第 35 卷第 2 期　　声　学　学　报　　Vol. 35, No. 2
2010 年 3 月　　ACTA　ACUSTICA　　Mar., 2010

献给马大猷教授 95 华诞

基于音色单元分布的音乐结构分析*

李相莲[2]　李明[1]　刘若伦[2]　颜永红[1]
(1　中国科学院声学研究所 中科信利语音实验室　北京　100190)
(2　山东大学威海分校　威海　264209)

2010 年 2 月 2 日收到
2010 年 2 月 11 日定稿

摘要　音乐的结构是音乐作品表达作者思想的一种重要形式，也是听众理解音乐作品内涵的有效途径。本文研究了基于音乐特征的音色单元建模方法，研究了在 Fisher 准则下，根据局部范围音色单元的分布，采用非监督聚类方法分析音乐的结构。实验结果证明了基于离散余弦变换的音色特征，用音色单元分布聚类算法进行音乐结构分析的有效性。

PACS 数：43.71, 43.75

Music structure analysis based on timbre unit distribution

LI Xianglian[2]　LI Ming[1]　LIU Ruolun[2]　YAN Yonghong[1]
(1　*ThinkIT Speech Lab, Chinese Academy of Sciences*　Beijing　100190)
(2　*Shandong University at Weihai*　Weihai　264209)

Received Feb. 2, 2010
Revised Feb. 11, 2010

Abstract　Music structure is not only an important form of the music works to express artists' ideas, but also is an effective way for listeners to understand the meaning of the music. This paper proposes a timbre unit modeling method based on musical features, using unsupervised clustering method to analyze music structure according to the distribution of local timbre units, with the amended Fisher rule. The experiment results show that the unsupervised clustering algorithm with DCT-based chroma feature is an effective way to analyze the music structure.

引言

音乐结构是音乐的重要表达方式，也是最重要的音乐语义之一。在音乐作品中，结构包括称之为织体的 "空间" 上的结构与称之为曲式的 "时间" 上的结构。目前对音乐的结构分析主要是分析它的曲式结构。音乐一般表现出较强的自相似性，因而有一些重复的类型以及显著的重复性结构 [1] 。歌曲结构一般由序曲、主歌、副歌、过渡段和尾曲组成，所以对音乐结构进行分析，一般是将一首乐曲进行分割，标注出各个段落。

音乐结构自动分析是音乐理解研究的一个重要领域，这在整个音乐分析领域属于基础研究，它在很多领域都有着重要的应用。它有助于计算机辅助音乐理解与分析，使音乐结构分析与乐理规则相结合，实现自动辨别曲式、体裁、曲风、音乐类型等 [1]；将分割后得到的各个段落作为确定歌曲的部分声纹，还可以用来区分不同的歌曲及同一首歌曲中不同的段落来进行歌曲识别 [2]；副歌可以作为一首歌曲的摘要，在音乐浏览器或者音乐检索系统中，它使得用户能够快速地预览一首歌曲的副歌部分，将之作为 "音频索引" 来找到期望得到的歌曲。通过仅仅用副歌部分进行匹配查询的歌曲，它还能提高音乐检索系统的效率和精度 [3]，同时还能减少查询歌曲与乐曲库进行匹配的空间 [2]；此外，音乐结构分析还有助于音乐标注、音乐流式传输 [4] 、在线音乐目录浏览、文档浏览、 MPEG-7 标准提出的用元数据来存储多媒体摘要信息 [5] 、结构感知音乐播放器 [7] 、主动音乐倾听、音乐目录的声音浏览、音乐创作、媒体压缩等 [6] 。

* 国家重点基础研究发展规划项目计划 (973 计划， 2004CB318106) 、国家自然科学基金 (10574140, 60535030) 、国家高技术研究发展计划 (863 计划， 2006AA010102 ， 2006AA01Z195) 和中日双边联合项目资助项目。

近几年，由于音乐结构分析的广泛应用及其对其他领域发展的促进作用，它已经成为一个重要的研究领域，研究的算法也层出不穷，在本文第一部分将讲到音乐结构分析技术的发展现状，第二部分给出本文提出的基于音色结构单元的自动聚类算法，第三部分将给出实验结果，第四部分得出结论。

1 音乐结构分析技术发展现状

根据音乐结构的表示形式，目前音乐结构分析方法主要有“状态”与“序列”表示两种方法[5−6,8]。“状态”定义为包含相同声音信息的邻近时间的集合，它与流行音乐中“部分”(如序曲、主歌、副歌和过渡段) 一词密切相关。大多数的研究者都用状态来表示音乐结构，将音频信号看成是一系列不同尺度的状态，对应于一首音乐的不同尺度的结构[5,7,9]。状态表示算法主要有：切割[10]、凝聚聚类或谱聚类算法[11−13]或隐马尔科夫模型[11,15]等。“序列”定义为一个连续时间的集合，与另一个连续时间的集合类似。它与流行音乐中的旋律或者和弦一词紧密相关。用序列的方法将得到一个比状态表示方法更高精度的表示。“序列”表示的算法主要有动态时间规整 (DTW) 或模式匹配算法[2,14−16], DTW 与分层的方法相结合，基于最大似然算法的方法等[6]。

根据选取的特征量的不同可以将音乐结构分析方法分成三类，分别为基于谱形状的分割，基于泛音切割与基于节奏和音高的分割。基于谱形状的分割用宽谱特征来评价结构相似性，在此类方法研究者几乎都采用 Mel-Frequency Cepstral Coefficient (MFCC) 特征量来描述谱的特征[17−19,10,6]，它实质上反映了音色的特征。MIT Media Laboratory Wei Chai 等人中用谱表示以及音高表示作为特征量[9]，结果表明该方法具有一定的前景，但它仅限于对某些歌曲结构的分析。基于泛音的切割侧重于泛音而非音色结构，主要采用 Chroma 描述 12 个同性质的音高类间的功率谱能量的分布[3,20]。2003 年，Goto 给出的 RefraiD 系统[3]，可检测副歌及转调后的副歌，使副歌检测得到很大改进。基于节奏和音高切割，一些研究者采用基于符号[21]的方法或者通过 N-gram[22]的统计分析的方法进行结构分析。微软亚洲研究院 Lu Lie 等人描述了一种基于 HMM 的方法来进行切割，他们除了使用 MFCC 特征之外，还使用了 Octave Constant-Q 带通滤波器来进行音高的选择[23]，实验结果表明，结合了 Constant Q Transform(CQT) 变换的 MFCC 要比单独使用 MFCC 的切割性能有所改善。

虽然以上的方法各有千秋，但最后得到的分割结果都存在了一些碎片，也即存在过分割的现象。为了解决这个问题，2006 年伦敦大学 Samer Abdallah 等人提出了一种 Duration 模型来减少音频分割中出现的碎片[3]与专家标注的结果更接近。本文提出的基于音色单元分布算法在减少音频分割出现的碎片问题上有了很大的成效，该算法将第二部分中详细讲述。

2 音色单元分布聚类算法描述

本文提出的算法的结构框图如图 1 所示。

将歌曲提取出特征量之后，可以用相似性矩阵的方法，也可以用聚类的方法来描述特征向量之间的相似性，分别对应于“序列”与“状态”的结构分析方法，进而来得到音乐结构的切分。由于相似性矩阵更多使用了歌曲局部的特征量信息，所以在最后结构切分结果中难免产生更多碎片片段。本文根据音色单元分布的特性，在整首歌曲范围内分析具有相似音色单元分布的片段，从而大幅减少了碎片的产生。

本文假定每首歌曲有一些基本的音色单元，这些音色单元可以通过非监督聚类的方法获得。为了能大致确定一些可能的结构边界，本文引入候选边界分析算法。本文认为相同的音乐结构单元应该具有基本一致的音色单元分布，因此在候选结构边界的辅助下，本文采用非监督聚类算法分析确定音乐结构边界。最后将根据一些先验常识对结构边界进行调整，得到最后的音乐结构边界。

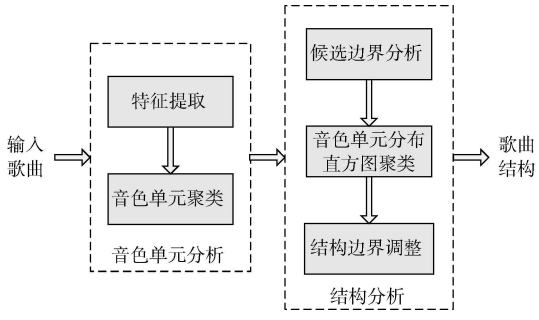

图 1　音乐结构分析算法结构框图

2.1 特征提取

特征提取方法首先将输入的音频信号进行分帧，可以分成固定长度的帧序列[4]，也可以根据乐理知识将音频信号分成最小音符长度的帧等非固定长度的帧序列，然后对每一帧提取特征向量。特征量的选择有很多种，可以选择其中的一种、两种或多种相结合。常用的几个特征量主要有 MFCC, CQT[23],

Constant-Q log-power Spectrum (CQS)[2], Chroma[3] 等，本文采用了对音色或者演奏乐器具有更强鲁棒性的基于离散余弦变换的 (DCT-Based) Chroma 特征 [24] 。

MFCC[3,7] 主要描述了音色信息，描述主旋律的信息比较少。而人们对流行歌曲重复性的感知，更多的依赖旋律的相似性而非音色的相似性，于是 Lu Lie[23] 等提出了基于音符的 CQT 特征，它能够将音乐信号表示为确切音乐音符的谱能量，其方法是通过中心频率成几何分布的滤波器组。

CQS 特征是将 CQT 值转化到对数功率尺度上进行计算[2], 其转换依据是西方音乐的音高范围在每个八度上也是呈 12 个对数等距离分布的音高。

一些最近的研究表明, 音乐结构分析问题更多的取决于泛音特征而非音色特性， Wakefield 及 Goto 采用 Chroma 特征来描述 12 个同性质的音高类之间的功率谱能量的分布。本文从对数频率尺度上的功率谱中提取 12 维的 Chroma 向量 [3], 表示某个八度上第 c 个 ($c = 1, \cdots, 12$) 音高类,

$$v_c(t) = \sum_{\mathrm{Oct}_L}^{\mathrm{Oct}_H} \int_{-\infty}^{+\infty} \mathrm{BPF}_{c,h}(f) \Psi_p(f,t) \mathrm{d}f, \qquad (1)$$

其中 Oct_L, Oct_H 为八度的范围，分别为 3 和 8, 两者覆盖的频率范围为 130 Hz 到 8 kHz 。 $\mathrm{BPF}_{c,h}(f)$ 是一个用汉宁窗定义的允许对数尺度频率通过的带通滤波器，具体请参考 [3] 。

本文采用了对音色变化具有更强鲁棒性的基于离散变换的 (DCT-Based) Chroma 特征 [24]，其与 Chroma 特征的计算区别在于, 在合并六个八度的每类半音之前, 先对 72 维 (六个八度, 每个八度分为十二个半音) 的数据进行 DCT 变换，然后一部分低维数据置为零，再做 DCT 逆变换，最后对所有八度上的每类半音求和，这样就得到了 12 维的特征。该算法主要依据是 DCT 变换后的低维能量和音色的相关度比较大, 将其置零再逆变换后将降低音色的影响。

2.2 音色单元聚类

我们假定歌曲由一些基本的音色单元组成，每个音色单元可以代表一系列在音色上相似的音频特征。每首歌曲的音色单元通过自动聚类的方法获得 [25] 。特征聚类描述了帧与帧之间的特征量的相似程度, 它描述的是信号的局部相似特性。在聚类过程中, 相似距离测度可以用欧式距离，余弦距离，基于结构的距离 [23,26], Kullback-Leibler (KL) 距离等 [2], 本文特征聚类中选用欧式距离, 并根据经验值, 将特征聚为 60 类左右。

2.3 候选边缘分析

在进行直方图聚类之前, 引入边缘检测算法, 其主要思想是处于边界点两边的音色单元分布差异很大或者很小。当两者的分布差别很大时, 说明两者属于不同的结构部分, 比如序曲和主歌; 当两者的分布差别很小时, 说明两者属于相同的结构部分, 比如副歌 A 和副歌 B 。该算法主要特点是根据某一时刻相邻窗长区域内分布的相似程度得到差异曲线，经过筛选后，保留最有可能的极值点作为直方图聚类过程中的边界辅助点，见图 2 。

图 2 中, 曲线代表的是分布状态差异曲线, 标号为 “+” 的点为筛选过后保留下的可能的边界点，纵向的直线代表歌曲真实的边界点。

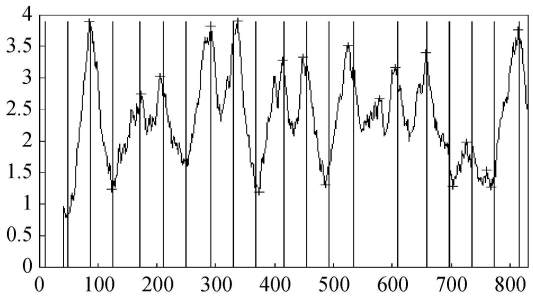

图 2 边缘检测曲线图

2.4 音色单元分布直方图聚类

根据特征量聚类的结果，将信号划分为具有 N 帧长度的窗序列，然后根据窗与窗之间的不同单元分布, 以窗为单位进行聚类, 这就是所谓的直方图聚类，它描述了音乐信号的局域整体特性。

根据边缘检测曲线, 在聚类过程中, 窗内状态量根据离保留住的极值点的距离选择合适的权值进行聚类，该权值表达式为：

$$W = \left[\frac{K_{\max}}{\overline{K}} \right]^{\lambda(|P_{\max} - \overline{P}| - L/2)}, \qquad (2)$$

其中， $K_{\max}$ 与 $\overline{K}$ 分别为此窗长内的状态差异曲线的最大值与平均值， $P_{\max}$ 与 $\overline{P}$ 分别为此窗长内保留的极值所在的时刻与此窗的中点所在的时刻, λ 为 [0,1] 范围内的参数值，需要根据经验确定。

本文直方图聚类中选用修正的 KL 距离，其计算表达式为：

$$d_{KL} = \frac{1}{N} \sum_{i=1}^{M} \left(x_i \left| \log \frac{x_i}{x_i'} \right| + x_i' \left| \log \frac{x_i'}{x_i} \right| \right). \qquad (3)$$

为了使聚类结果达到最优化，引入 Fisher 准则 [27], 它的主要思想是使类内距离尽可能小, 类间距离尽可能大。根据 Fisher 比来确定最佳聚类数目，得到最佳聚类结果， Fisher 比越大，表明类之间的

区分度就越大。特征之间的距离测度同样也选用修正的 KL 距离，实验结果证明选取修正的 KL 距离比其它距离效果要好。

2.5 结构边界调整

在直方图聚类的结果中仍然存在大约持续时间小于几秒的短小的片段，有的甚至更短，这在实际的音乐切分结果中很少出现。为此，采用短小片段合并的算法，将直方图聚类结果中小于一定时间阈值的短小片段，合并到与之距离最小的临近部分中去。

3 实验结果及分析

本文选用的数据集是 RWC Pop，包含 100 首 MP3 格式的日语歌曲 [31] 。

3.1 评价标准

本文选用两种评价标准来对结构分析结果进行评测，第一种评价标准是 K, average cluster purity (acp), average speaker purity (asp)[28] 。

acp, asp 分别代表在歌曲结构分割的过程中，由于可能的过分割与欠分割的错误得到的精度，其计算表达式分别为：

$$r_j^e=\sum_{i=1}^{Na} n_{ij}^2/(n_j^e)^2, acp=\frac{1}{N}\sum_{j=1}^{Ne} r_j^e\cdot n_j^e$$

$$r_i^a=\sum_{j=1}^{Ne} n_{ij}^2/(n_i^a)^2, asp=\frac{1}{N}\sum_{i=1}^{Na} r_i^a\cdot n_i^a \qquad (4)$$

其中，N 代表人工标注与切分结果中的所有帧的数目，N_a 表示人工标注状态的数目，N_e 表示切分结果中状态的数目，n_{ij} 表示属于人工结果中的第 i 状态，同时又属于切分结果中第 j 状态的数目，n_i^a 代表真实结果中属于状态 i 的总帧数，n_j^e 代表自动切分结果中属于状态 j 的总帧数。用 $K=\sqrt{\text{acp}\cdot\text{asp}}$ 作为最后的评价指标，K 值越大，说明切分结果与真实结果的匹配程度越高。

第二种评价标准是 Pairwise Precision (P_P), Pairwise Recall (R_P), Pairwise F-measure (F_P)[28−29], 其计算表达式为：

$$P_{p=}\frac{|M_e\cap M_a|}{|c|}, R_{p=}\frac{|M_e\cap M_a|}{|M_a|}, F_p=\frac{2\cdot P_p\cdot R_p}{P_p+R_p} \qquad (5)$$

其中，M_a 表示人工标注结果中，具有相同类标号的帧对的集合，同样地，M_e 表示在实验切分结果中，具有相同类标号的帧对的集合，$|\cdots|$ 表示相对应数的数目。

Pairwise Precision 表明了分割算法由于欠分割存在时的精度，而 Pairwise Recall 表明了由于过分割存在时算法的精度。

3.2 实验结果

首先给出在直方图聚类之前，有无边缘检测时，对 RWC 数据库前 32 首歌曲 K 值的统计结果如表 1 所示。可以看出，对边缘检测后，在直方图聚类时对窗内状态量进行加权处理，会使得最终结构切分结果有明显的改善，Chroma K 值提升幅度最大，为 16.4%，其次是 DCT-Based Chroma，为 16.1% 。但利用 Bayesian Information Criterion (BIC)[30] 算法来进一步确定结构边界时，在实验结果中并没有体现出更好的性能。

用前面提到的两种评价准则对 RWC 数据库的所有 100 首流行歌曲进行统计，其结果如表 2 所示。

从表 2 可以看出，基于 6 个八度的 DCT-Based Chroma 特征的自动聚类结果最好，其次是 Chroma，

表 1　有无边缘检测算法前后 K 值统计结果对比

K	Chroma	DCT-Based Chroma	CQT	CQS	MFCC
无边缘检测	0.5512	0.5644	0.5016	0.5040	0.5560
有边缘检测	0.6415	0.6552	0.5445	0.5660	0.6293

表 2　RWC 数据库 100 首流行歌曲自动聚类结构分析统计结果

特征	Chroma	DCT-Based Chroma(6Oct.)	DCT-Based Chroma(4Oct.)	CQT	CQS	MFCC
K	0.6311	0.6339	0.6286	0.5424	0.5698	0.627
Acp	0.6562	0.6673	0.6574	0.5245	0.5874	0.6464
Asp	0.6158	0.6074	0.6062	0.5827	0.5614	0.6175
F_P	0.5786	0.59	0.5841	0.4665	0.5037	0.5757
P_P	0.6178	0.635	0.6225	0.4633	0.5334	0.6096
R_P	0.5731	0.5689	0.5688	0.5369	0.5075	0.5758

然后依次是 4 个八度的 DCT-Based Chroma, MFCC, CQS, CQT 。由此可以证明，DCT-Based Chroma 特征，对音色的改变确实具有一定的鲁棒性。随机选取一首切分结果较好的歌曲，其自动聚类结果示意图如图 3 。

图 3 中，最上面是边缘检测曲线，中间是实验切分结果，相同的颜色代表同一类，最下面是人工标注结果，也是相同的颜色代表同一类。通过观察图可以看出，自动聚类得到的切分片段边界与真实人工标注结果的边界基本吻合，证明了边缘检测算法的必要性。在切分结果中，不可避免的出现了误分类的结果，从图 3 可以看出过分割现象的存在，但总体来看，自动聚类结构切分效果较好，没有出现较多的碎片。

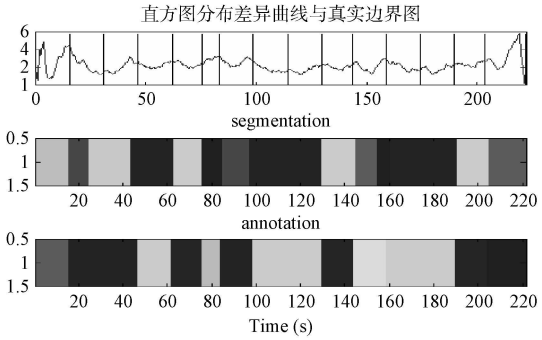

图 3 切分结果与标注结果对比图 ($K = 0.8258$)

4 结论

本文给出了基于音色单元的自动聚类的算法来分析音乐结构，实验结果表明了基于 6 个八度的 DCT-Based Chroma 特征的自动聚类结构分析统计结果最好，验证了边缘检测算法在确定结构最佳边界的有效性。同时实验结果还表明，自动聚类算法能避免更多切分碎片的产生。但在边缘检测算法及应用 BIC 算法如何更准确地确定结构边界点，仍有很大的提升空间。

参 考 文 献

1 Chen T. Music structure analysis and application. Master Thesis, Harbin Institute of technology, 2006

2 Abdallah S, Sandler M, Rhodes C, Casey M. Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, 2006; **65**(2): 485—515

3 Goto M. A chorus-section detecting method for musical audio signals. In ICASSP, Hong Kong, 2003: 437–440

4 Maddage N C, Xu C, Kankanhalli M S, Shao X. Content-based music structure analysis with applications to music semantic understanding. In: Proc. 12th annual ACM International Conference on Multimedia, 2004: 112—119

5 Peeters G, Laburthe A, Rodet X. Toward automatic music audio summary generation from signal analysis. In: Proc. ISMIR, 2002: 94—100

6 Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In: Proc. International Conference on Music Information Retrieval (ISMIR'07), 2007

7 Paulus J, Klapuri A. Music structure analysis using a probabilistic fitness measure and an integrated musicological model. In: Proc. International Conference on Music Information Retrieval (ISMIR'08), 2008: 369—374

8 Peeters G. Deriving musical structures from signal analysis for music audio summary generation: sequence and state approach. Lecture Notes in Computer Science, 2004

9 Chai W, Vercoe B. Structural analysis of musical signals for indexing and thumbnailing. In: Proc. Third ACM/IEEE-CS JCDL, 2003

10 Foote J. Automatic audio segmentation using a measure of audio novelty, In: Proc. International Conference on Multimedia and Expo (ICME'00), 2000; **1**) 2000.

11 Logan B, Chu S, Music summarization using key phrases. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'00), 2000; **1**

12 Cooper M, Foote J. Summarizing popular music via structural similarity analysis. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003: 127—130

13 Abdallah S, Nolan K, Sandler M, Casey M, Rhodes C. Theory and evaluation of a bayesian music structure extractor. In: Proc. ISMIR, London, UK, 2005: 420—425

14 Dannenberg R B, Hu N. Pattern discovery techniques for music audio. In: Proc. International Conference on Music Information Retrieval (ISMIR'02), Journal of New Music Research, 2003; **32**(2): 153—163

15 Aucouturier J J, Sandler M. Finding repeating patterns in acoustic musical signals: applications for audio thumbnailing. In: Proc. AES 22nd, International Conference on Virtual, Synthetic and Entertainment Audio, 2002: 412—421

16 Maddage N C, Xu C, Kankanhalli M S, Shao X. Content-based music structure analysis with applications to music semantic understanding. In: Proc. 12th annual ACM International Conference on Multimedia, 2004: 112—119

17 Aucouturier J J, Pachet F, Sandle M. The way it sounds: Timbre models for analysis and retrieval of polyphonic music signals. *IEEE Transactions of Multimedia*, 2005; **7**(6): 1028—1035

18 Foote J. Visualizing music and audio using self-similarity. *ACM Multimedia*, 1999; **1**: 77—80

19 Eckmann J P, Kamphorst S O, Ruelle D. Recurrence plots of dynamical systems. *Europhysics Letters*, 1987; **4**(9): 973—977

20 Wakefield G H. Mathematical representation of joint time-chroma distributions. In SPIE, 1999: 637—645

21 Orio N, Neve G. Experiments on segmentation techniques for music documents indexing. In: Proc. In 6th ISMIR, 2005: 624—627

22 Downie S, Nelson M. Evaluation of a simple and effective music information retrieval method. In: Proc. ACM SIGIR, 2000: 73—80

23 Lu L, Wang M, Zhang H. Repeating pattern discovery and structure analysis from acoustic music data. In: Proc. 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'04), 2004: 275—282

24 Muller M, Ewert S, Kreuzer S. Making chroma features more robust to timbre changes. In: Proc. IEEE ICASSP, 2009: 1877—1880

25 Linde Y, Buzo A, Gray R. An Algorithm for vector quantizer desig n. IEEE Transactions on Communications, 1980; **28**: 84—94

26 赵 阳, 周燕红, 许 洁萍. 基于不同距离测度的音乐结构分析研究. 第三届和谐人机环境联合学术会议 (HHME2007), 2007

27 王 飒, 郑 链. 基于 Fisher 准则和特征聚类的特征选择. 计算机应用, 2007; **27**(11)

28 Lukashevich H. Towards quantitative measures of evaluating song segmentation. In: Proc. 9th International Conference on Music Information Retrieval (ISMIR'08), 2008: 375—380

29 Paulus J, Klapuri A. Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm, IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, 2009: 1159—1170

30 DU Yunfeng, HU Wei, YANG Yonghong. Audio segmentation via tri-model bayesian information criterion. ICASSP 2007; **1**

31 Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, Ryuichi Oka. RWC Music Database: Popular, Classical, and Jazz Music Databases, Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002), 2002: 287—288