文章编号:1001 - 9081(2005)04 - 0968 - 03

乐音识别方法及应用

徐国庆[1,2],杨　丹[1],王彬洁[3],文俊浩[1]

(1. 重庆大学 软件学院,重庆 400030; 2. 武汉化工学院 计算机科学与工程学院,湖北 武汉 430074;
3. 武汉化工学院,湖北 武汉 430074)

(xu_guoqing@hotmail.com)

摘　要:通过研究乐音的声音和物理特性,提出一种识别乐音信号的方法,该方法实现在频域的精确定位,在基音频率检测上优于单一小波方法,在识别效果方面优于 DTW 方法,并开发了乐音识别和自动作曲系统。此方法可以为乐音识别提供参考。

关键词:乐音识别;端点检测;小波变换;快速傅立叶变换

中图分类号: TP319;TP391.4　　**文献标识码**:A

Method and the application of musical tone recognition

XU Guo-qing[1,2], YANG Dan[1], WANG Bin-jie[3], WEN Jun-hao[1]

(1. *Faculty of Software Engineering, Chongqing University, Chongqing* 400030, *China;*
2. *College of Computer Science and Engineering, Wuhan Institute of Chemical Technology, Wuhan Hubei* 430074, *China;*
3. *Wuhan Institute of Chemical Technology, Wuhan Hubei* 430074, *China*)

Abstract: This thesis investigate the musical and physical characters of the musical sound, and advances a new method to recognize the musical tone firstly. This method realizes the precise frequency resolution. It is more advantageous than the pitch detection method and the single wavelet method. It is also more effective than Dynamic Time Warping method. This method is ideal for the pitch recognition of the musical sound. It is first advanced in the technology domain and it has a high application value.

Key words: musical tone recognition; edge detection; DWT; FFT; melodize automatically

乐音识别是实现自动谱曲的基础,在音乐创作中有很重要的实用价值。但这方面的研究比较少,大多局限在乐音的音效处理、编辑等方面,如国外较著名的 CakeWalk,国内的作曲大师系列等。在音符录入方面,则一般使用 MIDI 键盘通过简单的映射实现。乐音的识别问题涉及到乐音的声学和物理学特性,其频域成分的提取在很大程度上决定了音符的性质。

本文针对乐音音符的识别进行研究,提出了一套识别方法,并成功地应用于智能作曲系统的开发。

1　乐音特性分析

一段连续的乐音是由诸多的单音构成的,从物理学角度看,单音主要由基频、振幅及倍频三个要素构成。乐器发出的乐音通过人耳的听觉系统反映到听觉神经中枢,引起听者的主观感觉。这种感觉形成心理学上的乐音三要素,即音调、响度、音色,这三个特性分别和三个客观上易于确定的物理量密切相关。乐音的这种特性使其能够用物理的方法进行分析和测量。具体地讲,单音的音调(音高)是这个单音的基频给人的主观感觉。一个单音的响度,也就是常说的音强,是这个单音的空气振动到达人耳处的能流给人的主观感觉。音色的形成比较复杂,是人脑对听觉感受的单音频谱(即各谐波成分比例)的主观感觉和判断。

乐音信号也是典型的时变信号,在一个音乐片段中包含多个不同频率的单音。但是乐音的频域组成具有其明显的平稳特性,就一个音符的发音来看,从开始发音直到乐音消失,其中的基音及泛音完全确定。频率成分不变(只是幅值逐渐减小)。也就是说,从频域角度来看,单一的音符是典型的平稳时不变信号。

2　乐音识别方法

2.1　端点检测

对于连续乐音,首先要对其进行端点检测,以分割出单音,分割单音的目的在于使连续乐音的识别转化为单音识别,并且分割单音后可以准确计算音符的时值。

在语音的端点检测方法中,比较有效的端点检测方法是 FRED (Feature-based Real-time Endpoint Detection, 基于语音特征的实时端点检测算法)算法[1],该算法基于两级端点检测方案,可以更好地适应环境的干扰和变化,提高端点检测的精度。其第二步算法主要用来区分清、浊音,由于乐音中的频率构成单一,并且单音的能量在持续期内呈一致振荡衰减,所以可以只使用第一级 FRED。

在使用之前要将乐音进行分帧,设置帧长为 m,分成 k 帧。逐帧求出其短时能量和过零率 s_{ij}、q_{ij}(下标表示第 i 帧的第 j 个采样),元素为布尔值:

$$s_{ij} = \begin{cases} 1 & abs(u_{ij} - u_{i(j+1)}) > \delta \\ 0 & abs(u_{ij} - u_{i(j+1)}) < \delta \end{cases}$$

$$q_{ij} = \begin{cases} 1 & u_{ij}u_{i(j+1)} < 0 \\ 0 & u_{ij}u_{i(j+1)} > 0 \end{cases}$$

收稿日期:2004 - 10 - 19;修订日期:2005 - 01 - 06

作者简介:徐国庆(1974 -),男,江苏徐州人,讲师,硕士研究生,主要研究方向:乐音识别; 杨丹(1962 -),男,教授,博士生导师,主要研究方向:科学与工程计算、软件工程; 王彬洁(1977 -),女,主要研究方向:钢琴艺术; 文俊浩(1969 -),男,副教授,博士,主要研究方向:软件过程与管理、数据仓库与数据挖掘.

构成短时能量和过零率矩阵：

$$A = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & & \vdots \\ s_{k1} & \cdots & s_{km} \end{bmatrix}, \quad Z = \begin{bmatrix} q_{11} & \cdots & q_{1m} \\ \vdots & & \vdots \\ q_{k1} & \cdots & q_{km} \end{bmatrix}$$

将短时能量和过零率矩阵求积：

$Edge = A * Z$

Edge 称为乐音的端点判断矩阵，其元素为乐音帧的短时突变状态。基于端点判断矩阵进行实际端点判定的过程简单表述如下：

（1）设置每次参与判定的帧数端点判定标志：0 表示无乐音，1 表示从无乐音到有乐音；2 表示从有乐音到无乐音；

（2）0 状态向 1 状态转换：此时必须再检测本帧后连续 t 帧的情况，如连续 t 帧均有乐音，则可确定本帧为端点开始点，否则是短时高频尖峰噪声；

（3）2 状态向 0 状态转换：此时必须再检测本帧后连续 t 帧的情况，如连续 t 帧均无乐音，则可确定本帧为端点结束点。

将确定的端点位置数据映射到乐音中就是乐音中的单音端点位置。用这种方法计算量不大，检测的效果很明显。

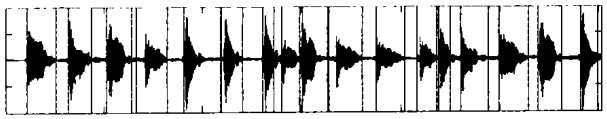

图 1　端点检测

2.2　乐音识别

人的听觉系统和听觉中枢并非是个严格的线性系统，采用反映人的耳朵的听觉特性的 Mel 频率倒谱系数 MFCC 能很好地提高识别性能。在频率为 1 000Hz 以下，听觉能力与频率呈线性关系，而在 1 000Hz 以上，听觉能力则与频率呈对数关系，Mel 频率和音频的关系为：

$$f_{mel} = 2595\lg\left(1 + \frac{f}{700}\right) \tag{1}$$

MFCC（Mel Frequency Cepstrum Coefficients，Mel 频率倒谱系数）计算过程[2]如图 2 所示：

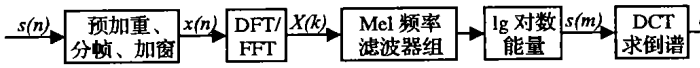

图 2　MFCC 计算过程

动态时间规整方法（Dynamic Time Warping，DTW）中利用 MFCC 参数进行对比识别，在进行特定人语音识别中效果比较理想。由于乐音的单音具有平稳周期时域特性，通过比较 FFT 和 DTW 方法的实际识别结果发现：FFT 识别单音的基频有极佳的频域定位（如图 3 所示）。

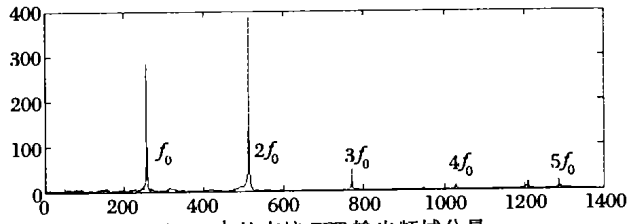

图 3　c^1 的直接 FFT 输出频域分量

从识别结果看来，频域中的最大能量成分并不是基频成分，而是 2 倍频成分。由于 FFT 的输出最高频率实际只有输入最高频率的一半，如果能够控制 FFT 的输入的时域信号中的最高频率为待识别乐音的 2 倍频到 4 倍频之间，然后再进行 FFT，那么输出量就只含有乐音的基音分量。

为了迅速将单音分解到基频所在的低频带，可以使用小波分解，工程界广为采用的 daubechies 小波[3]，其近似函数平滑性好，高频分解迅速。小波分解到尺度为 5 的近似函数可以比较清楚地从波形上反映出基音的周期情况[4]，另外对于 daubechies 小波，阶次越高，其低频近似函数的平滑性越好，但是由于滤波器个数是小波阶次的 2 倍，增加小波阶次意味着乘法计算量增加一倍，所以一般选取 4 阶 daubechies 小波。

对单音首先进行离散小波变换 DWT，分解乐音到低频尺度，分解的原则是把尽可能多的能量保留到分解尺度上，要使分解后的尺度分量只包含基音及其二倍频率分量，而不包含其他高倍频泛音分量。本文总结出下面的小波分解尺度计算公式：

$$j = \left[\log_2\left(\frac{f_s}{f_0}\right)\right] - 1 \tag{2}$$

其中 j 是分解尺度数，f_s 是信号采样频率，f_0 是信号中包含的待识别乐音的基音频率[5]，$[x]$ 为不超过 x 的最大整数。按照这个公式可以求出所有乐音的小波分解尺度。取采样频率 f_s 为 22 050Hz，以中音区为例计算结果如下：

表 1　中音区音符小波分解尺度

中音	c^1	d^1	e^1	f^1	g^1	a^1	b^1
f_0	261	293	329	349	392	440	494
j	5	5	5	4	4	4	4

对于中音区 c^1，采用 4 阶 daubechies 小波进行尺度为 5 的低通滤波分解，对低通分量实施 FFT，输出的频域识别结果如图 4，FFT 输出只含有一个基音分量，而不再含有各倍频泛音分量。

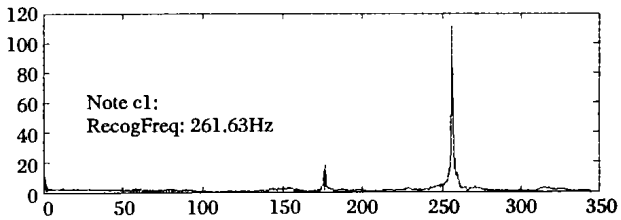

图 4　经尺度为 5 小波低通滤波后 c^1 的 FFT 输出

3　乐音识别应用和讨论

对中音区音阶（$c^1 \sim c^2$）采用上述乐音识别方法频率的识别结果如表 2。

表 2　音阶 $c^1 \sim c^2$ 识别结果及误差

识别音符	标准频率	识别频率	误差率(%)
c^1	261.63	261.2985	0.127
d^1	293.66	293.9221	0.089
e^1	329.63	329.7346	0.032
f^1	349.23	350.1726	0.27
g^1	392.00	392.8633	0.22
a^1	440.00	439.2440	0.172
b^1	493.88	494.7910	0.184
c^2	523.26	524.4255	0.223

音符识别的平均误差率为 0.164 6%，根据十二平均律[5]，连续两个半音之间的频率比为：

$$\frac{f_2}{f_1} = 2^{\frac{1}{12}} = 1.059463$$

即频率偏差率为 5.946 3%，远高于识别的平均误差率，所以很容易判断识别音符的标准频率。准确识别出基频以后，根据系统中的 Tones 数组中保存的基频和音名对照表，就

可以得到音符的音名。

表 3 音符时值识别及误差

序号	频率(Hz)	识别音名	检测时值	准确时值	误差(%)
1	261.566	c^1	616.78	555.56	+11.02
2	293.289	d^1	565.99	555.56	-1.88
3	330.756	e^1	431.75	555.56	-22.29
4	349.256	f^1	700.23	555.56	+26
5	391.594	g^1	588.72	555.56	-5.97
6	391.054	g^1	602.27	555.56	+8.41
7	393.449	g^1	275.74	277.78	-8.41
8	350.896	f^1	282.99	277.78	-0.73
9	330.194	e^1	413.6	555.56	-25.6
10	348.833	f^1	685.71	555.56	+23.43
11	348.403	f^1	595.01	555.56	-7.1
12	348.451	f^1	326.53	277.78	+17.55
13	329.352	e^1	312.02	277.78	+12.33
14	291.394	d^1	562.35	555.56	-1.22
15	261.047	c^1	573.25	555.56	-3.18
16	330.384	e^1	631.3	555.56	+13.63
17	392.010	g^1	446.72	555.56	+19.6

图 1 中的乐音为乐曲“洋娃娃和小熊跳舞”开始的 17 个音符的一段乐音，乐器为珠江牌立式钢琴。表 3 是乐音“洋娃娃和小熊跳舞”经过端点检测和音符识别所得到的频率、音名、时值。系统正确识别出了乐音中所有音符的音名。

对于时值项，其实际的含义是决定乐音的延时，即区分是全音符、二分音符、四分音符等。例子乐曲演奏的速度约为：

$\text{♩} = 108$， $\frac{4}{4}$

即四分音符时值为：555.56ms，八分音符时值为 277.78ms。在这个速度下，计算出时值识别误差。

总的来看，由于标准音符时值的时间间隔较大：全音和二分音符的时值相差 1111.13ms；二分音符和四分音符时值相差 555.56ms；四分音符和八分音符时值相差 277.78ms，所以就表 3 的误差来说，只要将误差容许范围设定为（-25%，+25%），就可以正确调整所有的时值。

系统正确识别出了所有的音符，结合调整的时值就可以完成谱曲。

4 结语

单一乐音的时域的平稳周期特性是进行音符识别的关键所在。本文提出的乐音识别方法，频域定位准确，识别的误差非常小，可以作为识别乐音的一个很好的方法。

参考文献：

[1] 李虎生，刘加，刘润生. 高性能汉语数码语音识别算法[J]. 北京：清华大学学报(自然科学版)，2000，40(1)：32-34.

[2] 蔡莲红，黄德智，蔡锐. 现代语音技术基础与应用.[M]. 北京：清华大学出版社，2003.236.

[3] DAUBECHIES I. Ten Lectures on Wavelets[M]. SIAM. Philadephia, 1992.

[4] 朱民雄，闻新，黄健群，等. 计算机语音技术[M]. 北京：北京航空航天大学出版社，2002.362-363.

[5] 缪天瑞. 律学[M]. 北京：人民音乐出版社，2002. 85-88，302-307.

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

[7] 李建平. 小波分析与信号处理——理论、应用及软件实现[M]. 重庆：重庆出版社，1997.

(上接第 967 页)

特点，直接利用各类像素的几何中心到窗口中心的距离作为窗口内各类像素分布的均匀值，并且根据织物图像上纹理呈大块状分布的特征，在进行初始分割之后通过确定种子区域把剩下未分割区域的像素直接指派到种子区域内，节省了计算时间，提升了算法速度。

图 3 展示了对一幅织物图像分割的详细算法过程，图 4 是另外三幅图像的分割结果。结果表明，本算法能够较好地实现各类织物图像的组织区域分割。

3 结语

通过实验发现，本文的算法在对大多织物图像的分割上都取得了比较好的分割效果。同时因为算法充分考虑了织物图像纹理性强的特点简化了计算，提高了分割速度，适合于在纺织业上应用。

参考文献：

[1] 叶齐祥，高文，王伟强，等. 一种融合颜色和空间信息的彩色图像分割算法[J]. 软件学报，2004，15(4).

[2] TOMINAGA S. Color Classification of Natural Color Images[J]. Color Research and Application, 1992, 17(4): 230-239.

[3] LIM YM, LEE SU. On the color image segmentation algorithm based on the thresholding and the fuzzy c-Means techniques[J]. Pattern recognition, 1990, 23(9): 935-952.

[4] CHANG CC, WANG LL. Color texture segmentation for clothing in a computer-aided fashion design system[J]. Image and Vision Computing, 1996, 14(9): 685-702.

[5] HARIS K, EFSTRATIADIS NS, MAGLAVERAS N, *et al.* Hybrid Image Segmentation Using Watersheds and Fast Region Merging[J]. IEEE Transactions On Image Processing, 1998, 7(12): 1684-1699.

[6] BOUKOUVALAS C, KITTLER J, MARIK R, *et al.* Color Grading of Randomly Textured Ceramic Tiles Using Color Histogram[J]. IEEE Transactions Industrial electronics, 2000, l: 219-226.

[7] CHEN KM, CHEN SY. Color texture segmentation using feature distributions[J]. Pattern Recogniation Letters, 2002, 23(7): 755-771.

[8] DENG YN, MANJUNATH BS. Unsupervised Segmentation of Color-Texture Regions in Images and Video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(8): 256-261.

[9] GRIGORESCU SE, PETKOV N, KRUIZINGA P. Comparison of Texture Features Based on Gabor Filters[J]. IEEE Transactions on Image processing, 2002, 11(10): 564-571.

[10] LU S, HERNANDEZ JE. Texture Segmentation by Clustering of Gabor Feature Vectors[A]. IEEE Proceedings of the International Conferece on Artificial Neural Networks I[C]. 1991. 683-687.

[11] DENG Y, MANJUNATH BS, SHIN H. Color image segmentation [A]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '99[C]. 1999. 446-451.