计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):158−164 [doi: 10.15888/j.cnki.csa.005994]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 DBN 的多特征融合音乐情感分类方法①

龚　安, 丁明波, 窦　菲

(中国石油大学(华东) 计算机与通信工程学院, 青岛 266580)

摘　要: 本文在音乐情感分类中的两个重要的环节: 特征选择和分类器上进行了探索. 在特征选择方面基于传统算法中单一特征无法全面表达音乐情感的问题, 本文提出了多特征融合的方法, 具体操作方式是用音色特征与韵律特征相结合作为音乐情感的符号表达; 在分类器选择中, 本文采用了在音频检索领域表现较好的深度置信网络进行音乐情感训练和分类. 实验结果表明, 该算法对音乐情感分类的表现较好, 高于单一特征的分类方法和 SVM 分类的方法.

关键词: 音乐情感分类; 深度学习; 深度置信网络; 音乐特征提取; 特征融合

引用格式: 龚安,丁明波,窦菲.基于 DBN 的多特征融合音乐情感分类方法.计算机系统应用,2017,26(9):158–164. http://www.c-s-a.org.cn/1003-3254/5994.html

Music Mood Classification Method Based on Deep Belief Network and Multi-Feature Fusion

GONG An, DING Ming-Bo, DOU Fei

(School of Computer &Communication, China University of Petroleum, Qingdao 266580, China)

Abstract: In the paper we explore the two important parts of music emotion classification: feature selection and classifier. In terms of feature selection, single feature cannot fully present music emotions in the traditional algorithm, which, however, can be solved by the multi-feature fusion put forward in this paper. Specifically, the sound characteristics and prosodic features are combined as a symbol to express music emotion. In the classifier selection, the deep belief networks are adopted to train and classify music emotions, which had a better performance in the area of audio retrieval. The results show that the algorithm performs better than the single feature classification and SVM classification in music emotion classification.

Key words: music mood classification; deep learning; deep belief network; music feature extraction; feature fusion

随着音乐推荐系统等数字音乐应用的蓬勃发展, 针对数字音乐分类问题的研究越来越受到重视, 而音乐情感分类作为音乐检索、音乐推荐等领域的基础问题, 逐渐成为该领域的重点研究对象. Krumhansl, C. L.于 1997 年第一次用数学模型描述音乐情感分类的问题[1], 从此开启了音乐情感自动分类的大门, 在该领域研究的主要方面有两个: 1)恰当音乐特征的选择; 2)选择合适的分类模型[2].

针对音乐特征选择问题, 许多学者对采用怎样的音乐特征才能够更好的表达音乐情感进行了探究, 比如 G. Tzanetakis 等人在音乐类型分类中用到的梅尔频率倒谱系数(MFCC)[3]、T. Li 等人研究中用到的频谱质心(SC)[4], T. Li and 等人的多贝西小波系数直方图(DWCH)[5], 文献[6-9]中用到的节拍直方图(BH), T. Fujishima 等人用到的和弦序列(CS)、音节轮廓(PCP)[10]等, 这些不同的特征在一定程度上体现了音乐所包含的内在情感, 但单一的音乐特征仍然存在情感表达不完整的问题, 在实际应用中导致分类正确率低. 在分类器选择中, 浅层分类器如 k-NN、SVM、贝叶斯分类器是音乐情感分类的常用分类器[11], 还有人工神

① 收稿时间: 2016-12-28; 采用时间: 2017-02-17

经网络(ANN)、自组织映射(SOM)、隐马尔科夫模型(HMM)和逻辑回归(LR)等也在该领域得到应用[12], 浅层分类器的分类效果很难满足人们的需要.

针对以上问题, 本文提出了一种特征融合的音乐情感分类算法: 从原始音频信号中先提取音色特征、韵律特征, 然后以这些特征作为原始的训练集, 在深度置信网络(DBN)训练模型中训练. 通过对音色特征、韵律特征的融合, 解决传统音乐情感分类中单个音频特征单一导致的音乐情感表达不准确问题, 同时利用深度置信网络在音频处理中的优势, 采用 4 层深度置信网络进行分类训练, 提高分类准确率.

1 音乐情感特征参数

大多数的音乐情感分类方法的研究是基于音乐信号本身提取的音乐特征[13], 音乐情感特征参数的提取与表达是音乐情感分类技术的基础内容, 我们可将音乐内容分为底层特征、中层特征和高层标签三个部分, 底层特征和中层特征是客观存在的符号特征, 而高层标签则是依赖于人的感官而特定标注的信息, 包括音乐情感、音乐流派等内容, 音乐的符号特征与高层标签之间存在的无法直接对应的问题称为语义间隙问题. 而音乐情感特征参数提取的主要目标就是提取合适的符号特征完成音乐情感的表达.

图 1 语义间隙

本文特征融合主要包含底层音色特征中的梅尔频率倒谱系数(MFCC)和中层韵律特征中的基音频率、共振峰、频带能量分布.

1.1 底层音色特征

底层特征是音乐情感分类系统的重要组成部分, 主要包括音色特征和时序特征, 其中音色特征作为音乐内容的基本元素, 既易于提取又具有良好的性能, 反映了音乐的声音品质, 具有相同音调的声音能够通过音色特征区分, 就像不同的绘画有不同的颜色一样, 不同的声源形成不同的谐波序列, 并产生不同的声色特征. 在实际应用中 SR、SC、MFCC 等底层特征广泛应用于音乐情感分类系统中.

梅尔频率倒谱系数(MFCC)是 Mermel 等人基于人的听觉和语言提出的, 在现阶段的音乐信息分类中, 由于 MFCC 相对于其他特征具有强抗噪性、高识别率的特点, 目前已成为音乐情感识别领域应用最为广泛的特征参数. MFCC 的具体计算步骤如下.

第一步: 对输入音频信号进行分帧、加窗处理, 然后进行傅里叶变换转变为频域信号, 得到频谱如下:

$$X(k)=\sum_{k=0}^{N-1}x(n)e^{-j\frac{2\pi nk}{N}},0\le k\le N-1 \quad (1)$$

x 表示输入信号, $x(n)$为输入信号在 n 处的信号强度, 若信号长度为 J, 则 n=0, 1, 2,..., J, ; N 表示离散傅里叶变换的点数.

第二步: 计算能量谱并将能量谱通过一组 Mel 尺度的三角滤波器组. 定义一个有 M(M=100)个滤波器的滤波器组, 采用的滤波器为三角滤波器, 中心频率为 $f(m)$, m=1, 2, 3,...M.

计算每个滤波器组输出的能量对数, 得到:

$$S(m)=\ln\left(\sum_{k=1}^{N-1}|X(k)|^2H_m(k)\right),0\le m\le M-1 \quad (2)$$

其中$H_m(k)$是三角滤波器的频率响应.

第三步: 经离散余弦变换(DCT)得到 MFCC 系数:

$$C(n)=\sum_{m=0}^{N-1}S(m)\cos(n\pi(m-0.5/m)),n=1,2,...,L \quad (3)$$

其中 n 为 MFCC 系数阶数, 将公式(2)所求的对数能量带入公式(3)计算离散余弦变换, 最终求得 L 阶梅尔倒谱特征参数. L 指的是 MFCC 系数阶数.

1.2 中层音色特征

韵律特征是音乐特征中另一个广泛应用于音乐分类的重要特征[14], 他是一种时间相关性的音频信息, 主要包括基音频率、振幅、发音持续时间、节奏等. 韵

律特征是情感分类的重要参考标准, 通常来说, 悲伤的音乐不会有舞曲一样的韵律, 而快节奏的音乐多带有积极欢快的情绪.

1.2.1 基音频率

音乐声音是由发音体发出的频率、振幅都不相同的振动组合而成的, 其中振动最低的频率音称为基音频率. 它决定了整个音乐片段的音高, 是音乐情感表达的一个重要特征. 音乐是一种有调音频, 基音的变化模式称为音调, 而不同的音乐情感有着不同的音调, 在具体的音乐情感识别中, 比如说同一句歌词由相同的人唱出, 它的基音频率会伴随着演唱环境、情绪状态、身体状况等音乐的不同而改变, 所以基音频率携带着非常重要的具有辨别音乐情感的的信息.

本系统中我们先对音乐片段的每一帧求取基音频率, 然后提取出基音频率的最大值、最小值、平均值、标准差、变化范围、上四分位数、中位数、下四分位数共计 8 维参数.

1.2.2 共振峰

共振峰频率简称共振峰, 是因共鸣作用而能量变强的频率成分, 它直接反映了音乐中声音的来源, 是反映声道物理特征的重要参数. 音频信号模型中把声道认为是一根具有分均匀的声管, 发声的过程就是不同位置的声管共鸣的过程共振峰与声道的形状和大小有关, 一种形状对应着一套共振峰, 音频的频率特性主要是由共振峰决定的, 当声音沿声管传播时, 其频谱形状就会随声管而改变. 而在音乐情感识别中, 如果音乐中包含喜、怒、哀、乐等不同情感信息, 那么音频的声道形状就会发生不同的变化, 所以共振峰是能够体现音乐情感变化的重要因素.

本文中利用峰值检测的方法计算出了共振峰的带宽和频率, 进而计算出第一共振峰的平均值、标准差、中位数、中位数所占的带宽, 第二共振峰的平均值、标准差、中位数、中位数所占的带宽, 第三共振峰的平均值, 、标准差、中位数、中位数所占的带宽, 共计 12 维参数.

1.2.3 频带能量分布

频带能量分布是指音频中某一频率范围内所蕴含能量的分布情况, 反映了音频信号的强度随频率变化的关系. 音频信号在频域上可以划分成若干个子带, 不同的音频信号在每个子带上的能量分布有所不同, 而对音乐情感起作用的信号主要分布在低频区, 是音乐信号的重要特征, 与音乐情感之间有很强的关联性. 例如悲伤的音乐一般较为舒缓而能量较小, 而慷慨激昂的音乐大多具有很强的能量, 在具体的研究中针对这一特性能够为音乐情感分类提高良好的借鉴.

本文中选取了 0-500 Hz 的频带能量平均值、500-1000 Hz 的频带能量平均值、1000-2000 Hz 的频带能量平均值、2000-3000 Hz 的频带能量平均值、3000-4000 Hz 的频带能量平均值、4000-5000 Hz 的频带能量平均值, 共计 6 维参数.

本文将以上 4 种特征共 40 维参数标签化处理后作为音乐情感特征, 在深度置信网络中训练, 完成音乐情感分类.

2 深度置信网络

深度置信网络(DBN)是由多个受限玻尔兹曼机(RBM)组成的概率生成模型[15], 在深度学习中具有重要的地位, 在近几年来已经成功的应用到自然语言理解[16]、语音识别[17]等领域.

2.1 受限玻尔兹曼机

受限玻尔兹曼机(RBM)是一个两层神经网络模型, 包含显层和隐层两层结构, 如图 2 所示, 底部的 v 层代表可见层, 顶部的 h 层代表隐层, 可见层与隐层采用全连接方式连接, 层内单元无连接, w 表示层间的连接权重.

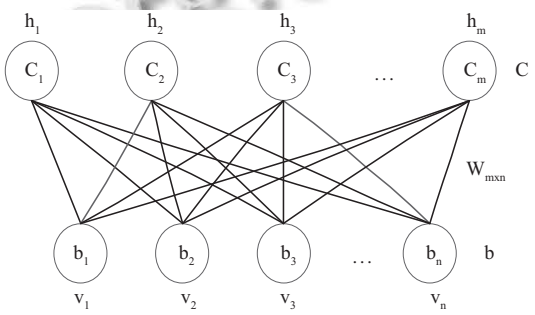

图 2 RBM 模型

用户提交的检索信息通过权重 w 与可见层 v 和隐层 h 相联系, 如果将显示单元分为两类(0 或 1), 那么当向量 $v(v_1,v_2,v_3,...)$ 作为可见层的输入时, 隐层单元 h_1 输出为 1 的概率为:

$$p(h_i=1|v)=\varphi\left(b_i+\sum\nolimits_i v_i w_{ij}\right) \tag{4}$$

这里 $\varphi(x)=1/(1+e^{-x})$, b 为隐层单元 j 的偏置.

当向量$h(h_1,h_2,h_3,...)$作为隐层输入时, 可见层单元v_i输出为 1 遵循概率:

$$p(v_i=1|h)=\varphi\left(a_i+\sum_i h_i w_{ij}\right) \quad (5)$$

这里 a_i 为可见层 i 的偏置.

2.2 深度置信网络模型

DBN 是一个通过多个 RBM 模型叠加在一起而生成的多层的神经网络模型. 由图 3 所示, 除第一个的输入为原始输入外, 其他层的输入均来源于上一层的输出, 由此可见 DBN 的叠加训练过程如下: 训练一个 RBM 之后, 将隐层单元的激活概率作为第二层 RBM 的输入值, 第二层 RBM 的激活概率作为第三层 RBM 的输入值, 之后以此类推, 直到整个神经网络训练完毕. 这个训练过程称为生成式预训练.

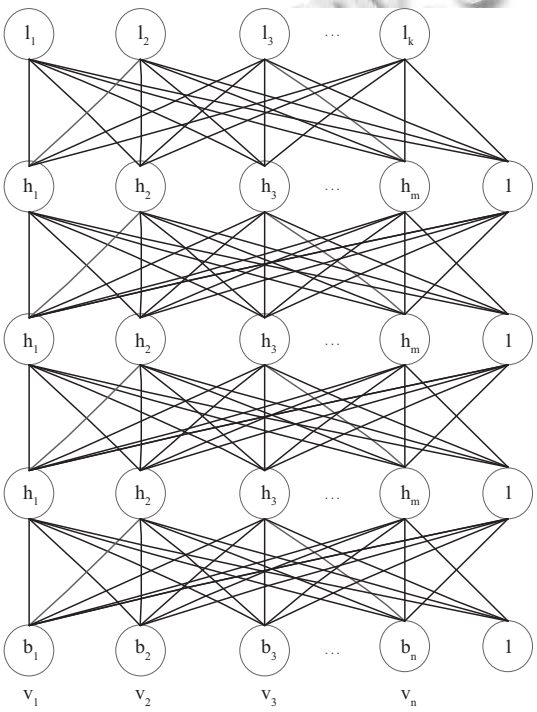

图 3 DBM 模型

在本文应用 DBN 做分类任务时, 采用生成式预训练和判别式方法结合使用的方式, 它通过有效地调整所有权值来改善网络的性能. 判别式精调的方式是在现有网络的最后一层上在增加一层节点, 如图 3 所示, 用来表示想要的输出或者训练数据提供的标签, 它与标准的前馈神经网络一样, 可以使用反向传播算法来调整或精调网络的权值. DBN 最后一层即标签层的内容, 根据不同的任务和应用来确定.

3 基于 DBN 的多特征融合分类方法

3.1 算法流程

本文实现音乐情感分类的流程如图 4 所示.

具体的实现步骤如下:

第一步: 音频预处理以及对相关特征的提取. 本文对所涉及的音乐音频分割为 30s 一段的音频片段, 并对分割完毕的音频进行 4 个特征共 40 个维度特征的提取.

第二步: 音乐特征融合. 对 40 个维度的特征进行整合、打包处理, 形成 40 维特征向量.

第三步: 训练. 将数据集分为训练集与测试集, 采用 4 层 DBN 深层网络训练.

第四步: 测试微调. 根据实验结果对权重、偏置等参数进行调整, 并评估实验结果.

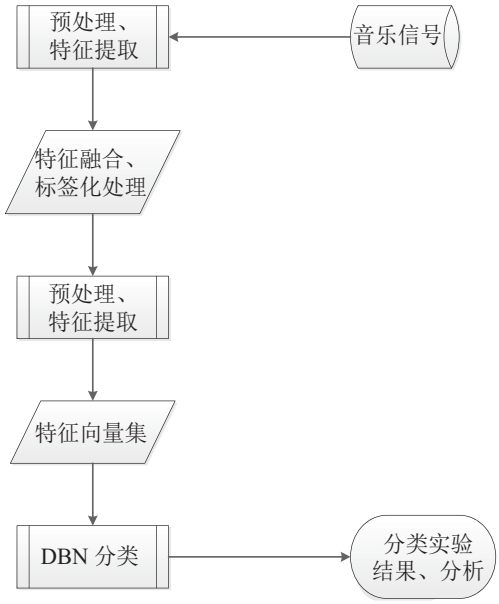

图 4 音乐情感分类流程图

3.2 多特征融合

在音乐分类中, 单特征方法能够较好的解决音乐类型、乐器等直观的分类类型, 但是对于复杂的音乐情感分类, 单特征很容易造成对对某些情感识别较好而对另外一些识别不好的情况, 针对这个问题, 本文使用了将音色特征中的 MFCC 与韵律特征中的基音频率、共振峰、频带能量分布四个在音乐情感分类中表现较好的特征相结合方法, 作为音乐情感表达特征.

如表 1 所示, 4 个种类的特征包括 14 维的 MFCC 特征, 8 维的基音频率特征, 12 维的共振峰特征, 6 维的

频带能量分布特征.

表 1 音乐特征参数

特征名称	MFCC	基音频率	共振峰	频带能量分布
特征维数	14	8	12	6

3.3 DBN 分类器的构建

本文的分类器使用了在音频处理中表现较好的 DBN, 并采用了预训练和判别式的方法, DBN 是由四个 RBM 分类器组成的深层结构, 每层的节点数为 40, 100, 100, 200, 10, 其中首层的 40 个节点对应 40 维的特征向量, 预训练的循环次数为 100 次, 微调的循环次数为 200 次, 最后一层将音乐分为 10 类情感.

图 5 为训练过程中分类误差与迭代次数的关系图, 从图中可以直观的看出随着迭代次数的不断增加, 误差呈现逐渐下降的趋势, 并且迭代次数在 0-100 次间误差下降幅度尤为明显, 在 150-200 次误差逐渐趋于平稳, 这个过程反映出深层的 DBN 相比浅层结构具有的更强的特征提取能力.

4 实验结果与分析

本文实验是在 Windows7 操作系统下, 使用了 NVIDIA TITAN X 显卡完成的, 使用了 Python 语言写的 Theano 库实现整个实验过程.

4.1 音乐库

为了评估音乐情感分类的方法, 本文使用了 MIREX 音乐情感库中的 903 首音乐, 外加根据百度音乐中的情感标签下载的 1397 首音乐, 共组成了 2200 首音乐, 并将 2000 首音乐作为训练集, 200 首作为测试集. 共使用了包括伤感、激情、安静、甜蜜、励志、寂寞、想念、浪漫、喜悦、轻松 10 类音乐情感, 其中安静与轻松、寂寞与想念为相似情感标签, 用以验证实际分类中本文方法对相似情感的分类情况.

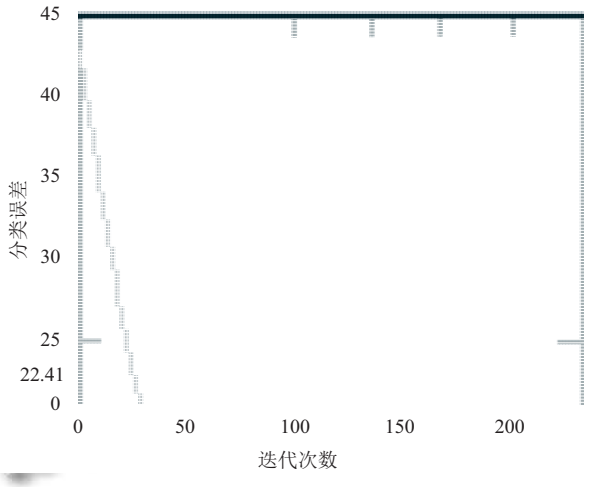

图 5 分类误差与迭代次数关系图

4.2 实验结果

本文将 2000 首共 10 类的音乐特征训练集随机分成总数相同的 10 组, 对 10 组分别进行训练, 最终得到 10 组的正确率如表 2 所示, 由 10 组的正确率可以得出平均分类正确率为 77.59%.

表 2 分组分类正确率

组别	1	2	3	4	5	6	7	8	9	10
正确率(%)	76.66	81.78	80.89	73.75	79.97	78.31	74.56	76.23	78.21	75.56

同时根据实验统计得到 10 类情感各自的分类正确率以及分类混淆情况, 如表 3 所示, 表中元素 a_{ij} 表示情感标签为 i 的音乐经 DBN 分类判断为 j 的概率, 由表中分析可得分类效果较好的是激情和喜悦两类, 正确率达到了 87.51% 以及 86.56%, 分类效果较差的是想念和寂寞, 正确率分别为 69.70% 和 70.53%, 可以直观的看出这两类相似情感在分类过程中产生了部分混淆.

表 3 分类详情

	伤感	激情	安静	甜蜜	励志	寂寞	想念	浪漫	喜悦	轻松
伤感	74.37	0	11.22	5.17	0	0	0	0	5.32	3.92
激情	1.07	87.51	0	0.65	3.35	3.05	0	2.01	2.36	0
安静	7.38	0.33	81.18	3.05	2.56	1.65	2.54	0.28	0	1.03
甜蜜	3.57	1.09	5.71	74.33	0	0.49	9.08	4.44	1.29	0
励志	1.38	5.22	0	3.67	78.97	0	1.53	3.16	6.07	0
寂寞	0	2.95	4.09	2.13	0	70.53	14.83	0.57	0	5.47
想念	7.70	0	0	1.07	0	13.21	69.70	0	1.56	6.76
浪漫	0.67	0	3.07	8.40	0	2.77	9.88	71.33	0	3.88
喜悦	0	0	0.72	0.37	7.09	0	0	4.21	86.56	1.05
轻松	2.30	0.63	0	0.51	3.11	3.01	5.31	3.64	0	81.49

4.3 方法对比

在相同的实验数据及实验环境下, 我们分别设计了单一特征音色特征(MFCC)与 SVM 结合, 组合韵律特征(基音频率、共振峰与频带能量分布)与 SVM 结合, 组合情感特征与 SVM 结合, 单一特征音色特征(MFCC)与 DBM 结合, 组合韵律特征(基音频率、共振峰与频带能量分布)与 DBN 结合五组对比实验, 实验结果如表 4 所示.

表 4 正确率对比

方法	正确率(%)
MFCC+SVM	41.33
韵律特征+SVM	38.1
情感特征+SVM	48.38
MFCC+DBN	69.67
韵律特征+DBN	63.51
情感特征+DBN	77.59

由表中数据可以看出, 单一特征无论是 MFCC 还是韵律特征与 SVM 结合的方法分类的正确率最低, 仅为 41.33% 和 38.1%, 而多特征融合结合 SVM 后分类正确率得到了提高, 达到 48.38%, 但限于 SVM 的简单 2 层结构, 正确率依然无法满足要求, 将分类器修改为 DBN 后, 得益于 DBN 强大的特征提取能力, 单一特征与之结合的分类正确率达到了 69.67% 和 63.51%, 而当单一特征替换为本文设计的情感特征后, 正确率得到了很大的提升, 达到 77.59%.

4.4 方法验证

为了能够验证本文方法在音乐情感分类中的普遍适用性, 我们在结束以上工作的同时, 将本文的方法运用到 allmusic 和 5sing 两个不同音乐库中进行训练并调优, 实验结果显示本文方法具有良好的适应性.

其中 allmusic 是一个依据情感分类进行检索的元数据音乐数据库, 包括了 288 类不同的情感标签, 我们使用了 300 个包括生气、高兴、伤心、放松 4 个标签的音乐片段, 通过训练得到 79.83% 的分类正确率; 5sing 是国内在线音乐平台, 支持用户自己上传原创、翻唱音乐并为音乐贴标签, 所以该音乐库中存在相同音乐因个人演唱情感不同而产生不同音乐标签的情况, 对验证本文方法有效性有一定参考价值, 我们同样选取其中 300 首 4 个情感标签的音乐进行训练分类, 最终取得 75.11% 分类正确率.

5 结束语

本文提出了一种基于 DBN 的多特征融合音乐情感方法, 将融合情感特征作为输入, 依托于 DBN 强大的特征提取能力, 针对输入的符号特征做二次特征提取并进行分类. 主要克服单一特征无法解决与音乐情感之间的语义间隙问题和音乐情感分类的分类器选择问题, 通过与单特征以及浅层 SVM 分类方法进行试验比较, 结果显示本文方法具有较高的分类正确率. 在后期音乐情感分类研究的过程中, 需进一步探索适合音乐特征表达的符号特征, 探究深度学习网络在音乐情感分类中的应用; 同时进一步考虑相近情感音乐之间的关联性, 采用合适的标注方法, 为情感相近的音乐提供个性化标签.

参考文献

1 Krumhansl CL. An exploratory study of musical emotions and psychophysiology. Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 1997, 51(4): 336–353. [doi: 10.1037/1196-1961.51.4.336]

2 Rajanna AR, Aryafar K, Shokoufandeh A, *et al*. Deep neural networks: A case study for music genre classification. Proc. of the 14th International Conference on Machine Learning and Applications (ICMLA). Miami, FL, USA. 2015. 655–660.

3 Tzanetakis G, Cook P. Musical genre classification of audio signals. IEEE Trans. on Speech and Audio Processing, 2002, 10(5): 293–302. [doi: 10.1109/TSA.2002.800560]

4 Li T, Ogihara M, Li Q. A comparative study on content-based music genre classification. Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. Toronto, Canada. 2003. 282–289.

5 Li T, Ogihara M. Toward intelligent music information retrieval. IEEE Trans. on Multimedia, 2006, 8(3): 564–574. [doi: 10.1109/TMM.2006.870730]

6 Feng YZ, Zhuang YT, Pan YH. Music information retrieval by detecting mood via computational media aesthetics. Proc. of IEEE/WIC International Conference on Web Intelligence. Halifax, NS, Canada. 2003. 235–241.

7 Yang D, Lee W. Disambiguating music emotion using software agents. Proc. of the 5th International Conference on Music Information Retrieval. Barcelona, Spain. 2004. 10–14.

8 Yang YH, Lin YC, Su YF, *et al*. A regression approach to

music emotion recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2008, 16(2): 448–457. [doi: 10.1109/TASL.2007.911513]

9 Chua BY, Lu GJ. Perceptual rhythm determination of music signal for emotion-based classification. Proc. of the 12th International Multi-Media Modelling Conference Proceedings. Beijing, China. 2006, 8.

10 Fujishima T. Realtime chord recognition of musical sound: A system using common lisp music. Proc. International Computer Music Association. Stanford, CA, USA. 1999. 464–467.

11 Gómez E, Herrera P. Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. Proc. of the 5th International Conference on Music Information Retrieval. Barcelona, Spain. 2004.

12 Basili R, Serafini A, Stellato A. Classification of musical genre: A machine learning approach. Proc. of the 5th International Conference on Music Information Retrieval. Barcelona, Spain. 2004.

13 Cuthbert MS, Ariza C. music21: A toolkit for computer-aided musicology and symbolic music data. Proc. of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010). Netherlands. 2010. 637–642.

14 Gouyon F, Dixon S. A review of automatic rhythm description systems. Computer Music Journal, 2005, 29(1): 34–54. [doi: 10.1162/comj.2005.29.1.34]

15 Yu D, Deng L. Deep learning and its applications to signal and information processing [Exploratory DSP]. IEEE Signal Processing Magazine, 2011, 28(1): 145–154. [doi: 10.1109/MSP.2010.939038]

16 Sarikaya R, Hinton GE, Deoras A. Application of deep belief networks for natural language understanding. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2014, 22(4): 778–784. [doi: 10.1109/TASLP.2014.2303296]

17 Yoshioka T, Gales MJF. Environmentally robust ASR front-end for deep neural network acoustic models. Computer Speech & Language, 2015, 31(1): 65–86.