基于听觉图像的音乐流派自动分类

李　锵，李秋颖，关　欣
(天津大学电子信息工程学院，天津 300072)

摘　要：音乐流派的自动分类是音乐信息检索系统的重要组成部分. 将听觉图像引入音乐流派的分类研究中，用听觉图像模型模拟人耳耳蜗结构，基于音乐流派分类研究常用的 GTZAN 数据库，将一维音频信号转换为二维听觉图像，对音乐听觉图像进行尺度不变特征转换(SIFT)及空间金字塔匹配(SPM)，从局部到整体地提取图像的纹理特征，最后采用 LibSVM 中线性核函数的支持向量机对音乐流派进行分类. 实验结果表明，与同样基于人耳耳蜗结构提出的美尔频率倒谱系数(MFCC)流派分类方法相比，基于听觉图像的流派分类正确率提高 15%.

关键词：音乐流派分类；听觉图像；尺度不变特征转换；空间金字塔匹配
中图分类号：TP18　　**文献标志码**：A　　**文章编号**：0493-2137(2013)01-0067-06

Automatic Music Genre Classification Based on Auditory Images

Li Qiang，Li Qiuying，Guan Xin
(School of Electronic Information Engineering，Tianjin University，Tianjin 300072，China)

Abstract：Automatic music genre classification is an important part of the music information retrieval system. The concept of “auditory image” is introduced into music genre classification in this paper. Auditory image model(AIM) converts the one-dimensional audio signal into two-dimensional auditory images by simulating the human ear cochlear structures for the commonly database of GTZAN. And then，the methods of scale invariant feature transformation(SIFT) and space pyramid matching(SPM) are used to extract image features from the part to the whole. And the linear kernel support vector machine is chosen for classification since the dimension of features was high. Experimental results show that the genre classification accuracy based on the auditory images can be 15% higher than the Mel-frequency cepstral coefficients(MFCC) which is also based on the cochlear structure of the human ear.

Keywords：music genre classification；auditory image；scale invariant feature transformation(SIFT)；space pyramid matching(SPM)

面对互联网上海量的音乐数据，对音乐信息的检索显得尤为重要. 目前绝大多数音乐数据库除了可以根据音乐名称或者艺术家姓名建立索引以外，还可以利用音乐的流派信息建立索引. 现有的音乐流派分类方法大多是在音乐数字符号的基础上提取音乐的音色、节奏和音高等内容，这些特征主要包括短时傅里叶变换(short time Fourier transform，STFT)系数、美尔频率倒谱系数(Mel-frequency cepstral coefficients，MFCC)、线性预测系数(linear prediction coefficients，LPC)、过零率(zero-crossings ratio)、最强节拍(strong beat)和最强节拍力度(strength of strong beat)[1-3]等，也有使用网络上对音乐进行标注的标签和音乐专家对音乐的评价作为特征进行音乐流派分类的.

听觉图像模型(auditory image model，AIM)[4]是剑桥大学 Roy Patterson 实验室开发的通过模拟人耳耳蜗的结构特征，将一维声音信号转化为二维听觉图像的时域模型. 根据声音听觉图像的不同，可用来区分声音的元音和辅音[5]，监测深海中障碍物的大小[6]，进行声音排序(sound ranking)[7]，还能对复合音中的音强度进行分析[8]. Ness 等[9]曾在听觉图像的基础上进行了古典作曲家分类和音乐情感分类的研究，

收稿日期：2012-08-27；**修回日期**：2012-09-20.
基金项目：国家自然科学基金资助项目(61101225，60802049)；天津大学自主创新基金资助项目(60302015).
作者简介：李　锵(1978—　　)，男，博士，副教授，liqiang@tju.edu.cn.
通讯作者：关　欣，guanxin@tju.edu.cn.

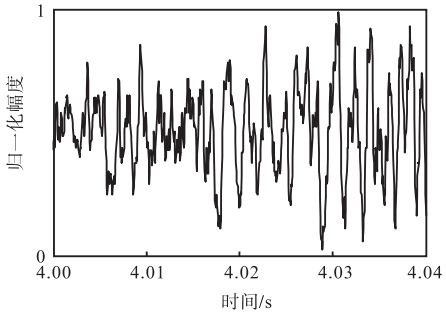
但没有进行音乐流派的分类.

本文首先利用听觉图像模型将音频信号转化为听觉图像,再采用尺度不变特征转换(scale invariant feature transformation, SIFT)[10-11]和空间金字塔匹配(space pyramid matching, SPM)[12]方法提取听觉图像的特征向量进行音乐流派分类,分类结果优于同样基于人耳耳蜗结构提取的美尔频率倒谱系数的分类结果.

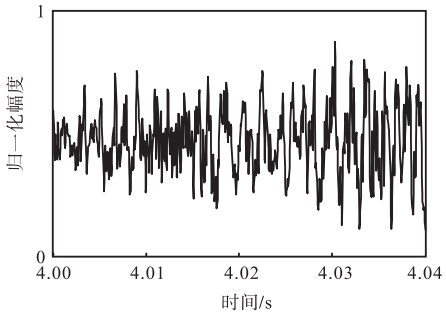
1 听觉图像模型

听觉图像模型,通过模拟人耳的听觉系统,经过耳蜗预处理、基底膜活动、神经活动模式、频点短时整合,最后得到稳定的听觉图像,具体过程如下所述.

耳蜗预处理(pre-cochlear processsing, PCP)过程就是利用带通滤波器来模拟外耳和中耳对音频信号的滤波功能. 以一首古典音乐为例,其原始音频波形如图 1(a)所示,经 PCP 预处理后的波形如图 1(b)所示,滤除超出人耳听觉频带范围的信号,便于后续分析.



(a) 原始音频信号



(b) PCP 波形

图 1 耳蜗预处理波形

Fig.1 Wave figures for PCP

基底膜活动(basilar membrane motion, BMM)部分就是仿照人耳耳蜗对音频信号的处理过程建立的耳蜗功能模块. 根据耳蜗基底膜上不同位置的细胞对不同频率声音的选择过程,将一维音频信号转换成多通道(multi-channel)的不同频带上的波形信号. 动态压缩的 Gammachirp(dynamic compressive Gammachirp, dcGC)滤波器级联结构和如图 2 所示的极零点滤波器级联结构(pole-zero filter cascade, PZFC)[13]都可以模拟人耳不同位置基膜上声音信号的幅度和时延.

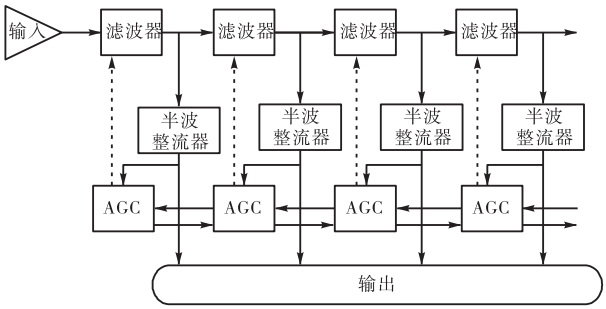
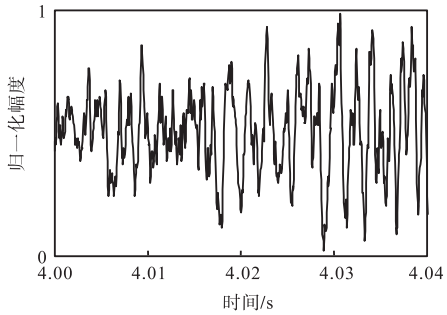


图 2 极零点滤波器级联结构

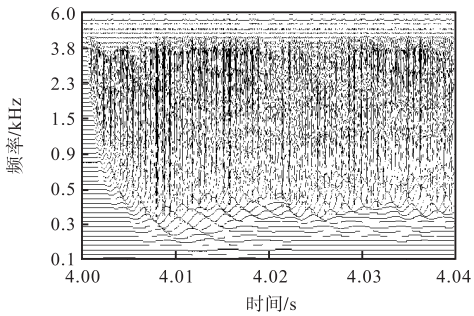
Fig.2 Structure chart of PZFC

图 2 中的自动增益环路对应于脑干中橄榄复合体的传出神经元对耳蜗外纤毛细胞活动的控制. 半波整流器用于保持所有带通信号的能量和精细时间结构. 以一首古典音乐为例,采用极零点滤波器的 BMM 过程如图 3 所示,其中图 3(a)为原始音频信号,图 3(b)为将原始音频转换为等效矩形带宽(equivalent rectangular bandwidth, ERB)刻度下不同频带的波形. 将时域滤波器中心频率 f 转换为 ERB 刻度下的频率关系式为

$$\mathrm{ERB}(f)=24.7(4.37f+1) \tag{1}$$



(a) 原始音频信号

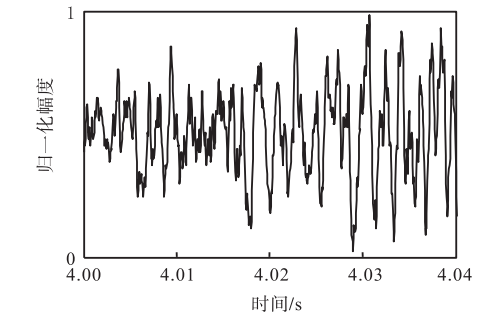


(b) 等效矩形带宽

图 3 将音频信号转化为多通道信号的 BMM 过程

Fig.3 Multi-channel figure for BMM

神经活动模式(neural activity pattern，NAP)模拟耳蜗内耳毛细胞，将 BMM 模块的响应信号进行半波整流、压缩和低通滤波，转换为耳蜗的神经活动. 半波整流用于模拟内耳毛细胞的响应过程. 压缩是为了模拟人耳耳蜗的压缩功能，对输入和输出信号起到平滑的作用. 低通滤波是为了减少随着频率增加和锁相环所造成的损失. 以一首古典音乐为例的 NAP 图如图 4 所示，增强了 BMM 的频谱信息和短时信息.



(a) 原始音频信号

(b) 等效矩形带宽

图 4　由 BMM 转换为 NAP 的过程

Fig.4　Figure for NAP from BMM

频点短时整合(strobe temporal integration，STI)是基于人耳的声音感知原理，利用频点检测技术检测出每条通道的峰值. 以一首古典音乐为例的单通道频点检测过程如图 5 所示.

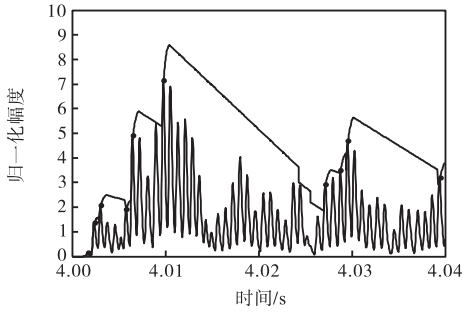


图 5　单一通道的频点检测过程

Fig.5　STI on a single channel

稳定的听觉图(stabilized auditory image，SAI)依据听觉皮层的二维结构及在听觉神经系统中的各种映射方式，将听觉神经上的信号转换为二维滑动图. 具体而言，即将由 STI 得到的波形显著峰值作为触发选通信号，与各通道的信号进行短时互相关运算，完成触发式时域融合，得到最终的听觉图像. 将时域 NAP 信号转换成时间间隔和频率维度上的稳定听觉图信号，以一首古典音乐为例的听觉图如图 6 所示，右侧部分是信号在频率上的分布，下侧部分是在时间间隔上的分布.

图 7 是将听觉图用图像的形式表现出来的听觉图像，可以看出听觉图像的模式及其纹理结构，通过观察不同流派的音乐听觉图像，发现不同流派的听觉图像在图像模式及纹理走向上都是不同的，提取图像的纹理结构特征可以作为音乐流派分类的基础.

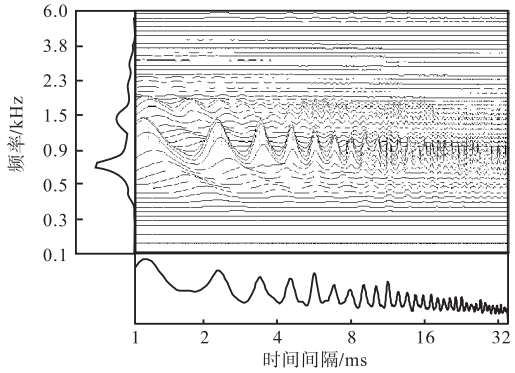


图 6　稳定听觉图

Fig.6　Stabilized auditory image

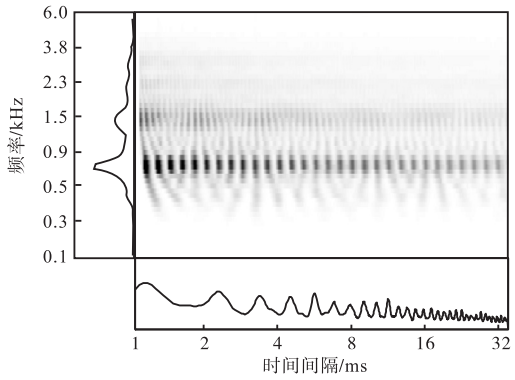


图 7　用图像的形式表示的听觉图

Fig.7　Image form of SAI

2　听觉图像的音乐流派自动分类

2.1　特征提取

本文采用尺度不变特征转换和金字塔匹配方法提取图像特征. 因为尺度不变特征转换可以更全面地提取图像的局部信息，能够更准确地描述图像所包

含的特征. 先将图像划分成相互重叠的图像块,提取各图像块的 SIFT 描述符,然后对提取的 SIFT 描述符进行稀疏编码,用少量的非零值表示 SIFT 描述符,再根据不同的空间金字塔匹配方法,对听觉图像在不同刻度上进行映射,这样就将局部特征整合到整体特征,用整体特征表示听觉图像更全面.

空间金字塔匹配方法主要有 3 种,即均方根值法(the square root of mean squared statistics)Sqrt、绝对值均值法(the mean of absolute values)Abs 和最大绝对值法(max pooling)max,如式(2) ~ (4)所示.

Sqrt:

$$z_j = \sqrt{\frac{1}{M}\sum_{i=1}^{M} u_{ij}} \tag{2}$$

Abs:

$$z_j = \frac{1}{M}\sum_{i=1}^{M}\left|u_{ij}\right| \tag{3}$$

max:

$$z_j = \max\left\{\left|u_{1j}\right|,\left|u_{2j}\right|,\cdots,\left|u_{3j}\right|\right\} \tag{4}$$

式中:u_{ij} 为 SIFT 描述符向量中第 i 行第 j 列的元素;M 为划分的区域内 SIFT 描述符的个数;z_j 为映射后的向量的第 j 个元素.

2.2 分类方法

对于音乐流派的自动分类,本文采用线性支持向量机的分类方法. 台湾大学林智仁副教授等开发设计的 LibSVM 中对于高维度的特征向量,线性支持向量机效果最佳.

3 实验及结果分析

3.1 实验环境及参数选择

本文实验采用国内外音乐流派分类研究中常用的 GTZAN 数据库,其中包括蓝调、古典、乡村、迪斯科、嘻哈、爵士、电子、流行、雷鬼和摇滚 10 种音乐流派的 1 000 首音乐,采样频率为 22.05 kHz. 实验中每首音乐选取时间长度为 5 s 的音乐片段对其进行听觉图像转换,AIM 的设置与文献[8, 14]相同,PZFC 中滤波器的频率范围选取为 40.00 ~ 0.85 f_s/2 Hz(f_s 为采样频率),每秒选取 50 帧. 为了能够更好地描述音乐信息,计算总帧数的图像均值作为每首音乐的听觉图像. 对于动态压缩的 Gammachirp 听觉滤波器频率范围选取 40 ~ 16 000 Hz,通道数选择 50 和 75 两种情况. 通过比较模拟人耳耳蜗的 PZFC 和 dcGC 滤波器级联结构,选择更适合音乐流派分类的滤波器.

不同听觉滤波器在相同图像块大小(图像块大小分别为 16 × 16、32 × 32 和 64 × 64),映射方法选择 max,采用线性支持向量机作为分类器,音乐流派的分类正确率如表 1 中每列所示. 相同的听觉滤波器在不同图像块大小、相同的 max 映射方法和线性支持向量机分类器下,音乐流派分类正确率如表 1 中每行所示. 相同的听觉滤波器在相同的 16 × 16 图像块、相同的线性分类器、不同的映射方法下,音乐流派分类正确率如表 2 中每行所示.

由表 1 和表 2 可见,采用音乐听觉图像对音乐流派进行分类最好的设置是听觉滤波器选择 PZFC,图像块大小选择 16 × 16,增大图像块的大小反而会降低正确率. 3 种空间金字塔匹配方法中,选取最大绝对值法能够达到最好的实验效果.

表 1 采用不同滤波器在不同图像块大小下的分类正确率

Tab.1 Accuracy of different filters and different sizes of images

%

基底膜活动模块	16 × 16	32 × 32	64 × 64
PZFC	62.6 ± 2.3	59.8 ± 3.4	55.6 ± 3.1
dcGC(50)	58.6 ± 5.4	56.4 ± 3.8	53.4 ± 3.9
dcGC(75)	57.6 ± 3.4	55.8 ± 2.6	53.4 ± 2.4

表 2 采用不同滤波器在不同的映射方法下的音乐流派分类正确率

Tab.2 Accuracy of different filters and different methods for mapping

%

基底膜活动模块	均方根值法	绝对值均值法	最大绝对值法
PZFC	59.4 ± 2.7	57.6 ± 5.4	62.6 ± 2.3
dcGC(50)	57.0 ± 3.5	55.4 ± 3.4	58.6 ± 5.4
dcGC(75)	53.6 ± 5.1	52.8 ± 3.8	57.6 ± 3.4

下面讨论在最佳分类效果下各流派的具体分类效果. 采用 PZFC 听觉滤波器,图像块大小选取 16 × 16,匹配方法选择最大绝对值法,分类器选择线性支持向量机的音乐流派分类结果如表 3 所示.

表 3 最佳参数下各音乐流派分类结果

Tab.3 Accuracy with the best parameters

%

音乐流派	蓝调	古典	乡村	迪斯科	嘻哈	爵士	电子	流行	雷鬼	摇滚
蓝调	48	0	12	2	2	2	4	2	6	10
古典	2	92	2	2	2	6	0	0	4	2
乡村	2	2	48	0	0	2	0	6	6	22
迪斯科	6	2	6	56	8	2	0	2	12	12
嘻哈	2	0	2	8	70	0	2	2	0	2
爵士	12	2	6	4	4	82	0	8	2	4
电子	4	0	0	6	2	0	84	6	0	2
流行	12	0	4	6	2	2	4	62	10	6
雷鬼	6	2	4	8	6	4	4	2	52	8
摇滚	6	0	16	8	4	0	2	10	8	32

以表 3 中的古典音乐为例，92%的古典音乐被认为是古典音乐，2%的被误判为乡村音乐，2%的被误判为迪斯科，2%的被误判为爵士，2%的被误判为雷鬼，故古典音乐分类的正确率为 92%. 由表 3 可知，古典、爵士和电子音乐的分类效果比较好，而摇滚音乐的正确率最低，因为古典、爵士和电子音乐的听觉图像的纹理特征和亮度特征较明显，而摇滚音乐的听觉图像的纹理特征不明显，容易误判为其他流派的音乐.

3.2　结果分析

Tzanetakis 等[3]用单一特征集在高斯分类器下的分类结果如表 4 所示，如 5 维的音阶特征的正确率为 23.0%，6 维的节拍特征的正确率为 28.0%，9 维的短时傅里叶变换特征的正确率为 45.0%，10 维基于人耳耳蜗结构得到的 MFCC 特征的分类效果只有 47.0%. 所有特征集的总和才只能得到 59.0%的正确率，而本文基于模拟人耳耳蜗结构得到的听觉图像的流派分类效果可以达到 62.6%，高于 MFCC 的分类效果，也高于任意其他单一特征集的分类效果，甚至比使用总特征集的效果还要好.

表 4　采用不同特征集的分类正确率

Tab.4　Accuracy for different feature sets

%

特征类型	正确率
随机	10.0
5 维的音阶特征	23.0
6 维的节拍特征	28.0
9 维的短时傅里叶变换特征	45.0
10 维的美尔频率倒谱系数特征	47.0
全部特征(30)	59.0
基于听觉图像的特征	62.6

Genussov 等[15]将“模糊映射(diffusion maps)”理论引入音乐流派自动分类系统，在从音乐符号中提取出的音色特征的基础上，选取 3 种不同规格的数据库验证实验效果，古典&电子是选取 GTZAN 中古典和电子两种音乐流派的数据作为两类分类的数据库，5 种流派库是 GTZAN 中蓝调、古典、电子、流行和雷鬼 5 种音乐流派的数据库. 本文也采用同样的数据库，将基于听觉图像的分类效果与采用“模糊映射”前后的音色特征的分类正确率作比较，比较结果如表 5 所示. 由表 5 所示，在 3 种不同的数据库下，采用听觉图像对音乐流派进行分类，优于基于“模糊映射”的音色特征的分类正确率.

Deshpande 等[16]采用的数据库是 52 首爵士、53 首古典和 52 首摇滚音乐组成的数据库，在 MFCC 和 STFT 的频谱图的基础上，对图像提取其纹理信息，采用 K-NN 分类器($k = 3$)时得到最好的实验结果是 75.00%，在同样规格的数据库下基于听觉图像可以得到 77.35%的正确率，高于 MFCC 和 STFT 的频谱图提取特征的分类效果.

表 5　不同规格数据库下的分类正确率

Tab.5　Accuracy in different databases

%

音乐流派	音色特征	模糊映射	听觉图像
古典&电子	87.99 ± 6.83	96.74 ± 3.75	99.00 ± 0.24
5 种流派库	49.89 ± 6.21	84.91 ± 4.88	85.63 ± 6.69
GTZAN	28.27 ± 3.93	56.55 ± 4.50	62.60 ± 2.30

通过以上比较结果可知，相较于 MFCC，听觉图像能够更好地模拟人耳耳蜗的结构，便于音乐流派的分类. 基于听觉图像的音乐流派分类结果优于单一特征集的分类效果.

4　结　语

本文将听觉图像引入音乐流派自动分类系统，用尺度不变特征转换和空间金字塔匹配方法提取图像特征向量，优于同样模拟人耳耳蜗的 MFCC 特征集，也优于任意单一特征集的分类效果. Gjerdigen 和 Perrot 曾做过一个实验，用 1 年的时间训练 52 名心理学专业的大学生去听音乐，培养他们的乐感，对于 250 ms 的音乐片段的音乐，对音乐流派判断结果的正确率为 40.0%左右. 本文从音乐中提取听觉图像的采样点仅为 35 ms 的时间长度，却能达到 62.6%的正确率.

参考文献：

[1] Aucouturier J J，Pachet F. Representing musical genre：A state of the art[J]. *Journal of New Music Research*，2003，32(1)：83-93.

[2] Tzanetakis G，Cook P. Marsyas：A framework for audio analysis[J]. *Organised Sound*，1999，4(3)：169-175.

[3] Tzanetakis G，Cook P. Musical genre classification of audio signals[J]. *IEEE Transactions on Speech and Audio Processing*，2002，10(5)：293-302.

[4] Bleeck S，Ives T，Patterson R D. Aim-mat：The auditory image model in MATLAB[J]. *Acta Acustica United with Acustica*，2004，90(4)：781-787.

[5] Patterson R D. Auditory images: How complex sounds are represented in the auditory system[J]. *Journal of the Acoustical Society of America*, 2000, 21(4): 183-190.

[6] Fox P D, Bleeck S, White P R, et al. Initial results on size discrimination of similar underwater objects using a human hearing model[C]//*Proceedings of the Institute of Acoustics*. St Albans, UK, 2007, 29(6): 233-239.

[7] Rehn M, Lyon R F, Bengio S, et al. Sound ranking using auditory sparse-code representations[C]// *ICML* 2009: *Workshop on Sparse Method for Music Audio*. Montreal, Canada, 2009: 118-120.

[8] Timothy I D, Patterson R D. Pitch strength decreases as F0 and harmonic resolution increase in complex tones composed exclusively of high harmonics[J]. *Journal of the Acoustical Society of America*, 2008, 123(5): 2670-2679.

[9] Ness S R, Walters T, Lyon R F. *Auditory Sparse Coding*[M]. Boca Raton, FL, USA: Music Data Mining, CRC Press, 2011.

[10] Lowe D G. Object recognition from local scale-invariant features[C]// *International Conference on Computer Vision*. Corfu, Greece, 1999: 1150-1157.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110.

[12] Yang Jianchao, Yu Kai, Gong Yihong, et al. Linear spatial pyramid matching using sparse coding for image classification[C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, USA, 2009: 1794-1801.

[13] Lyon R F. Machine hearing: An emerging field[J]. *IEEE Signal Processing Magazine*, 2010, 27(5): 131-139.

[14] Lyon R F, Rehn M, Bengio S, et al. Sound retrieval and ranking using sparse auditory representations[J]. *Neural Computation*, 2010, 9(22): 2390-2416.

[15] Genussov M, Cohen L. Musical genre classification of audio signals using geometric methods[C]//18^{th} *European Signal Processing Conference*(*EUSIPCO*-2010). Aalborg, Denmark, 2010: 497-501.

[16] Deshpande H, Nam U, Singh R. Classification of music signals in the visual domain[C]// *Proceedings of the COST G-6 Conference on Digital Audio Effects* (*DAFX*-01). Linerick, Ireland, 2001: DAFX-1-DAFX-4.