5G消息中的违规音乐识别方案

谢仪峒，杜刚，张晨，杜雪涛

（中国移动通信集团设计院有限公司，北京 100080）

【摘　要】　提出了识别5G消息音频内容中违规音乐的方案，可以应用在5G用户上传、下载多媒体内容的场景。基于色度能量归一化统计特征（CENS）、节奏特征的提取，使用动态时间规整算法（DTW）兼容速度差异，并进一步地提出了容错机制。该方案可识别5G消息中的违规音乐内容，满足实践中音频消息的内容安全管控需求，提升用户的使用体验。

【关键词】　5G；音乐匹配；CENS特征；节奏特征

1　引言

5G 消息突破了传统文本信息的局限性，将文字、图片、音视频、地理位置等信息有效融合，带来了多样化的便捷体验。其中，音视频内容中可能出现违规音频信息，一旦传播扩散将造成不良的社会影响，需要加强识别违规音乐的手段。

从当前的研究现状看，对于违规音乐内容的识别手段多聚焦于语音识别技术，即提取音频中的语音部分并将其转化成文本，以判定是否违规，该方法仅适用于音乐中包含语音的情况，无法对纯音乐进行有效识别。此外，违规音乐可能通过不同方式演绎，如不同乐器、不同演唱者、不同的声部结构等。已知一种违规音乐演绎方式并据此识别出其他演绎方式的违规音乐是具有挑战性的。

2　相关工作

2.1　音频分析相关技术

随着数字信号处理（DSP）技术的演进，音频特征可以从时域和频域两个维度分析。从时域维度上，可以分析信号幅度、相位随时间的变化，据此可以提取出过零率、短时能量、节奏特征等；从频域维度上，可以分析音乐的能量随频率变化的分布情况等，据此可以提取出相应的特征数据，包括色度特征（Chroma Feature）、色度能量归一化统计（CENS, Chroma Energy Normalized Statistics）特征、梅尔滤波器系数（MFC, Mel Filter Coefficients）、梅尔倒谱系数（MFCC, MeI-Freguency CeptraI Coefficients）等。

在时域特征中，节奏特征表示在时域上每个音符持续的时间构成的序列。在频域特征中，色度特征表示时间窗口内音高内容如何分布在十二个色度波段上，与音频信号的短时谐波含量密切相关。CENS 特征在局部上模糊色度特征，然后归一化处理。MFCC 特征是在 Mel 标度频率域提取出来的倒谱系数，Mel 标度描述了人耳对频率感知的非线性特性。

2.2　研究现状

在当前学者们的研究工作中，值得关注的成果不胜枚举。Joo S 等人采用音调序列分析的方式[1]，第一步进行音调候选估计，第二步进行音调序列识别，第三步进行平滑处理，对于单声部音乐，此种方法具有良好的效果，但是在和弦演奏的情况下，由于不同的声部音调、音色、节奏不同，结果的准确率不稳定，鲁棒性不足。Shih H H 等人使用基于 MFCC 特征训练的隐马尔可夫模型进行音符切分[2]，将信号划分类为“音符”或“静音”，然后使用混合高斯模型标记音高，此方案进行音高判断和前者具有相似的应用局限性。Wang A 提出了基于音乐指纹的搜索算法[3]，指纹令牌具有足够高的熵，以减少虚假匹配，这种匹配方法精确地对比时间轴上频率点，适用于识别同一首歌曲的情况，对于不同的演绎方式甚至演唱者不具备容错空间。

由此可见，对于哼唱或单声部演奏的音乐，常采用单声部旋律音高分析的研究路线，该路线优点是能提取出具体的乐谱旋律，缺点是仅分析了单声部的音高特征，和弦音乐效果不佳。对于原音乐识别，常采

用音乐指纹匹配作为研究路线，该路线优点是特征精简，对原音乐匹配精准，缺点是单当音频片段发生失真或非线性变化，音乐指纹也将演变成另一个截然不同的版本，从而失配。

违规音乐识别方案需要具备对音色、音调、节奏变化的鲁棒性，所以在违规音乐识别的场景中，音频特征的选型采用了 CENS 特征与节奏特征相结合的方式，以 CENS 特征模糊化音高信息，能够代表音乐在不同音阶上的能量分布特性，使用节奏特征代表人耳听力的感知效果，和 CENS 特征结合产生相互纠正的效应。

3 违规音乐识别方案

解析不同的违规音乐文件构建违规音乐数据库，以等长时间切片的音频存储到数据库，下文中称之为匹配样本 S。数据库中每个匹配样本包括 CENS 特征、节奏特征、音乐时长和名称等基本信息。其中，匹配样本的 CENS 特征表示为 SC，节奏特征表示为 ST。待识别的音频片段称为目标样本 Q，目标样本的 CENS 特征表示为 QC，节奏特征表示为 QT，目标样本和匹配样本默认时间等长。

3.1 特征提取

CENS 特征来自于色度特征的变换，色度特征体现为 12 维向量序列，每一个维度代表等律音阶的一个音高。将量化的色度特征向量序列与长为 w 的汉明窗进行分量卷积，然后基于参数 d 进行降采样，再次产生了一系列的向量，最后根据欧几里德范数进行归一化得到。

除此之外，仅依靠色度特征代表的声波能量分布情况并不能完全代表一首歌曲的全部信息，人耳对于不同乐曲的一个较高的区分度来自于音乐的节奏，需要结合音频的节奏特征进行违规性判定。

3.2 匹配策略

（1）CENS 特征匹配

CENS 特征具有时间分辨率属性，如图 1 所示，纵向排布的 12 维度对应 12 个八度等效的音高，而横向对应着每个音高在时间维度上的能量变化情况。每一个变化的单位是一个小方格，也是计算的基本单位，小方格的宽度即为时间分辨率。图中的分辨率是 500ms，即 500ms 内的能量分布变化会被忽略，从而消除了发音条件带来的无法预期的局部时间偏差。

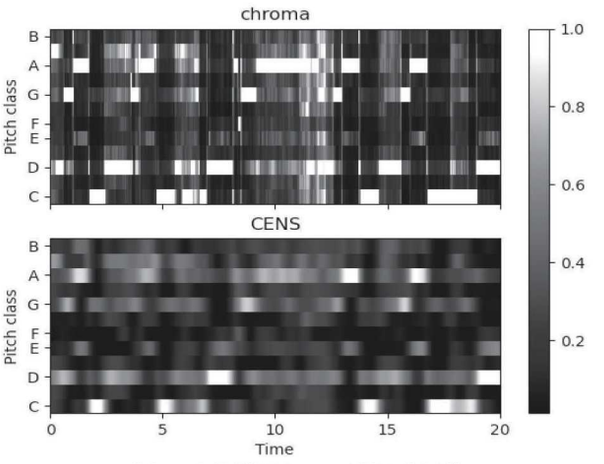

图1 色度特征及CENS特征示意图

CENS 特征纵向的一列代表音频在此单位时间内能量在不同音阶上的分布方式，也是检索过程中匹配的对象，在匹配样本中此列向量表示为 sc_i，匹配样本和目标样本时间定长，在时间分辨率不变的情况下，包含的向量数目均为常数 N，表示为 $SC=(sc_1, sc_2, \ldots, sc_N)$。$sc$ 由 12 个维度构成，即 $sc=(v_1, v_2, \ldots, v_{12})$。由于 CENS 特征经过归一化处理，所以 sc 向量是单位向量，几何上可以表示为在单位球体表面的一个点。

在时间维度上，对于 CENS 特征向量的对比具体化为计算 sc 和 qc 向量之间的差异，sc、qc 序列间的差异计算采用 DTW（Dynamic Time Warping）方案，以兼容局部节奏的差异性。在音阶维度上，对于向量 sc 和 qc 的对比具体化为计算其中元素 $v(j)$ 之间的差值。对比的结果称为样本距离。对于原歌曲相同位置的片段，样本距离趋近于 0，其他同乐谱不同演绎方式的歌曲，歌曲风格越相似、音色特点越相似，样本距离越小。

（2）节奏特征匹配

节奏特征 $T=(t_1, t_2, \ldots, t_M)$，体现为一维数组，$ST$ 和 QT 的匹配采用 DTW 算法。对于原歌曲相同位置的片段，样本距离接近于 0，其他同乐谱不同演绎方式的歌曲，节奏特征越相似，样本距离越小。

3.3 检索策略

匹配样本采集的过程中，设置片段的定长区间为 D，从违规音乐起始时刻开始，以固定步长 $Step$ 向后逐一切分，得到的音乐片段进行特征和其他属性提取，然后入库。

将目标样本和匹配样本的 CENS 特征和节奏特征进

行逐一对比，在比对过程中，如果CENS特征与节奏特征的样本距离均趋近于0，则停止对比，当前的匹配样本所属歌曲即为目标样本的来源，其所处于歌曲的位置也就是目标样本在原歌曲中的位置。否则，如果所有匹配样本得到的样本距离都不趋近于0，将所有样本的CENS特征和节奏特征产生的距离降序排列，分别得到列表*Lcens*和*Ltempo*，对于*Lcens*取*topK*，其中命中的歌曲对应在*Ltempo*中的节奏距离若不在*topK*之列，则进行剔除，在余下的命中歌曲中取*Lcens*和*Ltempo*的公共部分，并选取距离值最小的，即为判定的命中违规音乐。

3.4 容错机制

在CENS特征的匹配过程中，利用CENS特征的局部模糊化和DTW的动态偏差，保证了匹配方案的局部稳定性，但是对于不同的歌曲演奏版本，全曲的节奏快慢差异能达到10%甚至20%之多，在这种情况下，DTW的对比方案会在匹配结束的情况下产生不属于区间*D*的冗余内容。

基于以上考虑，要消除较大程度的全局节奏差异，需要对匹配样本进行区间D或者采样率调整。在长度调整的情况下，在原有匹配的流程中，添加两组匹配流程，其中一组将目标样本进行时域上的伸缩，以比例α减掉或增加匹配样本长度尾部内容，由此得到的*Lcens*和*Ltempo*会增加至三组，分别为(*Lcens*, *Ltempo*)、(*Lcens*1, *Ltempo*1)、(*Lcens*2, *Ltempo*2)。在检索的过程中，对三组结果的*topK*取并集，并用*Ltempo*的并集进行剔除。同理，在采样率调整的情况下，仍然添加两组匹配流程，其中两组的CENS特征生成过程中将(*w*, *d*)参数分别向上和向下调整，达到容错的目的。

4 实验分析

实验分两类共计三组：第一类测试出自于原曲的片段在原曲的Step递增过程中CENS特征距离和节奏特征距离产生的变化，共计一组；第二类测试第一组通过反复选取不同条件的目标样本，在匹配样本音乐数据库中进行检索，记录违规性判定的准确率，第二组通过改变匹配样本区间*D*进行容错方案的验证，同样记录判定准确率，共计两组。

4.1 实验条件

实验基于python3开发程序，在windows环境下执行，音乐样本库选型采用MySQL5.7。

（1）数据预备和特征提取。

一类实验选取电子琴版本回家，歌曲长度共计173秒，目标样本在歌曲中位置处于13秒，片段定长区间*D*为30秒。

二类实验选取30首不同的歌曲构建违规音乐数据库，采集音乐的基本信息和整曲特征，所有音乐总长度共计5639秒，片段定长区间*D*为30秒，*Step*为5秒，进行切片，得到切片数量总计955片。二类实验三选取的目标样本长度分别为：30s+15%、30s、30s-15%。

（2）匹配和检索。在*SC*和*QC*向量进行匹配的时候，DTW算法的最大容忍长度差异设置成16，超过16则不被允许。在结果判定过程中，原曲判定CENS特征的阈值为2，节奏特征的阈值为2，若没有命中原曲，需要从*Lcens*和*Ltempo*中进行比较筛选，候选范围*topK*取*K*=20。

（3）测试目标样本。用于测试的目标样本总数为16个，其中3个来自于原曲，其他的来自于不同演唱者、不同演绎方式（独奏、合唱）、不同乐器。目标样本的片段区间D仍为30秒。

4.2 实验结果

一类实验第一组结果如图2所示，在13~15秒的位置，CENS特征距离和节奏特征距离同时出现了最低点，在方案中，即可认定为出自于原曲并终止检索。图中的CENS特征距离在85~90秒位置也出现了低点，但是对应的节奏距离不在原曲判定的阈值范围内，故可以被舍弃。

二类实验第一组结果如表1所示，CENS特征取最小作为判定音频相似的依据，准确率75%，使用节奏特征作为唯一依据，正确率62.5%，而将两者结合能达到81.25%，说明此二种特征均能表现音频的一部分特点，但不足以完全判定。

二类实验第二组结果如表2所示，其中以30s+15%和30s-15%的特性都会遗漏一定的相似样本，是因为将目标样本进行长度调整可能造成原本完全合适的相似音乐缺少或冗余内容，但是二者检索到节奏差异较大的匹配样本不同，故能在结合到原区间方案的时候产生正确率的提高。

4.3 分析总结

从一类实验第一组的结果可以看出，结合CENS和节奏特征进行原曲判定尤其是片段的位置判定，具有较优秀的准确率与可靠性，进一步证明了将特征结

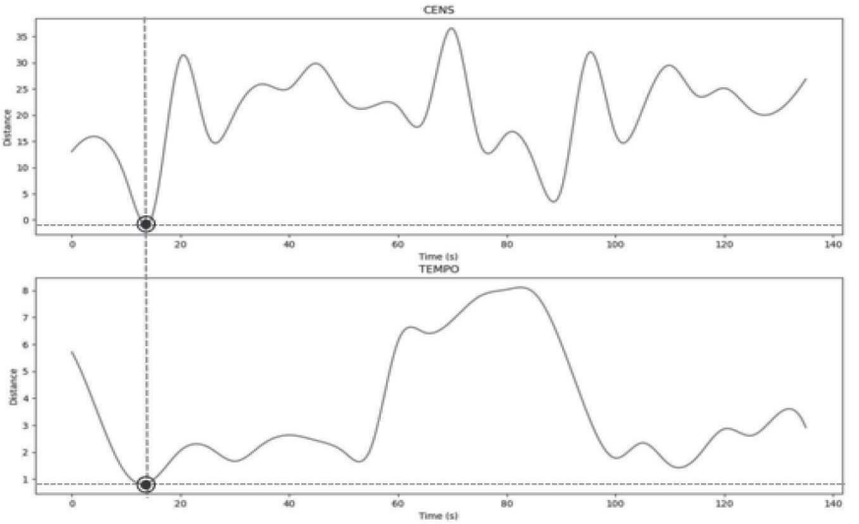

图2　CENS特征距离及节奏特征距离示意图

表1　判定特征对比实验准确率表

判定特征	CENS特征	节奏特征	二者结合特征
判定准确率	75.00%	62.50%	81.25%

表2　特征区间对比实验准确率表

特征区间	30s-15%	30s	30s+15%	三者结合
判定准确率	62.50%	81.25%	68.75%	87.50%

合是全方位解析音乐文件的一个重要方法。

从二类实验第一组的结果可以看出，基于CENS特征的违规音乐识别准确率稍显逊色，但是虚假识别的结果，与目标样本真正所属音乐的节奏特征和乐曲风格具备一定的相似度，这印证了CENS特征代表了不同音阶的能量分布特征的属性；基于节奏特征的音乐识别在准确率上有所提高，虚假识别结果的乐曲节奏感相似，但乐曲风格可以产生巨大的差异。而两种特征的结合在查准率上有着显著的提高。

从二类实验第二组的结果可以看出，将乐曲的全局节奏差异进行拉伸或者压缩，能够匹配到同一首歌曲不同演绎方式中的不同部分，将三种不同节奏的目标样本进行联合查询，能完善音乐检索的效果。

5　结束语

针对5G消息中的违规音乐提出了鲁棒的违规音乐识别方案，能够在单声部、多声部、多方式演绎场景中实现违规音乐的高效识别。在未来科研工作中，提高实时性和检索结果的广泛同类乐曲关联可以成为研究目标。

参考文献：

[1] Joo S, Jo S, Chang D Y. Yoo: Melody extraction from polyphonic audio signal MIREX 2009[J]. MIREX Audio Melody Extraction Contest Abstracts, 2009.

[2] Shih H H, Narayanan S S, Kuo C. Multidimensional humming transcription using a statistical approach for query by humming systems[C]//Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. IEEE, 2003.

[3] Wang A. An industrial-strength audio search algorithm[C]//ISMIR 2003, 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA, 2003. ★

作者简介

谢仪頔：硕士，现任中国移动通信集团设计院有限公司咨询设计师，主要研究方向为信息安全、计算机视觉和语音。

杜刚：高级工程师，博士，现任职于中国移动通信集团设计院有限公司，主要研究方向为信息安全、人工智能。

张晨：硕士，现任职于中国移动通信集团设计院有限公司，主要研究方向为信息安全、网络安全、内容安全、大数据分析。

杜雪涛：硕士，现任职于中国移动通信集团设计院有限公司，主要研究方向为信息安全、网络安全、内容安全。