

[19] 中华人民共和国国家知识产权局



[12] 发明专利申请公布说明书

[21] 申请号 200810083110.3

[43] 公开日 2008 年 8 月 6 日

[51] Int. Cl.
G10L 11/00 (2006.01)
G10L 19/00 (2006.01)

[11] 公开号 CN 101236742A

[22] 申请日 2008.3.3

[21] 申请号 200810083110.3

[71] 申请人 中兴通讯股份有限公司

地址 518057 广东省深圳市南山区高新技术
产业园科技南路中兴通讯大厦

[72] 发明人 刘开文 付中华

[74] 专利代理机构 北京康信知识产权代理有限责任
公司

代理人 尚志峰 吴孟秋

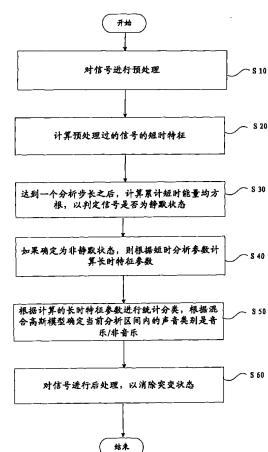
权利要求书 6 页 说明书 20 页 附图 3 页

[54] 发明名称

音乐/非音乐的实时检测方法和装置

[57] 摘要

本发明提供了一种音乐/非音乐的实时检测方法和装置，方法包括以下步骤：对信号进行预处理；计算预处理过的信号的短时特征；达到一个分析步长之后，计算累计短时能量均方根，以判定信号是否为静默状态；如果确定为非静默状态，则根据短时分析参数计算长时特征参数；根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及对信号进行后处理，以消除突变状态。本发明实现了音乐/非音乐的稳健而有效的实时检测或分割，与语音活动检测相结合，能够组成完整的声音活动检测方案。



1. 一种音乐/非音乐的实时检测方法，其特征在于，包括以下步骤：

对信号进行预处理；

计算预处理过的所述信号的短时特征；

达到一个分析步长之后，计算累计短时能量均方根，以判定所述信号是否为静默状态；

如果确定为非静默状态，则根据短时分析参数计算长时特征参数；

根据计算的所述长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及

对所述信号进行后处理，以消除突变状态。

2. 根据权利要求 1 所述的实时检测方法，其特征在于，对信号进行预处理具体包括：

入口参数控制、模型库加载、输入文件或数据格式处理，以及预加重、分帧加窗、参数和缓冲区初始化。

3. 根据权利要求 2 所述的实时检测方法，其特征在于，

入口参数控制包括设置语音信号或噪声信号检测得分的额外加分 spS 和 nsS；

模型库加载包括加载事先经过大量数据训练过的语音、噪音、音乐三者的统计模型，静音是以短时能量判断；

输入文件或数据格式处理采用 8kHz 采样 16 比特量化；

预加重系数取系数为 -0.80；

分帧加窗取帧长为 32 毫秒，256 个采样点；

参数和缓冲区初始化为帧移 10 毫秒，80 个采样点，窗函数采用 256 点的海明窗。

4. 根据权利要求 3 所述的实时检测方法，其特征在于，计算预处理过的所述信号的短时特征具体包括：

计算时域短时能量特征、幅度谱以及频谱特征、实倒谱、谱起伏程度参数、Mel 域子带能量以及短时调性强度特征，并标记当前帧的调性。

5. 根据权利要求 4 所述的实时检测方法，其特征在于，设加窗之后的各帧信号为 data，帧长为 N，

时域短时特征是指短时能量均方根，记为 feaRMS，则

$$\text{feaRMS} \text{ 定义: } \text{feaRMS} = \sqrt{\sum_{n=1}^N \text{data}^2(n)};$$

幅度谱定义： rawsp = |F(data)|，其中 F() 表示离散付立叶变换；

对数功率谱定义： logPowsp = log(|F(data)|^2)；

实倒谱定义： rcp = real(F^{-1}(\log(|F(data)|)))；

在对数幅度谱基础上，计算谱重心位置，记为 feaCenSP，谱重心定义为功率谱的某一频率，小于该频率的谱能量与大于该频率的谱能量相等；

在实倒谱基础上，计算谱起伏程度参数，记为 RcPr，其采用实倒谱 3 ~ 14 个系数绝对值之和与 1 ~ 2 系数绝对值之和的比值的对数；

计算 Mel 域子带能量，采用 40 个 Mel 域子带，用三角滤波器组计算每个子带内功率谱能量并取对数，最后对 40 个子带能量进行归一化和零均值化，得到的新矢量记为 spBP；

计算短时调性强度特征 $feaRcp = \sum_{k=l}^N |rcp(k)|$ ，其中 l 是求和起始点，该实施例选择 l=14；

标记当前帧的调性包括取 $xr=\max(rdp(1:N))$ ，如果 $xr>tonThres$ ，则标记为调性；否则标记为非调性，其中 tonThres 为调性门限，取 0.14。

6. 根据权利要求 5 所述的实时检测方法，其特征在于，达到一个分析步长之后，计算累计短时能量均方根，以判定所述信号是否为静默状态具体包括：

每个分析步长进行一次类别判定，队列长度取 100 帧，分析步长为 10 帧；

到达一个分析步长后，计算累计调性参数 trc，该参数是队列内所有帧的 feaRcp 之和；

进行静音判断，静音判断的依据是分析步长内最大的 feaRMS 参数，如果 $\max(\log(feaRMS)) < Thr_sil$ ，则是静音状态，否则为非静音状态，其中 Thr_sil 是静音检测门限，取 -3。

7. 根据权利要求 6 所述的实时检测方法，其特征在于，如果确定为非静默状态，则根据短时分析参数计算长时特征参数包括：

计算调性帧平均能量与平均能量之比 Deng：根据短时分析中得到的调性/非调性标记结果，对于队列内的所有帧信号，Deng 定义为 $Deng = \frac{\text{mean}_{i \in \Theta}(feaRMS(i))}{\text{mean}_{j \in All}(feaRMS(j))}$ ， Θ 是所有调性帧，All 是分析区间内所有帧；

计算调性强度特征 $trc = \text{sum}(feaRcp)$ ，其中 $\text{sum}()$ 是对整个缓冲队列内所有帧求和的求和函数；

计算 feaRMS 方差对数特征 logVRMS，

$$\log VRMS = \log(\text{var}(feaRMS / \text{mean}(feaRMS)))$$
；

计算谱重心方差对数特征 logVCenSpec，其是队列内所有 feaCenSP 的方差的对数；

计算谱起伏度对数方差特征 logVRcPr，其是队列内所有 RcPr 的方差的对数；

计算 4Hz 调制能量特征 f4Hzmelbp，包括：根据 Mel 域子带能量，对于队列内所有帧，计算各子带的 4Hz 调制能量，其中计算各子带的 4Hz 调制能量包括：采用 2 阶全极点滤波器，计算滤波器输出能量和原始能量之比，然后将 40 个子带的比值相加再取对数，得到所述 4Hz 调制能量，其中对于第 k 个子带，该子带的比值计算为 $R(k) = \frac{\text{sum}(\text{filter}(spBP_k))}{\text{sum}(spBP_k)}$ ，其中 $\text{sum}()$ 是对整个缓冲队列内所有帧求和的求和函数， $\text{filter}()$ 是滤波函数；

计算态范围特征 Dong，其是队列内最大能量和最小能量之比的对数，即 $Dong = \log\left(\max_{i \in All}(feaRMS(i)) / \min_{j \in All}(feaRMS(j))\right)$ ；

将计算的上述 7 种特征参数组成 7 维特征矢量，并对各特征参数进行平移和放缩处理，使其数值都分布在相近的范围内。

8. 根据权利要求 7 所述的实时检测方法，其特征在于，根据计算的所述长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐具体包括：

根据贝叶斯极大似然分类法则确定当前分析区间内的声音类别；

检测中，每类信号都用一个 GMM 来表示，或者用模型 λ 表示；

在 GMM 当中，高斯混合概率密度是 M 个高斯分量概率密度之和，用公式表示为 $p(\vec{x}|\lambda) = \sum_{i=1}^M p_i b_i(\vec{x})$ ，其中 \vec{x} 是一个 D 维随机变量， $b_i(\vec{x})$ 是第 i 个高斯分量的概率密度函数， p_i 是第 i 个高斯分量的权重， $i=1, \dots, M$ ；每个高斯分量的概率密度是一个 D 维高斯函数： $b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\right\}$ ，式中 $\vec{\mu}_i$ 和 Σ_i 表示第 i 个高斯分量的均值和协方差矩阵；所有的高斯分量权重之和满足归一化条件 $\sum_{i=1}^M p_i = 1$ ；

一个高斯混合概率密度用三种参数表示：各分量的权重，各分量的均值和协方差矩阵；这些参数放到一起，统称为模型参数，记为 $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i=1, \dots, M$ ；

首先对训练数据进行特征分析，得到音乐、语音、噪声三类训练数据的特征矢量集合，然后分别训练三组 64 个分量的 GMM 模型，包括 MUgmm、SPgmm、NSgmm，对应的模型用 $\lambda_{mu}, \lambda_{sp}, \lambda_{ns}$ 表示；GMM 模型的训练采用的期望值最大化算法；

在进行检测分类时，对于新输入的一个特征矢量 x ，在模型 λ 下的对数似然得分为 $Score_\lambda(x) = \log(p(x|\lambda))$ ，对于三个模型而言，最后的分类结果为

$CID(\mathbf{x}) = \arg \max (Score_{sp}(\mathbf{x}) + spS, Score_{ns}(\mathbf{x}) + nsS, Score_{mu}(\mathbf{x}))$, 其中 spS 是给语音模型的加分, nsS 是给噪声模型的得分。

9. 根据权利要求 8 所述的实时检测方法, 其特征在于, 对所述信号进行后处理, 以消除突变状态具体包括:

在音频类型改变持续 1s 才认为类型改变有效, 否则视为突变状态, 音频类型维持突变前的状态。

10. 一种音乐/非音乐的实时检测装置, 其特征在于, 包括:

预处理模块, 用于对信号进行预处理;

短时特征计算模块, 用于计算预处理过的所述信号的短时特征;

静默状态判断模块, 用于达到一个分析步长之后, 计算累计短时能量均方根, 以判定所述信号是否为静默状态;

长时特征计算模块, 用于如果确定为非静默状态, 则根据短时分析参数计算长时特征参数;

确定模块, 用于根据计算的所述长时特征参数进行统计分类, 根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐; 以及

后处理模块, 用于对所述信号进行后处理, 以消除突变状态。

音乐/非音乐的实时检测方法和装置

技术领域

本发明涉及通信领域，具体而言，涉及一种音乐/非音乐的实时检测方法和装置。

背景技术

在语音通信领域中，常常因为传输带宽的约束需要进行非连续传输（Discontinued Transmission, DTX），其中最关键的步骤就是语音活动检测（Voice Activity Detection, VAD）。随着多媒体业务的不断丰富，通信中除了语音和环境声学噪声之外，常常有彩铃等音乐信号加入，而在 DTX 传输条件下，一般的 VAD 会把部分音乐信号当作噪声进行处理，使得音乐信号无法正常传输，并且音乐信号的编码方式也有别于非音乐（包括噪声、语音、和静音等），因此必须及时检测出这些音乐信号，进而采用合适的编码算法进行传输。另外，噪声抑制（Noise Reduction, NR）中，如果音乐信号也采用非音乐的方式进行处理，会使音乐信号严重失真，因此也有必要做音乐/非音乐的判断。

在音频通信应用中，音乐/非音乐的检测难度在于音乐的多变以及语音中的噪声的多变。许多研究内容都分析了音乐与语音的差异，但因为音乐的多变使得这些差异只对部分音乐有效。通常语音的静默片断出现的概率大、能量变化大，但在节奏很快的音乐中也有类似现象；许多音乐的高频信息丰富，但在歌唱时也并非如此；音乐的基音频率要么变化小、要么突变，但和声和复调音乐使基音频率

的提取有时非常困难；音乐有节奏感，但却并非简单的周期重复。使问题更为棘手的是语音中包含的噪声，尤其是谐波噪声，这些谐波噪声在较短的时间内与乐声很像，只是因为持续时间长才成为噪声。

目前的音乐/非音乐分类方案主要存在以下不足：（1）从音频信号短时处理出发提取的短时特征仅仅利用了很少量的信息，不足以反映两类信号的差异。实际上在较短的时间内看，音乐、语音和噪声常常没有明显的界限；（2）长时分析缺乏有力的特征描述，要么时间片要求较长，例如对整个音频文件的分类，不适用于实时通信领域的要求，要么是从音频动态特征衍生出来新的统计特征，但其分类能力却无法保证；（3）常常需要获得音乐和语音及噪声的精细结构及其变化，对采样率和计算量要求较高，难以满足在嵌入式平台上应用；（4）所用的测试数据不充分，很难满足复杂的通信环境的要求。

在实现本发明过程中，发明人发现总之，实际应用对音乐/非音乐检测的要求是实时、稳健性、有效，能够为后续处理奠定基础，而目前的问题是——短时分析有用信息太少，难以反映两者的差异；长时分析可以较好地反映两者差异，但是计算量大，延时长；特征不够稳健，对音乐和语音及噪声内在的区别没有充分的反映。

发明内容

本发明旨在提供一种音乐/非音乐的实时检测方法和装置，能够解决现有的短时分析和长时分析上述各自存在的问题。

在本发明的实施例中，提供了一种音乐/非音乐的实时检测方法，包括以下步骤：对信号进行预处理；计算预处理过的信号的短时特征；达到一个分析步长之后，计算累计短时能量均方根，以判

定信号是否为静默状态；如果确定为非静默状态，则根据短时分析参数计算长时特征参数；根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及对信号进行后处理，以消除突变状态。

优选的，对信号进行预处理具体包括：入口参数控制、模型库加载、输入文件或数据格式处理，以及预加重、分帧加窗、参数和缓冲区初始化。

优选的，入口参数控制包括设置语音信号或噪声信号检测得分的额外加分 spS 和 nsS；模型库加载包括加载事先经过大量数据训练过的语音、噪音、音乐三者的统计模型，静音是以短时能量判断；输入文件或数据格式处理采用 8kHz 采样 16 比特量化；预加重系数取系数为 -0.80；分帧加窗取帧长为 32 毫秒，256 个采样点；参数和缓冲区初始化为帧移 10 毫秒，80 个采样点，窗函数采用 256 点的海明窗。

优选的，计算预处理过的信号的短时特征具体包括：计算时域短时能量特征、幅度谱以及频谱特征、实倒谱、谱起伏程度参数、Mel 域子带能量以及短时调性强度特征，并标记当前帧的调性。

优选的，设加窗之后的各帧信号为 data，帧长为 N，时域短时特征是指短时能量均方根，记为 feaRMS，则 feaRMS 定义：

$$feaRMS = \sqrt{\sum_{n=1}^N data^2(n)}; \text{ 幅度谱定义: } rawsp = |F(data)|, \text{ 其中 } F(\cdot) \text{ 表示}$$

离散付立叶变换；对数功率谱定义： $logPowsp = \log(|F(data)|^2)$ ；实倒谱定义： $rcp = real(F^{-1}(\log(|F(data)|)))$ ；在对数幅度谱基础上，计算谱重心位置，记为 feaCenSP，谱重心定义为功率谱的某一频率，小于该频率的谱能量与大于该频率的谱能量相等；在实倒谱基础上，计算谱起伏程度参数，记为 RcPr，其采用实倒谱 3~14 个系数绝对值之

和与 1~2 系数绝对值之和的比值的对数；计算 Mel 域子带能量，采用 40 个 Mel 域子带，用三角滤波器组计算每个子带内功率谱能量并取对数，最后对 40 个子带能量进行归一化和零均值化，得到的新矢量记为 spBP；计算短时调性强度特征 $feaRcp = \sum_{k=l}^N |rcp(k)|$ ，其中 l 是求和起始点，该实施例选择 l=14；标记当前帧的调性包括取 $xr = \max(rcp(1:N))$ ，如果 $xr > tonThres$ ，则标记为调性；否则标记为非调性，其中 tonThres 为调性门限，取 0.14。

优选的，达到一个分析步长之后，计算累计短时能量均方根，以判定信号是否为静默状态具体包括：每个分析步长进行一次类别判定，队列长度取 100 帧，分析步长为 10 帧；到达一个分析步长后，计算累计调性参数 trc，该参数是队列内所有帧的 feaRcp 之和；进行静音判断，静音判断的依据是分析步长内最大的 feaRMS 参数，如果 $\max(\log(feaRMS)) < Thr_sil$ ，则是静音状态，否则为非静音状态，其中 Thr_sil 是静音检测门限，取 -3。

优选的，如果确定为非静默状态，则根据短时分析参数计算长时特征参数包括：计算调性帧平均能量与平均能量之比 Deng：根据短时分析中得到的调性/非调性标记结果，对于队列内的所有帧信号，Deng 定义为 $Deng = \frac{\text{mean}(feaRMS(i))}{\text{mean}(feaRMS(j))}$ ， Θ 是所有调性帧，All 是分析区间内所有帧；计算调性强度特征 $trc = \text{sum}(feaRcp)$ ，其中 sum() 是对整个缓冲队列内所有帧求和的求和函数；计算 feaRMS 方差对数特征 logVRMS， $\log VRMS = \log(\text{var}(feaRMS / \text{mean}(feaRMS)))$ ；计算谱重心方差对数特征 logVCenSpec，其是队列内所有 feaCenSP 的方差的对数；计算谱起伏度对数方差特征 logVRcPr，其是队列内所有 RcPr 的方差的对数；计算 4Hz 调制能量特征 f4Hzmelbp，包括：根据 Mel 域子带能量，对于队列内所有帧，计算各子带的 4Hz 调制能量，其中计算各子带

的 4Hz 调制能量包括：采用 2 阶全极点滤波器，计算滤波器输出能量和原始能量之比，然后将 40 个子带的比值相加再取对数，得到 4Hz 调制能量，其中对于第 k 个子带，该子带的比值计算为 $R(k) = \frac{\text{sum}(\text{filter}(spBP_k))}{\text{sum}(spBP_k)}$ ，其中 sum () 是对整个缓冲队列内所有帧求和的求和函数，filter () 是滤波函数；计算态范围特征 Dong，其是队列内最大能量和最小能量之比的对数，即 $\text{Dong} = \log \left(\max_{i \in \text{All}} (\text{feaRMS}(i)) / \min_{j \in \text{All}} (\text{feaRMS}(j)) \right)$ ；将计算的上述 7 种特征参数组成 7 维特征矢量，并对各特征参数进行平移和放缩处理，使其数值都分布在相近的范围内。

优选的，根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐具体包括：根据贝叶斯极大似然分类法则确定当前分析区间内的声音类别；检测中，每类信号都用一个 GMM 来表示，或者用模型 λ 表示；在 GMM 当中，高斯混合概率密度是 M 个高斯分量概率密度之和，用公式表

示为 $p(\vec{x}|\lambda) = \sum_{i=1}^M p_i b_i(\vec{x})$ ，其中 \vec{x} 是一个 D 维随机变量， $b_i(\vec{x})$ 是第 i 个高斯分量的概率密度函数， p_i 是第 i 个高斯分量的权重， $i=1, \dots, M$ ；每个高斯分量的概率密度是一个 D 维高斯函数：

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \text{，式中 } \vec{\mu}_i \text{ 和 } \Sigma_i \text{ 表示第 } i \text{ 个高斯}$$

分量的均值和协方差矩阵；所有的高斯分量权重之和满足归一化条件 $\sum_{i=1}^M p_i = 1$ ；一个高斯混合概率密度用三种参数表示：各分量的权重，各分量的均值和协方差矩阵；这些参数放到一起，统称为模型参数，记为 $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i=1, \dots, M$ ；首先对训练数据进行特征分析，得到音乐、语音、噪声三类训练数据的特征矢量集合，然后分别训练三组 64 个分量的 GMM 模型，包括 MUgmm、SPgmm、NSgmm，对应的模型用 $\lambda_{mu}, \lambda_{sp}, \lambda_{ns}$ 表示；GMM 模型的训练采用的期望值最大化算

法；在进行检测分类时，对于新输入的一个特征矢量 x ，在模型 λ 下的对数似然得分为 $Score_{\lambda}(x) = \log(p(x|\lambda))$ ，对于三个模型而言，最后的分类结论为 $CID(x) = \arg \max(Score_{sp}(x) + spS, Score_{ns}(x) + nsS, Score_{mu}(x))$ ，其中 spS 是给语音模型的加分， nsS 是给噪声模型的得分。

优选的，对信号进行后处理，以消除突变状态具体包括：在音频类型改变持续 1s 才认为类型改变有效，否则视为突变状态，音频类型维持突变前的状态。

在本发明的实施例中，还提供了一种音乐/非音乐的实时检测装置，包括：预处理模块，用于对信号进行预处理；短时特征计算模块，用于计算预处理过的信号的短时特征；静默状态判断模块，用于达到一个分析步长之后，计算累计短时能量均方根，以判定信号是否为静默状态；长时特征计算模块，用于如果确定为非静默状态，则根据短时分析参数计算长时特征参数；确定模块，用于根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及后处理模块，用于对信号进行后处理，以消除突变状态。

本发明实现了音乐/非音乐的稳健而有效的实时检测或分割，与语音活动检测（Voice Activity Detection，VAD）相结合，能够组成完整的语音活动检测（Sound Activity Detection，SAD）方案。

附图说明

此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

图 1 示出了根据本发明实施例的音乐/非音乐的实时检测方法的流程图；

图 2 示出了根据本发明实施例的 M 个高斯分量的 GMM 拓扑图；

图 3 示出了根据本发明实施例的音乐/非音乐的实时检测装置的方框图；以及

图 4 示出了根据本发明优选实施例的音频传输系统的示意图。

具体实施方式

下面将参考附图并结合实施例，来详细说明本发明。

图 1 示出了根据本发明实施例的音乐/非音乐的实时检测方法的流程图，包括以下步骤：

步骤 S10，对信号进行预处理；

步骤 S20，计算预处理过的信号的短时特征；

步骤 S30，达到一个分析步长之后，计算累计短时能量均方根，以判定信号是否为静默状态；

步骤 S40，如果确定为非静默状态，则根据短时分析参数计算长时特征参数；

步骤 S50，根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及

步骤 S60，对信号进行后处理，以消除突变状态。

该实施例实现了音乐/非音乐的稳健而有效的实时检测或分割，与语音活动检测相结合，能够组成完整的声音活动检测方案。

该实施例把音频信号先分为语音、噪音、音乐和静音，然后语音、噪音和静音合并为非音乐类。因此在计算检测准确度时语音、噪音和静音之间相互误检算准确检测。方案各个部分的实施细节描述如下：

1) 信号预处理

信号预处理包括入口参数控制、模型库加载、输入文件（或数据）格式处理，以及预加重、分帧加窗、参数和缓冲区初始化等步骤。

在入口参数控制中可以设置语音信号或噪声信号检测得分的额外加分 spS 和 nsS。因为该方案采用的是极大似然得分的分类检测方法，而不同的应用领域中对语音、噪音、音乐三类信号的误检代价要求不同，因此可以通过这两个参数对最后的似然得分进行修正，以额外降低某类信号的误检率。

模型库加载是指加载事先经过大量数据训练过的语音、噪音、音乐三者的统计模型，静音是以短时能量判断。关于模型问题在下文统计模型和似然得分判断中说明。

另外，由于音频通信中通常采用 8kHz 采样 16 比特量化的单声道音频数据，因此该实施例以该格式化数据为处理对象。如果应用于其他格式数据，那么在此预处理阶段需要完成格式转换工作，例如降低采样率。

预加重系数与常规语音信号处理类似，取系数为 -0.80。帧长为 32 毫秒（256 个采样点，为了进行快速付立叶变换），帧移 10 毫秒

(80个采样点), 即帧间重叠22毫秒(172个采样点)。该帧长选择也符合常用的语音信号处理要求。窗函数采用256点的海明窗。

2) 计算时域短时能量特征、幅度谱以及频谱特征、实倒谱、谱起伏程度参数、Mel域子带能量以及短时调性强度特征, 并标记当前帧的调性。

时域短时特征是指短时能量均方根, 记为feaRMS。设加窗之后的各帧信号为data, 帧长为N, 则feaRMS定义如下:

$$feaRMS = \sqrt{\sum_{n=1}^N data^2(n)} \quad (1)$$

短时特征的计算主要基于幅度谱, 对数功率谱, 实倒谱, Mel域子带能量几个基本概念。设加窗之后的各帧信号为data, 则幅度谱定义如式(2):

$$rawsp = |F(data)| \quad (2)$$

其中F()表示离散付立叶变换。在计算出幅度谱之后, 对小于100Hz的低频数据进行衰减, 避免直流偏置和低频噪声对后续处理的影响。

对数功率谱定义如式(3)

$$logPows = \log(|F(data)|^2) \quad (3)$$

实倒谱定义如式(4)

$$rcp = real(F^{-1}(\log(|F(data)|))) \quad (4)$$

在对数幅度谱基础上，计算谱重心位置（记为 feaCenSP）。谱重心定义为功率谱的某一频率，小于该频率的谱能量与大于该频率的谱能量相等。

在实倒谱基础上，计算谱起伏程度参数，记为 RcPr，其采用实倒谱 3~14 个系数绝对值之和与 1~2 系数绝对值之和的比值的对数；（此部分为了计算达到一个分析步长之后的谱起伏度对数方差特征 logVRcPr）

计算 Mel 域子带能量，采用 40 个 Mel 域子带，用三角滤波器组计算每个子带内功率谱能量并取对数，最后对 40 个子带能量进行归一化和零均值化，得到的新矢量记为 spBP；（此部分为了计算达到一个分析步长之后的 4Hz 调制能量特征 f4Hzmelbp）

除了上述几种短时特征之外，还有一个基于实倒谱的重要特征，记为短时调性强度特征 feaRcp，定义如下：

$$feaRcp = \sum_{k=l}^N |rcp(k)| \quad (5)$$

其中 l 是求和起始点，该实施例选择 l=14。

求出短时特征后，还要标记当前帧的调性，如下方法：

```
xr=max ( rcp ( 1:N ) );
```

```
if xr>tonThres
```

 标记为调性；

```
else
```

标记为非调性；

end

其中 tonThres 为调性门限，取 0.14。

3) 达到一个分析步长之后，计算累计调性参数，接着判定是否静默音状态。

分析步长是进行信号类型判决的延时长度。由于从一帧信号中很难区分信号类型，因此必须进行长时特征分析，而长时特征一般在 1 秒钟甚至更长的时间内才具备分类能力。为了满足实时性的要求，该方案采用队列和分析步长相结合的办法，即每个分析步长进行一次类别判定，而判定所依据的长时特征则是在整个队列范围内提取的。这里队列长度取 100 帧，即 1 秒，分析步长为 10 帧，即 100 毫秒。使用队列可以保证长时特征的连续性。

到达一个分析步长后，计算累计调性参数 trc，该参数是队列内所有帧的 feaRcp 之和。

接着进行静音判断。静音判断的依据是分析步长内最大的 feaRMS 参数，具体判断方法如下：

if max (log (feaRMS)) <Thr_sil

静音状态；

else

非静音状态；

end

其中 Thr_sil 是静音检测门限，这里取-3。

4) 对非静音状态，根据短时分析参数计算 7 个长时特征参数。

如果当前分析步长内信号为非静默状态，则开始计算 7 个长时特征参数。

特征一：调性帧平均能量与平均能量之比 $Deng$ 。

根据短时分析中得到的调性/非调性标记结果，对于队列内的所有帧信号， $Deng$ 定义为所有调性帧的平均能量与整个队列内平均能量之比，即

$$Deng = \frac{\text{mean}_{i \in \Theta}(\text{feaRMS}(i))}{\text{mean}_{j \in All}(\text{feaRMS}(j))}, \quad (6)$$

Θ 是所有调性帧，All 是分析区间内所有帧

特征二：计算调性强度特征 $trc = \text{sum}(\text{feaRcp})$ ，其中 $\text{sum}()$ 是对整个缓冲队列内所有帧求和的求和函数。

特征三：feaRMS 方差对数特征 $\log VRMS$ 。

一般来说 feaRMS 对数方差特征就是对队列内所有帧的 feaRMS 求方差再求对数，但是为了避免特征受到信号能量的影响，对该特征进行能量归一化，用一个区间内的平均 feaRMS 来归一化，即

$$\log VRMS = \log(\text{var}(\text{feaRMS}/\text{mean}(\text{feaRMS}))) \quad (7)$$

特征四：谱重心方差对数特征 $\log VCenSpec$ 。

谱重心方差对数就是队列内所有 feaCenSP 的方差的对数。

特征五：谱起伏度对数方差特征 $\log VRcPr$.

谱起伏度对数方差就是队列内所有 $RcPr$ 的方差的对数。

特征六：4Hz 调制能量特征 $f4Hzmelbp$.

4Hz 调制能量是语音/音乐检测中常用的特征之一，一般说来语音的 4Hz 调制能量要大于音乐信号。这里的计算步骤是：在前文 Mel 域子带能量基础上，对于队列内所有帧，计算各子带的 4Hz 调制能量。我们采用了一个 2 阶全极点滤波器，计算滤波器输出能量和原始能量之比，然后将 40 个子带的比值相加再取对数，结果就是 4Hz 调制能量。以第 k 个子带为例，该子带的比值计算如下：

$$R(k) = \frac{\text{sum}(\text{filter}(spBP_k))}{\text{sum}(spBP_k)} \quad (8)$$

其中 $\text{sum}()$ 是求和函数，注意此处不是对一个分析区间内的所有帧求和，而是对整个缓冲队列内所有帧求和。式(8)中的 $\text{filter}()$ 是滤波函数。

特征七：动态范围特征 Dong

动态范围是队列内最大能量和最小能量之比的对数，即

$$Dong = \log \left(\max_{i \in All} (feaRMS(i)) / \min_{j \in All} (feaRMS(j)) \right) \quad (9)$$

计算完上述 7 种特征参数之后，将其组成 7 维特征矢量。为了避免各特征之间量纲差距太大，对各特征还进行了平移和放缩处理，使其数值都分布在相近的范围内。

5) 根据特征参数进行统计分类, 根据贝叶斯极大似然分类法则确定当前分析区间内的声音类别。

该采用的分类模型是 GMM, 分类方法是极大似然分类思想。在 GMM 当中, 高斯混合概率密度是 M 个高斯分量概率密度之和。

图 2 示出了根据本发明实施例的 M 个高斯分量的 GMM 拓扑图, 该方案采用的分类模型是 GMM, 该模型对文中提取的连续分布的特征矢量的分类能力较好, 用公式可以表示为:

$$p(\vec{x}|\lambda) = \sum_{i=1}^M p_i b_i(\vec{x}) \quad (10)$$

其中 \vec{x} 是一个 D 维随机变量, $b_i(\vec{x})$ 是第 i 个高斯分量的概率密度函数, p_i 是第 i 个高斯分量的权重, $i = 1, \dots, M$ 。每个高斯分量的概率密度实际上就是一个 D 维高斯函数:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \quad (11)$$

式中 $\vec{\mu}_i$ 和 Σ_i 表示第 i 个高斯分量的均值和协方差矩阵。所有的高斯分量权重之和满足归一化条件 $\sum_{i=1}^M p_i = 1$ 。

一个高斯混合概率密度可以用三种参数表示: 各分量的权重, 各分量的均值和协方差矩阵。为了后面描述方便, 我们把这些参数放到一起, 统称为模型参数, 记为:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \quad (12)$$

检测中, 每类信号都用一个 GMM 来表示, 或者用模型 λ 表示。

首先对训练数据进行特征分析，得到音乐、语音、噪声三类训练数据的特征矢量集合，然后分别训练三组64个分量的GMM模型，即 MUGmm, SPgmm, NSgmm，对应的模型用 $\lambda_{mu}, \lambda_{sp}, \lambda_{ns}$ 表示。GMM 模型的训练采用的期望值最大化(expectation-maximum, EM)算法，EM 算法的思想是不断地寻找新的模型参数 $\bar{\lambda}$ ，使其对应的似然值比旧的 λ 对应的似然值大，然后用 $\bar{\lambda}$ 替换 λ ，并不断地重复下去，直到达到某一收敛阈值。EM 为经典算法，此处不做赘述。注意全部 GMM 的训练都是事先完成的，在算法一开始时加载。

在进行检测分类时，对于新输入的一个特征矢量 x ，在模型 λ 下的对数似然得分为：

$$Score_{\lambda}(x) = \log(p(x|\lambda)) \quad (13)$$

则对于三个模型而言，最后的分类结论为：

$$CID(x) = \arg \max(Score_{sp}(x) + spS, Score_{ns}(x) + nsS, Score_{mu}(x)) \quad (14)$$

其中 spS 是给语音模型的加分， nsS 是给噪声模型的得分。可以根据不同的应用适当调整。

6) 后处理，消除突变状态。

得到当前分析步长内信号的判决类型之后，还需要进行后处理，以消除突变状态。在实际通信中音乐/非音乐持续时间很少短于 1s，因此在音频类型改变持续 1s 才认为类型改变有效，否则视为突变状态，音频类型维持突变前的状态。

图 3 示出了根据本发明实施例的音乐/非音乐的实时检测装置的方框图，包括：

预处理模块 10，用于对信号进行预处理；

短时特征计算模块 20，用于计算预处理过的信号的短时特征；

静默状态判断模块 30，用于达到一个分析步长之后，计算累计短时能量均方根，以判定信号是否为静默状态；

长时特征计算模块 40，用于如果确定为非静默状态，则根据短时分析参数计算长时特征参数；

确定模块 50，用于根据计算的长时特征参数进行统计分类，根据混合高斯模型确定当前分析区间内的声音类别是音乐/非音乐；以及

后处理模块 60，用于对信号进行后处理，以消除突变状态。

该实施例实现了音乐/非音乐的稳健而有效的实时检测或分割，与语音活动检测相结合，能够组成完整的语音活动检测方案。

检测效果评测

图 4 示出了根据本发明优选实施例的音频传输系统的示意图，该系统可以区分音乐和非音乐，并对音乐信号进行专门的编码，而对非音乐进行 VAD、增强，编码和 DTX 传输。

图 4 中的音乐/非音乐检测方法为该实施例的方法。

1) 测试数据

测试源数据来源由两部分构成，一部分数据来自 Slaney 和 Scheirer 搜集的数据库。该数据库中语音和音乐两个集合各有 20 分钟数据，其中每个集合中由 80 个 15 秒的音频片断组成。音频信号格式属于调频广播级（16 比特量化，单声道，22.05kHz 采样频率），包括了不同的广播站、内容风格、噪声水平等等。Scheier 等人尽力将所能收集到的各种音乐风格和语音风格数据整理到一起。对于语音，说话者包括男性和女性，有会议室的讨论也有电话交谈，有自然的静默也有不同背景噪声；对于音乐，其风格有爵士乐、流行乐、乡村音乐、萨尔萨舞曲、瑞格舞舞曲、交响乐、不同的西部风格音乐、不同类型的摇滚乐、新世纪音乐（new age music）等等，有单纯的音乐，也有带伴唱的。另一部分数据来自实际的彩铃信号和各种实际通信环境中的含噪语音，不仅包括了各种常见噪声，而且信噪比和语音音量也有很大差异。该部分数据包括 25 个音乐文件，共约 17 分钟，58 个含噪语音文件，共约 8.7 分钟。

根据上述的原始数据，我们拼接组成了专门的测试数据集合。数据总共分两个目录：CLN_mix 和 ZX_mix。CLN_mix 目录下包括了 20 个混合文件，每个混合文件由 Slaney 数据库中的两段音乐和两段语音拼接而成；ZX_mix 下包括了 11 个混合文件，每个文件由实测数据集合中的两段音乐和两段语音拼接而成。参照对象是人工标记的分类结果。

2) 测试结果（计算检测准确度（Accuracy）时语音、噪音和静音之间相互误检算准确检测，并把语音、噪音和静音作为非音乐）

不加消除突变

CLN_mix 目录:

Non-music	Music	spS	nsS
93.83%	91.06%	0	0
92.64%	91.96%	1	0
91.10%	92.12%	2	0

ZX_mix 目录

Non-music	Music	spS	nsS
90.36%	91.02%	0	0
91.28%	91.87%	1	0
92.87%	92.65%	2	0

加消除突变

CLN_mix 目录:

Non-music	Music	spS	nsS
95.96%	92.46%	0	0
94.27%	93.03%	1	0
92.06%	94.62%	2	0

ZX_mix 目录

Non-music	Music	spS	nsS
96.64%	92.42%	0	0
95.79%	93.07%	1	0
94.68%	94.09%	2	0

3) 测试结果说明

由于该方案不包括语音部分的噪声 VAD 检测，因此对语音间的停顿（噪声）未做检测，只有长时的静默或噪声才予以检测。测试数据尤其是含噪语音多属上述情况，因此噪声 Noise 检测成 Noise 得分很低，但 Noise 检测成非音乐的准确度达到 90% 以上。

经过对检测错误的音乐进行分析发现，出现检测错误的音乐片断大都是非常嘈杂的摇滚片断，这部分片断单纯用人耳也难以分辨（在一个队列长度内）。尽管如此，语音和音乐的平均检测成功率都在 90% 以上，说明了该方法的有效性。

该实施例实现了音乐/非音乐的稳健而有效的实时检测或分割，与语音活动检测相结合，能够组成完整的语音活动检测方案。

显然，本领域的技术人员应该明白，上述的本发明的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件结合。

以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

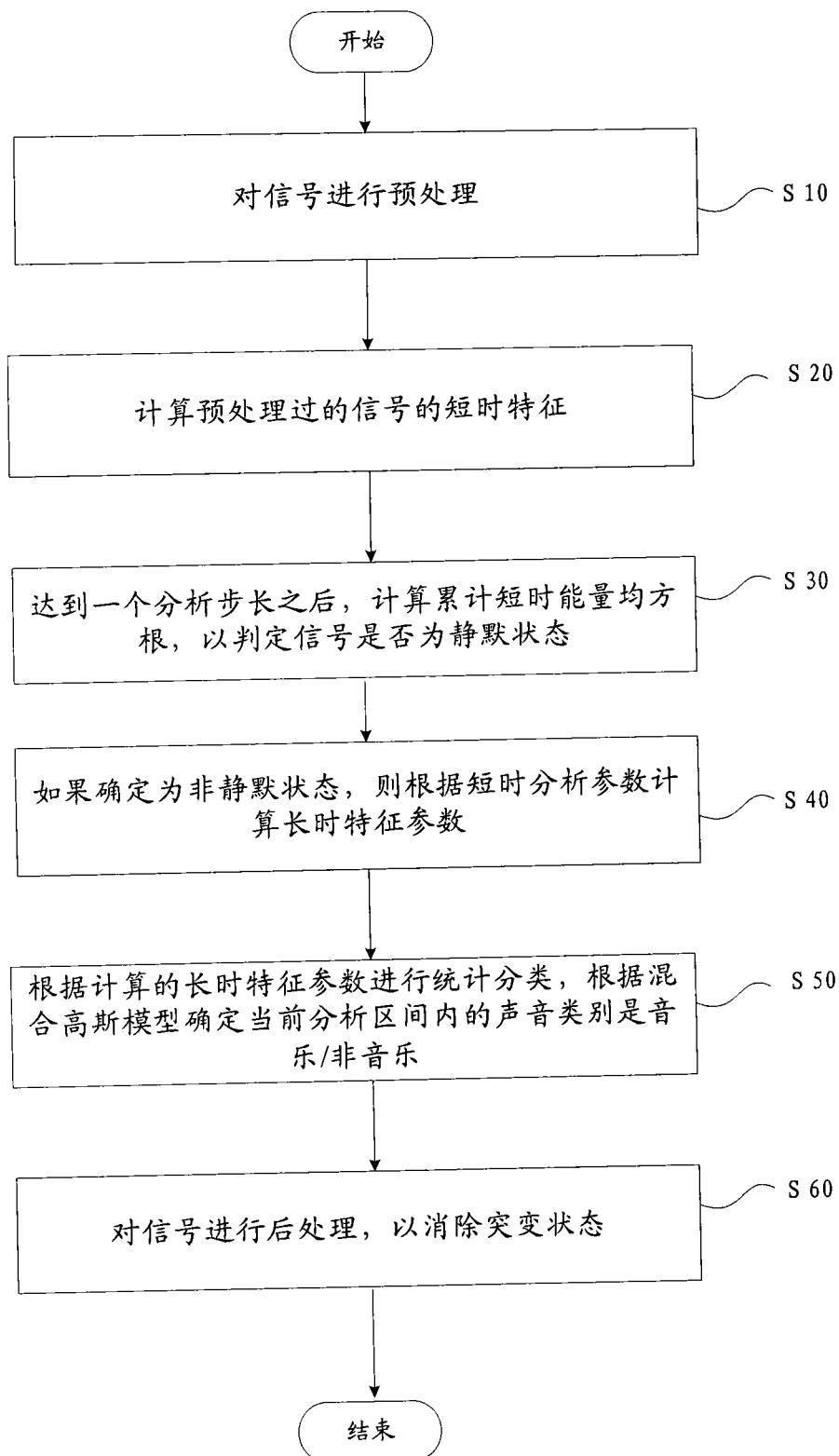


图 1

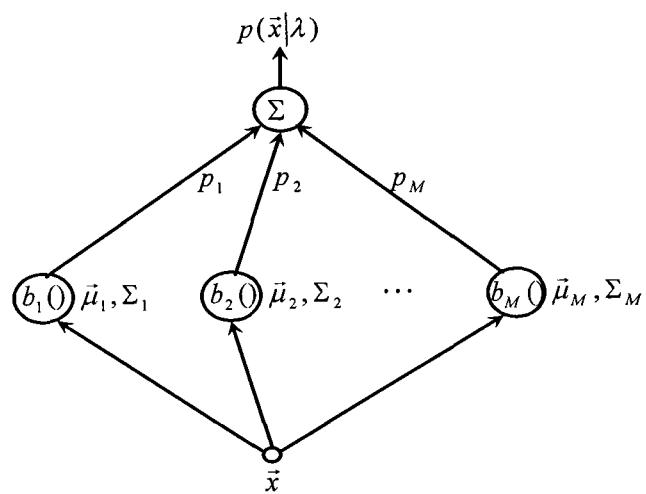


图 2

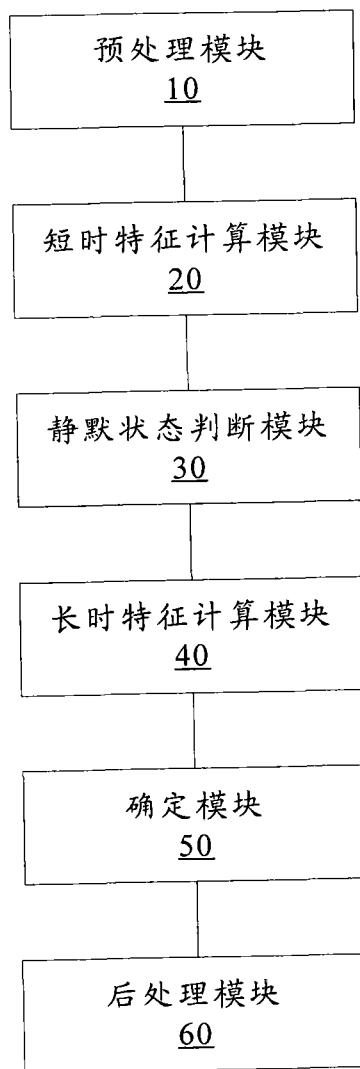


图 3

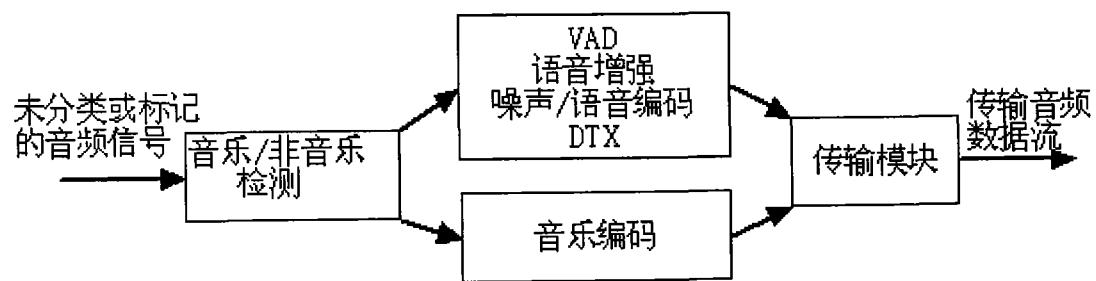


图 4