

一种音频信号分类方法和装置

本申请要求于 2013 年 8 月 6 日提交中国专利局、申请号为 201310339218.5, 发明名称为“一种音频信号分类方法和装置”的中国专利申请优先权, 上述专利的全部内容通过引用结合在本申请中。

5

技术领域

本发明涉及数字信号处理技术领域, 尤其是一种音频信号分类方法和装置。

背景技术

为了降低视频信号存储或者传输过程中占用的资源, 音频信号在发送端进行压缩处理后传输到接收端, 接收端通过解压缩处理恢复音频信号。

在音频处理应用中, 音频信号分类是一种应用广泛而重要的技术。例如, 在音频编解码应用中, 目前比较流行的编解码器是一种混合编解码。这种编解码器通常包含了一个基于语音产生模型的编码器(如 CELP)和一个基于变换的编码器(如基于 MDCT 的编码器)。在中低码率下, 基于语音产生模型的编码器可以获得较好的语音编码质量, 但对音乐的编码质量比较差, 而基于变换的编
15 码器能够获得较好的音乐编码质量, 对语音的编码质量又比较差。因此, 混合编解码器通过对语音信号采用基于语音产生模型的编码器进行编码, 对音乐信号采用基于变换的编码器进行编码, 从而获得整体最佳的编码效果。这里, 一个核心的技术就是音频信号分类, 或具体到这个应用, 就是编码模式选择。

20 混合编解码器需要获得准确的信号类型信息, 才能获得最优的编码模式选择。这里的音频信号分类器也可以被大致认为是一种语音/音乐分类器。语音识别率和音乐识别率是衡量语音/音乐分类器性能的重要指标。尤其对于音乐信号, 由于其信号特征的多样/复杂性, 对音乐信号的识别通常较语音困难。此外, 识别延时也是非常重要的指标之一。由于语音/音乐特征在短时上的模糊性, 通

常需要在一段相对长的时间区间内才能够较准确的识别出语音/音乐来。一般来说，在同一类信号中段时，识别延时越长，识别越准确。但在两类信号的过渡段时，识别延时越长，识别准确率反而降低。这在输入是混合信号（如有背景音乐的语音）的情况下尤为严重。因此，同时兼具高识别率和低识别延时是一个高性能语音/音乐识别器的必要属性。此外，分类的稳定性也是影响到混合编码器编码质量的重要属性。一般来说，混合编码器在不同类型编码器之间切换时会产生质量下降。如果分类器在同一类信号中发生频繁的类型切换，对编码质量的影响是比较大的，这就要求分类器的输出分类结果要准确而平滑。另外，在一些应用中，如通信系统中的分类算法，也要求其计算复杂度和存储开销要

5 个高性能语音/音乐识别器的必要属性。此外，分类的稳定性也是影响到混合编码器编码质量的重要属性。一般来说，混合编码器在不同类型编码器之间切换时会产生质量下降。如果分类器在同一类信号中发生频繁的类型切换，对编码质量的影响是比较大的，这就要求分类器的输出分类结果要准确而平滑。另外，在一些应用中，如通信系统中的分类算法，也要求其计算复杂度和存储开销要

10 尽可能的低，以满足商业需求。

ITU-T 标准 G.720.1 包含有一个语音/音乐分类器。这个分类器以一个主参数，频谱波动方差 `var_flux`，做为信号分类的主要依据，并结合两个不同的频谱峰度参数 `p1`, `p2`，做为辅助依据。根据 `var_flux` 对输入信号的分类，是通过在一个 FIFO 的 `var_flux` buffer 中，根据 `var_flux` 的局部统计量来完成的。

15 具体过程概述如下。首先对每一输入音频帧提取频谱波动 `flux`，并缓存在一个第一 buffer 中，这里的 `flux` 是在包括当前输入帧在内的最新的 4 帧中计算的，也可以有其它的计算方法。然后，计算包括当前输入帧在内的 N 个最新帧的 `flux` 的方差，得到当前输入帧的 `var_flux`，并缓存在第二 buffer 中。然后，统计第二 buffer 中包括当前输入帧在内的 M 个最新帧的 `var_flux` 中大于第一门限值的帧的个数 K。如果 K 与 M 的比值大于一个第二门限值，则判断当前输入帧为语音帧，否则为音乐帧。辅助参数 `p1`, `p2` 主要用于对分类的修正，也是对每一输入音频帧计算的。当 `p1` 和/或 `p2` 大于某第三门限和/或第四门限时，则直接判断当前输入音频帧为音乐帧。

20 的帧的个数 K。如果 K 与 M 的比值大于一个第二门限值，则判断当前输入帧为语音帧，否则为音乐帧。辅助参数 `p1`, `p2` 主要用于对分类的修正，也是对每一输入音频帧计算的。当 `p1` 和/或 `p2` 大于某第三门限和/或第四门限时，则直接判断当前输入音频帧为音乐帧。

这个语音/音乐分类器的缺点一方面对音乐的绝对识别率仍然有待提高，另一方面，由于该分类器的目标应用没有针对混合信号的应用场景，所以对混合信号的识别性能也还有一定的提升空间。

现有的语音/音乐分类器有很多都是基于模式识别原理设计的。这类分类器通常都是对输入音频帧提取多个特征参数（十几到几十不等），并将这些参数馈入一个或者基于高斯混合模型，或者基于神经网络，或者基于其它经典分类方法的分类器来进行分类的。

这类分类器虽然有较高的理论基础，但通常具有较高的计算或存储复杂度，实现成本较高。

10 发明内容

本发明实施例的目的在于提供一种音频信号分类方法和装置，在保证混合音频信号分类识别率的情况下，降低信号分类的复杂度。

第一方面，提供了一种音频信号分类方法，包括：

根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存
15 储于频谱波动存储器中，其中，所述频谱波动表示音频信号的频谱的能量波动；

根据音频帧是否为敲击音乐或历史音频帧的活动性，更新频谱波动存储器中存储的频谱波动；

根据频谱波动存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧。

20 在第一种可能的实现方式中，根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

若当前音频帧为活动帧，则将当前音频帧的频谱波动存储于频谱波动存储器中。

在第二种可能的实现方式中，根据当前音频帧的声音活动性，确定是否获
25 得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

若当前音频帧为活动帧，且当前音频帧不属于能量冲击，则将当前音频帧的频谱波动存储于频谱波动存储器中。

在第三种可能的实现方式中，根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

- 5 若当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中。

结合第一方面或第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式或第一方面的第三种可能的实现方式，在第四种可能的实现方式中，根据所述当前音频帧是否为敲击音乐，更新频谱波动存储器中存储的频谱波动包括：

若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。

结合第一方面或第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式或第一方面的第三种可能的实现方式，在第五种可能的实现方式中，根据所述历史音频帧的活动性，更新频谱波动存储器中存储的频谱波动包括：

如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且前一帧音频帧为非活动帧，则将频谱波动存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据；

- 20 如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且当前音频帧之前连续三帧历史帧不全都为活动帧，则将当前音频帧的频谱波动修正为第一值；

如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且历史分类结果为音乐信号且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

结合第一方面或第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式或第一方面的第三种可能的实现方式或第一方面的第四种可能的实现方式或第一方面的第五种可能的实现方式，在第六种可能的实现方式中，

根据频谱波动存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧包括：

获得频谱波动存储器中存储的频谱波动的部分或全部有效数据的均值；

当所获得的频谱波动的有效数据的均值满足音乐分类条件时，将所述当前

5 音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

结合第一方面或第一方面的第一种可能的实现方式或第一方面的第二种可能的实现方式或第一方面的第三种可能的实现方式或第一方面的第四种可能的实现方式或第一方面的第五种可能的实现方式，在第七种可能的实现方式中，该音频信号分类方法还包括：

10 获得当前音频帧的频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；其中，频谱高频带峰度表示当前音频帧的频谱在高频带上的峰度或能量锐度；频谱相关度表示当前音频帧的信号谐波结构在相邻帧间的稳定度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

15 根据所述当前音频帧的声音活动性，确定是否将所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；

其中，所述根据频谱波动存储器中存储的频谱波动的部分或全部数据的统计量，对所述音频帧进行分类包括：

分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，
20 值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈
25 值。

第二方面，提供了一种音频信号的分类装置，用于对输入的音频信号进行分类，包括：

存储确认单元，用于根据所述当前音频帧的声音活动性，确定是否获得并存储当前音频帧的频谱波动，其中，所述频谱波动表示音频信号的频谱的能量波动；

存储器，用于在存储确认单元输出需要存储的结果时存储所述频谱波动；

- 5 更新单元，用于根据语音帧是否为敲击音乐或历史音频帧的活动性，更新存储器中存储的频谱波动；

分类单元，用于根据存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧。

- 10 在第一种可能的实现方式中，所述存储确认单元具体用于：确认当前音频帧为活动帧时，输出需要存储当前音频帧的频谱波动的结果。

在第二种可能的实现方式中，所述存储确认单元具体用于：确认当前音频帧为活动帧，且当前音频帧不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

- 15 在第三种可能的实现方式中，所述存储确认单元具体用于：确认当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

- 20 结合第二方面或第二方面的第一种可能的实现方式或第二方面的第二种可能的实现方式或第二方面的第三种可能的实现方式，在第四种可能的实现方式中，所述更新单元具体用于若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。

- 25 结合第二方面或第二方面的第一种可能的实现方式或第二方面的第二种可能的实现方式或第二方面的第三种可能的实现方式，在第五种可能的实现方式中，所述更新单元具体用于：如果当前音频帧为活动帧，且前一帧音频帧为非活动帧时，则将存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据；或

如果当前音频帧为活动帧，且当前音频帧之前连续三帧不全都为活动帧时，则将当前音频帧的频谱波动修正为第一值；或

如果当前音频帧为活动帧，且历史分类结果为音乐信号且当前音频帧的频

谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

结合第二方面或第二方面的第一种可能的实现方式或第二方面的第二种可能的实现方式或第二方面的第三种可能的实现方式或第二方面的第四种可能的实现方式或第二方面的第五种可能的实现方式，在第六种可能的实现方式中，
5 所述分类单元包括：

计算单元，用于获得存储器中存储的频谱波动的部分或全部有效数据的均值；

判断单元，用于将所述频谱波动的有效数据的均值与音乐分类条件做比较，
10 当所述频谱波动的有效数据的均值满足音乐分类条件时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

结合第二方面或第二方面的第一种可能的实现方式或第二方面的第二种可能的实现方式或第二方面的第三种可能的实现方式或第二方面的第四种可能的实现方式或第二方面的第五种可能的实现方式，在第七种可能的实现方式中，
15 该音频信号分类装置还包括：

参数获得单元，用于获得当前音频帧的频谱高频带峰度、频谱相关度、浊音度参数和线性预测残差能量倾斜度；其中，频谱高频带峰度表示当前音频帧的频谱在高频带上的峰度或能量锐度；频谱相关度表示当前音频帧的信号谐波结构在相邻帧间的稳定度；浊音度参数表示当前音频帧与一个基音周期之前的
20 信号的时域相关度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

所述存储确认单元还用于，根据所述当前音频帧的声音活动性，确定是否将所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；

25 所述存储单元还用于，当存储确认单元输出需要存储的结果时存储所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

所述分类单元具体用于，分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据

的统计量将所述音频帧分类为语音帧或者音乐帧。

结合第二方面的第七种可能的实现方式，在第八种可能的实现方式中，所述分类单元包括：

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

第三方面，提供了一种音频信号分类方法，包括：

将输入音频信号进行分帧处理；

15 获得当前音频帧的线性预测残差能量倾斜度；所述线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

将线性预测残差能量倾斜度存储到存储器中；

根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类。

20 在第一种可能的实现方式中，将线性预测残差能量倾斜度存储到存储器中之前还包括：

根据所述当前音频帧的声音活动性，确定是否将所述线性预测残差能量倾斜度存储于存储器中；并在确定需要存储时将所述线性预测残差能量倾斜度存储于存储器中。

25 结合第三方面的或第三方面的第一种可能的实现方式，在第二种可能的实现方式中，预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差；所述根据存储器中预测残差能量倾斜度部分数据的统计量，对所

述音频帧进行分类包括:

将预测残差能量倾斜度部分数据的方差与音乐分类阈值相比较, 当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时, 将所述当前音频帧分类为音乐帧; 否则将所述当前音频帧分类为语音帧。

- 5 结合第三方面的或第三方面的第一种可能的实现方式, 在第三种可能的实现方式中, 该音频信号分类方法还包括:

获得当前音频帧的频谱波动、频谱高频带峰度和频谱相关度, 并存储于对应的存储器中;

- 其中, 所述根据存储器中预测残差能量倾斜度部分数据的统计量, 对所述
10 音频帧进行分类包括:

分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量, 根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧; 所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值。

- 15 结合第三方面的第三种可能的实现方式, 在第四种可能的实现方式中, 分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量, 根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧包括:

- 分别获得存储的频谱波动有效数据的均值, 频谱高频带峰度有效数据的均
20 值, 频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差;

- 当下列条件之一满足时, 将所述当前音频帧分类为音乐帧, 否则将所述当前音频帧分类为语音帧: 所述频谱波动有效数据的均值小于第一阈值; 或者频谱高频带峰度有效数据的均值大于第二阈值; 或者所述频谱相关度有效数据的均值大于第三阈值; 或者线性预测残差能量倾斜度有效数据的方差小于第四阈
25 值。

结合第三方面的或第三方面的第一种可能的实现方式, 在第五种可能的实现方式中, 该音频信号分类方法还包括:

获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率, 并存

储于对应的存储器;

其中, 所述根据存储器中预测残差能量倾斜度部分数据的统计量, 对所述音频帧进行分类包括:

分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量;

根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率, 将所述音频帧分类为语音帧或者音乐帧; 所述统计量指对存储器中存储的数据运算操作后获得的数据值。

结合第三方面的第五种可能的实现方式, 在第六种可能的实现方式中, 分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量包括:

获得存储的线性预测残差能量倾斜度的方差;

获得存储的频谱音调个数的均值;

根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率, 将所述音频帧分类为语音帧或者音乐帧包括:

当当前音频帧为活动帧, 且满足下列条件之一, 则将所述当前音频帧分类为音乐帧, 否则将所述当前音频帧分类为语音帧:

线性预测残差能量倾斜度的方差小于第五阈值; 或

频谱音调个数的均值大于第六阈值; 或

频谱音调个数在低频带上的比率小于第七阈值。

结合第三方面或第三方面的第一种可能的实现方式或第三方面的第二种可能的实现方式或第三方面的第三种可能的实现方式或第三方面的第四种可能的实现方式或第三方面的第五种可能的实现方式或第三方面的第六种可能的实现方式, 在第七种可能的实现方式中, 获得当前音频帧的线性预测残差能量倾斜度包括:

根据下列公式计算当前音频帧的线性预测残差能量倾斜度:

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

其中, $epsP(i)$ 表示当前音频帧第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。

结合第三方面的第五种可能的实现方式或第三方面的第六种可能的实现方式, 在第八种可能的实现方式中, 获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率包括:

统计当前音频帧在 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量作为频谱音调个数;

计算当前音频帧在 $0 \sim 4\text{kHz}$ 频带上频点峰值大于预定值的频点数量与 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量的比值, 作为频谱音调个数在低频带上的比率。

第四方面, 提供一种信号分类装置, 用于对输入的音频信号进行分类, 其包括:

15 分帧单元, 用于对输入音频信号进行分帧处理;

参数获得单元, 用于获得当前音频帧的线性预测残差能量倾斜度; 所述线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度;

存储单元, 用于存储线性预测残差能量倾斜度;

20 分类单元, 用于根据存储器中预测残差能量倾斜度部分数据的统计量, 对所述音频帧进行分类。

在第一种可能的实现方式中, 信号分类装置还包括:

存储确认单元, 用于根据所述当前音频帧的声音活动性, 确定是否将所述线性预测残差能量倾斜度存储于存储器中;

25 所述存储单元具体用于, 当存储确认单元确认需要确定需要存储时将所

述线性预测残差能量倾斜度存储于存储器中。

结合第四方面的或第四方面的第一种可能的实现方式，在第二种可能的实现方式中，预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差；

- 5 所述分类单元具体用于将预测残差能量倾斜度部分数据的方差与音乐分类阈值相比较，当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

- 结合第四方面的或第四方面的第一种可能的实现方式，在第三种可能的实现方式中，参数获得单元还用于：获得当前音频帧的频谱波动、频谱高频带峰度10 度和频谱相关度，并存储于对应的存储器中；

所述分类单元具体用于：分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值。

- 15 第四方面的第三种可能的实现方式，在第四种可能的实现方式中，所述分类单元包括：

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

- 20 判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

- 25 结合第四方面的或第四方面的第一种可能的实现方式，在第五种可能的实现方式中，所述参数获得单元还用于：获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率，并存储于存储器；

所述分类单元具体用于：分别获得存储的线性预测残差能量倾斜度的统计

量、频谱音调个数的统计量；根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的数据运算操作后获得的数据值。

- 5 第四方面的第五种可能的实现方式，在第六种可能的实现方式中，所述分类单元包括：

计算单元，用于获得线性预测残差能量倾斜度有效数据的方差和存储的频谱音调个数的均值；

- 10 判断单元，用于当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：线性预测残差能量倾斜度的方差小于第五阈值；或频谱音调个数的均值大于第六阈值；或频谱音调个数在低频带上的比率小于第七阈值。

- 15 结合第四方面或第四方面的第一种可能的实现方式或第四方面的第二种可能的实现方式或第四方面的第三种可能的实现方式或第四方面的第四种可能的实现方式或第四方面的第五种可能的实现方式或第四方面的第六种可能的实现方式，在第七种可能的实现方式中，所述参数获得单元根据下列公式计算当前音频帧的线性预测残差能量倾斜度：

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

- 20 其中，epsP(i)表示当前音频帧第 i 阶线性预测的预测残差能量；n 为正整数，表示线性预测的阶数，其小于等于线性预测的最大阶数。

- 结合第四方面的第五种可能的实现方式或第四方面的第六种可能的实现方式，在第八种可能的实现方式中，所述参数获得单元用于统计当前音频帧在 0~8kHz 频带上频点峰值大于预定值的频点数量作为频谱音调个数；所述参数获得单元用于计算当前音频帧在 0~4kHz 频带上频点峰值大于预定值的频点数量与 25 0~8kHz 频带上频点峰值大于预定值的频点数量的比值，作为频谱音调个数在低

频带上的比率。

本发明实施例根据频谱波动的长时统计量对音频信号进行分类，参数较少，识别率较高且复杂度较低；同时考虑声音活动性和敲击音乐的因素对频谱波动
5 进行调整，对音乐信号识别率更高，适合混合音频信号分类。

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。
10

图 1 为对音频信号分帧的示意图；

图 2 为本发明提供的音频信号分类方法的一个实施例的流程示意图；

图 3 为本发明提供的获得频谱波动的一个实施例的流程示意图；

图 4 为本发明提供的音频信号分类方法的另一个实施例的流程示意图；

15 图 5 为本发明提供的音频信号分类方法的另一个实施例的流程示意图；

图 6 为本发明提供的音频信号分类方法的另一个实施例的流程示意图；

图 7 至图 10 为本发明提供的音频信号分类的一种具体分类流程图；

图 11 为本发明提供的音频信号分类方法的另一个实施例的流程示意图；

图 12 为本发明提供的音频信号分类的一种具体分类流程图；

20 图 13 为本发明提供的音频信号的分类装置一个实施例的结构示意图；

图 14 为本发明提供的分类单元一个实施例的结构示意图；

图 15 为本发明提供的音频信号的分类装置另一个实施例的结构示意图；

图 16 为本发明提供的音频信号的分类装置另一个实施例的结构示意图；

图 17 为本发明提供的分类单元一个实施例的结构示意图；

图 18 为本发明提供的音频信号的分类装置另一个实施例的结构示意图；

图 19 为本发明提供的音频信号的分类装置另一个实施例的结构示意图。

具体实施方式

5 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

数字信号处理领域，音频编解码器、视频编解码器广泛应用于各种电子设备中，例如：移动电话，无线装置，个人数据助理（PDA），手持式或便携式计算机，GPS 接收机/导航器，照相机，音频/视频播放器，摄像机，录像机，监控设备等。通常，这类电子设备中包括音频编码器或音频解码器，音频编码器或者解码器可以直接由数字电路或芯片例如 DSP（digital signal processor）实现，或者由软件代码驱动处理器执行软件代码中的流程而实现。一种音频编码
10 器中，首先对音频信号进行分类，对不同类型的音频信号采用不同的编码模式
15 进行编码后，再将编码后码流传输给解码端。

一般的，音频信号在处理时采用分帧的方式，每一帧信号代表一定时长的音频信号。参考图 1，当前输入的需要分类的音频帧可以称为当前音频帧；当前
20 音频帧之前的任意一帧音频帧可以称为历史音频帧；按照从当前音频帧到历史音频帧的时序顺序，历史音频帧可以依次成为前一音频帧，前第二帧音频帧，前第三帧音频帧，前第 N 帧音频帧，N 大于等于四。

本实施例中，输入音频信号为 16kHz 采样的宽带音频信号，输入音频信号

以 20ms 为一帧进行分帧，即每帧 320 个时域样点。在提取特征参数前，输入音频信号帧首先经降采样为 12.8kHz 采样率，即 256 采样点每帧。后文中的输入音频信号帧均指降采样后的音频信号帧。

参考图 2，一种音频信号分类方法的一个实施例包括：

- 5 S101：将输入音频信号进行分帧处理，根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中，其中，频谱波动表示音频信号的频谱的能量波动；

10 音频信号分类一般按帧进行，对每个音频信号帧提取参数进行分类，以确定该音频信号帧属于语音帧还是音乐帧，以采用对应的编码模式进行编码。一个实施例中，可以在音频信号进行分帧处理后，获得当前音频帧的频谱波动，再根据当前音频帧的声音活动性，确定是否将该频谱波动存储于频谱波动存储器中；另一个实施例中，可以在音频信号进行分帧处理后，根据当前音频帧的声音活动性，确定是否将该频谱波动存储于频谱波动存储器中，在需要存储时再获得该频谱波动并存储。

- 15 频谱波动 flux 表示信号频谱的短时或长时能量波动，为当前音频帧与历史帧在中低频带频谱上对应频率的对数能量差的绝对值的均值；其中历史帧指当前音频帧之前的任意一帧。一个实施例中，频谱波动为当前音频帧与其历史帧在中低频带频谱上对应频率的对数能量差的绝对值的均值。另一个实施例中，频谱波动为当前音频帧与历史帧在中低频带频谱上对应频谱峰值的对数能量差
20 的绝对值的均值。

参考图 3，获得频谱波动的一个实施例包括如下步骤：

S1011：获得当前音频帧的频谱；

25 一个实施例中，可以直接获得音频帧的频谱；另一个实施例中，获得当前音频帧任意两个子帧的频谱，即能量谱；利用两个子帧的频谱的平均值得到当前音频帧的频谱；

S1012：获得当前音频帧历史帧的频谱；

其中历史帧指当前音频帧之前的任意一帧音频帧；一个实施例中可以为当前音频帧之前的第三帧音频帧。

S1013: 计算当前音频帧与历史帧分别在中低频带频谱上对应频率的对数能量差的绝对值的均值, 作为当前音频帧的频谱波动。

一个实施例中, 可以计算当前音频帧在中低频带频谱上所有频点的对数能量与历史帧在中低频带频谱上对应频点的对数能量之间差值的绝对值的均值;

- 5 另一个实施例中, 可以计算当前音频帧在中低频带频谱上频谱峰值的对数能量与历史帧在中低频带频谱上对应频谱峰值的对数能量之间差值的绝对值的均值。

中低频带频谱, 例如 $0 \sim f_s/4$, 或者 $0 \sim f_s/3$ 的频谱范围。

- 10 以输入音频信号为 16kHz 采样的宽带音频信号, 输入音频信号以 20ms 为一帧为例, 对每 20ms 当前音频帧分别做前后两个 256 点的 FFT, 两个 FFT 窗 50% 重叠, 得到当前音频帧两个子帧的频谱 (能量谱), 分别记做 $C^0(i)$, $C^1(i)$, $i=0, 1 \dots 127$, 其中 $C^x(i)$ 表示第 x 个子帧的频谱。当前音频帧第 1 子帧的 FFT 需要用到前一帧第 2 子帧的数据。

$$C^x(i) = \text{rel}^2(i) + \text{img}^2(i)$$

- 15 其中, $\text{rel}(i)$ 和 $\text{img}(i)$ 分别表示第 i 频点 FFT 系数的实部和虚部。当前音频帧的频谱 $C(i)$ 则由两个子帧的频谱平均得到。

$$C(i) = \frac{1}{2}(C^0(i) + C^1(i))$$

- 一个实施例中, 当前音频帧的频谱波动 flux 为当前音频帧与其 60ms 前的帧在中低频带频谱上对应频率的对数能量差的绝对值的均值, 在另一实施例中也可
20 可为不同于 60ms 的间隔。

$$\text{flux} = \frac{1}{42} \sum_{i=0}^{42} [10 \log(C(i)) - 10 \log(C_{-3}(i))]$$

其中 $C_{-3}(i)$ 表示当前当前音频帧之前的第三历史帧, 即在本实施例中当帧长为 20ms 时, 当前音频帧 60ms 以前的历史帧的频谱。在本文中类似 $X_{-n}()$ 的形式,

均表示当前音频帧的第 n 历史帧的参数 X ，当前音频帧可省略下角标 0 。 $\log(\cdot)$ 表示以 10 为底的对数。

在另一个实施例中，当前音频帧的频谱波动 flux 也可由下述方法得到，即，为当前音频帧与其 60ms 前的帧在中低频带频谱上对应频谱峰值的对数能量差的绝对值的均值，

$$\text{flux} = \frac{1}{K} \sum_{i=0}^K [10 \log(P(i)) - 10 \log(P_{-3}(i))]$$

其中 $P(i)$ 表示当前音频帧的频谱的第 i 个局部峰值能量，局部峰值所在的频点即为频谱上能量高于高低两相邻频点上能量的频点。 K 表示中低频带频谱上局部峰值的个数。

10 其中，根据当前音频帧的声音活动性，确定是否将该频谱波动存储于频谱波动存储器中，可以用多种方式实现：

一个实施例中，若音频帧的声音活动性参数表示音频帧为活动帧，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。

另一个实施例中，根据音频帧的声音活动性和音频帧是否为能量冲击，确定是否将所述频谱波动存储于存储器中。若音频帧的声音活动性参数表示音频帧为活动帧，且表示音频帧是否为能量冲击的参数表示音频帧不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储；另一个实施例中，若当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。例如，当前音频帧为活动帧，且当前音频帧、前一帧音频帧和前第二帧音频帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。

声音活动性标识 vad_flag 表示当前输入信号是活动的前景信号（语音，音

乐等)还是前景信号静默的背景信号(如背景噪声,静音等),由声音活动性检测器 VAD 获得。vad_flag=1 表示输入信号帧为活动帧,即前景信号帧,反之 vad_flag=0 表示背景信号帧。由于 VAD 不属本发明的发明内容,VAD 的具体算法在此不再详述。

5 声音冲击标识 attack_flag 表示当前当前音频帧是否属于音乐中的一个能量冲击。当当前音频帧之前的若干历史帧以音乐帧为主时,若当前音频帧的帧能量较其之前第一历史帧有较大跃升,且较其之前一段时间内音频帧的平均能量有较大跃升,且当前音频帧的时域包络较其之前一段时间内音频帧的平均包络也有较大跃升时,则认为当前当前音频帧属于音乐中的能量冲击。

10 根据所述当前音频帧的声音活动性,当当前音频帧为活动帧时,才存储当前音频帧的频谱波动;能够降低非活动帧的误判率,提高音频分类的识别率。

当如下条件满足时,attack_flag 置 1,即表示当前当前音频帧为一个音乐中的能量冲击:

$$\begin{cases} etot - etot_{-1} > 6 \\ etot - lp_speech > 5 \\ mode_mov > 0.9 \\ \log_max_spl - mov_log_max_spl > 5 \end{cases}$$

15 其中,etot 表示当前音频帧的对数帧能量;etot₋₁ 表示前一音频帧的对数帧能量;lp_speech 表示对数帧能量 etot 的长时滑动平均;log_max_spl 和 mov_log_max_spl 分别表示当前音频帧的时域最大对数样点幅度及其长时滑动平均;mode_mov 表示信号分类中历史最终分类结果的长时滑动平均。

20 以上公式的含义是,当当前音频帧之前的若干历史帧以音乐帧为主时,若当前音频帧的帧能量较其之前第一历史帧有较大跃升,且较其之前一段时间内音频帧的平均能量有较大跃升,且当前音频帧的时域包络较其之前一段时间内音频帧的平均包络也有较大跃升时,则认为当前当前音频帧属于音乐中的能量冲击。

对数帧能量 etot,由输入音频帧的对数总子带能量表示:

$$etot = 10 \log \left(\sum_{j=0}^{19} \left[\frac{1}{hb(i) - lb(i) + 1} \cdot \sum_{i=lb(j)}^{hb(j)} C(i) \right] \right)$$

其中, $hb(j), lb(j)$ 分别表示输入音频帧频谱中第 j 子带的高低频边界; $C(i)$ 表示输入音频帧的频谱。

当前音频帧的时域最大对数样点幅度的长时滑动平均 $mov_log_max_spl$ 只在活动声音帧中更新:

$$5 \quad \begin{aligned} & mov_log_max_spl = \\ & \begin{cases} 0.95 \cdot mov_log_max_spl_{-1} + 0.05 \cdot log_max_spl & log_max_spl > mov_log_max_spl_{-1} \\ 0.995 \cdot mov_log_max_spl_{-1} + 0.005 \cdot log_max_spl & log_max_spl \leq mov_log_max_spl_{-1} \end{cases} \end{aligned}$$

一个实施例中, 当前音频帧的频谱波动 $flux$ 被缓存在一个 FIFO 的 $flux$ 历史 buffer 中, 本实施例中 $flux$ 历史 buffer 的长度为 60 (60 帧)。判断当前音频帧的声音活动性和音频帧是否为能量冲击, 当当前音频帧为前景信号帧且当前音频帧及其之前的两帧均未出现属于音乐的能量冲击, 则将当前音频帧的频谱波动

10 $flux$ 存储于存储器中。

在缓存当前当前音频帧的 $flux$ 之前, 检查是否满足如下条件:

$$\begin{cases} vad_flag \neq 0 \\ attack_flag \neq 1 \\ attack_flag_{-1} \neq 1 \\ attack_flag_{-2} \neq 1 \end{cases}$$

若满足, 则缓存, 否则不缓存。

其中, vad_flag 表示当前输入信号是活动的前景信号还是前景信号静默的背景信号, $vad_flag=0$ 表示背景信号帧; $attack_flag$ 表示当前当前音频帧是否属于音乐中的一个能量冲击, $attack_flag=1$ 表示当前当前音频帧为一个音乐中的能量冲击。

15

上述公式的含义为: 当前音频帧为活动帧, 且当前音频帧、前一帧音频帧和前第二帧音频帧均不属于能量冲击。

20 S102: 根据音频帧是否为敲击音乐或历史音频帧的活动性, 更新频谱波动存储器中存储的频谱波动;

一个实施例中, 若表示音频帧是否属于敲击音乐的参数表示当前音频帧属于敲击音乐, 则修改频谱波动存储器中存储的频谱波动的值, 将频谱波动存储

器中有效的频谱波动值修改为小于等于音乐阈值的一个值，其中当音频帧的频谱波动小于该音乐阈值时该音频被分类为音乐帧。一个实施例中，将有效的频谱波动值重置为 5。即当敲击声响标识 `percus_flag` 被置为 1 时，`flux` 历史 `buffer` 中所有的有效缓冲数据均被重置为 5。这里，有效缓冲数据等价于有效频谱波动值。一般的，音乐帧的频谱波动值较低，而语音帧的频谱波动值较高。当音频帧属于敲击音乐时，将有效的频谱波动值修改为小于等于音乐阈值的一个值，则能提高该音频帧被分类为音乐帧的概率，从而提高音频信号分类的准确率。

另一个实施例中，根据当前音频帧的历史帧的活动性，更新存储器中的频谱波动。具体的，一个实施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且前一帧音频帧为非活动帧，则将频谱波动存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据。前一帧音频帧为非活动帧而当前音频帧为活动帧时，当前音频帧与历史帧的语音活动性不同，将历史帧的频谱波动无效化，则能降低历史帧对音频分类的影响，从而提高音频信号分类的准确率。

另一个实施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且当前音频帧之前连续三帧不全都为活动帧，则将当前音频帧的频谱波动修正为第一值。第一值可以为语音阈值，其中当音频帧的频谱波动大于该语音阈值时该音频被分类为语音帧。另一个实施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且历史帧的分类结果为音乐帧且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

如果当前音频帧的 `flux` 被缓存，且前一帧音频帧为非活动帧 (`vad_flag=0`)，则除被新缓存入 `flux` 历史 `buffer` 的当前音频帧 `flux` 以外，其余 `flux` 历史 `buffer` 中的数据全部重置为 -1 (等价于将这些数据无效化)。

如果 `flux` 被缓存入 `flux` 历史 `buffer`，且当前音频帧之前连续三帧不全都为活动帧 (`vad_flag=1`)，则将刚缓存入 `flux` 历史 `buffer` 的当前音频帧 `flux` 修正为 16，即是否满足如下条件：

$$\begin{cases} vad_flag_{-1} = 1 \\ vad_flag_{-2} = 1 \text{ 若不满足, 则将刚缓存入 flux 历史 buffer 的当前音频帧 flux} \\ vad_flag_{-3} = 1 \end{cases}$$

修正为 16;

如果当前音频帧之前连续三帧都为活动帧 ($vad_flag=1$), 则检查是否满足如下条件:

$$\begin{cases} mode_mov > 0.9 \\ flux > 20 \end{cases}$$

若满足, 则将刚缓存入 flux 历史 buffer 的当前音频帧 flux 修正为 20, 否则不做操作。

其中, $mode_mov$ 表示信号分类中历史最终分类结果的长时滑动平均; $mode_mov > 0.9$ 表示信号处于音乐信号中, 根据音频信号的历史分类结果将 flux 进行限制, 以降低 flux 出现语音特征的概率, 目的是提高判断分类的稳定性。

当当前音频帧之前连续三帧历史帧都为非活动帧, 当前音频帧为活动帧时, 或当前音频帧之前连续三帧不全都为活动帧, 当前音频帧为活动帧时, 此时处于分类的初始化阶段。在一个实施例中为了使分类结果倾向于语音 (音乐), 可以将当前音频帧的频谱波动修改为语音 (音乐) 阈值或接近于语音 (音乐) 阈值的数值。在另一个实施例中, 如果当前信号之前的信号是语音 (音乐) 信号, 则可以将当前音频帧的频谱波动修改为语音 (音乐) 阈值或接近于语音 (音乐) 阈值的数值以提高判断分类的稳定性。在另一个实施例中, 为了使分类结果倾向于音乐, 可以对频谱波动进行限制, 即可以修改当前音频帧的频谱波动使其不大于一个阈值, 以降低频谱波动判定为语音特征的概率。

敲击声响标识 $percus_flag$ 表示音频帧中是否有敲击声响存在。 $percus_flag$ 置 1 表示检测到敲击声响, 置 0 则表示没有检测到敲击声响。

当当前信号 (即包括当前音频帧和其若干历史帧在内的若干最新的信号帧) 在短时和长时均出现较尖锐的能量突起, 且当前信号不具有明显的浊音特征时, 若当前音频帧之前的若干历史帧以音乐帧为主, 则认为当前信号是一个敲击音乐; 否则, 进一步的若当前信号的每个子帧均不具有明显的浊音特征且当前信

号的时域包络较其长时平均也出现较明显的跃升变化时, 则也认为当前信号是一个敲击音乐。

敲击声响标识 `percus_flag` 通过如下步骤获得:

首先获得输入音频帧的对数帧能量 `etot`, 由输入音频帧的对数总子带能量表

5 示:

$$etot = 10 \log \left(\sum_{j=0}^{19} \left[\frac{1}{hb(i) - lb(i) + 1} \cdot \sum_{i=lb(j)}^{hb(j)} C(i) \right] \right)$$

其中, $hb(j), lb(j)$ 分别表示输入帧频谱第 j 子带的高低频边界, $C(i)$ 表示输入音频帧的频谱。

当满足如下条件时, `percus_flag` 置 1, 否则置 0。

$$10 \left\{ \begin{array}{l} etot_{-2} - etot_{-3} > 6 \\ etot_{-2} - etot_{-1} > 0 \\ etot_{-2} - etot > 3 \\ etot_{-1} - etot > 0 \\ etot_{-2} - lp_speech > 3 \\ 0.5 \cdot voicing_{-1}(1) + 0.25 \cdot voicing(0) + 0.25 \cdot voicing(1) < 0.75 \\ mode_mov > 0.9 \end{array} \right.$$

或

$$\left\{ \begin{array}{l} etot_{-2} - etot_{-3} > 6 \\ etot_{-2} - etot_{-1} > 0 \\ etot_{-2} - etot > 3 \\ etot_{-1} - etot > 0 \\ etot_{-2} - lp_speech > 3 \\ 0.5 \cdot voicing_{-1}(1) + 0.25 \cdot voicing(0) + 0.25 \cdot voicing(1) < 0.75 \\ voicing_{-1}(0) < 0.8 \\ voicing_{-1}(1) < 0.8 \\ voicing(0) < 0.8 \\ \log_max_spl_{-2} - mov_log_max_spl_{-2} > 10 \end{array} \right.$$

其中, $etot$ 表示当前音频帧的对数帧能量; lp_speech 表示对数帧能量 $etot$ 的长时滑动平均; $voicing(0), voicing_{-1}(0), voicing_{-1}(1)$ 分别表示当前输入音频帧第

一子帧和第一历史帧的第一，第二子帧的归一化开环基音相关度，浊音度参数 voicing 是通过线性预测分析得到的，代表当前音频帧与一个基音周期之前的信号的时域相关度，取值在 0~1 之间；mode_mov 表示信号分类中历史最终分类结果的长时滑动平均；log_max_spl₂ 和 mov_log_max_spl₂ 分别表示第二历史帧 5 的时域最大对数样点幅度，及其长时滑动平均。lp_speech 在每一活动声音帧中进行更新（即 vad_flag=1 的帧），其更新方法为：

$$lp_speech = 0.99 \cdot lp_speech_{-1} + 0.01 \cdot etot$$

以上两式的含义为：当当前信号（即包括当前音频帧和其若干历史帧在内的若干最新的信号帧）在短时和长时均出现较尖锐的能量突起，且当前信号不 10 具有明显的浊音特征时，若当前音频帧之前的若干历史帧以音乐帧为主，则认为当前信号是一个敲击音乐，否则进一步的若当前信号的每个子帧均不具有明显的浊音特征且当前信号的时域包络较其长时平均也出现较明显的跃升变化时，则也认为当前信号是一个敲击音乐。

浊音度参数 voicing，即归一化开环基音相关度，表示当前音频帧与一个基 15 音周期之前的信号的时域相关度，可以由 ACELP 的开环基音搜索中获得，取值在 0~1 之间。由于属现有技术，本发明不做详述。本实施例中当前音频帧的两个子帧各计算一个 voicing，求平均得到当前音频帧的 voicing 参数。当前音频帧的 voicing 参数也被缓存在一个 voicing 历史 buffer 中，本实施例中 voicing 历史 buffer 的长度为 10。

20 mode_mov 在每一活动声音帧且在该帧之前已出现连续 30 帧以上的声音活动帧时进行更新，更新方法为：

$$mode_mov = 0.95 \cdot move_mov_{-1} + 0.05 \cdot mode$$

其中 mode 为当前输入音频帧的分类结果，二元取值，“0”表示语音类别，“1”表示音乐类别。

25 S103: 根据频谱波动存储器中存储的频谱波动的部分或全部数据的统计量，将该当前音频帧分类为语音帧或者音乐帧。当频谱波动的有效数据的统计量满足语音分类条件时，将所述当前音频帧分类为语音帧；当频谱波动的有效数据的统计量满足音乐分类条件时，将所述当前音频帧分类为音乐帧。

此处的统计量为频谱波动存储器中存储的有效的频谱波动（即有效数据）做统计操作得到的值，例如统计操作可以为平均值或者方差。下面实施例中的统计量具有类似的含义。

一个实施例中，步骤 S103 包括：

- 5 获得频谱波动存储器中存储的频谱波动的部分或全部有效数据的均值；

当所获得的频谱波动的有效数据的均值满足音乐分类条件时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

例如，当所获得的频谱波动的有效数据的均值小于音乐分类阈值时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

- 10 一般的，音乐帧的频谱波动值较小，而语音帧的频谱波动值较大。因此可以根据频谱波动对当前音频帧进行分类。当然还可以采用其他分类方法对该当前音频帧进行信号分类。例如，统计频谱波动存储器中存储的频谱波动的有效数据的数量；根据该有效数据的数量，将频谱波动存储器由近端到远端划分出至少两个不同长度的区间，获得每个区间对应的频谱波动的有效数据的均值；
- 15 其中，所述区间的起点为当前帧频谱波动存储位置，近端为存储有当前帧频谱波动的一端，远端为存储有历史帧频谱波动的一端；根据较短区间内的频谱波动统计量对所述音频帧进行分类，若此区间内的参数统计量足够区分出所述音频帧的类型则分类过程结束，否则在其余较长区间中最短的区间内继续分类过程，并以此类推。在每个区间的分类过程中，根据每一个区间对应的分类阈值，
- 20 对所述当前音频帧进行分类，将所述当前音频帧分类为语音帧或者音乐帧，当频谱波动的有效数据的统计量满足语音分类条件时，将所述当前音频帧分类为语音帧；当频谱波动的有效数据的统计量满足音乐分类条件时，将所述当前音频帧分类为音乐帧。

- 在信号分类后，可以对不同的信号采用不同的编码模式进行编码。例如，
- 25 语音信号采用基于语音产生模型的编码器（如 CELP）进行编码，对音乐信号采用基于变换的编码器（如基于 MDCT 的编码器）进行编码。

上述实施例，由于根据频谱波动的长时统计量对音频信号进行分类，参数较少，识别率较高且复杂度较低；同时考虑声音活动性和敲击音乐的因素对频

谱波动进行调整，对音乐信号识别率更高，适合混合音频信号分类。

参考图 4，另一个实施例中，在步骤 S102 之后还包括：

5 S104：获得当前音频帧的频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度，将所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；频谱高频带峰度表示当前音频帧频谱在高频带上的峰度或能量锐度；频谱相关度表示信号谐波结构在相邻帧间的稳定度；线性预测残差能量倾斜度表示线性预测残差能量倾斜度表示输入音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

10 可选的，在存储这些参数之前，还包括：根据所述当前音频帧的声音活动性，确定是否将频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；如果当前音频帧为活动帧，则存储上述参数；否则不存储。

频谱高频带峰度表示当前音频帧频谱在高频带上的峰度或能量锐度；一个实施例中，通过下列公式计算频谱高频带峰度 ph ：

$$15 \quad ph = \sum_{i=64}^{126} p2v_map(i)$$

其中 $p2v_map(i)$ 表示频谱第 i 个频点的峰度，峰度 $p2v_map(i)$ 由下式得到

$$p2v_map(i) = \begin{cases} 20 \log(peak(i)) - 10 \log(vl(i)) - 10 \log(vr(i)) & peak(i) \neq 0 \\ 0 & peak(i) = 0 \end{cases}$$

其中 $peak(i) = C(i)$ ，如果第 i 频点是频谱的局部峰值，否则 $peak(i)=0$ 。 $vl(i)$ 和 $vr(i)$ 分别表示第 i 个频点的高频侧和低频侧与之最临近的频谱局部谷值 $v(n)$ 。

$$20 \quad peak(i) = \begin{cases} C(i) & C(i) > C(i-1), C(i) > C(i+1) \\ 0 & else \end{cases}$$

$$v = \forall C(i) \quad C(i) < C(i-1), C(i) < C(i+1)$$

当前音频帧的频谱高频带峰度 ph 也被缓存在一个 ph 历史 buffer 中，本实施例中 ph 历史 buffer 的长度为 60。

频谱相关度 cor_map_sum 表示信号谐波结构在相邻帧间的稳定度，其通过以下步骤获得：

首先获得输入音频帧 $C(i)$ 的去底频谱 $C'(i)$ 。

$$C'(i) = C(i) - \text{floor}(i)$$

5 其中， $\text{floor}(i)$, $i=0,1,\dots,127$ ，表示输入音频帧频谱的谱底。

$$\text{floor}(i) = \begin{cases} C(i) & C(i) \in v \\ vl(i) + (i - \text{idx}[vl(i)]) \cdot \frac{vr(i) - vl(i)}{\text{idx}[vr(i)] - \text{idx}[vl(i)]} & \text{else} \end{cases}$$

其中， $\text{idx}[x]$ 表示 x 在频谱上的位置， $\text{idx}[x]=0,1,\dots,127$ 。

然后在每两个相邻的频谱谷值之间，求输入音频帧与之前一帧的去底频谱的互相关 $\text{cor}(n)$ ，

$$10 \quad \text{cor}(n) = \frac{\left(\sum_{i=\text{lb}(n)}^{\text{hb}(n)} C'(i) \cdot C'_{-1}(i) \right)^2}{\left(\sum_{i=\text{lb}(n)}^{\text{hb}(n)} C'(i) \cdot C'(i) \right) \cdot \left(\sum_{i=\text{lb}(n)}^{\text{hb}(n)} C'_{-1}(i) \cdot C'_{-1}(i) \right)}$$

其中， $\text{lb}(n), \text{hb}(n)$ 分别表示第 n 个频谱谷值区间（即位于两个相邻得谷值之间的区域）的端点位置，即限定该谷值区间的两个频谱谷值的位置。

最后，通过下列公式计算输入音频帧的频谱相关度 cor_map_sum ：

$$\text{cor_map_sum} = \sum_{i=0}^{127} \text{cor}(\text{inv}[\text{lb}(n) \leq i, \text{hb}(n) \geq i])$$

15 其中， $\text{inv}[f]$ 表示函数 f 的反函数。

线性预测残差能量倾斜度 epsP_tilt 表示输入音频信号的线性预测残差能

量随线性预测阶数的升高而变化的程度。可以通过下列公式计算获得:

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

其中, $epsP(i)$ 表示第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。例如一个实施例中, $n=15$ 。

5 则步骤 S103 可以被以下步骤替代:

S105: 分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量, 根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧; 所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值, 运算操作可以包括求均值, 求方差等操作。

10 一个实施例中, 该步骤包括:

分别获得存储的频谱波动有效数据的均值, 频谱高频带峰度有效数据的均值, 频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差;

当下列条件之一满足时, 将所述当前音频帧分类为音乐帧, 否则将所述当前音频帧分类为语音帧: 所述频谱波动有效数据的均值小于第一阈值; 或者频谱高频带峰度有效数据的均值大于第二阈值; 或者所述频谱相关度有效数据的均值大于第三阈值; 或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

一般的, 音乐帧的频谱波动值较小, 而语音帧的频谱波动值较大; 音乐帧的频谱高频带峰度值较大, 语音帧的频谱高频带峰度较小; 音乐帧的频谱相关度的值较大, 语音帧的频谱相关度值较小; 音乐帧的线性预测残差能量倾斜度的变化较小, 而语音帧的线性预测残差能量倾斜度的变化较大。而因此可以根据上述参数的统计量对当前音频帧进行分类。当然还可以采用其他分类方法对该当前音频帧进行信号分类。例如, 统计频谱波动存储器中存储的频谱波动的有效数据的数量; 根据该有效数据的数量, 将存储器由近端到远端划分出至少

两个不同长度的区间，获得每个区间对应的频谱波动的有效数据的均值、频谱高频带峰度有效数据的均值、频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；其中，所述区间的起点为当前帧频谱波动的存储位置，近端为存储有当前帧频谱波动的一端，远端为存储有历史帧频谱波动的一端；

- 5 根据较短区间内的上述参数的有效数据的统计量对所述音频帧进行分类，若此区间内的参数统计量足够区分出所述音频帧的类型则分类过程结束，否则在其余较长区间中最短的区间内继续分类过程，并以此类推。在每个区间的分类过程中，根据每一个区间对应的分类阈值，对所述当前音频帧进行分类，当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分
- 10 类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

在信号分类后，可以对不同的信号采用不同的编码模式进行编码。例如，语音信号采用基于语音产生模型的编码器（如 CELP）进行编码，对音乐信号采

15 用基于变换的编码器（如基于 MDCT 的编码器）进行编码。

上述实施例中，根据频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度的长时统计量对音频信号进行分类，参数较少，识别率较高且复杂度较低；同时考虑声音活动性和敲击音乐的因素对频谱波动进行调整，根据当前音频帧所处信号环境，对频谱波动进行修正，提高分类识别率，适合混

20 合音频信号分类。

参考图 5，音频信号分类方法的另一个实施例包括：

S501：将输入音频信号进行分帧处理；

音频信号分类一般按帧进行，对每个音频信号帧提取参数进行分类，以确

25 定该音频信号帧属于语音帧还是音乐帧，以采用对应的编码模式进行编码。

S502：获得当前音频帧的线性预测残差能量倾斜度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

一个实施例中，线性预测残差能量倾斜度 epsP_tilt 可以通过下列公式计

算获得:

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

其中, $epsP(i)$ 表示第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。例如一个实施例中, $n=15$ 。

5 S503: 将线性预测残差能量倾斜度存储到存储器中;

可以将线性预测残差能量倾斜度存储到存储器中。一个实施例中, 该存储器可以为一个 FIFO 的 buffer, 该 buffer 的长度为 60 个存储单位 (即可存储 60 个线性预测残差能量倾斜度)。

10 可选的, 在存储线性预测残差能量倾斜度之前, 还包括: 根据所述当前音频帧的声音活动性, 确定是否将线性预测残差能量倾斜度存储于存储器中; 如果当前音频帧为活动帧, 则存储线性预测残差能量倾斜度; 否则不存储。

S504: 根据存储器中预测残差能量倾斜度部分数据的统计量, 对所述音频帧进行分类。

15 一个实施例中, 预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差; 则步骤 S504 包括:

将预测残差能量倾斜度部分数据的方差与音乐分类阈值相比较, 当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时, 将所述当前音频帧分类为音乐帧; 否则将所述当前音频帧分类为语音帧。

20 一般的, 音乐帧的线性预测残差能量倾斜度值变化较小, 而语音帧的线性预测残差能量倾斜度值变化较大。而因此可以根据线性预测残差能量倾斜度的统计量对当前音频帧进行分类。当然还可以结合其他参数采用其他分类方法对该当前音频帧进行信号分类。

另一个实施例中, 步骤 S504 之前还包括: 获得当前音频帧的频谱波动、频谱高频带峰度和频谱相关度, 并存储于对应的存储器中。则步骤 S504 具体为:

25 分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差

能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值。

进一步的，分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧包括：

分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

一般的，音乐帧的频谱波动值较小，而语音帧的频谱波动值较大；音乐帧的频谱高频带峰度值较大，语音帧的频谱高频带峰度较小；音乐帧的频谱相关度的值较大，语音帧的频谱相关度值较小；音乐帧的线性预测残差能量倾斜度值变化较小，而语音帧的线性预测残差能量倾斜度值变化较大。而因此可以根据上述参数的统计量对当前音频帧进行分类。

另一个实施例中，步骤 S504 之前还包括：获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率，并存储于对应的存储器。则步骤 S504 具体为：

分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量；

根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；所述统计量指对存储器中存储的数据运算操作后获得的数据值。

进一步的，分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量包括：获得存储的线性预测残差能量倾斜度的方差；获得存储的

频谱音调个数的均值。根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧包括：

当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：

线性预测残差能量倾斜度的方差小于第五阈值；或

频谱音调个数的均值大于第六阈值；或

频谱音调个数在低频带上的比率小于第七阈值。

其中，获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率包括：

统计当前音频帧在 0~8kHz 频带上频点峰值大于预定值的频点数量作为频谱音调个数；

计算当前音频帧在 0~4kHz 频带上频点峰值大于预定值的频点数量与 0~8kHz 频带上频点峰值大于预定值的频点数量的比值，作为频谱音调个数在低频带上的比率。一个实施例中，预定值为 50。

频谱音调个数 N_{tonal} 表示当前音频帧中的 0~8kHz 频带上频点峰值大于预定值的频点个数。一个实施例中，可以通过如下方式获得：对当前音频帧，统计其在 0~8kHz 频带上频点峰值 $p2v_map(i)$ 大于 50 的个数，即为 N_{tonal} ，其中， $p2v_map(i)$ 表示频谱第 i 个频点的峰度，其计算方式可以参考上述实施例的描述。

频谱音调个数在低频带上的比率 ratio_Ntonal_lf 表示频谱音调个数与低频带音调个数的比值。一个实施例中，可以通过如下方式获得：对当前音频帧，统计其在 0~4kHz 频带上 $p2v_map(i)$ 大于 50 的个数， $N_{\text{tonal_lf}}$ 。 ratio_Ntonal_lf 即为 $N_{\text{tonal_lf}}$ 与 N_{tonal} 的比值， $N_{\text{tonal_lf}}/N_{\text{tonal}}$ 。其中， $p2v_map(i)$ 表示频谱第 i 个频点的峰度，其计算方式可以参考上述实施例的描述。另一个实施例中，分别获得存储的多个 N_{tonal} 的均值和存储的多个 $N_{\text{tonal_lf}}$ 的均值，计算 $N_{\text{tonal_lf}}$ 的均值与 N_{tonal} 的均值的比值，作为频谱音调个数在低频带上的比率。

本实施例中，根据线性预测残差能量倾斜度的长时统计量对音频信号进行分类，同时兼顾了分类的鲁棒性和分类的识别速度，分类参数较少但结果较为准确，复杂度低、内存开销低。

5 参考图 6，音频信号分类方法的另一个实施例包括：

S601：将输入音频信号进行分帧处理；

S602：获得当前音频帧的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

10 频谱波动 flux 表示信号频谱的短时或长时能量波动，为当前音频帧与历史帧在中低频带频谱上对应频率的对数能量差的绝对值的均值；其中历史帧指当前音频帧之前的任意一帧。频谱高频带峰度 ph 表示当前音频帧频谱在高频带上的峰度或能量锐度。频谱相关度 cor_map_sum 表示信号谐波结构在相邻帧间的稳定度。线性预测残差能量倾斜度 epsP_tilt 表示线性预测残差能量倾斜度表示输入音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度。这
15 几个参数的具体计算方法参照前文实施例。

进一步的，可以获得浊音度参数；浊音度参数 voicing 表示当前音频帧与一个基音周期之前的信号的时域相关度。浊音度参数 voicing 是通过线性预测分析得到的，代表当前音频帧与一个基音周期之前的信号的时域相关度，取值在 0~1 之间。由于属现有技术，本发明不做详述。本实施例中当前音频帧的
20 个子帧各计算一个 voicing，求平均得到当前音频帧的 voicing 参数。当前音频帧的 voicing 参数也被缓存在一个 voicing 历史 buffer 中，本实施例中 voicing 历史 buffer 的长度为 10。

S603：分别将所述频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于对应的存储器；

25 可选的，在存储这几个参数之前，还包括：

一个实施例，根据所述当前音频帧的声音活动性，确定是否将所述频谱波动存储频谱波动存储器中。若当前音频帧为活动帧，则将当前音频帧的频谱波动存储于频谱波动存储器中。

另一个实施例，根据音频帧的声音活动性和音频帧是否为能量冲击，确定是否将所述频谱波动存储于存储器中。若当前音频帧为活动帧，且当前音频帧不属于能量冲击，则将当前音频帧的频谱波动存储于频谱波动存储器中；另一个实施例中，若当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。例如，当前音频帧为活动帧，且当前音频帧其前一帧以及历史第二帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。

声音活动性标识 `vad_flag` 和声音冲击标识 `attack_flag` 的定义和获得方式参照前述实施例的描述。

可选的，在存储这些参数之前，还包括：

根据所述当前音频帧的声音活动性，确定是否将频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；如果当前音频帧为活动帧，则存储上述参数；否则不存储。

15 S604：分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值，运算操作可以包括求均值，求方差等操作。

可选的，在步骤 S604 之前，还可以包括：

20 根据所述当前音频帧是否为敲击音乐，更新频谱波动存储器中存储的频谱波动；一个实施例中，若当前音频帧为敲击音乐，将频谱波动存储器中有效的频谱波动值修改为小于等于音乐阈值的一个值，其中当音频帧的频谱波动小于该音乐阈值时该音频被分类为音乐帧。一个实施例中，若当前音频帧为敲击音乐，则将频谱波动存储器中有效的频谱波动值重置为 5。

25 可选的，在步骤 S604 之前，还可以包括：

根据当前音频帧的历史帧的活动性，更新存储器中的频谱波动。一个实施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且前一帧音频帧为非活动帧，则将频谱波动存储器中已存储的除当前音频帧的频谱波动之

外的其他频谱波动的数据修改为无效数据。另一个实施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且当前音频帧之前连续三帧不全都为活动帧，则将当前音频帧的频谱波动修正为第一值。第一值可以为语音阈值，其中当音频帧的频谱波动大于该语音阈值时该音频被分类为语音帧。另一个实
5 施例中，如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且历史帧的分类结果为音乐帧且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

例如，如果当前音频帧前一帧为非活动帧 ($\text{vad_flag}=0$)，则除被新缓存入 flux 历史 buffer 的当前音频帧 flux 以外，其余 flux 历史 buffer 中的数据全部重
10 置为-1 (等价于将这些数据无效化)；如果当前音频帧之前连续三帧不全都为活动帧 ($\text{vad_flag}=1$)，则将刚缓存入 flux 历史 buffer 的当前音频帧 flux 修正为 16；如果当前音频帧之前连续三帧都为活动帧 ($\text{vad_flag}=1$)，且历史的信号分类结果长时平滑结果为音乐信号且当前音频帧 flux 大于 20，则将缓存的当前音频帧的频谱波动修改为 20。其中，活动帧以及历史的信号分类结果长时平滑结果的
15 计算可以参考前述实施例。

一个实施例中，步骤 S604 包括：

分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。
20

一般的，音乐帧的频谱波动值较小，而语音帧的频谱波动值较大；音乐帧
25 的频谱高频带峰度值较大，语音帧的频谱高频带峰度较小；音乐帧的频谱相关度的值较大，语音帧的频谱相关度值较小；音乐帧的线性预测残差能量倾斜度值较小，而语音帧的线性预测残差能量倾斜度值较大。而因此可以根据上述参数的统计量对当前音频帧进行分类。当然还可以采用其他分类方法对该当前音频帧进行信号分类。例如，统计频谱波动存储器中存储的频谱波动的有效数据

的数量；根据该有效数据的数量，将存储器由近端到远端划分出至少两个不同长度的区间，获得每个区间对应的频谱波动的有效数据的均值、频谱高频带峰度有效数据的均值、频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；其中，所述区间的起点为当前帧频谱波动的存储位置，近端为

5 存储有当前帧频谱波动的一端，远端为存储有历史帧频谱波动的一端；根据较短区间内的上述参数的有效数据的统计量对所述音频帧进行分类，若此区间内的参数统计量足够区分出所述音频帧的类型则分类过程结束，否则在其余较长区间中最短的区间内继续分类过程，并以此类推。在每个区间的分类过程中，根据每一个区间对应的分类阈值，对所述当前音频帧分类进行分类，当下

10 列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

在信号分类后，可以对不同的信号采用不同的编码模式进行编码。例如，

15 语音信号采用基于语音产生模型的编码器（如 CELP）进行编码，对音乐信号采用基于变换的编码器（如基于 MDCT 的编码器）进行编码。

本实施例中，根据频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度的长时统计量进行分类，同时兼顾了分类的鲁棒性和分类的识别速度，分类参数较少但结果较为准确，识别率较高且复杂度较低。

20

一个实施例中，在将上述频谱波动 flux、频谱高频带峰度 ph、频谱相关度 cor_map_sum 和线性预测残差能量倾斜度 epsP_tilt 存储于对应的存储器之后，可以根据存储的频谱波动的有效数据的数量，采用不同判断流程进行分类。如果声音活动性标识置为 1，即当前音频帧为活动声音帧，则，检查存储的频谱波

25 动的有效数据的个数 N。

存储器中存储的频谱波动中有效数据的个数 N 的值不同，判断流程也不同：

（1）参考图 7，若 N=60，则分别获得 flux 历史 buffer 中全部数据的均值，记为 flux60，近端 30 个数据的均值，记为 flux30，近端 10 个数据的均值，记为 flux10。分别获得 ph 历史 buffer 中全部数据的均值，记为 ph60，近端 30

个数据的均值, 记为 ph_{30} , 近端 10 个数据的均值, 记为 ph_{10} 。分别获得 cor_map_sum 历史 buffer 中全部数据的均值, 记为 $cor_map_sum_{60}$, 近端 30 个数据的均值, 记为 $cor_map_sum_{30}$, 近端 10 个数据的均值, 记为 $cor_map_sum_{10}$ 。并分别获得 $epsP_tilt$ 历史 buffer 中全部数据的方差, 记为 $epsP_tilt_{60}$, 近端 30 个数据的方差, 记为 $epsP_tilt_{30}$, 近端 10 个数据的方差, 记为 $epsP_tilt_{10}$ 。获得 voicing 历史 buffer 中数值大于 0.9 的数据的个数 $voicing_cnt$ 。其中, 近端为存储有当前音频帧对应的上述参数的一端。

首先检查 $flux_{10}$, ph_{10} , $epsP_tilt_{10}$, $cor_map_sum_{10}$, $voicing_cnt$ 是否满足条件: $flux_{10} < 10$ 或 $epsP_tilt_{10} < 0.0001$ 或 $ph_{10} > 1050$ 或 $cor_map_sum_{10} > 95$, 并且 $voicing_cnt < 6$, 若满足, 则将当前音频帧分类为音乐类型 (即 $Mode=1$)。否则, 检查 $flux_{10}$ 是否大于 15 且 $voicing_cnt$ 是否大于 2, 或者 $flux_{10}$ 是否大于 16, 若满足, 则将当前音频帧分类为语音类型 (即 $Mode=0$)。否则, 检查 $flux_{30}$, $flux_{10}$, ph_{30} , $epsP_tilt_{30}$, $cor_map_sum_{30}$, $voicing_cnt$ 是否满足条件: $flux_{30} < 13$ 且 $flux_{10} < 15$, 或 $epsP_tilt_{30} < 0.001$ 或 $ph_{30} > 800$ 或 $cor_map_sum_{30} > 75$, 若满足, 则将当前音频帧分类为音乐类型。否则, 检查 $flux_{60}$, $flux_{30}$, ph_{60} , $epsP_tilt_{60}$, $cor_map_sum_{60}$ 是否满足条件: $flux_{60} < 14.5$ 或 $cor_map_sum_{30} > 75$ 或 $ph_{60} > 770$ 或 $epsP_tilt_{10} < 0.002$, 并且 $flux_{30} < 14$ 。若满足, 则将当前音频帧分类为音乐类型, 否则分类为语音类型。

(2) 参考图 8, 如果 $N < 60$ 且 $N \geq 30$, 则分别获得 $flux$ 历史 buffer, ph 历史 buffer 和 cor_map_sum 历史 buffer 中近端 N 个数据的均值, 记为 $flux_N$, ph_N , $cor_map_sum_N$, 并同时得到 $epsP_tilt$ 历史 buffer 中近端 N 个数据的方差, 记为 $epsP_tilt_N$ 。检查 $flux_N$, ph_N , $epsP_tilt_N$, $cor_map_sum_N$ 是否满足条件: $flux_N < 13 + (N-30)/20$ 或 $cor_map_sum_N > 75 + (N-30)/6$ 或 $ph_N > 800$ 或 $epsP_tilt_N < 0.001$ 。若满足, 则将当前音频帧分类为音乐类型, 否则为语音类型。

(3) 参考图 9, 如果 $N < 30$ 且 $N \geq 10$, 则分别获得 $flux$ 历史 buffer, ph 历史 buffer 和 cor_map_sum 历史 buffer 中近端 N 个数据的均值, 记为 $flux_N$, ph_N 和 $cor_map_sum_N$, 并同时得到 $epsP_tilt$ 历史 buffer 中近端 N 个数据的方差, 记为 $epsP_tilt_N$ 。

首先检查历史分类结果的长时滑动平均 mode_mov 是否大于 0.8。若是, 则检查 fluxN, phN, epsP-tiltN, cor-map-sumN 是否满足条件: $\text{fluxN} < 16 + (N-10)/20$ 或 $\text{phN} > 1000 - 12.5 \times (N-10)$ 或 $\text{epsP-tiltN} < 0.0005 + 0.000045 \times (N-10)$ 或 $\text{cor-map-sumN} > 90 - (N-10)$ 。否则, 获得 voicing 历史 buffer 中数值大于 0.9 的数据的个数 voicing_cnt, 并检查是否满足条件: $\text{fluxN} < 12 + (N-10)/20$ 或 $\text{phN} > 1050 - 12.5 \times (N-10)$ 或 $\text{epsP-tiltN} < 0.0001 + 0.000045 \times (N-10)$ 或 $\text{cor-map-sumN} > 95 - (N-10)$, 并且 voicing_cnt < 6。如果满足上面两组条件中的任一组, 则将当前音频帧分类为音乐类型, 否则为语音类型。

(4) 参考图 10, 如果 $N < 10$ 且 $N > 5$, 则分别获得 ph 历史 buffer, cor-map-sum 历史 buffer 中近端 N 个数据的均值, 记为 phN 和 cor-map-sumN。以及 epsP-tilt 历史 buffer 中近端 N 个数据的方差, 记为 epsP-tiltN。同时获得 voicing 历史 buffer 中近端 6 个数据中数值大于 0.9 的数据的个数 voicing_cnt6。

检查是否满足条件: $\text{epsP-tiltN} < 0.00008$ 或 $\text{phN} > 1100$ 或 $\text{cor-map-sumN} > 100$, 并且 voicing_cnt < 4。若满足, 则将当前音频帧分类为音乐类型, 否则为语音类型。

(5) 如果 $N \leq 5$, 则以前一音频帧的分类结果作为当前音频帧的分类类型。

上述实施例为根据频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度的长时统计量进行分类的一种具体分类流程, 本领域技术人员可以理解的是, 可以使用别的流程进行分类。本实施例中的分类流程可以应用在前述实施例中的对应步骤, 例如作为图 2 的步骤 103、图 4 的步骤 105 或者图 6 中的步骤 604 的具体分类方法。

参考图 11, 一种音频信号分类方法的另一个实施例包括:

S1101: 将输入音频信号进行分帧处理;

S1102: 获得当前音频帧的线性预测残差能量倾斜度、频谱音调个数和频谱音调个数在低频带上的比率;

线性预测残差能量倾斜度 epsP-tilt 表示输入音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度; 频谱音调个数 Ntonal 表示当前音频帧

中的 0~8kHz 频带上频点峰值大于预定值的频点个数；频谱音调个数在低频带上的比率 ratio_Ntonal_lf 表示频谱音调个数与低频带音调个数的比值。具体计算参照前述实施例的描述。

S1103: 分别将线性预测残差能量倾斜度 epsP_tilt、频谱音调个数和频谱
5 音调个数在低频带上的比率存储到对应的存储器中；

当前音频帧的线性预测残差能量倾斜度 epsP_tilt、频谱音调个数各自被缓存入各自的历史 buffer 中，本实施例中这两个 buffer 的长度也均为 60。

可选的，在存储这些参数之前，还包括：根据所述当前音频帧的声音活动性，确定是否将所述线性预测残差能量倾斜度、频谱音调个数和频谱音调个数
10 在低频带上的比率存储于存储器中；并在确定需要存储时将所述线性预测残差能量倾斜度存储于存储器中。如果当前音频帧为活动帧，则存储上述参数；否则不存储。

S1104: 分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量；所述统计量指对存储器中存储的数据运算操作后获得的数据值，运
15 算操作可以包括求均值，求方差等操作。

一个实施例中，分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量包括：获得存储的线性预测残差能量倾斜度的方差；获得存储的频谱音调个数的均值。

S1105: 根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计
20 量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；

一个实施例中，该步骤包括：

当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：

线性预测残差能量倾斜度的方差小于第五阈值；或
25 频谱音调个数的均值大于第六阈值；或
频谱音调个数在低频带上的比率小于第七阈值。

一般的，音乐帧的线性预测残差能量倾斜度值较小，而语音帧的线性预测

残差能量倾斜度值较大；音乐帧的频谱音调个数较多，而语音帧的频谱音调个数较少；音乐帧的频谱音调个数在低频带上的比率较低，而语音帧的频谱音调个数在低频带上的比率较高（语音帧的能量主要集中在低频带上）。而因此可以根据上述参数的统计量对当前音频帧进行分类。当然还可以采用其他分类方法

5 对该当前音频帧进行信号分类。

在信号分类后，可以对不同的信号采用不同的编码模式进行编码。例如，语音信号采用基于语音产生模型的编码器（如 CELP）进行编码，对音乐信号采用基于变换的编码器（如基于 MDCT 的编码器）进行编码。

10 上述实施例中，根据线性预测残差能量倾斜度、频谱音调个数的长时统计量和频谱音调个数在低频带上的比率对音频信号进行分类，参数较少，识别率较高且复杂度较低。

一个实施例中，分别将线性预测残差能量倾斜度 epsP_tilt 、频谱音调个数 Ntonal 和频谱音调个数在低频带上的比率 ratio_Ntonal_lf 存储到对应的 buffer 后，获得 epsP_tilt 历史 buffer 中所有数据的方差，记为 epsP_tilt60 。
15 获得 Ntonal 历史 buffer 中所有数据的均值，记为 Ntonal60 。获得 Ntonal_lf 历史 buffer 中所有数据的均值，并计算该均值与 Ntonal60 的比，记为 ratio_Ntonal_lf60 。参考图 12，根据如下法则进行当前音频帧的分类：

如果声音活动性标识为 1（即 $\text{vad_flag}=1$ ），即当前音频帧为活动声音帧，
则，则检查是否满足条件： $\text{epsP_tilt60}<0.002$ 或 $\text{Ntonal60}>18$ 或
20 $\text{ratio_Ntonal_lf60}<0.42$ ，若满足，则将当前音频帧分类为音乐类型（即 $\text{Mode}=1$ ），否则为语音类型（即 $\text{Mode}=0$ ）。

上述实施例为根据线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率进行分类的一种具体分类流程，本领域技术人员可以理解的是，可以使用别的流程进行分类。本实施例中的分类流程
25 可以应用在前述实施例中的对应步骤，例如作为图 5 的步骤 504 或图 11 步骤 1105 的具体分类方法。

本发明是一种低复杂度低内存开销的音频编码模式选择方法。同时兼顾了分类的鲁棒性和分类的识别速度。

与上述方法实施例相关联，本发明还提供一种音频信号分类装置，该装置可以位于终端设备，或网络设备中。该音频信号分类装置可以执行上述方法实施例的步骤。

- 5 参考图 13，本发明一种音频信号的分类装置的一个实施例，用于对输入的音频信号进行分类，其包括：

存储确认单元 1301，用于根据所述当前音频帧的声音活动性，确定是否获得并存储当前音频帧的频谱波动，其中，所述频谱波动表示音频信号的频谱的能量波动；

- 10 存储器 1302，用于在存储确认单元输出需要存储的结果时存储所述频谱波动；

更新单元 1303，用于根据语音帧是否为敲击音乐或历史音频帧的活动性，更新存储器中存储的频谱波动；

- 15 分类单元 1304，用于根据存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧。当频谱波动的有效数据的统计量满足语音分类条件时，将所述当前音频帧分类为语音帧；当频谱波动的有效数据的统计量满足音乐分类条件时，将所述当前音频帧分类为音乐帧。

一个实施例中，存储确认单元具体用于：确认当前音频帧为活动帧时，输出需要存储当前音频帧的频谱波动的结果。

- 20 另一个实施例中，存储确认单元具体用于：确认当前音频帧为活动帧，且当前音频帧不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

另一个实施例中，存储确认单元具体用于：确认当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

- 25 一个实施例中，更新单元具体用于若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。

另一个实施例中，更新单元具体用于：如果当前音频帧为活动帧，且前一帧音频帧为非活动帧时，则将存储器中已存储的除当前音频帧的频谱波动之外

的其他频谱波动的数据修改为无效数据；或，如果当前音频帧为活动帧，且当前音频帧之前连续三帧不全都为活动帧时，则将当前音频帧的频谱波动修正为第一值；或，如果当前音频帧为活动帧，且历史分类结果为音乐信号且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，

5 第二值大于第一值。

参考图 14，一个实施例中，分类单元 1303 包括：

计算单元 1401，用于获得存储器中存储的频谱波动的部分或全部有效数据的均值；

判断单元 1402，用于将所述频谱波动的有效数据的均值与音乐分类条件做

10 比较，当所述频谱波动的有效数据的均值满足音乐分类条件时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

例如，当所获得的频谱波动的有效数据的均值小于音乐分类阈值时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

上述实施例，由于根据频谱波动的长时统计量对音频信号进行分类，参数

15 较少，识别率较高且复杂度较低；同时考虑声音活动性和敲击音乐的因素对频谱波动进行调整，对音乐信号识别率更高，适合混合音频信号分类。

另一个实施例中，音频信号分类装置还包括：

参数获得单元，用于获得当前音频帧的频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；其中，频谱高频带峰度表示当前音频帧的频谱在高频

20 带上的峰度或能量锐度；频谱相关度表示当前音频帧的信号谐波结构在相邻帧间的稳定度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

该存储确认单元还用于，根据所述当前音频帧的声音活动性，确定是否存

25 储所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

该存储单元还用于，当存储确认单元输出需要存储的结果时存储所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

该分类单元具体用于，分别获得存储的频谱波动、频谱高频带峰度、频谱

相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧。当频谱波动的有效数据的统计量满足语音分类条件时，将所述当前音频帧分类为语音帧；当频谱波动的有效数据的统计量满足音乐分类条件时，将所述当前音频帧分类为音乐帧。

5 一个实施例中，该分类单元具体包括：

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，
10 否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

上述实施例中，根据频谱波动、频谱高频带峰度、频谱相关度和线性预测
15 残差能量倾斜度的长时统计量对音频信号进行分类，参数较少，识别率较高且复杂度较低；同时考虑声音活动性和敲击音乐的因素对频谱波动进行调整，根据当前音频帧所处信号环境，对频谱波动进行修正，提高分类识别率，适合混合音频信号分类。

20 参考图 15，本发明一种音频信号的分类装置的另一个实施例，用于对输入的音频信号进行分类，其包括：

分帧单元 1501，用于对输入音频信号进行分帧处理；

参数获得单元 1502，用于获得当前音频帧的线性预测残差能量倾斜度；其中，线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶
25 数的升高而变化的程度；

存储单元 1503，用于存储线性预测残差能量倾斜度；

分类单元 1504，用于根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类。

参考图 16，音频信号的分类装置还包括：

存储确认单元 1505，用于根据所述当前音频帧的声音活动性，确定是否将所述线性预测残差能量倾斜度存储于存储器中；

则该存储单元 1503 具体用于，当存储确认单元确认需要确定需要存储时将
5 将所述线性预测残差能量倾斜度存储于存储器中。

一个实施例中，预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差；

所述分类单元具体用于将预测残差能量倾斜度部分数据的方差与音乐分类
阈值相比较，当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时，
10 将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

另一个实施例中，参数获得单元还用于：获得当前音频帧的频谱波动、频谱
高频带峰度和频谱相关度，并存储于对应的存储器中；

则该分类单元具体用于：分别获得存储的频谱波动、频谱高频带峰度、频谱
相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据
15 的统计量将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对
存储器中存储的有效数据运算操作后获得的数据值。

参考图 17，具体的，一个实施例中，分类单元 1504 包括：

计算单元 1701，用于分别获得存储的频谱波动有效数据的均值，频谱高频
带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜
20 度有效数据的方差；

判断单元 1702，用于当下列条件之一满足时，将所述当前音频帧分类为音
乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小
于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频
谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数
25 据的方差小于第四阈值。

另一个实施例中，参数获得单元还用于：获得当前音频帧的频谱音调个数
和频谱音调个数在低频带上的比率，并存储于存储器；

则该分类单元具体用于：分别获得存储的线性预测残差能量倾斜度的统计

量、频谱音调个数的统计量；根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的数据运算操作后获得的数据值。

5 具体的该分类单元包括：

计算单元，用于获得线性预测残差能量倾斜度有效数据的方差和存储的频谱音调个数的均值；

判断单元，用于当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：线性预测残差能量倾斜度的方差小于第五阈值；或频谱音调个数的均值大于第六阈值；或
10 频谱音调个数在低频带上的比率小于第七阈值。

具体的，参数获得单元根据下列公式计算当前音频帧的线性预测残差能量倾斜度：

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

15 其中， $epsP(i)$ 表示当前音频帧第 i 阶线性预测的预测残差能量； n 为正整数，表示线性预测的阶数，其小于等于线性预测的最大阶数。

具体的，该参数获得单元用于统计当前音频帧在 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量作为频谱音调个数；所述参数获得单元用于计算当前音频帧在 $0 \sim 4\text{kHz}$ 频带上频点峰值大于预定值的频点数量与 $0 \sim 8\text{kHz}$ 频带上频点
20 峰值大于预定值的频点数量的比值，作为频谱音调个数在低频带上的比率。

本实施例中，根据线性预测残差能量倾斜度的长时统计量对音频信号进行分类，同时兼顾了分类的鲁棒性和分类的识别速度，分类参数较少但结果较为准确，复杂度低、内存开销低。

25 本发明一种音频信号的分类装置的另一个实施例，用于对输入的音频信号

进行分类，其包括：

分帧单元，用于将输入音频信号进行分帧处理；

参数获得单元，用于获得当前音频帧的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；其中，频谱波动表示音频信号的频谱的能量波动，频谱高频带峰度表示当前音频帧的频谱在高频带上的峰度或能量锐度；
5 频谱相关度表示当前音频帧的信号谐波结构在相邻帧间的稳定度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

存储单元，用于存储频谱波动、频谱高频带峰度、频谱相关度和线性预测
10 残差能量倾斜度；

分类单元，用于分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据有效数据的统计量将所述音频帧分类为语音帧或者音乐帧；其中，所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值，运算操作可以包括求均值，求方
15 差等操作。

一个实施例中，音频信号的分类装置还可以包括：

存储确认单元，用于根据所述当前音频帧的声音活动性，确定是否存储当前音频帧的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

20 存储单元，具体用于当存储确认单元输出需要存储的结果时，存储频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度。

具体的，一个实施例中，存储确认单元根据所述当前音频帧的声音活动性，确定是否将所述频谱波动存储于频谱波动存储器中。如果当前音频帧为活动帧，则存储确认单元输出存储上述参数的结果；否则输出不需要存储的结果。另一个
25 实施例中，存储确认单元根据音频帧的声音活动性和音频帧是否为能量冲击，确定是否将所述频谱波动存储于存储器中。若当前音频帧为活动帧，且当前音频帧不属于能量冲击，则将当前音频帧的频谱波动存储于频谱波动存储器中；另一个实施例中，若当前音频帧为活动帧，且包含当前音频帧与其历史帧在内

的多个连续帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。例如，当前音频帧为活动帧，且当前音频帧其前一帧以及历史第二帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中；否则不存储。

5 一个实施例中，分类单元包括：

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

- 判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，
- 10 否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

- 当前音频帧的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能
- 15 量倾斜度的具体计算方式，可以参照上述方法实施例。

进一步的，该音频信号的分类装置还可以包括：

- 更新单元，用于根据语音帧是否为敲击音乐或历史音频帧的活动性，更新存储器中存储的频谱波动。一个实施例中，更新单元具体用于若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。另一个实施例
- 20 中，更新单元具体用于：如果当前音频帧为活动帧，且前一帧音频帧为非活动帧时，则将存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据；或，如果当前音频帧为活动帧，且当前音频帧之前连续三帧不全都为活动帧时，则将当前音频帧的频谱波动修正为第一值；或，如果当前音频帧为活动帧，且历史分类结果为音乐信号且当前音频帧的频谱波动大
- 25 于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

本实施例中，根据频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度的长时统计量进行分类，同时兼顾了分类的鲁棒性和分类的识别

速度，分类参数较少但结果较为准确，识别率较高且复杂度较低。

本发明一种音频信号的分类装置的另一个实施例，用于对输入的音频信号进行分类，其包括：

5 分帧单元，用于对输入音频信号进行分帧处理；

参数获得单元，用于获得获得当前音频帧的线性预测残差能量倾斜度、频谱音调个数和频谱音调个数在低频带上的比率；其中，线性预测残差能量倾斜度 epsP_tilt 表示输入音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；频谱音调个数 N_{tonal} 表示当前音频帧中的 0~8kHz 频带上频点峰值大于预定值的频点个数；频谱音调个数在低频带上的比率 ratio_Ntonal_lf 表示频谱音调个数与低频带音调个数的比值。具体计算参照前述实施例的描述。

存储单元，用于存储线性预测残差能量倾斜度、频谱音调个数和频谱音调个数在低频带上的比率；

15 分类单元，用于分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量；根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的数据运算操作后获得的数据值。

具体的，该分类单元包括：

20 计算单元，用于获得线性预测残差能量倾斜度有效数据的方差和存储的频谱音调个数的均值；

判断单元，用于当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：线性预测残差能量倾斜度的方差小于第五阈值；或频谱音调个数的均值大于第六阈值；或
25 频谱音调个数在低频带上的比率小于第七阈值。

具体的，参数获得单元根据下列公式计算当前音频帧的线性预测残差能量倾斜度：

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

其中, $epsP(i)$ 表示当前音频帧第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。

具体的, 该参数获得单元用于统计当前音频帧在 $0 \sim 8\text{kHz}$ 频带上频点峰值
5 大于预定值的频点数量作为频谱音调个数; 所述参数获得单元用于计算当前音频帧在 $0 \sim 4\text{kHz}$ 频带上频点峰值大于预定值的频点数量与 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量的比值, 作为频谱音调个数在低频带上的比率。

上述实施例中, 根据线性预测残差能量倾斜度、频谱音调个数的长时统计
10 量和频谱音调个数在低频带上的比率对音频信号进行分类, 参数较少, 识别率较高且复杂度较低。

上述音频信号的分类装置可以与不同的编码器相连接, 对不同的信号采用不同的编码器进行编码。例如, 音频信号的分类装置分别与两个编码器连接, 对语音信号采用基于语音产生模型的编码器 (如 CELP) 进行编码, 对音乐信号
15 采用基于变换的编码器 (如基于 MDCT 的编码器) 进行编码。上述装置实施例中的各个具体参数的定义和获得方法可以参照方法实施例的相关描述。

与上述方法实施例相关联, 本发明还提供一种音频信号分类装置, 该装置可以位于终端设备, 或网络设备中。该音频信号分类装置可以由硬件电路来实现
20 现, 或者由软件配合硬件来实现。例如, 参考图 18, 由一个处理器调用音频信号分类装置来实现对音频信号的分类。该音频信号分类装置可以执行上述方法实施例中的各种方法和流程。该音频信号分类装置的具体模块和功能可以参照上述装置实施例的相关描述。

图 19 的设备 1900 的一个例子是编码器。设备 100 包括处理器 1910 和存储

器 1920。

存储器 1920 可以包括随机存储器、闪存、只读存储器、可编程只读存储器、非易失性存储器或寄存器等。处理器 1920 可以是中央处理器 (Central Processing Unit, CPU)。

5 存储器 1910 用于存储可执行指令。处理器 1920 可以执行存储器 1910 中存储的可执行指令，用于：

设备 1900 的其它功能和操作可参照上面图 3 至图 12 的方法实施例的过程，为了避免重复，此处不再赘述。

10 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，所述的存储介质可为磁碟、光盘、只读存储记忆体 (Read-Only Memory, ROM) 或随机存储记忆体 (Random Access Memory, RAM) 等。

15

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或
20 一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为

单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，
5 也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

以上所述仅为本发明的几个实施例，本领域的技术人员依据申请文件公开的可以对本发明进行各种改动或变型而不脱离本发明的精神和范围。

权 利 要 求

1、一种音频信号分类方法，其特征在于，包括：

根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中，其中，所述频谱波动表示音频信号的频谱的能量波动；

5 根据音频帧是否为敲击音乐或历史音频帧的活动性，更新频谱波动存储器中存储的频谱波动；

根据频谱波动存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧。

2、根据权利要求1所述的方法，其特征在于，根据当前音频帧的声音活动
10 性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

若当前音频帧为活动帧，则将当前音频帧的频谱波动存储于频谱波动存储器中。

3、根据权利要求1所述的方法，其特征在于，根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

15 若当前音频帧为活动帧，且当前音频帧不属于能量冲击，则将当前音频帧的频谱波动存储于频谱波动存储器中。

4、根据权利要求1所述的方法，其特征在于，根据当前音频帧的声音活动性，确定是否获得当前音频帧的频谱波动并存储于频谱波动存储器中包括：

20 若当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击，则将音频帧的频谱波动存储于频谱波动存储器中。

5、根据权利要求1至4所述的任一方法，其特征在于，根据所述当前音频帧是否为敲击音乐，更新频谱波动存储器中存储的频谱波动包括：

若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。

25 6、根据权利要求1至4所述的任一方法，其特征在于，根据所述历史音频帧的活动性，更新频谱波动存储器中存储的频谱波动包括：

如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且前一帧音频

帧为非活动帧，则将频谱波动存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据；

如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且当前音频帧之前连续三帧历史帧不全都为活动帧，则将当前音频帧的频谱波动修正为第一

5 值；

如果确定当前音频帧的频谱波动存储于频谱波动存储器中，且历史分类结果为音乐信号且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值大于第一值。

7、根据权利要求 1-6 所述的任一方法，其特征在于，根据频谱波动存储器
10 中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧包括：

获得频谱波动存储器中存储的频谱波动的部分或全部有效数据的均值；

当所获得的频谱波动的有效数据的均值满足音乐分类条件时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

15 8、根据权利要求 1-6 所述的方法，其特征在于，还包括：

获得当前音频帧的频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；其中，频谱高频带峰度表示当前音频帧的频谱在高频带上的峰度或能量锐度；频谱相关度表示当前音频帧的信号谐波结构在相邻帧间的稳定度；线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而
20 变化的程度；

根据所述当前音频帧的声音活动性，确定是否将所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器中；

其中，所述根据频谱波动存储器中存储的频谱波动的部分或全部数据的统计量，对所述音频帧进行分类包括：

25 分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频

谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

9、一种音频信号的分类装置，用于对输入的音频信号进行分类，其特征在于，包括：

存储确认单元，用于根据所述当前音频帧的声音活动性，确定是否获得并存储当前音频帧的频谱波动，其中，所述频谱波动表示音频信号的频谱的能量波动；

存储器，用于在存储确认单元输出需要存储的结果时存储所述频谱波动；

10 更新单元，用于根据语音帧是否为敲击音乐或历史音频帧的活动性，更新存储器中存储的频谱波动；

分类单元，用于根据存储器中存储的频谱波动的部分或全部有效数据的统计量，将所述当前音频帧分类为语音帧或者音乐帧。

10、根据权利要求 9 所述的装置，其特征在于，所述存储确认单元具体用于：确认当前音频帧为活动帧时，输出需要存储当前音频帧的频谱波动的结果。

11、根据权利要求 9 所述的装置，其特征在于，所述存储确认单元具体用于：确认当前音频帧为活动帧，且当前音频帧不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

12、根据权利要求 9 所述的装置，其特征在于，所述存储确认单元具体用于：确认当前音频帧为活动帧，且包含当前音频帧与其历史帧在内的多个连续帧都不属于能量冲击时，输出需要存储当前音频帧的频谱波动的结果。

13、根据权利要求 9-12 所述的任一装置，其特征在于，所述更新单元具体用于若当前音频帧属于敲击音乐，则修改频谱波动存储器中已存储的频谱波动的值。

25 14、根据权利要求 9-12 所述的任一装置，其特征在于，所述更新单元具体用于：如果当前音频帧为活动帧，且前一帧音频帧为非活动帧时，则将存储器中已存储的除当前音频帧的频谱波动之外的其他频谱波动的数据修改为无效数据；或

如果当前音频帧为活动帧，且当前音频帧之前连续三帧不全都为活动帧时，则将当前音频帧的频谱波动修正为第一值；或

如果当前音频帧为活动帧，且历史分类结果为音乐信号且当前音频帧的频谱波动大于第二值，则将当前音频帧的频谱波动修正为第二值，其中，第二值
5 大于第一值。

15、根据权利要求 9-14 所述的任一装置，其特征在于，所述分类单元包括：
计算单元，用于获得存储器中存储的频谱波动的部分或全部有效数据的均值；

判断单元，用于将所述频谱波动的有效数据的均值与音乐分类条件做比较，
10 当所述频谱波动的有效数据的均值满足音乐分类条件时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

16、根据权利要求 9-14 所述的任一装置，其特征在于，还包括：

参数获得单元，用于获得当前音频帧的频谱高频带峰度、频谱相关度、浊
音度参数和线性预测残差能量倾斜度；其中，频谱高频带峰度表示当前音频帧
15 的频谱在高频带上的峰度或能量锐度；频谱相关度表示当前音频帧的信号谐波
结构在相邻帧间的稳定度；浊音度参数表示当前音频帧与一个基音周期之前的
信号的时域相关度；线性预测残差能量倾斜度表示音频信号的线性预测残差能
量随线性预测阶数的升高而变化的程度；

所述存储确认单元还用于，根据所述当前音频帧的声音活动性，确定是否
20 将所述频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度存储于存储器
中；

所述存储单元还用于，当存储确认单元输出需要存储的结果时存储所述频
谱高频带峰度、频谱相关度和线性预测残差能量倾斜度；

所述分类单元具体用于，分别获得存储的频谱波动、频谱高频带峰度、频
25 谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据
的统计量将所述音频帧分类为语音帧或者音乐帧。

17、根据权利要求 16 所述的任一装置，其特征在于，所述分类单元包括：

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰

度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

- 判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一
- 5 阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

18、一种音频信号分类方法，其特征在于，包括：

将输入音频信号进行分帧处理；

- 10 获得当前音频帧的线性预测残差能量倾斜度；所述线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度；

将线性预测残差能量倾斜度存储到存储器中；

根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类。

- 15 19、根据权利要求 18 所述的方法，其特征在于，将线性预测残差能量倾斜度存储到存储器中之前还包括：

根据所述当前音频帧的声音活动性，确定是否将所述线性预测残差能量倾斜度存储于存储器中；并在确定需要存储时将所述线性预测残差能量倾斜度存储于存储器中。

- 20 20、根据权利要求 18 或 19 所述的方法，其特征在于，预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差；所述根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类包括：

- 25 将预测残差能量倾斜度部分数据的方差与音乐分类阈值相比较，当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时，将所述当前音频帧分类为音乐帧；否则将所述当前音频帧分类为语音帧。

21、根据权利要求 18 或 19 所述的方法，其特征在于，还包括：

获得当前音频帧的频谱波动、频谱高频带峰度和频谱相关度，并存储于对

应的存储器中；

其中，所述根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类包括：

分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的有效数据运算操作后获得的数据值。

22、根据权利要求 21 所述的方法，其特征在于，分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量，根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧包括：

分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

当下列条件之一满足时，将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

23、根据权利要求 18 或 19 所述的方法，其特征在于，还包括：

获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率，并存储于对应的存储器；

其中，所述根据存储器中预测残差能量倾斜度部分数据的统计量，对所述音频帧进行分类包括：

分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量；

根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为语音帧或者音乐帧；所述统计量指对存储器中存储的数据运算操作后获得的数据值。

24、根据权利要求 23 所述的方法，其特征在于，分别获得存储的线性预测

残差能量倾斜度的统计量、频谱音调个数的统计量包括:

获得存储的线性预测残差能量倾斜度的方差;

获得存储的频谱音调个数的均值;

根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率, 将所述音频帧分类为语音帧或者音乐帧包括:

当当前音频帧为活动帧, 且满足下列条件之一, 则将所述当前音频帧分类为音乐帧, 否则将所述当前音频帧分类为语音帧:

线性预测残差能量倾斜度的方差小于第五阈值; 或

频谱音调个数的均值大于第六阈值; 或

10 频谱音调个数在低频带上的比率小于第七阈值。

25、根据权利要求 18-24 所述的任一方法, 其特征在于, 获得当前音频帧的线性预测残差能量倾斜度包括:

根据下列公式计算当前音频帧的线性预测残差能量倾斜度:

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

15 其中, $epsP(i)$ 表示当前音频帧第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。

26、根据权利要求 23-24 所述的任一方法, 其特征在于, 获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率包括:

统计当前音频帧在 0~8kHz 频带上频点峰值大于预定值的频点数量作为频谱音调个数;

计算当前音频帧在 0~4kHz 频带上频点峰值大于预定值的频点数量与 0~8kHz 频带上频点峰值大于预定值的频点数量的比值, 作为频谱音调个数在低频带上的比率。

27、一种信号分类装置, 用于对输入的音频信号进行分类, 其特征在于,

包括:

分帧单元, 用于对输入音频信号进行分帧处理;

参数获得单元, 用于获得当前音频帧的线性预测残差能量倾斜度; 所述线性预测残差能量倾斜度表示音频信号的线性预测残差能量随线性预测阶数的升高而变化的程度;

存储单元, 用于存储线性预测残差能量倾斜度;

分类单元, 用于根据存储器中预测残差能量倾斜度部分数据的统计量, 对所述音频帧进行分类。

28、根据权利要求 27 所述的装置, 其特征在于, 还包括:

10 存储确认单元, 用于根据所述当前音频帧的声音活动性, 确定是否将所述线性预测残差能量倾斜度存储于存储器中;

所述存储单元具体用于, 当存储确认单元确认需要确定需要存储时将所述线性预测残差能量倾斜度存储于存储器中。

29、根据权利要求 27 或 28 所述的装置, 其特征在于,

15 预测残差能量倾斜度部分数据的统计量为预测残差能量倾斜度部分数据的方差;

所述分类单元具体用于将预测残差能量倾斜度部分数据的方差与音乐分类阈值相比较, 当所述预测残差能量倾斜度部分数据的方差小于音乐分类阈值时, 将所述当前音频帧分类为音乐帧; 否则将所述当前音频帧分类为语音帧。

20 30、根据权利要求 27 或 28 所述的装置, 其特征在于, 参数获得单元还用于: 获得当前音频帧的频谱波动、频谱高频带峰度和频谱相关度, 并存储于对应的存储器中;

所述分类单元具体用于: 分别获得存储的频谱波动、频谱高频带峰度、频谱相关度和线性预测残差能量倾斜度中有效数据的统计量, 根据所述有效数据的统计量将所述音频帧分类为语音帧或者音乐帧; 所述有效数据的统计量指对
25 存储器中存储的有效数据运算操作后获得的数据值。

31、根据权利要求 30 所述的装置, 其特征在于, 所述分类单元包括:

计算单元，用于分别获得存储的频谱波动有效数据的均值，频谱高频带峰度有效数据的均值，频谱相关度有效数据的均值和线性预测残差能量倾斜度有效数据的方差；

- 判断单元，用于当下列条件之一满足时，将所述当前音频帧分类为音乐帧，
5 否则将所述当前音频帧分类为语音帧：所述频谱波动有效数据的均值小于第一阈值；或者频谱高频带峰度有效数据的均值大于第二阈值；或者所述频谱相关度有效数据的均值大于第三阈值；或者线性预测残差能量倾斜度有效数据的方差小于第四阈值。

- 32、根据权利要求 27 或 28 所述的装置，其特征在于，所述参数获得单元
10 还用于：获得当前音频帧的频谱音调个数和频谱音调个数在低频带上的比率，并存储于存储器；

- 所述分类单元具体用于：分别获得存储的线性预测残差能量倾斜度的统计量、频谱音调个数的统计量；根据所述线性预测残差能量倾斜度的统计量、频谱音调个数的统计量和频谱音调个数在低频带上的比率，将所述音频帧分类为
15 语音帧或者音乐帧；所述有效数据的统计量指对存储器中存储的数据运算操作后获得的数据值。

33、根据权利要求 32 所述的装置，其特征在于，所述分类单元包括：

计算单元，用于获得线性预测残差能量倾斜度有效数据的方差和存储的频谱音调个数的均值；

- 20 判断单元，用于当当前音频帧为活动帧，且满足下列条件之一，则将所述当前音频帧分类为音乐帧，否则将所述当前音频帧分类为语音帧：线性预测残差能量倾斜度的方差小于第五阈值；或频谱音调个数的均值大于第六阈值；或频谱音调个数在低频带上的比率小于第七阈值。

- 34、根据权利要求 27-33 所述的任一装置，其特征在于，所述参数获得单元
25 根据下列公式计算当前音频帧的线性预测残差能量倾斜度：

$$epsP_tilt = \frac{\sum_{i=1}^n epsP(i) \cdot epsP(i+1)}{\sum_{i=1}^n epsP(i) \cdot epsP(i)}$$

其中, $epsP(i)$ 表示当前音频帧第 i 阶线性预测的预测残差能量; n 为正整数, 表示线性预测的阶数, 其小于等于线性预测的最大阶数。

35、根据权利要求 32-33 所述的任一装置, 其特征在于, 所述参数获得单元用于统计当前音频帧在 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量作为频谱音调个数; 所述参数获得单元用于计算当前音频帧在 $0 \sim 4\text{kHz}$ 频带上频点峰值大于预定值的频点数量与 $0 \sim 8\text{kHz}$ 频带上频点峰值大于预定值的频点数量的比值, 作为频谱音调个数在低频带上的比率。