ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.7, No.10, April 2011,pp.2384-2386

E-mail: eduf@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

一种新的哼唱音符音高划分方法

杨剑锋,冯寅

(厦门大学 智能科学与技术系,福建 厦门 361005)

摘要:哼唱音符音高的准确划分,对哼唱音乐检索系统识别率的提高起着很大的作用。目前,大部分的哼唱音乐检索系统都是采用能量划分的方法,在很大程度上并不能对哼唱波形文件顺利完成单音切割,因此,论文提出的一种新的音符音高划分方法,在基于一般能量划分的基础上,采用基于倍音列的音高识别模型对划分结果进行二次划分、规整,最终实现哼唱音符音高的划分。实验表明,该划分方法能够有效地实现哼唱音符音高的准确划分。

关键词:哼唱;音符;能量切割;检索;倍音列模型

中图分类号:TP311 文献标识码:A 文章编号:1009-3044(2011)10-2384-03

A New Method of Note Segmentation for Humming Music

YANG Jian-feng, FENG Yin

(Cognitive Science Department, Xiamen 361005, China)

Abstract: The accuracy of note segmentation of humming music is very important for the recognition rate of humming music retrieval system. Currently, most humming music retrieval systems, adopting the method of energy segmentation, can not successfully complete the single-tone segmentation of humming music. Therefore, a new method's proposed in this paper, which based on the traditional energy segmentation method and overtone series model. The experiments show that the new method proposed in this paper can effectively completes the single-tone segmentation of humming music.

Key words: humming; note; energy segmentation; retrieval; overtone series model

随着信息技术的飞快发展,网络上的信息资源也呈爆炸性的增长。基于文本的信息检索技术已经相对比较成熟,然而对于基于内容的音频等多媒体信息检索的研究仍寥寥无几,单纯的文本或数值信息的检索方式已经远远不能满足应用的需求。因此,研究一种更为自然、方便、人性化的音频信息检索方式对于信息检索技术的发展是一项非常有应用价值的工作。其中,哼唱音符划分技术是基于内容音乐检索系统研究的重中之重[1-2]。本文旨在传统音符划分的基础上,采用一种新颖的音高识别方法进一步对划分的结果进行二次划分,最终有效地实现了哼唱音符音高的划分。

1 基于倍音列的音高模型

在介绍倍音模型之前,需要先引入基本乐音听觉能力的定义[3]。

1.1 基本乐音听觉能力定义

哼唱者 A 是具备基本乐音听觉能力的,如果 A 能通过他(或她)的听觉判明同时播放的任意二个音或哼唱旋律中依次播放的前后二个音是否为:

1)同度(或相差若干八度)音程关系

2)相差半音音程或更大音程的非同度关系且辨明哪一个音的音高更高(或低)

1.2 单音哼唱定义

设哼唱者 A 具备基本乐音听觉能力,称哼唱 S_H 为单音哼唱,如果 S_H 可由 A 的听觉,判断为仅包含唯一一个确定音高且为单音节的哼唱。

于是,我们可以把任一哼唱 H 视为一个单音哼唱序列 $S_{H1}, S_{H2},\cdots,S_{HN}(N\geqslant 1)$。需要说明的是不少歌手对其哼唱往往会实施一些技术性的处理。例如,在半音到全音音程范围内上下颤动以润色某一哼唱音等等。这里,我们不考虑这些情况。因为具备这种能力的哼唱者同样有能力在其哼唱时尽可能地去掉这样的哼唱技术而不影响旋律哼唱的正确性。

根据倍音列理论[4],一个有确定音高的乐音是一种复合音。它由基音及其倍音(也称谐音、泛音)组成。基音的音高就是人耳听到的这个乐音的音高(基频音高)。而其它的倍音成分决定这个音的音色。任意一个有确定音高的人声哼唱音就是一种复合音。其中,基音及其倍音构成这个复合音的倍音列(也称泛音列)。它们分别称为:1 倍音(基音)、2 倍音、3 倍音、……一般地,若一个复合音 T 的音高频率是 F_T,则它的 K 倍音的音高频率就是 $K*F_T$,这里 K 为正整数。

一个人声哼唱的复合音中的基音成分所占比例并非总是最大。有时,最大者有可能是复合音倍音列的 3 倍音或 5 倍音等。我们在处理这个问题时,是基于这样一个假设,即:只要所哼唱的音可被具有如定义 1.1 所述的基本乐音听觉能力的哼唱者判明有确定的音高,那么,均认为这个音的成分只能由音高频率为其基音频率的整数倍的倍音列所组成。

收稿日期:2011-01-22

作者简介:杨剑锋(1986-),男,福建泉州人,厦门大学智能科学与技术系,硕士,主要研究方向为基于内容的音乐检索;冯寅(1963-),男,福建福州人,副教授,博士,主要研究方向为算法作曲,计算机音乐和自然语言处理。

1.3 半音音阶音高名组合定义

设 T_D 是一个音域中所有半音音阶的音高名组成的集合。P_N 是集合 TD 中的一个音高名。F_{PN} 是表 1 中对应音高名 PN 的标准音高频率，称实数区间[FPN*1/$\sqrt[24]{2}$,FPN*$\sqrt[24]{2}$]，为音高名 P_N 的十二平均律频率界定区间，简称 P_N 的频率界定区间。

对以 44kHz 为采样率的哼唱时域信号波形，以每 2048 个点为一个窗口长度，相邻窗口重叠为 1024 个点。作 FFT 变换后得一窗口序列。其中，每一窗口描述了哼唱信号在这个窗长瞬间的幅频特性。设 $A_w(F)$为哼唱信号在窗口 w 中频率值为 F 的振幅取值。假定这个哼唱信号正好是一个单音哼唱 S_H。w 是处于单音哼唱 S_H 所对应的信号波形中的一个窗口。记作 $w \in S_H$。单音哼唱 S_H 的音高名是集合 T_D 中的一个元素。设乐音 T_{c1} 的音高名为 c_1，其音高频率为 F_{c1}，则表 1 描述了其倍音列的前 6 个倍音的音高频率 $f_K(c_1)$对应的音高名 P_{NK} 及该音高名在 12 平均律下的音高频率 $f_{12K}(c_1)$，K=1,2,3,4,5,6。

表 1 乐音 T_{c1} 倍音列的前 6 个倍音的音高频率，对应的音高名及该音高名 12 平均律下的音高频率

倍音列	1 倍音	2 倍音	3 倍音	4 倍音	5 倍音	6 倍音
倍音的音高频率 $f_K(c_1)$	F_{c1}	2 *F_{c1}	3* F_{c1}	4* F_{c1}	5 *F_{c1}	6* F_{c1}
倍音的音高名 P_{NK}	c_1	c_2	g_2	c_3	e_3	g_3
P_{NK}在 12 平均律下的音高频率 $f_{12K}(c_1)$	F_{c1}	$2*F_{c1}$	$2^{\frac{7}{12}}*F_{c1}$	$4*F_{c1}$	$2^{2\frac{1}{3}}*F_{c1}$	$2^{2\frac{7}{12}}*F_{c1}$

这里需要注意的是乐音 T_{c1} 的 3 倍音、5 倍音和 6 倍音的音高频率和相应的音高名在 12 平均律下的音高频率有差异（表现为非整倍数的无理数倍数关系），但其误差很小。事实上，人耳的中耳和内耳的共振关系已将那些偏离整数关系的音程纳入到整数关系中。若以 12 平均律下的音高频率作为相应音高名的标准音高频率，则这些倍音列的音高频率应还在相应音高名的频率界定区间内。一般地，设有单音哼唱 S_H。因其有唯一的音高，不妨设其音高名为 P。令单音哼唱 S_H 的 6 个倍音的相应音高名为 P_K，K=1,…,6。在 12 平均律下它们的音高频率分别为 $f_{12}(P_K)$。相应的界定区间 $D(f_{12}(P_K))=[f_{12}(P_K)*\frac{1}{2^{24}\sqrt{2}}, f_{12}(P_K)*\sqrt[24]{2})$。其中，1 倍音（基音）频率 $f_{12}(P_1)$就是 P 的音高频率。设 w 为单音哼唱 S_H 信号波形中的任一窗口，记：$S_{AK}(S_H, P)=\sum_{w\in S_{H,F}\in D(f_{12}(P_K))} A_w(F)$为出现在单音哼唱 S_H 的信号波形中所有窗口 w 中的所有频率值处于频率界定区间 $D(f_{12}(P_K))$的振幅总和。简称单音哼唱 S_H 的 K 倍音振幅和，其中，$A_w(F)$为窗口 w 中频率值为 F 的振幅取值。我们把单音哼唱 S_H 中的 1~6 倍音振幅和相加，得：

$$S_6(S_H, P)=\sum_{K=1}^{6} S_K(S_H, P) \tag{1-1}$$

我们认为，S_H 的音高名 P 必为使式(1-1)的值达到最大的音高名集合 T_D 中的一个元素。我们把此法称为基于倍音列的音高计算模型。

2 音高切割法

音高切割法，要基于一个假设，那就是，我们假定任何一个哼唱输入旋律中的音，必须持续一定的时长，这样才能使人们能够凝听得到它的音高[3]。换一句话说，如果我们直接对一个哼唱输入波形文件做 FFT 所得到的窗口序列[5]：

$$W_1W_2\cdots W_N \tag{2-1}$$

序列(2-1)可以划分成 M 个子序列：

$$W_{11}W_{12}\cdots W_{1R1} \tag{2-2}$$

$$W_{11}W_{12}\cdots W_{1R2} \tag{2-3}$$

…

$$W_{11}W_{12}\cdots W_{1RM} \tag{2-M}$$

这里，$W_1W_2\cdots W_N = W_{11}W_{12}\cdots W_{1R1}\ W_{11}W_{12}\cdots W_{1R2}\cdots W_{11}W_{12}\cdots W_{1RM}$

上面所述窗口序列(2-1)的最后一个窗口下标 $N=M_{RM}$。

这 M 个窗口序列的每一个序列描述哼唱输入波形文件的一个或多个具有相同音高的音构成的音块。其中，

第一个音块是窗口序列：$W_{11}W_{12}\cdots W_{1R1}$

第二个音块是窗口序列：$W_{11}W_{12}\cdots W_{1R2}$

……

第 M 个音块是窗口序列：$W_{11}W_{12}\cdots W_{1RM}$

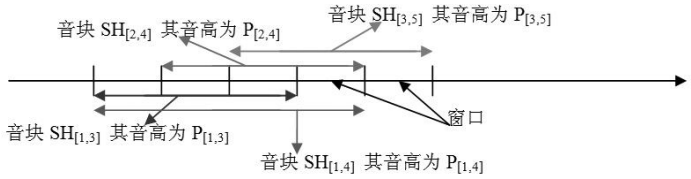
2.1 音高划分举例

我们假定，一个音块，最少包含 k(k>=3)个窗口以使得人可以凝听得到一个确定的音高的最短发音时长。用公式(1-1)计算在 k 个相邻窗口中使 $S_6(S_H, P)$达到最大值的音高频率。这里假设单音哼唱 SH 包含 k 个相邻窗口。（考虑到窗口重叠，1 个窗口的时长是 40ms，那么 3 个窗口大概是 80ms 左右）。我们通过此法可在 k 个相邻窗口中，估算出一个音块的一个基音音高 P。一个音块最少应包含三个相邻的窗口。

假设 $S_6(S_H[1,3], P[1,3])$是一个音块的最大振幅值，这里，P[1,3]是在这个音块中计算出来的音高。$S_H[1,3]$表示由第 1 到第 3 窗口构成的音块。

假设 $S_6(S_H[2,4], P[2,4])$是一个音块的最大振幅值，这里，P[2,4]是在这个音块中计算出来的音高。$S_H[2,4]$表示由第 2 到第 4 窗口构成的音块。

假设 $S_6(S_H[1,4], P[1,4])$是一个音块的最大振幅值，这里，P[1,4]是在这个音块中计算出来的音高。$S_H[1,4]$表示由第 1 到第 4 窗口构成的音块。



1) 音块合并 1：如果 P[1,3]=P[2,4]=P[1,4], 则音块 SH[1,3]和音块 S_H[2,4]可合并为音块 S_H[1,4]. 且该音块的音高是 P[1,4]=P[1,3]

2) 音块合并 2：如果 P[1,3]≠P[2,4]但 P[1,3]=P[1,4]且 P[2,4]≠P[3,5] , 则音块 S_H[1,3]和音块 S_H[2,4]可合并为音块 S_H[1,4]. 且该音块的音高是 P[1,4]=P[1,3]

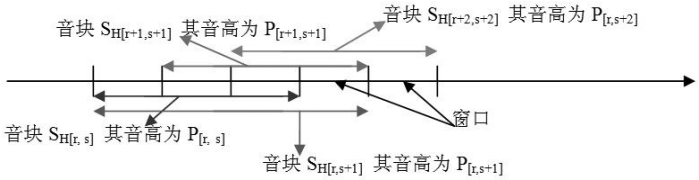
3) 音块边界确认：如果 P[1,3]≠P[2,4]且 P[1,3]≠P[1,4]则 SH[1,3]被认为是一个边界确认音块，且该音块的音高是 P[1,3]

2.2 音块划分一般情形

一般地，假设 $S_6(S_H[r, s], P[r, s])$是一个音块的最大振幅值，这里，P[r, s]是在这个音块中计算出来的音高。S_H[r, s]表示由第 r 到第 s 窗口构成的音块。这里，s−r≥2。

假设 $S_6(S_H[r+1,s+1], P[r+1,s+1])$是一个音块的最大振幅值，这里，P[r+1,s+1]是在这个音块中计算出来的音高。S_H[r+1,s+1]表示由第 r+1 到第 s+1 窗口构成的音块。

假设 $S_6(S_H[r,s+1], P[r,s+1])$是一个音块的最大振幅值,这里，P[r,s+1]是在这个音块中计算出来的音高。S_H[r,s+1]表示由第 r 到第 s+1 窗口构成的音块。



1) 音块合并 1：如果 P[r, s]=P[r+1,s+1]=P[r,s+1], 则音块 S_H[r, s]和音块 S_H[r,s+1]可合并为音块 S_H[r,s+1]. 且该音块的音高是 P[r,s+1]=P[r, s]

2) 音块合并 2：如果 P[r, s]≠P[r+1,s+1]但 P[r, s]=P[r,s+1]且 P[r+1,s+1]≠P[r+2,s+2] , 则音块 S_H[r, s]和音块 S_H[r+1,s+1]可合并为音块 S_H[r,s+1]. 且该音块的音高是 P[r,s+1]=P[r, s]

3) 音块边界确认：如果 P[r, s]≠P[r+1,s+1]且 P[r, s]≠P[r,s+1], 则 S_H[r, s]被认为是一个边界确认音块，且该音块的音高是 P[r, s]

一个音块可以通过音块合并扩展其窗口数。最后，通过音块边界确认形成最终音块。一旦一个音块边界获得确认，就开始其下一个音块的确认过程，直到整个哼唱输入波形文件做 FFT 所得到的窗口序列被划分成 M 个音块，通过这样音高划分音块的过程就得以完成。

3 实验结果

能量划分的方法[6]虽然能够在一定程度上对哼唱音乐实现较为简单的划分，且具有运算速度快，算法简单等优点，但是，还远远不能满足哼唱音乐检索系统中对哼唱音符划分准确率的要求，如图 1 所示，该实验结果是采用能量划分的方法得到的。通过对哼唱音符波形文件的回放可知该划分方法没有很好地实现音符的划分。而图 2 结果是采用本论文的划分方法得到的结果，可以发现，在能量划分方法没有划分到位的位置，该方法实现了很好的划分。

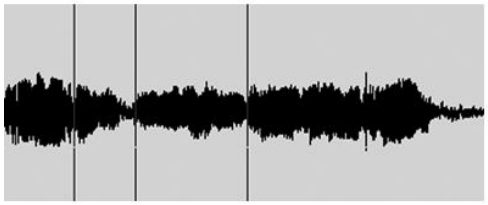


图 1　传统划分方法

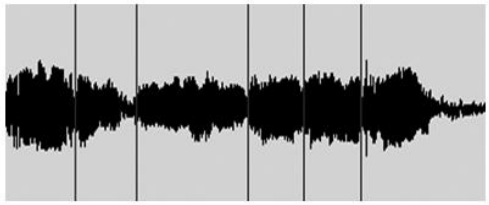


图 2　改进型的划分方法

4 结论

传统的能量划分方法虽然具有算法简单，运算速度快，容易实现等优点[8]，但是对于音符连续、发音急促等情况则不能很好的划分，本文提出的哼唱音符音高划分方法，首先介绍了基于倍音列音高识别模型，然后在能量划分的基础上，采用倍音列音高识别进行音符的再次划分、归并、规整，实验结果也表明了该方法的有效性，最终能够得到一个较为满意的哼唱音符划分结果。

参考文献：

[1] 李明.基于哼唱的音乐检索研究[D].北京:中国科学院声学研究所,2005.
[2] 张宝华,张品.基于旋律的音乐检索[J].电声基础,2005:4−7.
[3] 李政缪.中国传统律学[M].福建:福建教育出版社,2008.
[4] 郭红波.基于内容的音乐检索关键技术的研究[D].西安:西北大学,2007.
[5] 李扬,吴亚栋,刘宝龙. 一种新的近似旋律匹配方法及其在哼唱检索系统中的应用[J].计算机研究与发展,2003,40(11):1554−1560.
[6] 王小凤.基于内容的音乐检索关键技术研究[D].西北大学,2008.
[7] 赵力.语音信号处理[M].北京:机械工业出版社,2003.
[8] Yi peng Li, De liang wang. Separation of singing voice from music accompaniment for monaural recording. IEEE transactions on audio, vol.15, No.4, May 2007.