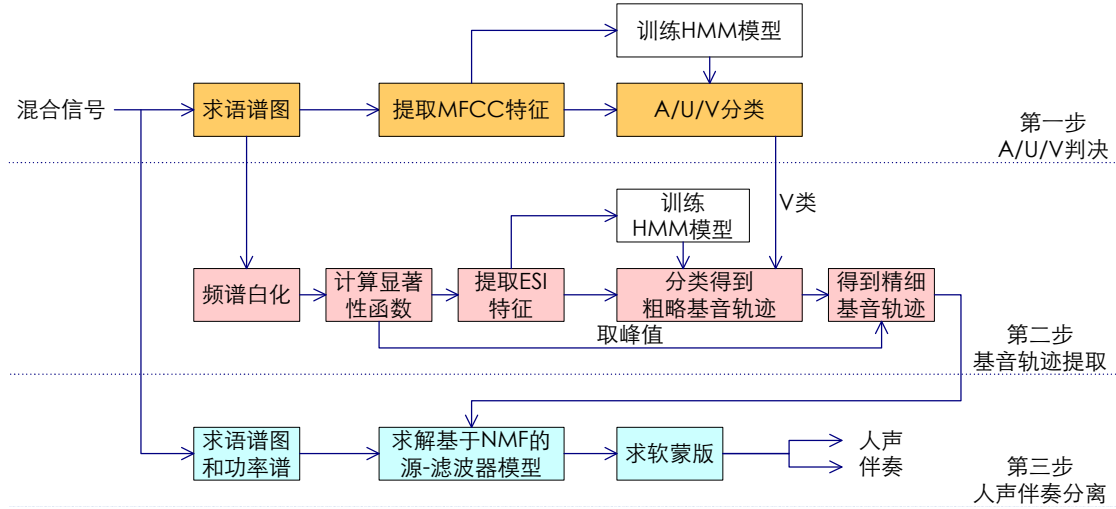


图3.1 人声伴奏分离系统的流程图

本文提出的人声与伴奏分离系统由三个阶段组成：A/U/V 判决、基音轨迹提取、分离。前两步借鉴了台湾派的方法[15][16]，即基于 HMM 的模式分类方法，其中第二步基音轨迹提取使用由 Klapuri 的“显著性函数”导出的一种特征替换了台湾派原本使用的两种特征。第三步分离采用了 Durrieu 的源-滤波器模型和软蒙版法[19]，并做了一定的改进。系统的详细流程图如图 3.1。

训练 HMM 模型时使用了 MIR-1K 数据库[15]。该数据库由 1000 段音频组成，采样率为 16 kHz，量化精度为 16 bit。它们节选自 110 首流行歌曲，由 19 名业余人士（11 男 8 女）演唱，每段音频长度为 4 秒至 13 秒不等，总长为 133 分钟。每段音频由人声和伴奏两个声道组成，两个声道的能量近似相等。该数据对每帧的 A/U/V 属性进行了人工标注，帧长为 40 ms，帧移为 20 ms；对于浊音帧，还标出了人声的基频。基频是用 midi 代码表示的，其单位为半音数。midi 代码 f_{midi} 与以赫兹为单位的频率 f_{Hz} 的转换关系为：

$$f_{\text{midi}} = 69 + 12 \log_2(f_{\text{Hz}} / 440) \quad (3-1)$$

系统采用 Matlab 语言实现。

3.2 利用 MFCC 特征进行 A/U/V 判决

3.2.1 MFCC 特征的提取

A/U/V 判决步骤使用的特征是 Mel 频率倒谱系数 (Mel-frequency cepstral coefficients, MFCC)。这是在语音识别中常用的一种特征,它与同态处理中的倒谱系数类似,主要刻画了频谱的包络,即声道的特征。由于人在发清音和浊音时声道形状不同,而这与乐器发出乐音时“声道”的形状又不同,所以伴奏、清音、浊音的频谱包络就会很不相同。这就是 MFCC 特征能够区分伴奏、清音、浊音的理论基础。

计算一帧信号的 MFCC 系数的步骤如下[27]:

1. 加窗,做 FFT;
2. 取一组在 Mel 尺度上均匀分布的三角滤波器,求出各频带内的能量;
3. 把各频带内能量值取对数,构成一个序列;
4. 对上述序列做离散余弦变换 (DCT);
5. 取变换后序列的前若干个值,这就是该帧信号的 MFCC 系数。

MFCC 与倒谱系数的区别有如下两点:第一,它的滤波器是在 Mel 尺度上均匀分布的,低频段较密集,高频段较稀疏,这有助于分辨出低频段较密集的谐波。第二,它的第二步变换是 DCT 而不是 IFFT,这是由于 DCT 的结果是实数,且具有良好的“能量压缩”特性,使得只取较少个系数就能保持大部分信息。

事实上,上述计算步骤只是一个概要,其中有许多细节可以由人们自己决定。下面介绍我们采用的具体细节。

我们使用的信号采样率为 16 kHz,帧移为 20 ms (320 个采样点),帧长为 40 ms (640 个采样点)。窗函数选择旁瓣较低的汉明窗。FFT 长度为 1024 个采样点。Mel 尺度滤波器组由网上找到的 UIUC 大学的一个 melfb 函数[28]生成,频带数取为 40,滤波器组的幅频响应如图 3.2 所示。求各频带内的能量时并没有把 FFT 的结果取模方,而只取了模。这样做并不违反 MFCC 的定义,事实上,在当前的研究中,这里平方和不平方的做法都很常见。做 DCT 之后,抛弃第 1 个系数(代表直流分量),而取第 2 至第 13 个系数做为 MFCC 系数。

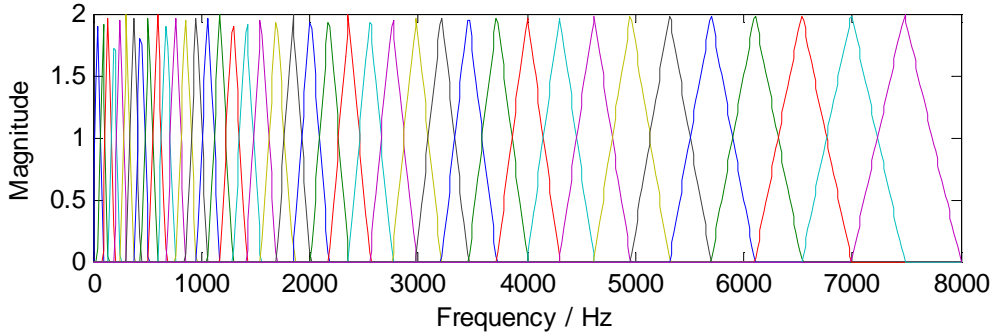


图3.2 Mel尺度滤波器组的幅频响应

MFCC 特征就是在 MFCC 系数的基础上扩展出来的。在 12 个 MFCC 系数的基础上，首先添加一维“帧总能量”，即加窗语音帧中各样本的平方和，构成 13 维特征。帧总能量对于区分 A/U/V 也是有用的——通常，仅有伴奏的地方能量也相对小一些。然后，在整段语音信号的范围内，把每一维都归一化至零均值、单位方差。这一操作称为“倒谱均值归一化”（cepstral mean normalization），其目的是消除不同音乐的录音环境对 MFCC 特征造成的整体性影响。之后，对 13 维的 MFCC 特征求一阶、二阶差分，得到最终的 39 维 MFCC 特征。由于人声有调幅、调频特性，而乐音没有，所以人声特征的时域稳定性较差，其差分项绝对值也就较大，这也有助于区分 A/U/V。

差分的具体求法也有很多种。如用 t 表示帧号， x 表示特征， Δx 表示特征的差分，则 Δx 可以有如下三种定义（若右边某项的下标超出范围，则令该项为 0）：

$$\Delta x(t) = x(t) - x(t-1) \quad (3-2)$$

$$\Delta x(t) = x(t+1) - x(t) \quad (3-3)$$

$$\Delta x(t) = \frac{1}{2} [x(t+1) - x(t-1)] \quad (3-4)$$

理论上讲，采用哪一种定义的效果应该是差不多的。但我们实验发现，采用第三种定义可以莫名其妙地把判决准确率提升 4%，所以我们采用了第三种定义。

3.2.2 HMM 模型的建立与使用

一个 HMM 模型由状态 i 、状态的初始概率 π_i 、状态间的转移概率 A_{ij} 、状态的输出概率密度函数 $p_i(o)$ 四个要素组成。在上述记号中， A_{ij} 里的 i 表示起始状态， j 表示终止状态； $p_i(o)$ 中的 i 表示状态， o 表示输出的特征矢量，在本模型中就是 MFCC 特征。在我们的模型中，状态有 A、U、V 三个，即 $i \in \{A, U, V\}$ 。状态的初始概率由数据库中各段音频的第一帧的标注确定，状态 i 的初始概率 π_i

等于第一帧标注为 i 的音频段数除以总音频段数。状态间转移概率由数据库中相邻两帧的转移次数统计而得，设数据库中由状态 i 转移至状态 j 的总次数为 n_{ij} ，则 $A_{ij} = n_{ij} / \sum_k n_{ik}$ 。状态的输出概率密度函数 $p_i(o)$ 用高斯混合模型（GMM）建模。对于每个状态，用数据库中标注为此状态的那些帧的 MFCC 特征进行训练，分量数取为 32，协方差阵取为对角型，最大迭代次数为 20。训练直接通过 Matlab 中的 `gmdistribution.fit` 函数完成。

HMM 模型训练完成后，需要对某段音频中的各帧进行 A/U/V 判决时，只需将此段音频各帧的 MFCC 特征使用 HMM 模型进行 Viterbi 解码即可。

3.2.3 A/U/V 判决的效果

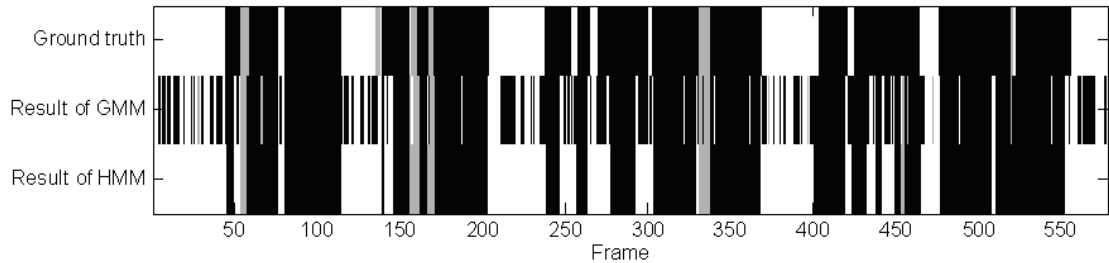


图3.3 A/U/V判决的效果。黑色代表浊音，灰色代表清音，白色代表伴奏。

图 3.3 展示了用 HMM 模型进行 A/U/V 判决的效果。最上面一行是数据库中的标注，中间一行是只用 GMM 分类的结果，最下面一行是用 HMM 进行 Viterbi 解码的结果。可以看到，只使用 GMM 时，对于浊音帧的判决效果比较好，但对伴奏帧的判决结果较差，且沿时间轴判决结果在三个类别之间剧烈地跳变；引入 HMM 之后，对伴奏帧的判决效果变好了，且整体上判决结果更为平滑，与标注的真实类别也更接近。

3.3 利用 ESI 特征提取基音轨迹

基音轨迹提取也是采用了 HMM 模式分类的方法，它利用的特征称为 ESI 特征（energy at semitones of interest）。台湾派在文献[16]中提出了两种 ESI 特征，一种是直接从语谱图中计算出来的，而另一种是对语谱进行“归一化谐波叠加”后计算的。他们之所以应用两种 ESI 特征，是因为其中任何一种特征都不能足够好地评估每个基频的可能性。而我们则只采用了一种 ESI 特征，它的基础是 Klapuri

用“谐波叠加法”得到的“显著性函数”[4]。我们发现，只用这一种 ESI 特征就能达到与台湾派可相媲美的结果。

3.3.1 用谐波叠加法计算显著性函数

3.3.1.1 频谱白化

在进行谐波叠加之前，首先要进行“频谱白化”[4]，以抑制音色对基音提取的影响。音色由频谱上能量的粗略分布情况来代表，具体来说，可以像求 MFCC 特征的前两步那样，用一个滤波器组各频带内的能量来描述。之后按频带进行能量压缩：设某频带内信号总能量为 A^2 （幅度的平方平均值为 A ），则对此频带内的信号乘以系数 $A^{-2/3}$ 。从宏观上看，这相当于对幅度语谱进行三次方根压缩，从而抑制了整体上能量的不均匀；从微观上看，由于每个频带都覆盖了 FFT 谱上连续若干个点，所以有谐波处的能量与无谐波处的能量之比并没有被压缩。从这一点上讲，频谱白化优于三次方根压缩。

图 3.4 是一段歌曲的原始幅度语谱图、白化幅度语谱图和三次方根压缩后的幅度语谱图。在白化语谱图上可以清晰地看到原始语谱图上看不到的高频分量，而且白化语谱图上的谐波与背景的对比度比三次方根压缩后的高。

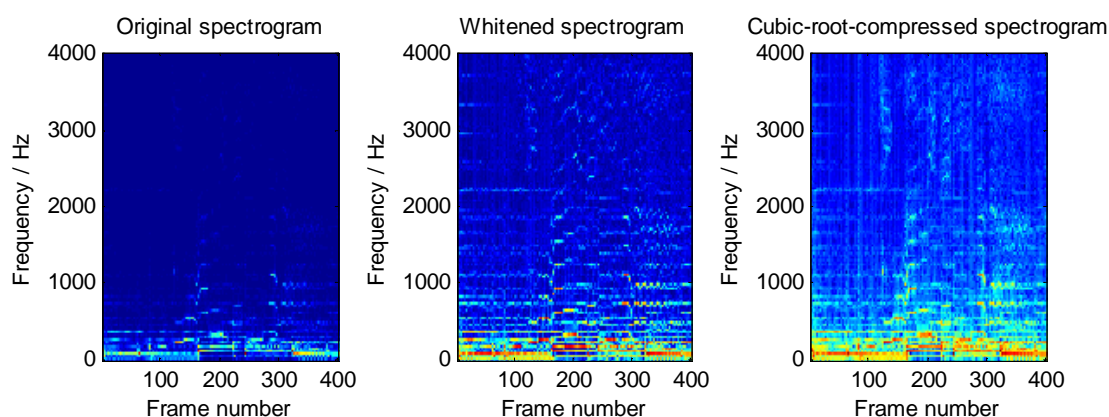


图3.4 一段歌曲的原始、白化、三次方根压缩幅度语谱图。浅色代表高能量。

3.3.1.2 计算显著性函数

显著性函数具有两个自变量，一是时间，一是频率。由于显著性函数是逐帧计算的，因此在下面的记号中我将省略表示时间的下标。

设某一帧的（连续）幅度频谱为 $M(f)$ ，则基频 f_0 的显著性 $s(f_0)$ 定义为：

$$s(f_0) = \sum_{k=1}^K g(f_0, k) M(kf_0) \quad (3-5)$$

即显著性为 f_0 及其各次谐频处的幅度值的加权和。 K 为考虑的最高谐波次数，本文中取为 20。由于显著性定义式中需要任意连续频率处的幅度谱值，而 FFT 的频域分辨率不足以满足这个要求，所以 $M(kf_0)$ 是在 kf_0 周围 50Hz 频带内的频谱上加三角窗后积分求得的。

显著性函数中还有一个至关重要的加权函数 $g(f_0, k)$ 。文献[4]提出的加权函数形式如下：

$$g(f_0, k) = \frac{f_0 + \alpha}{kf_0 + \beta} \quad (3-6)$$

这种函数形式是通过机器学习的结果拟合出来的。它是有道理的：它的分母表明，频率越高的谐波权重越小，因为高频谐波往往较弱，易受噪声影响；它的分子表明，频率越高的基频权重越大，这是为了防止算法过于偏好低基频，而低基频对应的往往是伴奏而不是人声。加权函数中的参数 α 和 β （均具有频率的量纲）是在一个训练集上优化得到的。文献[4]给出了一些不同的帧长下 α 和 β 的最优取值，如表 3.1。从表中数据来看， β 可以取为常数。 α 似乎应与帧长成正比，但并没有什么道理支持这一论断。本文采用的帧长为 40 ms，与表中的第一行比较接近，故也取 $\alpha = 27\text{Hz}$ ， $\beta = 320\text{Hz}$ 。

表3.1 加权函数中参数取值与帧长的关系

帧长 / ms	α / Hz	β / Hz
46	27	320
93	52	320

图 3.5 显示了一段音频的显著性函数谱，其纵轴为 midi 代码，即频率按对数分布。从其中红色的部分已经能够看出基音轨迹。从图中还可以看出，在真实基频的倍频和半频处，显著性函数也显示出峰值，但不如真实基频处的峰高。

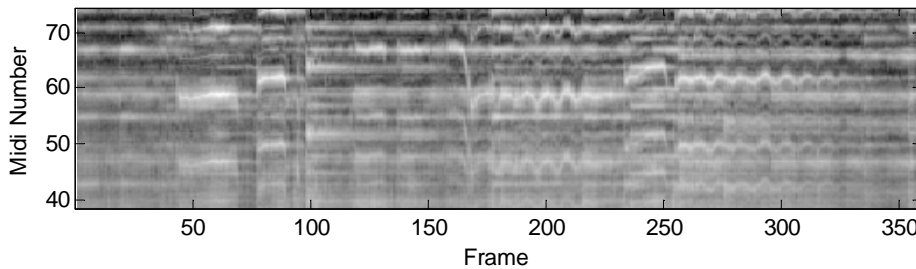


图3.5 一段歌曲的显著性语谱图。浅色代表高能量。

3.3.2 ESI 特征的计算和 HMM 模型的建立

根据 MIR-1K 数据库中人声基频的范围，本研究把基频提取的范围定在 midi 代码 38.5 ~ 74.5（即 75.6Hz ~ 604.5Hz），精确度为 0.1 个半音。因此，在计算显著性函数时，频率自变量就取在上述 midi 代码范围内，步长为 0.1 个半音。如果直接把每帧的显著性函数当作特征，则维数太高（361 维！），所以，我们把 midi 代码 $n-1 \sim n+1$ 范围内的显著性函数值加三角窗后求和，作为 midi 代码 n 处的能量（下式中显著性函数 s 的自变量为 midi 代码）：

$$E(n) = \sum_{i \in A} s(i)w(i-n) \quad (3-7)$$

$$A = \{n-1, n-0.9, \dots, n+0.9, n+1\} \quad (3-8)$$

$$w(x) = \begin{cases} 1-|x|, & |x| < 1 \\ 0, & |x| \geq 1 \end{cases} \quad (3-9)$$

将 midi 代码 39 ~ 74 各处的能量组成一个 36 维向量，并归一化至各元素之和为 1，就得到了 ESI 特征。与用来刻画声道响应的 MFCC 特征相对，ESI 特征主要用来刻画声门激励的频率信息。图 3.6 是一帧典型的 ESI 特征，可以看到在基频及其倍频、半频处都有峰值，但还是基频处的峰值最高。

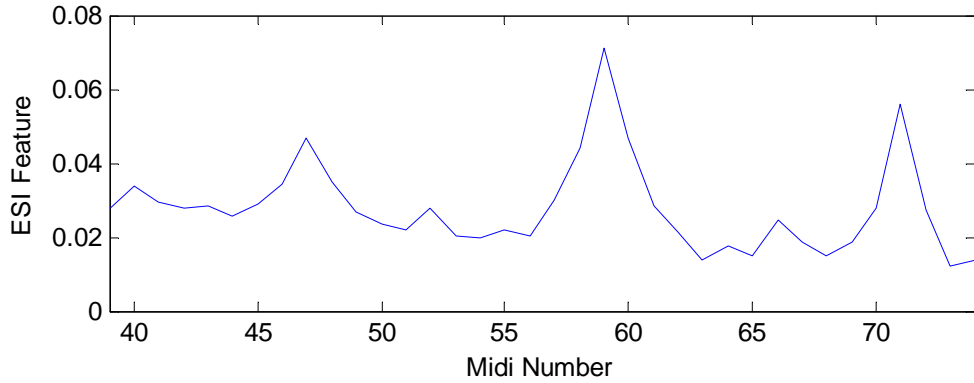


图3.6 一帧典型的ESI特征。该帧基频的midi代码为59.2。

HMM 模型就以 midi 代码 39 ~ 74 为状态，数据库中每一帧的真实状态定为标注 midi 代码四舍五入后的整数。初始概率、转移概率、输出概率的训练方法与 A/U/V 判决时基本相同，只是这里“每一段”音频指的不再是数据库中完整的一段音频，而只是其中的一个浊音段。由于状态数多，每种状态对应的帧数就少，训练输出概率的 GMM 模型时分量数仅取为 8，且若训练不成功则逐次减半至训练成功为止。

用 HMM 模型对一段浊音进行解码,可以得到粗略的基音轨迹 $n_1(t)$ (用 midi 代码表示, t 为帧号), 其精确度为 1 个半音。为了得到精确到 0.1 个半音的基音轨迹, 最后还需要在显著性函数谱上粗略基音轨迹 ± 0.5 个半音的范围内取最大值, 最大值对应的 midi 代码就是最终的精细基音轨迹 $n_2(t)$ (下式中频率均用 midi 代码表示):

$$n_2(t) = \arg \max_{n_1(t)-0.5 \leq n \leq n_1(t)+0.5} s(n) \quad (3-10)$$

对图 3.5 所示的显著性函数进行基音轨迹提取的结果如图 3.7 所示。

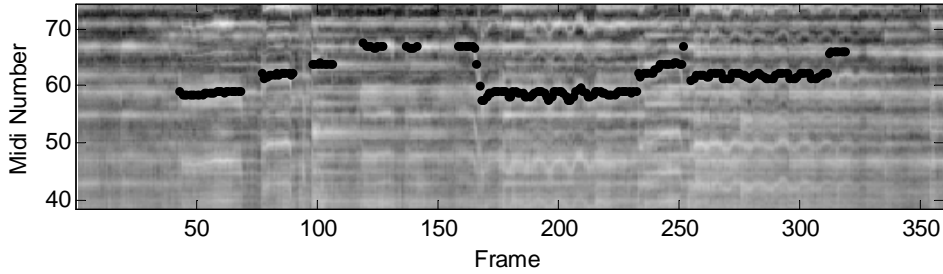


图3.7 基音轨迹的提取结果。黑线表示基音轨迹。

3.4 利用源-滤波器模型分离人声与伴奏

3.4.1 “功率谱”的两种含义

在本节中, 将会常常用到“功率谱”这个词, 而这个词可以表示两种不同的含义。为避免混淆, 在此首先对这两种含义加以辨析。

“功率谱”的一种简单的定义就是 STFT 语谱图的模方。这种功率谱是针对确定性信号而言的, 本文称为“观测功率谱”。观测功率谱用 S 表示, 若用 X 表示 STFT 语谱图, 则有 $S = |X|^2$ 。

“功率谱”的另一种定义是在讨论随机信号时使用的。设 x 为一随机语音信号, X 为其语谱, 则 X 是一个随机复数矩阵。为简便起见, 假设 X 的各个元素是相互独立的; 它的每一个矩阵元 X_{ft} 服从均值为 0、协方差为 D_{ft} 的旋转对称复高斯分布。 X_{ft} 的概率密度函数为:

$$p(X_{ft} | D_{ft}) = \frac{1}{\pi D_{ft}} \exp \left(-\frac{|X_{ft}|^2}{D_{ft}} \right) \quad (3-11)$$

随机信号的“功率谱”就是指其语谱各元素的协方差组成的矩阵。本文称这种功率谱为“本质功率谱”，并用 \mathbf{D} 表示。

两种功率谱之间的关系可以这样理解：随机语音信号及其语谱可以看成随机过程，其“本质功率谱” \mathbf{D} 是这个随机过程的参数。而确定性语音信号、其语谱及其“观测功率谱” \mathbf{S} 都是对这个随机过程的一次观测结果。

这两种功率谱的区别在 0 综述 Durrieu 工作时并未强调出来，读者在读下文时需注意。

3.4.2 源-滤波器模型的建立

Durrieu 在文献[19]中提出的源-滤波器模型，事实上是对“本质功率谱” \mathbf{D} 进行非负矩阵分解：

$$\mathbf{D} = (\mathbf{B}_F \mathbf{A}_F) \times (\mathbf{B}_K \mathbf{A}_K) + (\mathbf{B}_M \mathbf{A}_M) \quad (3-12)$$

这里“ \times ”号表示两个矩阵对应元素相乘，所有的矩阵元均非负。公式右端的三个括号，分别代表了人声的声门的功率谱、人声的声道功率谱响应和伴奏的功率谱。 \mathbf{B}_F 的每一列是某一特定基频下的声门激励的功率谱，而 \mathbf{A}_F 中的比例系数则用来选择基频； \mathbf{B}_K 的每一列是某种声道响应的功率谱，可以理解成一种元音，而 \mathbf{A}_K 则用来选择一种元音或几种元音的过渡； \mathbf{B}_M 的每一列代表乐器可以发出的一种功率谱，而 \mathbf{A}_M 的矩阵元把这些功率谱进行线性组合。

我在这个模型的基础上提出了一点改进，即考虑了声道响应功率谱的平滑性。在上面的模型中，矩阵 \mathbf{B}_K 的每一列就是一帧完整的功率谱，其自由度是相当高的。实验发现，用此模型求解出的 \mathbf{B}_K ，每一列都具有许多微小的毛刺。声道的响应代表着功率谱的包络，所以我们希望 \mathbf{B}_K 尽可能平滑。为此，我把原来的矩阵 \mathbf{B}_K 进一步拆分成两个非负矩阵的乘积 $\mathbf{C}_K \mathbf{B}_K$ ，其中 \mathbf{C}_K 是一些“平滑”的带通声道响应单元，而新的 \mathbf{B}_K 则是这些响应单元的组合系数。事实上，Durrieu 在后来的一篇文献[23]中也做了这一点改进。

改进后的源-滤波器模型的数学表达式为：

$$\mathbf{D} = (\mathbf{B}_F \mathbf{A}_F) \times (\mathbf{C}_K \mathbf{B}_K \mathbf{A}_K) + (\mathbf{B}_M \mathbf{A}_M) \quad (3-13)$$

3.4.3 根据源-滤波器模型估计本质功率谱

3.4.3.1 目标函数

估计“本质功率谱” \mathbf{D} 的目标是使观测到 \mathbf{X} 的似然值最大。

给定“本质功率谱” \mathbf{D} ，观测到语谱 \mathbf{X} 的似然值为：

$$L(\mathbf{X} | \mathbf{D}) = \prod_{f,t} p(X_{ft} | D_{ft}) = \prod_{f,t} \frac{1}{\pi D_{ft}} \exp\left(-\frac{S_{ft}}{D_{ft}}\right) \quad (3-14)$$

最大化此似然值，可以等价于最小化如下的目标函数（式中常数项已省略）：

$$\text{cost}(\mathbf{S} | \mathbf{D}) = -\log L(\mathbf{X} | \mathbf{D}) = \sum_{f,t} \left(\frac{S_{ft}}{D_{ft}} + \log D_{ft} \right) \quad (3-15)$$

3.4.3.2 约束条件

源-滤波器模型公式(3-13)中右端的各个矩阵并不是完全自由的。其中 \mathbf{B}_F 和 \mathbf{C}_K 两个矩阵是完全固定的，而 \mathbf{A}_F 是部分固定的；还有一些矩阵有归一化要求。下面分别讨论。

\mathbf{B}_F 的每一列是某一特定基频下的声门激励的功率谱，这些功率谱是根据 KLGLOTT88 模型[21]计算出来的，详见附录 B。

\mathbf{A}_F 用来选取 \mathbf{B}_F 中的基频。在测出基音轨迹之后，可以要求仅能选取实测基频周围某一小范围内的基频，而其它系数均为 0。由于迭代求解各矩阵时使用的更新规则是乘性的，被强制置 0 的系数将永远为 0。

\mathbf{C}_K 是一组“平滑”的带通声道响应单元，不妨取为汉宁窗的形状。

另外， \mathbf{B}_F 、 \mathbf{C}_K 、 \mathbf{B}_M 的每一列作为功率谱的基，最好是归一化的，即每一列的和均为 1。 \mathbf{B}_M 各列的幅度可以转移到 \mathbf{A}_M 的各行中。代表声道响应的 \mathbf{B}_K 的各列最好也归一化，因为其幅度可以转移到 \mathbf{A}_K 的各行中。而 \mathbf{A}_K 的各列最好也归一化，因为其幅度可以转移到 \mathbf{A}_F 的各列中。

3.4.3.3 迭代算法

源-滤波器模型中各矩阵的求解可以通过迭代算法来完成。迭代公式如下[20]：

$$\mathbf{Q}_F = \mathbf{C}_K \mathbf{B}_K \mathbf{A}_K / \mathbf{D} \quad (3-16)$$

$$\mathbf{P}_F = \mathbf{S} \times \mathbf{Q}_F / \mathbf{D} \quad (3-17)$$

$$\mathbf{A}_F \leftarrow \mathbf{A}_F \times (\mathbf{B}_F^T \mathbf{P}_F) / (\mathbf{B}_F^T \mathbf{Q}_F) \quad (3-18)$$

$$\mathbf{Q}_K = \mathbf{B}_F \mathbf{A}_F / \mathbf{D} \quad (3-19)$$

$$\mathbf{P}_K = \mathbf{S} \times \mathbf{Q}_K / \mathbf{D} \quad (3-20)$$

$$\mathbf{A}_K \leftarrow \mathbf{A}_K \times (\mathbf{B}_K^T \mathbf{C}_K^T \mathbf{P}_K) / (\mathbf{B}_K^T \mathbf{C}_K^T \mathbf{Q}_K) \quad (3-21)$$

$$\mathbf{B}_K \leftarrow \mathbf{B}_K \times (\mathbf{C}_K^T \mathbf{P}_K \mathbf{A}_K^T) / (\mathbf{C}_K^T \mathbf{Q}_K \mathbf{A}_K^T) \quad (3-22)$$

$$\mathbf{A}_M \leftarrow \mathbf{A}_M \times [\mathbf{B}_M^T (\mathbf{S} / \mathbf{D}^2)] / [\mathbf{B}_M^T (\mathbf{I} / \mathbf{D})] \quad (3-23)$$

$$\mathbf{B}_M \leftarrow \mathbf{B}_M \times [(\mathbf{S} / \mathbf{D}^2) \mathbf{A}_M^T] / [(\mathbf{I} / \mathbf{D}) \mathbf{A}_M^T] \quad (3-24)$$

上面公式中，两个矩阵并列书写代表矩阵乘法，“ \times ”号表示两个矩阵对应元素相乘，“/”号代表两个矩阵对应元素相除，平方代表矩阵每个元素分别平方， $\mathbf{1}$ 代表全1矩阵（不是单位矩阵）。在每次迭代中，只要新的矩阵与原矩阵不完全相同，目标函数就会减小。

迭代的初值，对于仅有归一化限制的矩阵 \mathbf{B}_K 、 \mathbf{A}_K 、 \mathbf{B}_M 、 \mathbf{A}_M 可以采用随机初值；对于仅有某些元素可以非零的矩阵 \mathbf{A}_F ，可以把允许非零的位置全置为1。

迭代时各个矩阵的更新顺序是无关紧要的。需要注意的是，每更新完一个矩阵，都要重新计算 \mathbf{D} 的值，才能保证下一次更新时目标函数减小。更新操作可能导致有归一化约束的矩阵不再满足归一化要求，需要重新归一化。

3.4.4 求软蒙版

估计出似然值最大的“本质功率谱” \mathbf{D} 之后，我们还要进一步估计出软蒙版，并由此求出人声和伴奏的语谱，才能重新合成出人声和伴奏。蒙版的估计可以使用最大似然估计（MLE）和最小均方误差估计（MMSE，即“维纳滤波”），二者的结果是一样的。

首先讲述最大似然估计。“本质功率谱” \mathbf{D} 可以分为两部分，即人声本质功率谱 \mathbf{D}_V 和伴奏本质功率谱 \mathbf{D}_M ：

$$\mathbf{D}_V = (\mathbf{B}_F \mathbf{A}_F) \times (\mathbf{C}_K \mathbf{B}_K \mathbf{A}_K) \quad (3-25)$$

$$\mathbf{D}_M = \mathbf{B}_M \mathbf{A}_M \quad (3-26)$$

设人声和伴奏的语谱分别为 \mathbf{X}_V 和 \mathbf{X}_M ，并对它们施加“保守性”约束

$\mathbf{X}_V + \mathbf{X}_M = \mathbf{X}$ 。我们要在已知 \mathbf{D}_V 和 \mathbf{D}_M 的条件下，估计出似然值最大的 \mathbf{X}_V 和 \mathbf{X}_M 。由于我们假设了语谱中各时频单元是相互独立的，所以这个估计可以对每个时频单元独立地进行。在下面的叙述中将省略频率和时间下标。

对一个时频单元， X_V 和 X_M 的总似然值为

$$\begin{aligned} p(X_V | D_V) p(X_M | D_M) &= \frac{1}{\pi D_V} \exp\left(-\frac{S_V}{D_V}\right) \cdot \frac{1}{\pi D_M} \exp\left(-\frac{S_M}{D_M}\right) \\ &= C \cdot \exp\left(-\frac{|X_V|^2}{D_V} - \frac{|X_M|^2}{D_M}\right) \\ &= C \cdot \exp\left(-\frac{|X_V|^2}{D_V} - \frac{|X - X_V|^2}{D_M}\right) \\ &= C' \cdot \exp\left(-\frac{D_V + D_M}{D_V D_M} \left|X_V - \frac{D_V}{D_V + D_M} X\right|^2\right) \end{aligned} \quad (3-27)$$

此式可以看成是随机变量 X_V 的一个概率密度函数，对应的分布是均值为 $(\frac{D_V}{D_V + D_M} X, 0)$ 的旋转对称复高斯分布。显然，其最大值在 $X_V = \frac{D_V}{D_V + D_M} X$ 时取得，故 X_V 和 X_M 的最大似然估计为：

$$\hat{X}_V = \frac{D_V}{D_V + D_M} X \quad (3-28)$$

$$\hat{X}_M = \frac{D_M}{D_V + D_M} X \quad (3-29)$$

下面我们再从最小均方误差估计（维纳滤波）的角度来求解 X_V 和 X_M 。 X_V 的最小均方误差估计，应该是在 $X_V + X_M = X$ 条件下 X_V 的条件均值，即

$$\begin{aligned} \hat{X}_V &= E(X_V | X_V + X_M = X) \\ &= \frac{\int p(X_V | D_V) p(X_M | D_M) X_V dX_V}{\int p(X_V | D_V) p(X_M | D_M) dX_V} \end{aligned} \quad (3-30)$$

上面已经计算出， $p(X_V | D_V) p(X_M | D_M)$ 相当于 X_V 的一个概率密度函数，对应的分布是均值为 $(\frac{D_V}{D_V + D_M} X, 0)$ 的旋转对称复高斯分布。那么上面的条件均值计算结果自然也就是

$$\hat{X}_V = \frac{D_V}{D_V + D_M} X \quad (3-31)$$

类似地

$$\hat{X}_M = \frac{D_M}{D_V + D_M} X \quad (3-32)$$

我们看到，最大似然估计和最小均方误差估计（维纳滤波）给出了相同的软蒙版：

$$W_V = \frac{D_V}{D_V + D_M} \quad (3-33)$$

$$W_M = \frac{D_M}{D_V + D_M} \quad (3-34)$$

这两个蒙版都是实数，因此分离后人声和伴奏的相位谱均与混合信号的相位谱相同。用此蒙版乘以混合信号语谱就得到了人声和伴奏各自的语谱，然后再经重叠相加法就可以合成出时域的人声和伴奏信号。

3.4.5 算法中一些参数的选择

分离步骤选取的帧长、帧移与之前的步骤相同，分别为 40 ms 和 20 ms。由于需要重新合成，FFT 长度必须等于帧长，即 40 ms（640 个采样点）。分析和合成时均采用正弦窗（即汉宁窗的平方根），这样一来能够实现完美重建；二来合成窗的两端衰减至零，能够保证帧间的平滑过渡；三来分析窗与正弦窗相比，旁瓣更低，合成后人声各谐波之间杂音小。

源-滤波器模型中各矩阵的大小取法如下：取基音提取步骤可能给出的基频作为声门激励的基频，即 midi 代码 38.5 到 74.5，步长 0.1，共 361 个值。这样 \mathbf{B}_F 就有 361 列。 \mathbf{C}_K 的列数（即平滑声道响应单元的个数）取为 30， \mathbf{B}_K 列数（即声道响应的基的个数）取为 9， \mathbf{B}_M 的列数（即伴奏功率谱基的个数）取为 20。后面三个参数的具体取值对算法性能影响不大，一般取为远小于 FFT 长度一半的值。

其它参数： \mathbf{A}_F 中允许非零的元素对应的基频与实测基频的差距不超过 0.2 个半音。迭代次数定为 50 次。

第4章 性能评价

4.1 A/U/V 判决的评价

A/U/V 判决是在 MIR-1K 数据库[15]上进行的。该数据库已在 0 开头介绍过。测试采用二重交叉检验（two-fold cross validation）的方法，即把整个数据库分成两个规模相近的子集，在子集 1 上训练出的 HMM 用来对子集 2 进行测试，反之亦然。两个子集包含的歌手名如表 4.1。所有训练和测试均在人声与伴奏能量混合比为 0 dB 的条件下进行。

表4.1 MIR-1K数据库两个子集包含的歌手列表
(正体表示男歌手, 斜体表示女歌手)

子集	歌手名	歌手数	片断数	总时长
1	Ani Kenshin abjones <i>amy annar ariel</i> bobon bug davidson	9 (5 男 4 女)	487	64.7 分钟
2	fdps geniusturtle <i>heycat jmzen khair leon stool tammy titon yifen</i>	10 (6 男 4 女)	513	68.6 分钟

判决的结果用混淆矩阵表示，性能由分类准确率来衡量。分类准确率定义为判决结果与真实类别相同的帧数占总帧数的比例。我们把我们系统的性能与提出 HMM 分类算法的台湾派[15]的系统性能比较，如图 4.1：

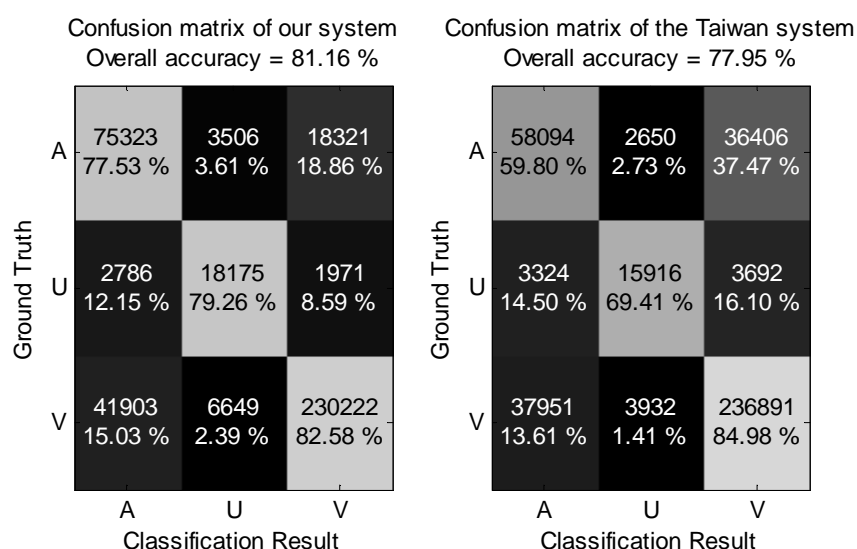


图4.1 我们的系统（左）与台湾派系统（右）的混淆矩阵比较
(每个小格中的数字表示帧数，百分比为该格帧数在本行中所占比例)

尽管我们的系统与台湾派系统采用的是同样的方法，但是我们的分类准确率更高（前文说过，这很莫名其妙地归功于差分公式(3-4)的选择）。另外，台湾派的分类结果明显偏向 V 类，而我们的系统分类结果平衡。

4.2 基音轨迹提取的评价

基音轨迹提取的评价也在 MIR-1K 数据库上进行，同样采用二重交叉检验的方式。评价的指标有 raw-pitch accuracy 和 overall accuracy[22]。二者的定义为：

$$\text{Raw - pitch accuracy} = \frac{\text{V类基音提取正确的帧数}}{\text{判决结果为V类的帧数}} \quad (4-1)$$

分母错了。应该是标注为V的帧数

$$\text{Overall accuracy} = \frac{\text{A、U类判决正确的帧数} + \text{V类基音提取正确的帧数}}{\text{总帧数}} \quad (4-2)$$

其中，“V 类基音提取正确”指的是真实类别和判决结果均为 V 类，且提取出的基频与真实基频相差不超过 1 个半音。

我们用这两个指标把我们的系统和台湾派系统[16]进行了比较。在我们的系统中，V 类基音提取正确的帧数为 191979，结合 4.1 节的混淆矩阵，可以算得 raw-pitch accuracy 为 ~~76.63%~~^{68.87%}，overall accuracy 为 71.57%。台湾派系统的 raw-pitch accuracy 约为 71%（从[16]中图上读出），overall accuracy 为 71.10%。~~台湾派系统 raw-pitch accuracy 较低的原因就是 A/U/V 判决过于偏向 V 类。~~从 overall accuracy 上来看，两个系统的性能几乎一致。但是，在 4.3 节中我们将看到，这个准确率是远远不够的，A/U/V 判决和基音提取的准确率是整个系统性能的瓶颈。

另外，我们统计了我们的系统提取的基频与真实基频的误差分布，如图 4.2。可以看出，半频错误略多于倍频错误，但总体上比较平衡。另外，1 ~ 4 个半音的误差数目稍多。

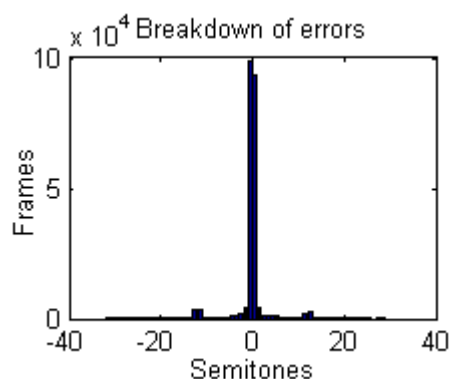


图4.2 基频提取的误差分布

4.3 人声伴奏分离的评价

4.3.1 评价标准

人声与伴奏分离到目前为止还没有一个普适而全面的评价标准。有一些评价标准是仅适用于某一类算法的，如王德良等人提出的“能量损失比”和“噪声残余比”[8]就仅适用硬蒙版法。Vincent 等人[24]总结了当前评价标准的缺陷，并提出了一系列不仅赖于算法类型，且声称能够反映人的主观感觉的评价标准：声源-失真比、声源-干扰比、声源-噪声比、声源-伪像比（source-to-distortion ratio (SDR), source-to-interferences ratio (SIR), sources-to-noise ratio (SNR), sources-to-artifacts ratio (SAR)）。后三者分别评价了算法对干扰、噪声、伪像的抑制能力，而 SDR 则是对这三种能力的综合评价。其思路是通过正交投影的方法，把分离出的信号分解成声源、干扰、噪声、伪像四种成分，然后比较四者的能量。在本研究中，由于没有噪声，所以没有噪声项和 SNR。分解的过程中有一个至关重要的概念是容许失真种类，它可以包括时不变增益、时不变滤波器、时变增益、时变滤波器四种类型。具体类型和参数（如滤波器阶数等）的选取取决于应用。

Vincent 等人的评价标准的思路接近人的认知方式。如果确实能够把分离结果分解成名副其实的声源、干扰、噪声、伪像四项，则这组评价标准就可以说是一种普适而全面的评价标准。但是，实验发现这组评价具有两点缺陷：一是容许失真种类及其参数的选取过于随意；二是分解出来的伪像项中仍包括相当大的声源和干扰成分。也就是说，正交投影法并没有真正把分离结果按人的听觉认知分解开来。这样，讨论各项的比例也就没有什么意义了。

鉴于此，本文使用了最朴素的客观评价标准。这种标准也称为声源-失真比（SDR），它相当于在不容许任何种类失真时的情形下 Vincent 等人的 SDR。设声源 i 的真实信号为 s_i ，分离出来的声源 i 的信号为 \hat{s}_i ，则 SDR 定义如下：

$$\text{SDR}_i = 10 \log_{10} \frac{\|s_i\|^2}{\|s_i - \hat{s}_i\|^2} \quad (4-3)$$

这种评价标准在不同的文献中有不同的称呼。Durrieu 的文献[19]中 SDR 的含义与本文相同，在芬兰派的两篇文献[2][3]中，这种 SDR 分别被称为 SNR 和 VAR（voice-accompaniment ratio）。

SDR 衡量的是在分离后的信号，纯净信号与失真的能量比。由于各声源混合时能量比可能不同，有时只比较分离后的 SDR 会显得不公平。为此，定义分离前混合信号中 x 中各声源的 SDR 为：

$$\text{SDR}_{i0} = 10 \log_{10} \frac{\|s_i\|^2}{\|s_i - x\|^2} \quad (4-4)$$

这相当于把混合信号当作分离出的声源。两个 SDR 之差称为 SDR 增益：

$$\Delta \text{SDR}_i = \text{SDR}_i - \text{SDR}_{i0} \quad (4-5)$$

SDR 增益衡量的是分离前后 SDR 的变化。

SDR 有一些奇妙的性质。在本文使用的算法中，分离后的两个声源之和仍等于混合信号。用 v 、 m 、 \hat{v} 、 \hat{m} 分别表示原始人声、原始伴奏、分离出的人声、分离出的伴奏，则有：

$$v + m = \hat{v} + \hat{m} \quad (4-6)$$

这意味着分离出的人声中的失真能量等于分离出的伴奏中的失真能量：

$$\|v - \hat{v}\|^2 = \|m - \hat{m}\|^2 = E_d \quad (4-7)$$

另记

$$\|v\|^2 = \|m - x\|^2 = E_v \quad (4-8)$$

$$\|m\|^2 = \|v - x\|^2 = E_m \quad (4-9)$$

则有

$$\text{SDR}_v = 10 \log_{10} \frac{E_v}{E_d} = 10 \log_{10} \frac{E_m}{E_d} - 10 \log_{10} \frac{E_m}{E_v} = \Delta \text{SDR}_m \quad (4-10)$$

$$\text{SDR}_m = 10 \log_{10} \frac{E_m}{E_d} = 10 \log_{10} \frac{E_v}{E_d} - 10 \log_{10} \frac{E_v}{E_m} = \Delta \text{SDR}_v \quad (4-11)$$

即：分离后人声的 SDR 等于伴奏 SDR 的增益，分离后伴奏的 SDR 等于人声 SDR 的增益。又，当原始人声与原始伴奏以 1:1 的能量比混合时， $E_v = E_m$ ，此时 SDR_v 、 SDR_m 、 ΔSDR_v 、 ΔSDR_m 四个数完全相等。

除此之外，为了在 MIR-1K 数据库上与台湾派的分离结果进行比较，我们也采用了台湾派提出的 SDR：

$$\text{SDR}_i = 10 \log_{10} \frac{\langle s_i, \hat{s}_i \rangle^2}{\|s_i\|^2 \|\hat{s}_i\|^2 - \langle s_i, \hat{s}_i \rangle^2} = 10 \log_{10} (\cot^2 \theta) \quad (4-12)$$

式中尖括号表示内积， θ 为 s_i 与 \hat{s}_i 的夹角。为避免与前面提出的 SDR 混淆，下文称这种新的 SDR 为“台式 SDR”，而前面的 SDR，由于 Vincent 和 Durrieu 都来自法国，称为“法式 SDR”。

法式 SDR 和台式 SDR 没有确定的大小关系。放在 Vincent 等人的大框架下来看，法式 SDR 是不容许任何失真情形下的 SDR，而台式 SDR 是容许有时不变增益失真情形下的 SDR。两种 SDR 的关系示意图如下：

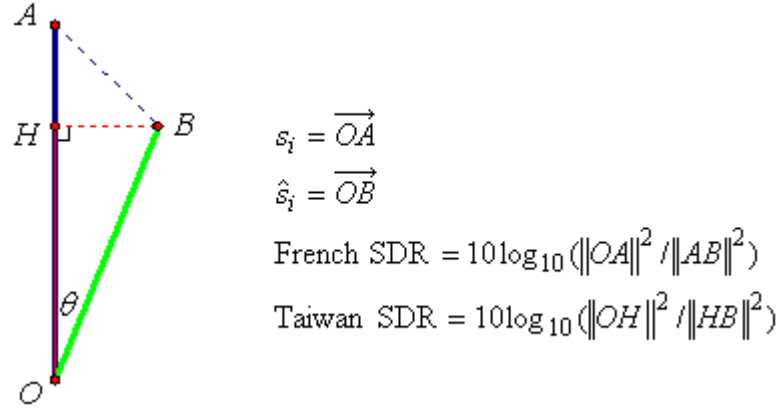


图4.3 法式SDR与台式SDR关系示意图

对于台式 SDR，可以用与法式 SDR 相同的方法定义分离前的各声源的 SDR 和 SDR 增益。

4.3.2 测试集

为了方便声源分离的研究者们评价算法性能并互相比较，需要一些大规模的、有标注的数据库。当前这样的数据库有 MIR-1K[15]、MIREX、RWC[25]等等。我们在 MIR-1K 数据库上进行了测试。与文献[15]中一样，在人声伴奏能量混合比为-5 dB、0 dB、5 dB 时分别进行了测试。为了反映 A/U/V 判决和基音提取性能对整体性能的影响，我们分别使用数据库中标注的基音轨迹和前两个步骤实测的基音轨迹进行测试。在子集 1 上进行测试时，使用的是在子集 2 上训练的 HMM，反之亦然。所有的 HMM 都是在 0 dB 的混合比下训练的。评价指标采用了法式和台式两种 SDR。采用台式 SDR 测得的结果可以直接与[15]比较，采用法式 SDR 测得的结果可以作为以后其它研究者比较的基准。

Durrieu[19]测试时没有采用上述数据库。为了与他的实验结果比较，我们也取得了他的数据库中的部分数据，具体包括：

表4.2 Durrieu数据库的组成

子集	歌曲代号	歌曲名称	歌手性别	长度
A	Ab1	Bearlin – Roads	男	14 秒
	At1	Tamy – Que Pena Tanto Faz	女	13 秒
B	Bb1 ~ Bb4	Bent Out of Shape	男	2 分 40 秒
	Bc1 ~ Bc4	Chevalier Bran	男	4 分 56 秒
	Bp1 ~ Bp4	Le Pub	男	4 分 36 秒
	Bs1 ~ Bs3	Schizosonic	男	3 分 08 秒
	Bu1 ~ Bu4	Into the Unknown	男	2 分 38 秒
C	Ch1 ~ Ch5	Shame	女	4 分 20 秒
	Ci1 ~ Ci4	Silence	女	4 分 12 秒
	Cl1 ~ Cl4	We Are In Love	女	3 分 42 秒
	Cm1 ~ Cm5	Matter of Time	女	4 分 37 秒
	Cu1 ~ Cu4	Sunrise	女	3 分 16 秒

我们未能取得 B 集的全部。数据库中长度超过 1 分钟的歌曲被切成 1 分钟长的段落（最后一段除外），若最后一段中没有人声，则舍弃。“歌曲代号”的最后一位就是段落的序号。C 集中 Ci、Cl、Cu 三首歌有人工标注的基音轨迹。

原数据库为双声道、44100 Hz，我们把它降采样至单声道、16000 Hz 以与我们的实验环境匹配。

4.3.3 测试结果

4.3.3.1 在 MIR-1K 数据库上的测试结果

在 MIR-1K 数据库上分离性能的测试结果如表 4.3。其中的数值是在所有音频段落上加权平均的结果，权重为音频长度，这一点与台湾派的处理方式相同。表中数据并不精确满足(4-10)(4-11)两式，这是由合成时的边界效应（即对音频两端的处理方式）引起的。

表4.3 在MIR-1K数据库上分离性能的测试结果

混合比	法式 SDR / dB				台式 SDR / dB			
-5 dB		分离前	标注基音	实测基音		分离前	标注基音	实测基音
	人声	-5.00	6.53	2.56	人声	-5.00	5.34	-0.97
	伴奏	5.00	9.24	7.50	伴奏	5.00	8.68	6.65
0 dB		分离前	标注基音	实测基音		分离前	标注基音	实测基音
	人声	-0.00	9.32	6.56	人声	-0.00	8.70	5.31
	伴奏	0.00	9.24	6.51	伴奏	-0.00	8.68	5.65
5 dB		分离前	标注基音	实测基音		分离前	标注基音	实测基音
	人声	5.00	11.87	9.65	人声	5.00	11.53	9.09
	伴奏	-5.00	6.82	4.62	伴奏	-5.00	5.99	3.64

下面把我们的结果与台湾派的结果[15]进行比较。台湾派采用的算法分成与本文相同的三个步骤，第一步 A/U/V 判决的方法也与本文相同，第二步基音轨迹提取采用的是 Dressler 的方法[17]，而第三步分离则是采用了王德良等人的硬蒙版法[7]。文献[15]中报告了在 MIR-1K 数据库上，人声伴奏混合比分别为-5 dB、0 dB、5 dB，分别采用理想蒙版、标注基音、实测基音时的平均人声台式 SDR 增益，如表 4.4。表中也计算了我们在这后两种条件下取得的增益。

表4.4 在MIR-1K数据库上人声台式SDR增益的对比（单位：dB）

（注：台湾派论文中没有给出标注基音和实测基音的情况下的具体数值，

表中数值是从图上估读出来的）

混合比	台湾派			我们	
	理想蒙版	标注基音	实测基音	标注基音	实测基音
-5 dB	10.62	7.5	-0.5	10.34	4.03
0 dB	8.36	6.0	0.9	8.70	5.31
5 dB	5.82	3.0	0.2	6.53	4.09

从表中可以看出，无论是在标注基音还是实测基音的情况下，我们的系统的性能都远远超过台湾派。尤其值得注意的是，我们的系统在标注基音情况下的性能已经逼近甚至超过了台湾派在使用理想蒙版的性能。这说明，Durrieu 的基于源-滤波器模型的软蒙版法可以打破硬蒙版法的性能上界。

另一方面，从我们自己的系统性能来看，可以得到两点结论：1)尽管 HMM 都是在 0 dB 下训练的，但±5 dB 的混合比波动对系统整体性能影响不大；2)实测

基音时的 SDR 增益比标注基音时要低不少,这说明前两个步骤还有很大的改进余地。

4.3.3.2 在 Durrieu 数据库上的测试结果

由于我们的方法与台湾派的差别较大,且台湾派的性能太差,我们需要找一种与我们方法相似且性能较好的方法进行比较。Durrieu 的方法[19]就很合适。这篇文章中的方法没有进行 A/U/V 判决。其基音提取也是根据源-滤波器模型进行的:在 A_F 不受限的情况下进行迭代,把迭代得到的 A_F 中的较大的值作为基频的候选,然后用动态规划方法得到平滑的基音轨迹。此后的分离步骤,除了没有引入 C_K 以平滑声道响应以外,与我们的方法只有参数选取上的不同。

文献[19]中给出了在 A、B、C 三个子集上,分离后人声和伴奏的平均法式 SDR。这里的平均没有加权,每段音频的权重都一样。由于 Durrieu 数据库中大部分数据没有标注基音轨迹,所以只能使用实测的基音轨迹。另外,在这篇文献的网页[29]上,还可以找到具体几段音频的分离性能。我们针对同样的数据进行了测试,与 Durrieu 的结果对比列于表 4.5。

表4.5 在Durrieu数据库上的法式SDR对比（单位：dB）

歌曲代号	分离前		Durrieu		我们 ¹	
	人声	伴奏	人声	伴奏	人声	伴奏
Ab1	-5.37	5.37	6.2	11.6	3.44	8.76
At1	0.51	-0.51	11.5	9.2 ²	4.17	3.66
Bb2	0.01	-0.01	5.5	5.6	8.46	8.45
Bc3	-6.79	6.79	1.5	8.3	2.72	9.50
Ci2	0.28	-0.28	8.6	8.4	5.17	4.89
Cm3	-4.72	4.72	8.0	12.7	4.52	9.24
A 集平均	-2.43	2.43	8.2	10.8	3.80	6.21
B 集平均 ³	-7.37	7.37	2.4	8.5	1.68	9.03
C 集平均 ⁴	-6.30	6.30	2.7	9.1	1.37	7.66

注1: 由于我们迭代过程采取随机初值，所以对于单个音频，不同次的运行结果之间可能存在0.2 dB左右的波动，表中列出的是某一次运行的结果。

注2: 这个数字摘自网页[29]。它与(4-10)(4-11)两式相差较大，疑有误。下载数据后实测值为10.99。

注3: 由于我们没有取得B集的全部，这一行不具有可比性。

注4: 这里B、C集的平均指的是对该集中所有音频的平均，不仅限于表中列出的音频，下同。

从表中看，我们的系统在 B 集的两段音频上的性能优于 Durrieu 的系统，但在 A、C 集上的表现逊色不少。事实上，B 集中有大段的时间是纯伴奏，Durrieu 的系统没有进行 A/U/V 判决，所以在这些段落也分离出了“人声”，这是他的系统性能略差的原因。而我们的系统性能差的主要原因在于基音提取不正确，尤其是在 A 集上，Durrieu 基音提取的表现近乎完美。为了验证这一结论，我们对 C 集中有标注的三首歌曲用标注的基音轨迹进行了分离，与[19]中的结果比较。另外，我们对在 A 集上实测的基音轨迹进行了手工修正，作为标注重新对 A 集进行了分离，并把性能与网页[29]上的数据比较。比较结果如表 4.6:

表4.6 在Durrieu数据库上使用标注基音时的法式SDR对比（单位：dB）

歌曲代号	分离前		Durrieu		我们	
	人声	伴奏	人声	伴奏	人声	伴奏
Ab1	-5.37	5.37	6.2	11.6	6.42	11.71
At1	0.51	-0.51	10.8 ¹	10.6 ¹	10.92	10.41
Ci2	0.28	-0.28	10.6	10.4	11.66	11.38
A 集平均	-2.43	2.43	8.2	10.8	8.67	11.06
C 集平均	-5.91	5.91	3.5	9.7	8.26	14.16

注1: 这两个数字与(4-10)(4-11)两式相比也有一定误差。下载数据后实测值分别为10.90和10.36。

从此表中看，我们系统的性能在 A 集上与 Durrieu 的系统近似相同，而在 C 集上超过了 Durrieu 系统的性能。这证实了我们的系统的性能瓶颈确实在于 A/U/V 判决和基音轨迹提取。使用标注基音轨迹时，我们的系统与 Durrieu 的系统的差别仅在于 C_K 的有无和参数的不同，这二者都可能是我们的系统性能更好的原因。

4.4 效率评价

值得强调的是，我们的系统与 Durrieu 的系统尽管方法相似，但我们的系统处理速度要远远超过 Durrieu 的系统。我们取得了 Durrieu 的 Python 程序与我们的 Matlab 程序进行效率比较。比较时使用的电脑配置为：DELL 640m 笔记本，1.66 GHz 主频，1.5 GB 内存，Matlab 版本为 7.6.0 (2008a)，Python 版本为 2.6。比较的材料为 Durrieu 数据库中的音频 At1（长度为 13 秒）。比较时 Durrieu 程序中源-滤波器模型中矩阵大小等参数已调至与我们的程序相同。

在这段音频上，我们的程序用的总时间为 12 秒，其中 A/U/V 判决用时 1 秒，对源-滤波器模型进行 50 次迭代用时 9 秒，其余时间用于计算 KLGLOTT88 模型等预处理。Durrieu 的程序在基音提取和分离两个步骤中需要对源-滤波器模型各进行一轮迭代（每轮迭代 50 次），而每轮迭代的时间竟长达 9.5 分钟。

显然，这几十倍的效率差别不能仅仅归因于编程语言的速度不同。我们在实现时考虑了效率上的优化，如用把 \mathbf{A}_F 用稀疏矩阵实现，等等。这些考虑对程序的高效率有很大的贡献。

第5章 讨论

从第4章的性能评价结果可以看出，我们系统的 A/U/V 判决和基音提取性能还不够好。我们为改善这两部分性能进行了一些尝试，尽管没有达到改善性能的效果，但思路也许会对以后的研究有帮助。5.1 节和 5.2 节讨论了这样的两种尝试。

另外，我们现在所用的源-滤波器模型只能分离人声中的浊音部分，而对清音部分，我们没有做任何处理。在 5.3 节中我们简要叙述了一下清音分离的可能思路，供以后的研究参考。

5.1 A/U/V 判决和基音提取能否同时进行

A/U/V 判决和基音提取两个步骤都是用 HMM 模式分类实现的，基音提取是在第一步得到的 V 类上的进一步分类。这很容易让人想到：把两个步骤合到一起，直接以 A、U 和各种基频为类别，同时使用 MFCC 和 ESI 两种特征进行模式分类可以吗？

我们进行了这方面的实验：把 V 类拆分成基频 midi number 从 40 ~ 72 的 33 个类，与 A、U 两类并列，共 35 个类。每个类分别训练自己的 MFCC GMM 和 ESI GMM。但实验结果并不理想。我们观察到 A/U/V 判决的结果明显地偏向 V 类。由于 A、U 类分类正确帧数的大量减少，A/U/V 判决的准确率降低到 75.48%，而基音提取的 raw-pitch accuracy 和 overall accuracy 分别降低到 64.54% 和 63.36%。

我们认为 A/U/V 判决的结果明显偏向 V 类的原因在于对 V 类的表示详细程度远远超过了 A、U 两类，与三类的帧数占数据库总帧数的比例不相称。因此，我们又把 A 类拆分成 12 类，拆分的依据是 ESI 特征中最大值对应的 midi 代码除以 12 的余数。这样做的效果仍不理想。尽管 A、V 两类的判决结果不再明显有偏，但有一半的 U 类帧被错判为 A 或 V 类。这表明 U 类的表示太过简略。此时的性能测试数据为：A/U/V 判决准确率 76.67%，基音提取 raw-pitch accuracy 75.10%，overall accuracy 68.22%。

在上面的实验中，A、V 两类的每个子类所用的 MFCC GMM 是不同的。事实上，由于 A/U/V 判决并不需要用到 ESI 特征，所以可以让所有的 A 类共用同一

个 MFCC GMM，所有的 V 类也共用同一个 MFCC GMM。但这样做之后 A/U/V 判决结果又明显偏向 V 类（原因不明），各项指标均有所下降。

为什么引入 ESI 特征之后 A/U/V 判决性能反而变差了呢？有两种可能：一种是因为 A、U、V 三类被拆分；一种是 ESI 特征对 A/U/V 判决起了反作用。为验证到底是哪一种原因，我们把 Viterbi 译码时 ESI GMM 的输出对数似然值全部设为 0。此时，A/U/V 判决准确率竟达到了 82.58%，且无偏。由此看来，的确是 ESI 特征对 A/U/V 判决起了反作用，A/U/V 判决和基音提取似乎没有必要做成一体化的。当然，最后这种情况下 A/U/V 的判决率高并不能说明什么，因为它与正文中的 A/U/V 判决相比，差别仅在于 A、V 两个类别被莫名其妙地拆分掉了，而这些子类还是共用相同的 MFCC GMM。

5.2 用 GMM 模型建模 ESI 特征是否合适

我们对 ESI 特征进行过这样的思考：一帧的基频主要体现在 ESI 特征中的峰值上。如果我们在训练 ESI GMM 时只采用三个峰值点及其周围点共 9 维特征的话，则可以降低复杂度，甚至也可以像 MFCC 特征那样，引入帧间差分。但是，我们发现这样做的效果一塌糊涂，没有几帧的基频能够正确提取。

我们观察了某一个基频非常显著的帧的情况。在这一帧中，基频对应的 ESI 值达到了 0.07，而此基频对应的 ESI GMM 的各个分量的均值中，此基频对应的 ESI 值均在 0.05 左右。尽管按理来说这一帧的基频应该被认为是比 GMM 中的更显著，但是，高斯分布的意义就在于过大、过小都不行——对于 0.05 的均值来说，0.07 和 0.03 的意义是一样的。这就造成当仅考虑 9 维特征时，这一帧与 GMM 的匹配程度不高。而当考虑全部维时，恰恰是 ESI 特征中那些值较小的维使得二者能够匹配。上面的情况说明，用 GMM 模型建模 ESI 特征并不是最合适的选择。

我们还探讨了 Durrieu 提取基音轨迹的方法[19]。在不对 A_F 施加限制的情况下迭代得到的 A_F 与我们使用的“显著性函数”一样，在是基频的可能性比较大的地方显示出峰值。也就是说，这样的 A_F 可以看成是一种广义的“显著性函数”。如果能找到一种比较好的显著性函数，以及一种比较合理的加权方式来权衡显著性函数和基音轨迹的平滑性，那么在“显著性函数”上进行动态规划，也可能是提取基音轨迹的一种有效方法。当然，也可以用 HMM 中的状态转移概率矩阵来代替平滑性罚分。使用动态规划与使用 HMM 分类相比，可以省去训练的步骤，更直接地利用显著性函数包含的信息。

5.3 关于清音的分离

尽管清音在语音中所占比例不大，但清音的分离在人声伴奏分离中也是相当重要的。人声中清音的缺失会影响可理解度，而伴奏中清音的残留会使人感觉仿佛整个音节都残留在伴奏中。本文提出的系统的第一步 A/U/V 判决之所以把清音 U 也作为一类，就是为了后续的处理，尽管由于时间限制，我们没能实现。在此，我们将综述一下清音分离的方法。

台湾派[15]采用的是基于 CASA 和 GMM 分类的方法。他们首先把清音段像王德良[7]那样分成时频单元，然后以理想蒙版为真实分类，使用单个时频单元的 MFCC 特征，对“人声为主”和“伴奏为主”两类时频单元训练 GMM 模型，用以分类。文献[15]中报告说，在 0 dB 的混合比下，这种方法可以把 86% 的时频单元正确分类，但是作为硬蒙版法，其性能仍然受到理想蒙版的限制。

Durrieu 在文献[23]中，也在源-滤波器模型的框架下考虑了清音的分离问题。他在矩阵 \mathbf{B}_F （声门激励）中增加了一个所有值均相等的列，表示发清音时类似于白噪声的声门激励。但是对于声道部分他没有做任何修改，即认为清音与浊音共用相同的声道响应。由于这一假设不甚合理，这种方法分离出的清音往往含有伴奏，而清音没有分离干净。

事实上，在 Durrieu 的源-滤波器模型中，可以在矩阵 \mathbf{B}_K 中也增加几列，用来逼近清音的声道响应。但是，由于一段音频中清音帧数往往太少，如果完全让迭代算法自行求 \mathbf{B}_K 的话，存在训练数据不够的问题。因此， \mathbf{B}_K 中对应于清音声道响应的那些列可能需要提前训练。

第6章 结论

本文研究了从单声道音频中分离人声和伴奏的问题。我们借鉴了台湾派的方法，使用 HMM 分类进行 A/U/V 判决及基音提取，又借鉴 Durrieu 的源-滤波器模型，用软蒙版法进行人声和伴奏的分离。在两个国际公开数据库上的性能测试结果表明，从分离效果方面来看，我们的系统与台湾派系统相比，性能有大幅度的提高，而与 Durrieu 完全基于源-滤波器模型的系统相比还有一定差距；而从效率方面来看，我们的系统远远优于 Durrieu 的系统。

本文系统性能的瓶颈在于 A/U/V 判决和基音提取。鉴于此，我们对提高这两个步骤性能的方法进行了讨论，并探讨了在广义显著性函数上进行动态规划的可行性。此外，我们还讨论了人声中清音部分的分离方法。提高 A/U/V 判决和基音提取性能并实现清音的分离，是我们进一步研究的方向。

插图索引

图 2.1 谐波蒙版与理想蒙版的比较	5
图 2.2 CASA 蒙版与理想蒙版的比较	6
图 3.1 人声伴奏分离系统的流程图	13
图 3.2 Mel 尺度滤波器组的幅频响应	15
图 3.3 A/U/V 判决的效果	16
图 3.4 一段歌曲的原始、白化、三次方根压缩幅度语谱图	17
图 3.5 一段歌曲的显著性语谱图	18
图 3.6 一帧典型的 ESI 特征	19
图 3.7 基音轨迹的提取结果	20
图 4.1 我们的系统与台湾派系统的混淆矩阵比较	26
图 4.2 基频提取的误差分布	27
图 4.3 法式 SDR 与台式 SDR 关系示意图	30
图附 1 KLGLOTT88 模型的声门激励波形	58
图附 2 基频为 425Hz 的声门激励功率谱	59

表格索引

表 3.1 加权函数中参数取值与帧长的关系	18
表 4.1 MIR-1K 数据库两个子集包含的歌手列表	26
表 4.2 Durrieu 数据库的组成	31
表 4.3 在 MIR-1K 数据库上分离性能的测试结果	32
表 4.4 在 MIR-1K 数据库上人声台式 SDR 增益的对比	32
表 4.5 在 Durrieu 数据库上的法式 SDR 对比	33
表 4.6 在 Durrieu 数据库上使用标注基音时的法式 SDR 对比	34

参考文献

- [1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization", *Neural Information Processing Systems*, 2001.
- [2] M. Ryyänän, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription", *IEEE Int'l Conf. on Multimedia & Expo*, 2008.
- [3] T. Virtanen, A. Mesaros, M. Ryyänän, "Combining pitch-based inference and non-negative matrix spectrogram factorization in separating vocals from polyphonic music", *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Sep. 2008.
- [4] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes", *Proc. of 7th Int'l Society for Music Info. Retrieval*, pp. 216-221, 2006.
- [5] M. P. Ryyänän and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music", *Computer Music Journal*, vol. 32, no. 3, pp. 72-86, fall 2008.
- [6] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE Trans. on Signal Proc.*, vol. 52, no. 7, pp. 1830-1847, Jul. 2004.
- [7] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings", *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 4, pp. 1475-1487, May 2007.
- [8] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", *IEEE Trans. on Neural Networks*, vol. 15, no. 5, pp. 1135-1150, Sep. 2004.
- [9] Y. G. Zhang and C. S. Zhang, "Separation of voice and music by harmonic structure stability analysis", *IEEE Int'l Conf. on Multimedia & Expo*, 2005.
- [10] Y. G. Zhang and C. S. Zhang, "Separation of music signals by harmonic structure modeling", *Neural Information Processing Systems*, 2006.
- [11] Z. Y. Duan, Y. G. Zhang, C. S. Zhang, and Z. W. Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling", *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 16, no. 4, pp. 766-778, May 2008.
- [12] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems", *IEEE Int'l Conf. on Acoustics, Speech and Signal Proc.*, pp. 1352-1355, 1983.
- [13] J. W. Xu and J. C. Principe, "A pitch detector based on a generalized correlation function", *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 16, no. 8, pp. 1420-1431, Nov. 2008.

- [14] A. de Cheveigné and H. Kawahara, “YIN, a fundamental frequency estimator for speech and music”, *Journal of the Acoustic Society of America*, vol. 111, no. 4, pp. 1917-1930, Apr. 2002.
- [15] C. L. Hsu and J. S. R. Jang, “On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset”, *IEEE Trans. Audio, Speech and Language Proc.*, vol. 18, no. 2, pp. 310-319, Feb. 2010.
- [16] C. L. Hsu, L. Y. Chen, J. S. R. Jang, and H. J. Li, “Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement”, *Proc. of 10th Int’l Society for Music Info. Retrieval*, pp. 201-206, 2009.
- [17] K. Dressler, “An auditory streaming approach on melody extraction”, *2nd Music Info. Retrieval Evaluation Exchange*, 2006.
- [18] G. E. Poliner et al., “Melody transcription from music audio: approaches and evaluation”, *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 4, pp. 1247-1256, May 2007.
- [19] J.-L. Durrieu, G. Richard, and B. David, “An iterative approach to monaural music mixture de-soloing”, *IEEE Int’l Conf. on Acoustics, Speech and Signal Proc.*, pp. 105-108, 2009.
- [20] J.-L. Durrieu, G. Richard, and B. David, “Singer melody extraction in polyphonic signals using source separation methods”, *IEEE Int’l Conf. on Acoustics, Speech and Signal Proc.*, pp. 169-172, 2008.
- [21] D. Klatt and L. Klatt, “Analysis, synthesis, and perception of voice quality variations among female and male talkers”, *Journal of the Acoustic Society of America*, vol. 87, no. 2, pp. 820-857, 1990.
- [22] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, “Source/filter model for unsupervised main melody extraction from polyphonic audio signals”, *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 3, pp. 564-575, Mar. 2010.
- [23] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, “Main instrument separation from stereophonic audio signals using a source/filter model”, *17th European Signal Proc. Conf.*, Aug. 2009.
- [24] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation”, *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.
- [25] M. Goto, “Development of the RWC music database”, *Proc. of 18th Int’l Congress on Acoustics*, pp. I-553-556, Apr. 2004.
- [26] [Online] <http://www.cs.tut.fi/sgn/arg/music/tuomasv/pitchnmf/>
- [27] [Online] http://en.wikipedia.org/wiki/Mel-frequency_cepstral_coefficient
- [28] [Online] http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/code/melfb.m
- [29] [Online] <http://perso.telecom-paristech.fr/~durrieu/en/icassp09/>

致 谢

感谢欧智坚老师对我耐心指导、热情鼓励，并提供了大量的参考文献。

感谢许肇凌提供 **MIR-1K** 数据库，并与作者讨论基音提取的性能评价标准。

感谢 **Durrieu** 向作者提供了源-滤波器模型的程序用作参考，并与作者讨论模型的实现和改进。

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。

签 名： 王赞 日 期： 2010.6.30

附录 A 外文资料的书面翻译

通过谐波估计和非负语谱分解分离人声与伴奏

Tuomas Virtanen, Annamaria Mesaros, Matti Ryynänen

摘要

本文提出了一种从音乐中分离人声和伴奏的新方法。在谐波估计的基础上，该方法首先在幅度语谱图上创建一个 0/1 蒙版，标示人声信号的谐波分量存在的时频单元。然后，对于蒙版标示的不属于人声的部分进行非负矩阵分解（NMF），以获得伴奏的模型。NMF 可以预测人声部分中噪声的大小，这使得即使在人声与噪声在时、频域均重叠时也能将它们分离开来。仿真表明，对于商业音乐和合成音乐，与利用正弦模型的参考算法相比，本文提出的算法的分离性能分别提高了 1.3 dB 和 1.8 dB，听觉质量也有了显著提高。通过该方法分离出来的人声还用于与文本歌词对齐的试验，结果也比参考方法要好。

关键词：声源分离，非负矩阵分解，无监督学习，基频估计

一、引言

在许多音频分析任务中，声源的分离是一个关键阶段，因为来自真实世界的录音往往含有多个声源。人耳具有很强的听辨出各个声源的能力。在对混合音频的计算分析中，往往也需要同样的能力。例如，在自动语音识别的现有算法中，加性干扰是一个重要的限制因素。

许多现有的单声道声源分离算法，或者利用了谐波估计，或者利用了语谱分解技术。谐波估计算法（简短综述见 2.1 节）利用了声音的谐波结构，对随时间变化的基频进行估计，并把它用于分离。另一方面，语谱分解技术（见 2.2 节）

利用声源的冗余性，将输入信号分解为若干重复的分量之和，并把每一个分量分配给一个声源。

本文提出的复合系统同时使用了谐波估计和无监督语谱分解，以达到更好的分离人声和伴奏的效果。第3章中提出的复合系统首先估计人声信号的基频，然后生成一个0/1蒙版，它覆盖了含有人声成分的时频区域。接下来，对非人声区进行非负语谱分解。由于伴奏声源的冗余性，由这一步可以得到人声区中伴奏分量的估计，从原始信号中减去由此估计出的伴奏即得到人声。第4章的仿真表明，这样做可以取得更好的分离质量。第5章把该系统分离出来的人声用于与文本歌词对齐的试验，证明它优于以往的算法。

二、背景

大多数已有的声源分离算法，都是基于谐波估计或非负语谱分解中的一者。下面两小节将对它们做简要的综述。

2.1 谐波估计

语音中的浊音以及有音调的乐器发出的声音，都是由谐波分量组成的，这些分量的频率约为基频 f_0 的整数倍。对于这种声音，一种有效的模型是正弦模型，即把每个分量建模在一个频率、幅度、相位均随时间变化的正弦波。

估计正弦模型参数的方法有很多。一种比较稳健的方法是，先估计出目标声源基频随时间的变化情况，再把这些基频用于更精确地提取每个谐波分量的参数。可以认为目标人声在混合信号中具有最显著的谐波结构，而提取最显著基频的算法有多种，如[1]和[2]。谐波频率可以认为是基频的整数倍，但Fujihara等[3]对此进行了改进，他们取功率谱上基频整数倍附近处的最大值所对应的频率为谐波频率。谐波幅度和相位可以直接采用语谱图中相应的幅度和相位值。

当每个谐波分量在每一帧中的频率、幅度、相位都估计出来之后，就可以通过插值获得平滑的幅度和相位轨迹。例如，Fujihara等[3]采用二次插值来处理相位。最后，这些正弦形式的谐波分量相加，就可以得到人声信号的估计。

上述过程可以得到比较好的结果，尤其是伴奏在谐波频率处能量不高时。但它的缺陷就是，在谐波频率处，所有的能量都被分配给目标声源。尤其是在

音乐信号中，各个声源的频率常常成简单的整数比，这使得它们有很多谐波分量具有相同的频率。另外，不具有谐波结构的声音在高频段往往也具有较高的能量，其中的一部分就会与目标人声的谐波分量重叠。这使得谐波幅度的估计值偏大，分离出来的人声频谱失真。Goto[2]曾经利用人声频谱的先验分布阐述过这个现象。

2.2 语谱分解

最近，语谱分解技术，如非负矩阵分解（NMF）及其推广，在声源分离问题中取得了很好的效果[4]。这些算法利用了声音在一段时间区间内的冗余性：通过把信号分解为一系列重复的频谱基之和，可以把每个声源分别用一个基的集合来代表。

这些算法通常在一个与相位无关的时频域表示，如幅度语谱上进行操作。记输入信号的幅度语谱为 \mathbf{X} ，其各个元素为 $\mathbf{X}_{k,m}$ ，其中下标 $k = 1, \dots, K$ 表示离散频率值， $m = 1, \dots, M$ 表示帧号。在 NMF 中，语谱被近似地表示为两个非负矩阵的乘积 $\mathbf{X} \approx \mathbf{S}\mathbf{A}$ ，其中 \mathbf{S} 的各列为各个频谱基， \mathbf{A} 的每行代表一个基在各帧中的强度。通过最小化 \mathbf{X} 和乘积 $\mathbf{S}\mathbf{A}$ 之间的某种误差，并限制 \mathbf{S} 和 \mathbf{A} 的元为非负，可以快速估计出 \mathbf{S} 和 \mathbf{A} 。一种常用的误差称为“偏离度”（divergence）：

$$D(\mathbf{X}||\mathbf{S}\mathbf{A}) = \sum_{k=1}^K \sum_{m=1}^M d(\mathbf{X}_{k,m}, [\mathbf{S}\mathbf{A}]_{k,m}) \quad (1)$$

其中偏离度函数 d 定义为

$$d(p, q) = p \log(p/q) - p + q. \quad (2)$$

计算出基以后，对应于目标声源的那些基就可以被检测出来，以进行进一步分析。上述方法存在的一个问题是，它只能学习到并分离出混合信号中有重复的频谱。如果目标声音的一部分在混合信号中只出现一次，那么它很可能并不能被很好地分离出来。

与音乐中的伴奏相比，人声的频谱通常更多样化。人声信号短时频谱的精细结构由其基频决定，而频谱的轮廓则取决于因素，即歌词。实际中，这二者都是随时间变化的。尤其是当输入信号很短时，这些性质使得学习所有频谱基非常困难。

Raj 等[5]曾经探讨过上述问题。他利用音乐中手动标注的非人声段训练得到一组伴奏的频谱，然后把伴奏频谱固定，学习人声部分的频谱。Ozerov 等[6]使用

过类似的方法，他们首先把信号分为人声段和非人声段，并用非人声段改进事先训练好的背景模型。这些方法需要时域上的非人声段，即只有伴奏没有人声的部分。

三、我们提出的复合系统

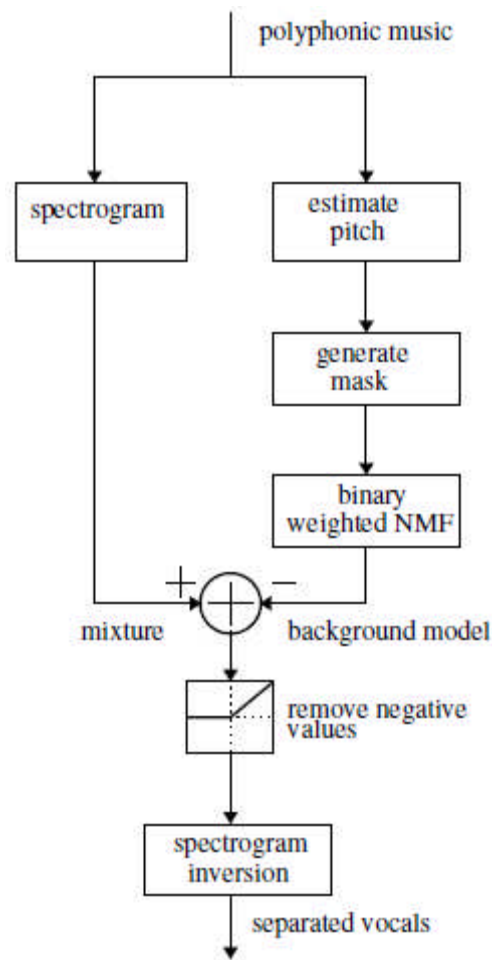


图1 本文提出的系统的框图，详细说明见正文。

为了克服仅利用谐波估计或仅利用无监督学习的方法的局限性，我们提出了一种复合系统，它发挥了两种方法的优点。图 1 是系统的框图。在右边的分支里，首先用谐波估计和 0/1 蒙版找到含有人声成分的区域，详见 3.1 节。然后对于剩下的非人声区进行非负矩阵分解，以学习伴奏的模型，详见 3.2 节。这一阶段同

时也预测出人声区内伴奏的语谱。从人声区中减去预测出的伴奏，并将余下的语谱逆变换回时域，就得到了人声信号的估计，详见 3.3 节。

3.1 由谐波分析求 0/1 蒙版

首先提取出输入信号中随时间变化的人声基频。本研究的主要对象是音乐信号，我们发现 Ryyänen 和 Klapuri 提出的旋律转写算法[7]给出了很好的基频提取结果。为了得到随时间变化的基频的准确估计，我们以基频显著性函数[7]在量化后的音调值附近的极大值作为准确的基频值。此算法每隔 20 ms 给出一个基频估计值。

根据估计出的基频，可以预测出人声在时频域上所占的区域。我们发现，基频提取算法的精确度足够好，以至于我们可以直接用基频的整数倍作为谐波的频率。NMF 在通过帧长为 N 的 DFT 得到的幅度语谱上进行操作。语谱图的频率轴由一系列离散频率 $f_s k / N$ 组成，其中 $k = 0, \dots, N/2$ ，因为只有奈奎斯特频率一半以内的频率才被使用。在每一帧中，每个谐波频率预测值周围的一定频率范围被标记为人声区。在我们的系统中，谐波频率周围 50 Hz 的带宽被标记为人声区，即如果某个离散频率处在以谐波频率为中心、宽度为 50 Hz 的频带内，它就被标记为人声区。在 $N = 1764$ 时，谐波频率周围会有 2~3 个频带被标记为人声区，具体的数值要看谐波频率与离散频率值的对齐情况。在实际中，谐波频率周围人声区的带宽取值至少应与窗长有关，在我们的系统中，窗长取为 40 ms。基频估计这一步可以预测出每个时频单元内是否有人声。对于清音段，所有频带均被标记为非人声区。

对每一帧进行上述过程之后，我们就得到了一个 $K \times M$ 的 0/1 蒙版 \mathbf{W} ，其中每个元素代表了相应时频单元是否属于人声区（0 表示人声区，1 表示非人声区）。0/1 蒙版的一个例子见图 2。

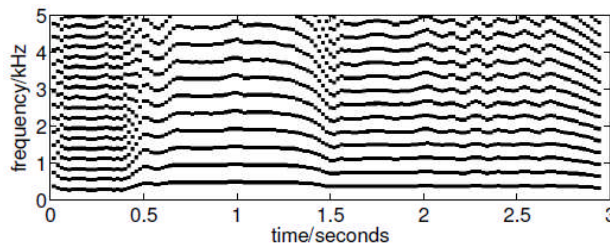


图2 估计出的0/1蒙版举例。黑色代表人声区。

3.2 0/1 加权的非负矩阵分解

利用 0/1 蒙版中标记为 1 的区域，即非人声区，可以训练出噪声模型。这个模型就是 NMF 中的模型，即噪声的幅度语谱被分解在两个矩阵——频谱基矩阵 \mathbf{S} 和强度矩阵 \mathbf{A} 。模型的计算是通过最小化观测到的语谱 \mathbf{X} 与模型 \mathbf{SA} 的偏离度进行的。人声区（蒙版中标记为 0 的区域）在计算中被忽略，即计算偏离度时，求和不包括这些区域。这使得在有人声的时间段内的非人声时频单元的信息也能得到利用。人声段内的非人声区域使得人声区内伴奏分量的预测成为可能。

计算背景模型时，要最小化如下的加权偏离度函数：

$$D_{\mathbf{W}}(\mathbf{X}||\mathbf{SA}) = \sum_{k=1}^K \sum_{m=1}^M \mathbf{W}_{k,m} d(\mathbf{X}_{k,m}, [\mathbf{SA}]_{k,m}) \quad (3)$$

它等价于

$$D_{\mathbf{W}}(\mathbf{X}||\mathbf{SA}) = D(\mathbf{W} \otimes \mathbf{X} || \mathbf{W} \otimes (\mathbf{SA})) \quad (4)$$

其中 \otimes 表示对应元素相乘。

要最小化加权偏离度函数，可以先用随机正值初始化 \mathbf{S} 和 \mathbf{A} ，然后交替使用如下两条乘法更新规则：

$$\mathbf{S} \leftarrow \mathbf{S} \otimes \frac{(\mathbf{W} \otimes \mathbf{X} \oslash \mathbf{SA}) \mathbf{A}^T}{\mathbf{WA}^T} \quad (5)$$

$$\mathbf{A} \leftarrow \mathbf{A} \otimes \frac{\mathbf{S}^T (\mathbf{W} \otimes \mathbf{X} \oslash \mathbf{SA})}{\mathbf{S}^T \mathbf{W}} \quad (6)$$

这里带圈的除法和分数线都表示对应元素相除。可以一直使用这两条更新规则直至收敛。在我们的研究中我们发现，迭代 30 次已经足以达到比较好的分离效果。

算法的收敛性可以证明如下。将加权偏离度写成如下形式：

$$D(\mathbf{W} \otimes \mathbf{X} || \mathbf{W} \otimes (\mathbf{SA})) = \sum_{m=1}^M D(\mathbf{W}_m \mathbf{x}_m || \mathbf{W}_m \mathbf{S} \mathbf{a}_m) \quad (7)$$

其中 \mathbf{W}_m 是对角矩阵，其对角元为 \mathbf{W} 的第 m 列； \mathbf{x}_m 和 \mathbf{a}_m 分别代表 \mathbf{X} 和 \mathbf{A} 的第 m 列。

在求和式(7)中，各帧的偏离度是相互独立的。因此，可以对每一帧分别导出更新规则。对于第 m 帧，(7)式右边可以写成

$$D(\mathbf{W}_m \mathbf{x}_m || \mathbf{W}_m \mathbf{S} \mathbf{a}_m) = D(\mathbf{y}_m || \mathbf{B}_m \mathbf{a}_m) \quad (8)$$

其中 $\mathbf{y}_m = \mathbf{W}_m \mathbf{x}_m$ ， $\mathbf{B}_m = \mathbf{W}_m \mathbf{S}$ 。对上式，可直接应用 Lee, Seung 在[8]中给出的更新规则：

$$\mathbf{a}_m \leftarrow \mathbf{a}_m \otimes \frac{\mathbf{B}_m^T (\mathbf{y}_m \oslash (\mathbf{B}_m \mathbf{a}_m))}{\mathbf{B}_m^T \mathbf{1}} \quad (9)$$

其中 $\mathbf{1}$ 是一个 $K \times 1$ 的全 1 向量。Lee, Seung [8] 证明了，在更新规则(9)的作用下，(8)式中的偏离度是单调不增的。将 $\mathbf{y}_m = \mathbf{W}_m \mathbf{x}_m$ ， $\mathbf{B}_m = \mathbf{W}_m \mathbf{S}$ 两式代回(9)式，即得

$$\mathbf{a}_m \leftarrow \mathbf{a}_m \otimes \frac{\mathbf{S}^T \mathbf{W}_m (\mathbf{x}_m \odot (\mathbf{S} \mathbf{a}_m))}{\mathbf{S}^T \mathbf{W}_m} \quad (10)$$

上式就是(6)式对 \mathbf{A} 的每一列的表达，因此(3)式中的加权偏离度在更新规则(6)的作用下是单调不增的。类似地，交换 \mathbf{S} 与 \mathbf{A} 的地位，并利用矩阵转置将(3)式写成

$$D_{\mathbf{W}}(\mathbf{X} \parallel \mathbf{S} \mathbf{A}) = D_{\mathbf{W}^T}(\mathbf{X}^T \parallel \mathbf{A}^T \mathbf{S}^T) \quad (11)$$

再利用同上的证明过程即可得到的更新规则(5)。

3.3 将人声语谱逆变换回时域

按下式可以求出人声信号的幅度语谱 \mathbf{V} ：

$$\mathbf{V} = [\max(\mathbf{X} - \mathbf{S} \mathbf{A}, 0)] \otimes (\mathbf{1} - \mathbf{W}), \quad (12)$$

其中 $\mathbf{1}$ 是一个 $K \times M$ 的全 1 矩阵。 $\mathbf{X} - \mathbf{S} \mathbf{A}$ 这步操作从混合信号中减去估计出的背景声音，而且我们发现，通过取最大值操作将这一步的结果限制为非负是有好处的。乘以 $\mathbf{1} - \mathbf{W}$ 是为了只允许人声区内幅度值为正。伴奏的幅度语谱可以通过 $\mathbf{X} - \mathbf{V}$ 算得。

图 3 显示了一段混合音乐信号以其从中分离出来的人声和伴奏的语谱图。人声中的浊音部分在语谱图中显示为随时间变化的梳状，它们基本都从伴奏中被除去了。

把幅度语谱与原始混合信号的相位语谱结合，即得到复数语谱。时域人声信号可以用重叠相加法求得。分离出的人声信号的例子可以参见 <http://www.cs.tut.fi/~tuomasv/demopage.html>。

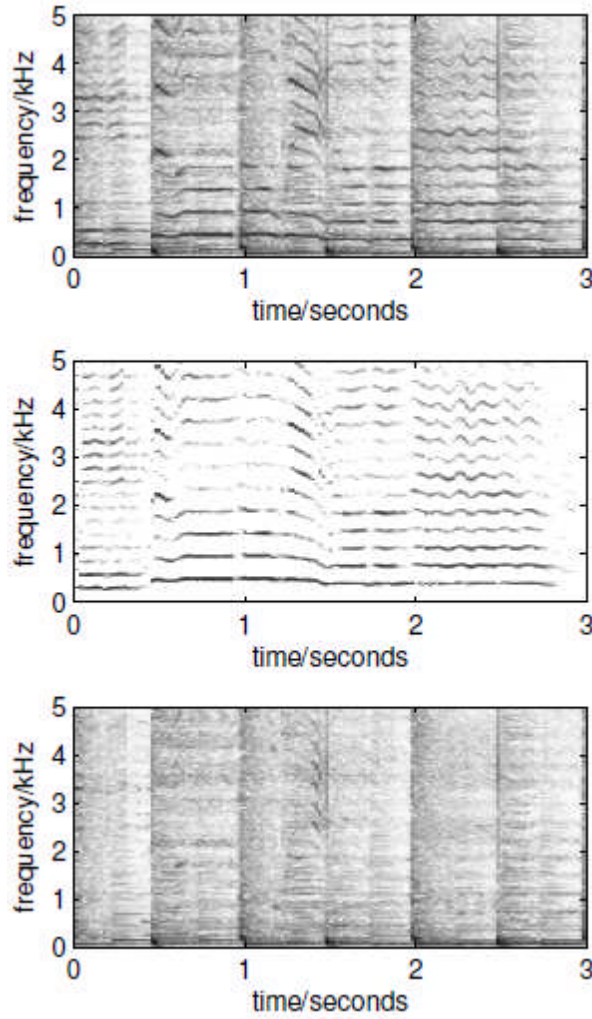


图3 语谱图分离举例：上为混合信号，中为分离出的人声，下为分离出的伴奏。颜色越深，对应时频单元的幅度越大。

3.4 讨论

对于 NMF 中基的数目（即矩阵 S 的列数），我们取了不同的值进行实验。此数目取决于输入信号的长度及复杂性，一般基的数目不必太多（10 ~ 20），迭代 10 ~ 30 次即能达到较好的效果。而且，方法对这些参数的具体取值似乎并不敏感。另一方面，我们发现如果基的数目和迭代次数太多，分离质量反而下降。这可能是由伴奏模型的过度适应（overfitting）引起的，也可能是因为伴奏模型学习到了求蒙版时伴奏中未被觉察到的人声成分。上述现象受蒙版的结构影响很大：如果某一帧中只有较少的频带被标记为人声，那么就很可能导致分离质量降低。对于最优 0/1 蒙版和 NMF 参数的进一步分析是后续研究的一个课题。

由于迭代次数较少，我们提出的方法运行相当快，在一台 1.9 GHz 的台式机上，总计算时间小于输入信号的长度。

除了 NMF 以外，更复杂的模型（如允许频谱随时间变化的模型，见[9,10]）也可以与 0/1 加权矩阵共同使用，但实际中发现 NMF 模型已经足够了。也可以扩展这个模型使它能够学习到人声的频谱基（如[5]），但这需要相对较长的人声信号，使得每一个音调/音素组合在信号中都出现多次。

四、仿真

我们使用两组音乐信号对我们提出的复合方法进行了定量评价。第一组测试集包含了 65 段歌曲，总长度约 38 分钟。每段歌曲是由人声信号与 MIDI 伴奏混合而成的。各段歌曲的人声-伴奏比统一调节为 -5 dB。

第二组测试集是从一张卡拉 OK DVD (Finnkidz 1, Svenska Karaokefabriken Ab, 2004) 上的九首歌中节选出来的。这张 DVD 包含了每首歌的带人声版和纯伴奏版。二者在采样点级别上是同步的，所以将伴奏从带人声版中减去，就可以得到纯人声。含有多个并发人声的段落（如倍频和声）被手动标记出来，不用于测试。这样，我们得到了大约 20 分钟的音频，段落长度从 10 秒钟至几分钟不等。DVD 测试集中的平均人声-伴奏为 -4.0 dB。

每个段落都用我们提出的方法和下面的参考方法进行了处理。所有的方法均使用了相同的旋律转写算法，即 Ryyänen 和 Klapuri 在[7]中提出的算法。所有方法均采用 40 ms 的窗长，相邻的窗之间有 50% 的重叠。谐波分量数均设为 60，使用 0/1 蒙版的方法均使用了相同的蒙版。NMF 中频谱基的数量为 20，迭代次数为 30。

- 正弦建模法：谐波分量的幅度和相位是通过计算加窗语音和与该谐波同频的复指数信号的互相关估计的。在合成正弦谐波信号时，相位采用二次插值，幅度采用线性插值。
- 0/1 蒙版法：不从人声区中减去伴奏成分，直接用式 $V = X \otimes (I - W)$ 计算人声幅度语谱。
- 我们提出的复合方法：除了原始方法之外，还测试了计算人声幅度语谱 V 时不乘以蒙版，即用式 $V = \max(X - SA, 0)$ 计算的方法。后者标记为“复合方法*”。

分离质量用分离出的人声中的人声-伴奏比来衡量，其计算公式为：

$$\text{VAR}[\text{dB}] = 10 \log_{10} \frac{\sum_n s(n)^2}{\sum_n (s(n) - \hat{s}(n))^2}, \quad (13)$$

其中 $s(n)$ 代表参考人声信号， $\hat{s}(n)$ 代表分离出的人声信号。对每段歌曲分别计算人声-伴奏比，并按各段歌曲的长度进行加权平均，就得到平均人声-伴奏比。表 1 展示了各种方法对两组测试集的分离性能。

表1 各被测方法的平均人声-伴奏比

方法	测试集	
	1（合成）	2（卡拉 OK）
复合方法	2.1 dB	4.9 dB
正弦建模法	0.3 dB	3.6 dB
0/1 蒙版法	-0.8 dB	2.9 dB
复合方法*	2.1 dB	4.6 dB

从结果可以看出，复合方法的分离性能优于正弦建模和 0/1 蒙版两种参考方法。所有的方法都能显著提高混合信号的人声-伴奏比（两个测试集中原始的人声-伴奏比分别为-5.0 dB 和-4.0 dB）。听一下分离出的信号可以发现，大多数误差，尤其是在合成测试集上，都在由于旋律转写的错误导致的。复合方法分离结果的听觉质量显著优于参考方法。复合方法*的性能与复合方法相比，在测试集 1 上是相同的，在测试集 2 上稍有逊色，这表明计算人声幅度语谱 \mathbf{V} 时乘以蒙版会稍微提高性能。

五、在音频-文本对齐中的应用（略）

六、结论

我们提出了一种从音乐中分离人声和伴奏的新方法。它结合使用了两种功能强大的技术：谐波估计和无监督非负矩阵分解。根据人声基频的估计，该方法能够利用输入信号幅度语谱中的非人声区学习到伴奏的模型，这使得我们能够从人声区中减去伴奏的成分。我们使用真实商业音乐和合成音乐对分离效果进行了测

试，新方法的性能显著优于参考方法。新方法也被用于人声与文本歌词的对齐，测试表明它能够小幅度提高现有方法的性能。

参考文献

- [1] M. Wu, D. Wang, and G. J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.
- [2] M. Goto, “A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals,” *Speech Communication*, vol. 43, no. 4, 2004.
- [3] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, “Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals,” in *IEEE International Symposium on Multimedia*, San Diego, USA, 2006.
- [4] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.
- [5] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, “Separating a foreground singer from background music,” in *International Symposium on Frontiers of Research on Speech and Music*, Mysore, India, 2007.
- [6] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, “Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, 2007.
- [7] M. Ryyänen and A. Klapuri, “Automatic transcription of melody, bass line, and chords in polyphonic music,” *Computer Music Journal*, vol. 32, no. 3, 2008, to appear.
- [8] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proceedings of Neural Information Processing Systems*, Denver, USA, 2000, pp. 556–562.
- [9] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Proceedings of the 5th International Symposium on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, September 2004.
- [10] T. Virtanen, “Separation of sound sources by convolutive sparse coding,” in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

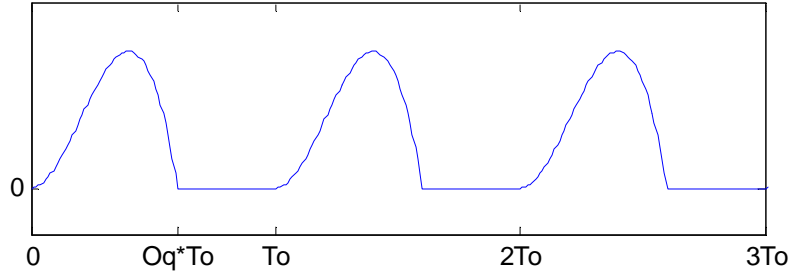
- [11] H. Fujihara and M. Goto, “Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection,” in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Las Vegas, USA, 2008.
- [12] Cambridge University Engineering Department. The Hidden Markov Model Toolkit (HTK), <http://htk.eng.cam.ac.uk/>.

书面翻译对应的原文索引

T. Virtanen, A. Mesaros, M. Ryyänen, “Combining pitch-based inference and non-negative matrix spectrogram factorization in separating vocals from polyphonic music”, *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Sep. 2008.

附录 B KLGLOTT88 模型

KLGLOTT88 模型提出的声门激励波形如下：



图附1 KLGLOTT88模型的声门激励波形

该波形有两个参数：基音周期 T_0 和开放商 Oq ，开放商就是一个周期内发声的时间占周期的比例，即波形的占空比。发声时间段内的波形用一个三次多次式来表达：

$$x(t) = \frac{t^2}{T_0 Oq^2} - \frac{t^3}{T_0^2 Oq^3}, t \in [0, OqT_0] \quad (\text{附-1})$$

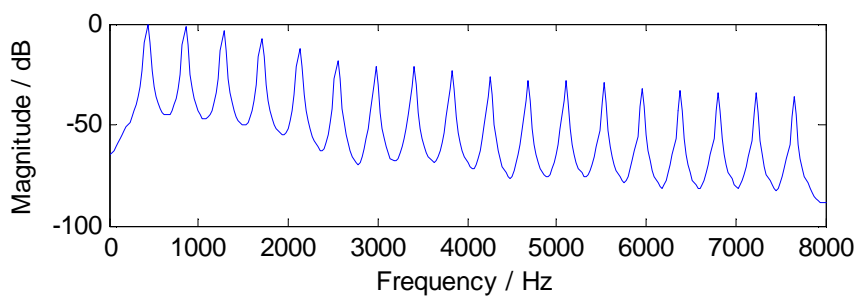
将某基频的声门激励波形加窗（与求语谱图时的窗函数相同）、做 FFT 并取模方，就得到了矩阵 \mathbf{B}_F 中的一列。 Oq 的具体取值对算法性能影响不大，我们取为 0.25。

事实上，我们并不是用上面的方法生成矩阵 \mathbf{B}_F 的。这是因为声门激励波形 $x(t)$ 在 $T = OqT_0$ 处有一个尖点，因此它在频域上不是带限的，会造成频谱混叠。我们没有直接根据(附-1)式计算 $x(t)$ ，而是模仿 Durrieu 在文献[22]的附录中提出的方法，首先求出 $x(t)$ 的傅里叶级数，舍去直流分量和超过 Nyquist 频率的那些项，再在时域上合成 $x(t)$ 。 $x(t)$ 的傅里叶级数如下：

$$x(t) = \sum_{h=0}^{\infty} c_h \exp\left(\frac{i2\pi h}{T_0} t\right), t \in [0, T_0] \quad (\text{附-2})$$

$$c_h = T_0 Oq \times \left[\frac{\exp(-i2\pi h Oq)}{(i2\pi h Oq)^2} + 2 \times \frac{1 + 2\exp(-i2\pi h Oq)}{(i2\pi h Oq)^3} - 6 \times \frac{1 - \exp(-i2\pi h Oq)}{(i2\pi h Oq)^4} \right] \quad (\text{附-3})$$




需要注意的是，文献[22]的附录中给出的傅里叶级数对应的是 $x(t)$ 的二阶导数而不是 $x(t)$ 本身。但是这没有关系，因为二者的傅里叶级数只是差一个频率的平方因子，相当于一个包络，这个包络的逼近可以交给声道部分去完成。



图附2 基频为425Hz的声门激励功率谱

KLGLOTT88 模型生成的 \mathbf{B}_F 矩阵的每一列都呈现一个有包络的梳状，例如如图附 2。事实上，这个包络并没有什么必要，也可以以理想冲激串作为声门激励波形，重复上面的步骤生成 \mathbf{B}_F 。文献[22]中说使用 KLGLOTT88 模型比使用理想冲激串的效果更好，但我们的实验没有观测到这一点。

综合论文训练记录表

学生姓名	王赞	学号	2006011130	班级	无 65
论文题目	单声道音频中人声与伴奏的分离				
主要内容以及进度安排	<p>主要内容:</p> <p>实现一个单声道音频人声与伴奏分离系统并进行性能测试。系统由以下几个环节组成:</p> <ol style="list-style-type: none"> 1) 伴奏、清音、浊音 (A/U/V) 的判决; 2) 基音轨迹提取; 3) 人声与伴奏的分离。 <p>进度安排:</p> <p>第 1~3 周: 文献调研、开题;</p> <p>第 4~6 周: 探索用芬兰派分离人声与伴奏, 效果不理想;</p> <p>第 7~10 周: 实现用 HMM 分类法进行 A/U/V 判决;</p> <p>第 11~12 周: 实现用 HMM 分类法进行基音提取;</p> <p>第 13~14 周: 探索用 CASA 方法分离人声与伴奏, 效果不理想;</p> <p>第 14~15 周: 用 Durrieu 的方法分离人声与伴奏, 达到了很好的效果;</p> <p>第 16 周: 优化系统, 进行性能测试;</p> <p>第 17 周: 撰写论文, 准备答辩。</p> <p>指导教师签字: </p> <p>考核组组长签字: </p> <p>2010 年 3 月 19 日</p>				
中期考核意见	<p>王赞同学积极认真开展工作, 进展顺利, 取得了很好的中期成果, 预期能按时完成论文工作。</p> <p>考核组组长签字: </p> <p>2010 年 4 月 27 日</p>				

指导教师评语	<p>单声道音频的信号分离是信号处理的一个难点问题。论文实现了一个完整的单声道人声伴奏分离系统。它采用 HMM 分类方法进行伴奏、清音、浊音的判决和基音轨迹的提取，并对人声建立了一种基于非负矩阵分解的源-滤波器模型，使用软蒙版法分离人声与伴奏。在两个国际公开数据库上的性能测试表明，该分离系统的性能达到了人声伴奏分离的前沿水平，并具有更快的处理速度。王赞同学工作认真，学习能力和动手水平突出，论文书写流畅，结构清楚，图表规范，很好完成了本科综合论文训练。</p> <p>指导教师签字: <u>张智强</u></p> <p>2010年6月30日</p>
评阅教师评语	<p>王赞同学的论文对音乐中的人声与伴奏的分离技术进行了研究，论文详述了硬蒙版法和软蒙版法的技术特点，研究了 AM-V 判决和基音提取的算法，最终采用软蒙版法实现了人声与伴奏的分离，取得了满意的效果。该论文书写规范，论文清楚，理论分析严谨，实验数据充分，具有很好的参考价值，是一篇优秀的本科综合训练论文。</p> <p>评阅教师签字: <u>张智强</u></p> <p>2010年6月30日</p>
答辩小组评语	<p>王赞同学出色完成了毕业设计工作。论文工作量大，难度高，成果突出。在论文答辩中思路清楚，回答问题正确。同意通过论文答辩，并推荐参评优秀本科论文。</p> <p>答辩小组组长签字: <u>张智强</u></p> <p>2010年7月1日</p>

总成绩: 93

教学负责人签字: 张智强

2010年7月5日