



## (12) 发明专利

(10) 授权公告号 CN 1920947 B

(45) 授权公告日 2011.05.11

(21) 申请号 200610113117.6

US 20050246178 A1, 2005.11.03, 全文.

(22) 申请日 2006.09.15

US 6973184 B1, 2005.12.06, 全文.

(73) 专利权人 清华大学

CN 1357136 A, 2002.07.03, 全文.

地址 100084 北京市 100084-82 信箱

CN 13347549 A, 2002.05.01, 全文.

(72) 发明人 张斌 窦维蓓

US 7027982 B2, 2006.04.11, 全文.

(51) Int. Cl.

CN 1175854 A, 1998.03.11, 全文.

G10L 15/00(2006.01)

审查员 隋欣

G10L 15/08(2006.01)

G10L 15/02(2006.01)

G10L 11/00(2006.01)

G10L 19/00(2006.01)

G10L 19/08(2006.01)

G10L 19/12(2006.01)

G10L 19/14(2006.01)

(56) 对比文件

JP 7-334195 A, 1995.12.22, 全文.

JP 3121094 B2, 2000.12.25, 全文.

JP 2001-128171 A, 2001.05.11, 全文.

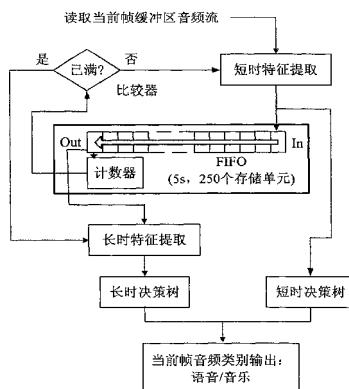
权利要求书 2 页 说明书 9 页 附图 2 页

(54) 发明名称

用于低比特率音频编码的语音 / 音乐检测器

(57) 摘要

本发明属于音频信号识别技术领域，其特征在于，在提取短时特征向量利用短时决策树进行语音信号和音乐信号判别后，还利用一个 FIFO 来进行长时特征向量的特征提取和长时决策树的判断。其中，短时特征向量包括短时能量函数和短时频谱分布函数；长时特征向量包括能量方差、能量过中值率、低能量比率、频谱分布变化率和低频谱分布比率；所述短时频谱分布参数是经过设定的多分辨率小波分析滤波器结合降采样来进行子带分解的。实验证明，本发明测试得到的平均误检率为 0.8%。



1. 用于低比特率音频编码的语音 / 音乐检测器, 其特征在于, 所述检测器是在一个数字集成电路上实现的, 包含如下 6 个模块 :

模块 (1), 短时特征提取 : 输入信号是当前帧缓冲区输出的以帧长为单位的音频流, 经过计算, 得到短时音频特征向量, 该短时音频特征向量包括两个分量, 短时能量函数  $E[n]$  与短时频谱分布参数 SP, 分别如下 (a)、(b) 所述 :

(a), 短时能量函数  $E[n]$  :

$$E[n] = \log_{10} \left( \sum_{n=1}^N (x[n])^2 \right),$$

$x[n]$  是离散化的输入音频信号,  $N$  是计算短时能量所取的音频信号片段的样点数,  $N = F_s \times$  帧长,  $F_s$  为音频采样率, 单位是 kHz, 帧长的单位是时间 ms ;

(b), 短时频谱分布参数 SP :

首先, 在设定的采样率下, 把每帧音频信号按设定的技术进行多分辨率子带分解, 得到频带由低到高的 1 级子带, 用  $1, 2, \dots, 1$  表示, 所述 1 级子带是通过阶数与设定级数相对应的 Daubechies 小波构建的分析滤波器组对原信号进行滤波后得到的, 其次, 按下式计算短时频谱分布参数 SP :

$$SP_{21}[n] = E_2[n] - E_1[n],$$

$$SP_{31}[n] = E_3[n] - E_1[n],$$

...

$$SP_{11}[n] = E_1[n] - E_1[n].$$

$E_1[n], E_2[n], \dots, E_1[n]$  分别为各子带的短时能量函数 ;

从而得到短时特征向量  $F_s[n]$  :

$$F_s[n] = (E[n], SP_{21}[n], SP_{31}[n], \dots, SP_{11}[n])^\top;$$

模块 (2), 先进先出存储器 : 即 FIFO, 顺次排列的若干存储单元, 所存储序列的长度单位是秒, 存储单元数 =  $\frac{\text{秒长}}{\text{帧长}} \times 1000$ , 该序列以帧为单位接收从短时特征提取模块输出的每帧的  $E_1[n], E_2[n], \dots, E_1[n]$  ;

模块 (3), 比较器 : 输入是 FIFO 中已占用的存储单元的数量, 即计数器的输出, 与预设的 FIFO 长度比较判断 FIFO 是否已满, 若 FIFO 未满, 该比较器便向所述短时特征提取模块输出允许短时特征输出的信号 ;

模块 (4), 长时特征提取 : 设有一个控制信号输入端, 接收允许输出长时特征的信号, 还有一个数据输入端, 从 FIFO 输入  $E_n[n-i]$ ,  $i = 0, 1, \dots, N-1$ ,  $i$  是用 FIFO 内采样点序号表示的帧长序号, 所述长时特征提取模块在接收到所述比较器输出的 FIFO 已满的信号后, 计算长时特征向量, 其中包括 :

(c), 能量方差  $Var_E[n]$  :

$$Var_E[n] = \frac{1}{N-1} \sum_{i=0}^{N-1} (E[n-i] - \bar{E}[n])^2,$$

$\bar{E}[n]$  为短时能量函数的平均值,

$$\bar{E}[n] = \frac{1}{N} \sum_{i=0}^{N-1} E[n-i].$$

(d), 能量过中值率  $CR_{E_{med}}$  :

$$CR_{E_{med}}[n] = \frac{1}{2} \sum_{i=0}^{N-2} (\text{sgn}(E[n-i] - E_{med}) - \text{sgn}(E[n-i-1] - E_{med})),$$

$E_{med}$  是短时能量函数的中值, 在  $E[n-N+1]E$  到  $E[n]$  之间选取,  $\text{sgn}(x)$  为符号函数,

$$\text{sgn}(x) = \begin{cases} 1, & \text{如果 } x \geq 0 \\ -1, & \text{如果 } x < 0 \end{cases};$$

(e), 低能量比率  $R_{E_{low}}$  :

$$R_{E_{low}}[n] = \frac{\sum_{i=0}^{N-1} (E[n-i] < E_{th})}{N},$$

$E_{th}$  为低能量阈值, 取 -3.7;

(f), 频谱分布变化率 SF :

$$SF[n] = \sum_{i=0}^{N-2} \|S[n-i] - S[n-i-1]\|,$$

$\|\cdot\|$  为 2 范数,  $\|x\| = x^T x$ ;

(g), 低频谱分布比率  $R_{S_{Plow}}$  :

$$R_{S_{Plow}}[n] = \frac{\sum_{i=0}^{N-1} (E_{UV}[n-i] < E_{UVth})}{N},$$

$E_{UV}[n]$  函数定义为:

$E_{UV}[n] = \log_{10}(\text{未取对数的清音部分对应子带的短时能量之和})$

$- \log_{10}(\text{未取对数的浊音部分对应子带的短时能量之和})$ ,

所述清音部分对应子带与浊音部分对应子带之间有一个共同的过渡区;

$E_{UVth}$  为低能量阈值, 取 -2.5;

从而得到长时特征的特征向量:

$$F_L[n] = (Var_E[n], CR_{E_{med}}[n], R_{E_{low}}[n], SF[n], R_{S_{Plow}}[n])^T;$$

模块 (5), 短时决策树:一个二值决策树, 判断从短时特征提取模块接收的短时特征向量是语音还是音乐信号的, 该决策树上各节点的阈值是预先通过对大量样本的训练得到的, 是已知值, 而且每一个节点用一个为该节点设定的上限阈值来判断一个短时特征分量, 满足阈值判断规则, 则沿着左侧树枝往下前进到下一个节点, 或遇到端点做出判断; 否则, 则沿着右侧的树枝往下前进到下一个节点, 或是遇到端点做出判断; 从而最后对是语音信号还是音乐信号来做出判断, 并输出;

模块 (6), 长时决策树:一个二值决策树, 判断从长时特征提取模块接收的长时特征向量是语音信号还是音乐信号的, 判断方法与短时决策树同。

## 用于低比特率音频编码的语音 / 音乐检测器

### 技术领域

[0001] 本发明涉及音频信号的处理和分类算法，及其计算机实现。本发明属于音频信号处理和模式识别领域。

### 背景技术

[0002] 传统的高质量音频编码通过时频变换，将音频信号变换到频域进行量化编码，并结合心理声学理论，实现信号冗余度的去除。这种方法对于所有的音频信号均使用类似的信号处理方法，没有对不同类型的信号区分对待。虽然在这些编码器中引入窗切换，用于改善瞬态信号的编码质量，但也没有更多地利用不同音频信号本身的特点。在移动通信的音频编码中，为了节省传输带宽和嵌入式实现，音频编码向低比特率、低复杂度方向发展。在这种情况下，已经不是高质量音频编码，使用传统高质量音频编码方案的问题逐渐凸现。对所有音频信号采用同样的信号处理方法，将导致在低比特率下音质的大大下降。因此，有必要对于不同类型的音频信号，充分挖掘它们的特殊性，分别构建适于各种类型音频信号的低比特率编码器，并把它们封装在一起。在实际编码过程中，首先对信号类型进行识别，然后调用对应的编码函数对其进行编码。

[0003] 从理论上讲，对于各种不同类型的音频信号，分别定制特殊的编码器，编码的效果必定很好。随着音频信号分类的细化，编码效果也将相应提高。然而，如果将音频类型设定得过多，在实现上也是不经济的。这必将导致类型的识别过于复杂，同时也增大了整个编码器的复杂度和存储空间占用。所以，通常将音频信号分为语音和音乐两类，使用低比特率的语音编码器（如 CELP）对语音信号进行编码，使用通用音频编码器（如 MPEG-AAC）对音乐信号进行编码。由于前者通过建立语音发声模型，充分利用了语音的特点，对于语音信号的编码效率很高，加之其技术已经相当成熟，故可以通过在通用音频编码器上扩展语音编码器使其语音编码质量得到很大提高。类似地，也可以通过在低比特率的语音编码器上扩展通用音频编码器使其宽带音乐的编码质量得到提高。必须同时客观地看到，这种类型识别的加入，势必增加编码器的复杂度。首先，类型识别算法将消耗 CPU 时间。同时，不同类型编码器的结合还引入了不同编码器之间的音频数据连接问题而引起处理上的麻烦。

[0004] 2005 年 3GPP 组织提出的超宽带自适应多速率音频编码器（AMR-WB+）正是基于低比特率 的语音编码器上扩展通用音频编码器的思想提出的。它是 AMR-WB 的宽带版本，其主要应用领域定位在第三代移动通讯设备。它的主要特点是工作在中低码率，并且有低的复杂度和延时。AMR-WB+ 是从语音编码器发展过来的。它结合参数编码和变换编码，支持 16/24/32/48kHz 的采样率，码率设定在 7.75kbps 到 54kbps 之间的范围，可以满足移动音频通讯的不同的质量要求。该编码器的重要特征，是它根据输入音频信号是语音还是音乐，用不同的方式进行编码，以在最大程度上减小码率，保证编码质量。AMR-WB+ 内部有两种编码模式，即基于代数码本激励线性预测语音编码器 ACELP (Algebraic Code Excited LinearPrediction) 和变换激励编码 TCX (Transform Coded Excitation)。两者有不同的适用范围。由于是时域预测编码器，ACELP 适合于语音和瞬态信号的编码。而 TCX 是变换

编码,因而更适合于典型音乐信号的编码。在该编码器方案中,根据输入信号的特点,合理地在两个编码模式之间进行选择,对于最终编码的效果具有重要的影响。

[0005] AMR-WB+ 标准设计了复杂的编码模式切换方案。其中包括闭环 (close-loop) 选择和开环 (open-loop) 选择。闭环选择通过试验的方法,选取最好的一种编码模式。它会分别调用 ACELP 和 TCX 编码函数对音频信号进行编码,并比较编码结果的平均信噪比,选取信噪比较高者作为最终的编码模式。显然,这样的选择方法是很准确的,但它运算量非常大。相反,开环选择直接通过分析音频信号的特征来选择编码模式,虽然精度低,但运算量减小了很多。所以,相比而言,开环模式选择在运算量上更有优势,更适用于移动通信的场合。而且通过合理选择特征和参数,开环模式选择同样也能达到较高的精度。

[0006] 遗憾的是,AMR-WB+ 语音 / 音乐检测器主要是基于单帧 (256 样点) 频谱的分布来进行的,准确度较低。这主要原因是它用到的单帧频谱分布特征等参数缺乏对语音 / 音乐足够的区分度。事实上,这种短时音频特征本身对于音频类型识别不具有足够的信息量。例如,给出一个 20ms 的信号波形,很难通过信号分析确定它是语音还是音乐。事实上,我们可以从人耳的听觉特性上获得新的方法。Balabko (1999) 指出,人耳识别出语音的关键是探测到信号频谱的缓慢变化,而不是瞬时的频谱值。通过频谱分析,可以发现,如果音频信号子带能量有低于 16Hz 的低频调制,人耳就有很大可能会把这种信号识别为语音。最早的数据在上世纪 30 年代就已经出现了 (Dudley, 1939)。有趣的是,人耳的听觉系统对于 4Hz 左右,也就是平均音节速率的调制频率最为敏感。人耳可以很容易地将该信号识别为语音,尽管可能这是一种听不懂的语言。基于以上分析,可以知道,通过对一段较长时间的音频信号低频调制频率的分析,可以实现较高精度的语音 / 音乐识别。而在这方面,短时音频特征无能为力,必须借助于长时特征。AMR-WB+ 标准中的特征除了单帧频谱的分布特征外,也有少量长时特征,但充其量也只用到了 12 个子带 4 帧和 16 帧 (100 ~ 300ms) 能量标准差,根本无法覆盖语音所特有的低频调制频率这一重要信息。所以引入长时音频特征,也就是处理音频数据对象长度 大于 1 秒的音频特征,是提高开环模式选择的必要手段。

[0007] 然而,长时特征的最大弱点在于它的计算延时。通常需要等待 1 秒以上缓冲足够的音频数据后才能得到长时特征的值。在音频编码中,实时性要求高,因而长时特征的使用受到很大限制。事实上,诸多长时特征已经广泛用于音频信息检索 (Audio Information Retrieval) 中。由于音频信息检索对于实时性要求不高,所以长时特征不仅未受限制反而倍受欢迎。所以,本发明提出把长时特征和短时特征结合起来,进行音频类型的识别,解决了长时特征的计算延时问题。本发明设置了一段 5 秒长的 FIFO 进行数据缓冲,通过包含当前帧在内的前 5 秒的音频数据来计算长时特征。只要 FIFO 是满的,就几乎没有计算延时。这时,检测器主要根据长时特征来进行音频类型识别。而当刚开始编码,FIFO 未满的时候,长时特征是无法计算,检测器就根据短时特征来进行音频类型识别,也没有延时。如上所述,检测器的精度由长时特征来保证。

[0008] 依据 FIFO 的操作原理,每编码一帧 PCM 音频数据, FIFO 中只压入一个新数据,并推出一个旧数据,而不是更新所有数据。所以,当 FIFO 中既有语音又有音乐的数据时,长时特征并不单纯地针对语音或音乐信号进行计算,不一定能提供有关当前帧是语音还是音乐的准确信息,可能会引起一些误判。这会出现在语音和音乐切换的过程中。但通常,音频信号不可能在语音和音乐之间以很高频率频繁切换,所以由于上述原因导致的检测误差是很

低的。这表明,本发明可以对一般的音频信号保持较高的分类精度。

[0009] 发明内容

[0010] 本发明的目的在于,提出一种适合低码率音频实时编码的语音 / 音乐检测器。

[0011] 本发明的特征在于,采用短时音频特征和长时音频特征相结合的方法,在维持低计算复杂度的基础上,获得较高的语音 / 音乐检测准确率。

[0012] 1、本发明的特征在于,所述检测器是在一个数字集成电路上实现的,包含如下 6 个模块:

[0013] 模块 (1),短时特征提取:输入信号是当前帧缓冲区输出的以帧长为单位的音频流,经过计算,得到短时音频特征向量,该短时音频特征向量包括两个分量,短时能量函数  $E[n]$  与短时频谱分布参数 SP,分别如下 (a)、(b) 所述:

[0014] (a),短时能量函数  $E[n]$ :

$$[0015] E[n] = \log_{10} \left( \sum_{n=1}^N (x[n])^2 \right),$$

[0016]  $x[n]$  是离散化的输入音频信号,  $N$  是计算短时能量所取的音频信号片段的样点数,  $N = F_s \times$  帧长,  $F_s$  为音频采样率,单位是 kHz,帧长的单位是时间 ms;

[0017] (b),短时频谱分布参数 SP:

[0018] 首先,在设定的采样率下,把每帧音频信号按设定的技术进行多分辨率子带分解,得到频带由低到高的 1 级子带,用 1, 2, ..., 1 表示,所述 1 级子带是通过阶数与设定级数相对应的 Daubechies 小波构建的分析滤波器组对原信号进行滤波后得到的。其次,按下式计算短时频谱分布参数 SP:

[0019]  $SP_{21}[n] = E_2[n] - E_1[n]$ ,

[0020]  $SP_{31}[n] = E_3[n] - E_1[n]$ ,

[0021] [0022]  $SP_{11}[n] = E_1[n] - E_1[n]$

[0022]  $E_1[n], E_2[n], \dots, E_1[n]$  分别为各子带的短时能量函数;

[0023] [0024] 从而,得到短时特征向量  $F_s[n]$ :

[0024]  $F_s[n] = (E[n], SP_{21}[n], SP_{31}[n], \dots, SP_{11}[n])^\top$ ;

[0025] 模块 (2),先进先出存储器:即 FIFO,顺次排列的若干存储单元,所存储序列的长度单位是秒, 存储单元数 =  $\frac{\text{秒长}}{\text{帧长}} \times 1000$ ,该序列以帧为单位接收从短时特征提取输出的每帧的  $E_1[n], E_2[n], \dots, E_1[n]$ ;

[0026] 模块 (3),比较器:输入是 FIFO 中已占用的存储单元的数量,即计数器的输出,与预设的 FIFO 长度比较判断 FIFO 是否已满,若 FIFO 未满,该比较器便向所述短时特征提取部分输出允许短时特征输出的信号;

[0027] 模块 (4),长时特征提取:设有一个控制信号输入端,接收允许输出长时特征的信号,还有一个数据输入端,从 FIFO 输入  $E_n[n-i]$ ,  $i = 0, 1, \dots, N-1$ ,  $i$  是用 FIFO 内采样点序号表示的帧长序号,所述长时特征提取部分在接收到所述比较器输出的 FIFO 已满的信号后,计算长时特征向量,其中包括:

[0028] (c),能量方差  $Var_E[n]$ :

[0029]  $Var_E[n] = \frac{1}{N-1} \sum_{i=0}^{N-1} (E[n-i] - \bar{E}[n])^2,$

[0030]  $\bar{E}[n]$  为短时能量函数的平均值,

[0031]  $\bar{E}[n] = \frac{1}{N} \sum_{i=0}^{N-1} E[n-i].$

[0032] (d), 能量过中值率  $CR_{Emed}$ :

[0033]  $CR_{Emed}[n] = \frac{1}{2} \sum_{i=0}^{N-2} (\text{sgn}(E[n-i] - E_{med}) - \text{sgn}(E[n-i-1] - E_{med})),$

[0034]  $E_{med}$  是短时能量函数的中值, 在  $E[n-N+1]$  到  $E[n]$  之间选取,  $\text{sgn}(x)$  为符号函数,

[0035] [0037]  $\text{sgn}(x) = \begin{cases} 1, & \text{如果 } x \geq 0 \\ -1, & \text{如果 } x < 0 \end{cases}.$

[0036] (e), 低能量比率  $R_{Elow}$ :

[0037] [0039]  $R_{Elow}[n] = \frac{\sum_{i=0}^{N-1} (E[n-i] < E_{th})}{N},$

[0038]  $E_{th}$  为低能量阈值, 取 -3.7;

[0039] (f), 频谱分布变化率 SF:

[0040]  $SF[n] = \sum_{i=0}^{N-2} \|S[n-i] - S[n-i-1]\|,$

[0041]  $\|\cdot\|$  为 2 范数,  $\|x\| = x^T x$ ;

[0042] (g), 低频谱分布比率  $R_{SPLow}$ :

[0043]  $R_{SPLow}[n] = \frac{\sum_{i=0}^{N-1} (E_{UV}[n-i] < E_{UVth})}{N},$

[0044]  $E_{UV}[n]$  函数定义为:

[0045]  $E_{UV}[n] = \log_{10}(\text{未取对数的清音部分对应子带的短时能量之和})$

[0046]  $-\log_{10}(\text{未取对数的浊音部分对应子带的短时能量之和})$

[0047] 所述清音部分对应子带与浊音部分对应子带之间有一个共同的过渡区;

[0048]  $E_{UVth}$  为低能量阈值, 取 -2.5;

[0049] 从而得到长时特征的特征向量:

[0050]  $F_L[n] = (Var_E[n], CR_{Emed}[n], R_{Elow}[n], SF[n], R_{SPLow}[n])^T;$

[0051] 模块 (5), 短时决策树:是一个二值决策树, 判断从短时特征提取部分接收的短时特征向量是语音还是音乐信号的, 该决策树上各节点的阈值是预先通过对大量样本的训练得到的, 是已知值, 而且每一个节点同一个为该节点设定的阈值来判断一个短时特征分量, 满足阈值判断规则, 则沿着左侧树枝往下前进到下一个节点, 或遇到端点做出判断; 否则, 则沿着右侧的树枝往下前进到下一个节点, 或是遇到端点做出判断; 从而最后对是语音信号还是音乐信号来做出判断, 并输出;

[0052] 模块 (6), 长时决策树:是一个二值决策树, 判断从长时特征提取部分接收的长时特征向量是语音信号还是音乐信号的, 判断方法与短时决策树同。

[0053] 本发明在自建的数据库上经过测试，具有较高的检测精度。本数据库的组成为：

[0054] 1. 语音数据库。

[0055] 本语音数据库共包含 55 个语音片断。其内容为正常语速的汉语朗读，每个片断的长度大约为 40 分钟，并由不同的人朗读。其中有 27 个片断为男声，28 个片断为女声。音频采样率为 16kHz，采样精度为 16bit。整个数据库包含的语音数据长度为 38 小时 33 分 14 秒。

[0056] 2. 音乐数据库。

[0057] 本音乐数据库共包含 693 首音乐片断。其内容涉及非常广泛的范围，囊括了各种音乐体裁和流派。其中包括：

[0058] 1) 中国民乐 14 首；

[0059] 2) 中文歌曲 184 首；

[0060] 3) 古典音乐 32 首；

[0061] 4) 歌剧戏剧 2 首；

[0062] 5) 英文歌曲 158 首；

[0063] 6) 环境音乐 16 首；

[0064] 7) 爵士乐 95 首；

[0065] 8) 现代器乐 26 首；

[0066] 9) 新世纪音乐 123 首；

[0067] 10) 摇滚乐 43 首。

[0068] 这些音乐片断的长度从 1 分钟到 10 多分钟不等。原始数据是 MP3 格式，44.1kHz 或 48kHz 采样，16bit 精度，立体声。为了使之与语音数据库匹配，并模拟移动通信的情况，将其转换成 PCM 格式，并降采样至 16kHz，保留 16bit 精度，下混至单声道。整个数据库包含的音乐数据长度为 47 小时 36 分 27 秒。

[0069] 决策树的学习通过选取上述数据库中 5% 的样本来进行。测试过程则面向整个数据库。测试结果如下：

[0070] 表 1 测试结果

[0071]

	总音频帧数	误检音频帧数	误检率
语音	6939680	90281	1.3%

[0072]

音乐	8569343	32498	0.4%
平均	15509023	122779	0.8%

[0073] 其中音频帧长度为 20ms，在 16kHz 的采样率下每帧样点数为 320。本检测器测试得到的平均误检率为 0.8%，非常令人满意。

## 附图说明

[0074] 图 1，本发明的处理框图；

[0075] 图 2，多分辨率子带分解示意图；

[0076] 图 3，短时特征决策树图；

[0077] 图 4，长时特征决策树。

[0078] 具体实现方式

[0079] 在图 1 中,缓冲区音频数据直接取自音频编码器输入的 PCM 数据缓冲区。其长度等于编码器输入音频信号的帧长。在 16kHz 采样率和 20ms 帧长的情况下,每帧 PCM 音频信号包含 320 个采样点。短时音频特征即基于此 320 个音频数据计算。

[0080] 图中的 FIFO 用于存储短时音频数据,为长时特征提取器提供输入数据。本发明的长时特征从包含当前帧在内的前 5 秒钟的短时特征中提取,在 20ms 帧长的情况下,该 FIFO 的长度为 250 个存储单元,每个存储单元存储一帧的短时音频特征,共计 250 个短时音频特征向量。FIFO 计数器和比较器用于切换分类所使用的特征。当 FIFO 未满时,检测器使用短时音频特征进行音频类型识别。当 FIFO 已满时,检测器使用长时音频特征进行音频类型识别。最终,检测器给出当前音频帧所属的类别。该结果是二值的:语音或音乐。

[0081] 本发明的特征还在于,通过反复实验,提取了几项突出反映语音和音乐信号特点的音频特征,对于语音和音乐具有较高的区分度。

[0082] 以下是本发明采用的 4 个短时音频特征。

[0083] 1. 短时能量函数 (E)。

[0084] 该特征直接描述音频响度随时间的变化。其计算方式为:

$$[0085] E[n] = \sum_{n=1}^N (x[n])^2,$$

[0086] 其中,  $x[n]$  是离散化的输入音频信号,N 是计算包含的音频信号样点数。N 的选择应满足使所包含的实际音频信号长度为音频编码器的一帧,例如 20ms。所以 N 与音频采样率  $F_s$  有关,在帧长为 20ms 的情况下,

$$[0087] N = F_s \times 20\text{ms}.$$

[0088] 由于人耳的听觉特性,响度不是与信号的幅度成正比,而是与信号的幅度成对数关系。所以,将上面计算的结果取对数,能更好地反映信号的响度:

$$[0089] E[n] = \log_{10} \left( \sum_{n=1}^N (x[n])^2 \right),$$

[0090] 在下文的短时能量函数中,如果不加特别说明,均是使用上式包含对数的  $E[n]$ 。

[0091] 2. 短时频谱分布参数 (SP)。

[0092] 为了描述信号的瞬时频谱分布,本发明使用了频谱分布参数。首先,将信号进行多分辨率子带分解。在 16kHz 的采样率下,有效的频带为 8kHz。通过 3 级分解,我们可以获得 4 个子带的信号  $x_1, x_2, x_3, x_4$ ,其频带分别为  $0 \sim 1\text{kHz}, 1 \sim 2\text{kHz}, 2 \sim 4\text{kHz}, 4 \sim 8\text{kHz}$ ,采样率分别为  $2\text{kHz}, 2\text{kHz}, 4\text{kHz}, 8\text{kHz}$ ,如图 2 所示。

[0093] 其中,HPF 是高通滤波器,LPF 为低通滤波器,↓ 为 2 倍降采样。两者是通过 5 阶 Daubechies 小波构建的分析滤波器对。然后,计算每个子带内的短时能量函数  $E_1[n], E_2[n], E_3[n], E_4[n]$ .

[0094] 定义如下短时频谱分布参数:

$$[0095] SP_{21}[n] = E_2[n] - E_1[n],$$

$$[0096] SP_{31}[n] = E_3[n] - E_1[n],$$

$$[0097] SP_{41}[n] = E_4[n] - E_1[n].$$

[0098] 这几个参数描述了信号短时频谱的粗略形状。

[0099] 上面的 4 个短时音频特征构成了短时特征的特征向量：

[0100]  $F_s[n] = (E[n], SP_{21}[n], SP_{31}[n], SP_{41}[n])^\top$ .

[0101] 短时特征提取的同时, 将未取对数的短时能量  $E[n]$  和未取对数的子带能量  $E_1[n]$ ,  $E_2[n]$ ,  $E_3[n]$ ,  $E_4[n]$  同时压入 FIFO 中, 用于计算长时特征。

[0102] 以下是本发明采用的 5 个长时音频特征。

[0103] 1. 能量方差 ( $Var_E$ )。

[0104] 由于语言自身的特点, 语音信号具有显著的响度跳变, 或称能量跳变, 即, 在字与字之间有停顿, 在句与句之间也有间歇。这些停顿和间歇在能量上均表现为很低的值, 而在发音时则能量较高。而且, 这一跳变是有一定频率范围的, 在一般的对话条件下其变化的频率一般在数赫兹范围内 (也就是上文说到的调制频率), 人们不会刻意拖长发音来改变这个频率。与此相反, 对于音乐信号, 一般通过乐器发声因而有所不同。由于乐器声音的持续性, 通常能量保持在一个较高的水平上, 出现很低能量的情况很少。即使是音乐中出现人声, 也总是存在一个宽带音频的背景音, 平滑了人声的响度变化。故而一般来说, 语音信号的能量变化比音乐信号大得多。能量方差就是一个从幅度上描述这种变化的音频特征。

[0105] 能量方差按下式计算：

$$[0106] Var_E[n] = \frac{1}{N-1} \sum_{i=0}^{N-1} (E[n-i] - \bar{E}[n])^2,$$

[0107] 其中  $E[n]$  为短时能量函数,  $N$  是计算包含的帧数,  $\bar{E}[n]$  为短时能量函数的平均值, 其计算公式为：

$$[0108] \bar{E}[n] = \frac{1}{N} \sum_{i=0}^{N-1} E[n-i].$$

[0109] 一般语音信号具有很高的能量标准差, 而音乐信号的能量标准差较低。

[0110] 2. 能量过中值率 ( $CR_{E_{med}}$ )。

[0111] 仅仅依靠能量标准差并不能完全描述语音信号的能量跳变。它只是通过从幅度上表达能量的变化, 却不能体现能量变化的频率。能量过中值率则是一个用于描述能量变化频率特性的有效音频特征。首先我们计算出能量的中值, 然后计算能量的过中值率。如果低能量帧和高能量帧频繁切换, 便会导致较高的能量过中值率。

[0112] 首先计算  $E[n-N+1]$  到  $E[n]$  之间, 短时能量函数的中值  $E_{med}$ 。

[0113] 然后计算过中值率：

$$[0114] CR_{E_{med}}[n] = \frac{1}{2} \sum_{i=0}^{N-2} (\text{sgn}(E[n-i] - E_{med}) - \text{sgn}(E[n-i-1] - E_{med})),$$

[0115] 其中,  $E[n]$  为短时能量函数,  $N$  是计算包含的帧数,  $\text{sgn}(x)$  为符号函数,

[0116]

$$\text{sgn}(x) = \begin{cases} 1, & \text{如果 } x \geq 0 \\ -1, & \text{如果 } x < 0 \end{cases}.$$

[0117] 实验证明, 语音信号的能量过中值率较低, 而音乐信号的能量过中值率较高。

[0118] 3. 低能量比率 ( $R_{E_{low}}$ )。

[0119] 原则上, 通过上面两个特征, 已经能够很好地描述语音信号能量跳变的特点。但是我们可以更进一步地挖掘语音信号的特点, 提取更多具有区分度的特征。事实上, 语音信号

不仅具有高能量跳变,而且其静音成分出现频繁,其对应于字与字、句与句之间的间歇。所以,通过统计在一定时间间隔内低能量帧出现的比率,可以有效地区分出语音和音乐信号。

[0120] 低能量比率通过下式计算:

$$[0121] R_{E\text{low}}[n] = \frac{\sum_{i=0}^{N-1} (E[n-i] < E_{\text{th}})}{N},$$

[0122] 其中  $E_{\text{th}}$  为低能量阈值,取为 -3.7, N 是计算包含的帧数。

[0123] 一般语音信号的低能量比率较高,而音乐信号的低能量比率较低。

[0124] 4. 频谱分布变化率 (SF)。

[0125] 该特征通过统计频谱分布的变化幅度,来区分语音和音乐。由于语音是各个音素的有序组合,而各个音素的频谱分布是不同的,所以在一段时间间隔内,其频谱变化的幅度较大。相反,对于音乐来说,虽然也有曲调和旋律的突变,但是也有很多部分是变化较为平缓的信号。这是其频谱分布变化率将较低。因此,采用这个特征也是有必要的。

[0126] 首先,将短时频谱分布参数组成向量:

$$[0127] S[n] = (E_1[n], E_2[n], E_3[n], E_4[n])^T,$$

[0128] 然后,按下式计算频谱分布变化率。

$$[0129] SF[n] = \sum_{i=0}^{N-2} \|S[n-i] - S[n-i-1]\|,$$

[0130] 其中,  $\|\cdot\|$  为 2 范数,  $\|x\| = x^T x$ , N 是计算包含的帧数。

[0131] 5. 低频谱分布比率 ( $R_{SP\text{low}}$ )。

[0132] 语音信号除了具有上述能量高低跳变外,还具有清音和浊音之间不断变化的特点。由于语言的特点,一般人们在说话时,清音和浊音是很频繁地切换的。所谓清音,就是发声时声带不振动的声音,其时域信号有较强的随机性,频谱较宽,具有噪声的性质。而浊音则是发声时声带振动的声音,时域信号更为规则,作傅立叶变换后能够获得具有谐波结构的频谱。利用这种谐波结构固然可以获得较高的浊音检测准确率,但是其运算复杂度较高,不适合用于实时音频编码。本发明采用了一种基于检测频谱能量分布变化的方法,来以较低的运算复杂度实现清音浊音切换的检测。

[0133] 首先计算能反映清音浊音各自频谱特点的函数。由于实验显示,清音的能量多集中在 2kHz 到 8kHz 区域,而浊音的能量多集中在 0 ~ 4kHz 区域,所以定义函数:

$$[0134] E_{UV}[n] = \log_{10}(E_2[n] + E_3[n] + E_4[n]) - \log_{10}(E_1[n] + E_2[n]),$$

[0135] 注意,此处的  $E_1[n]$ ,  $E_2[n]$ ,  $E_3[n]$ ,  $E_4[n]$  是 4 个子带信号的短时能量函数,没有经过取对数。然后计算低频谱分布比率:

$$[0136] R_{SP\text{low}}[n] = \frac{\sum_{i=0}^{N-1} (E_{UV}[n-i] < E_{UV\text{th}})}{N},$$

[0137] 其中  $E_{UV\text{th}}$  为低能量阈值,取为 -2.5, N 是计算包含的帧数。

[0138] 一般语音信号的低频谱分布比率较高,而音乐信号的低频谱分布比率较低。

[0139] 上面的 5 个长时音频特征构成了长时特征的特征向量:

$$[0140] F_L[n] = (Var_E[n], CR_{E\text{med}}[n], R_{E\text{low}}[n], SF[n], R_{SP\text{low}}[n])^T.$$

[0141] 本发明的特征还在于,采用了运算复杂度较低的决策树作为分类器。在音频实时

编码器中,如果采用较为复杂的模式分类器,如混合高斯模型、k-最近邻、人工神经网络、支持向量机等,虽然可以提高分类的精确度,但是随之带来的运算复杂度代价太大,是编码器所不能接受的。事实上,在编码器中追求过高的分类精度并没有太大意义,因为决定编码器效率最关键的还是编码模块。而采用一种简便的方法实现分类,并保证适当精度,则更能够适应编码器的需要。决策树能够很好地满足这样的需求。决策树学习完成后,在输入音频特征的情况下,在计算机上只需要用简单的逻辑判断就能实现分类。

[0142] 对于 4 个短时特征,通过对一定量样本的统计学习构建如下决策树,见图 3。

[0143] 其中,三角形表示树的节点,实心圆表示数的端点。在每个节点处,都有一个规则,格式为“ $\times k < \text{Thr}$ ”,表示的是特征向量中的第 k 个特征分量与设定阈值 Thr 之间比较大小。在每个端点处,都有一个值,M 或 S,是决策的结果。M 表示音乐,S 表示语音。当输入音频特征后,在每个节点处进行规则的判断,如果规则满足,则沿着左侧的树枝往下前进到下一个节点,或遇到端点(树枝的末端)做出判断;如果规则不满足,则沿着右侧的树枝往下前进到下一个节点,或遇到端点做出判断。

[0144] 对于长时音频特征向量,构建如下决策树,见图 4,判断方法与短时音频特征所用的决策树相同。

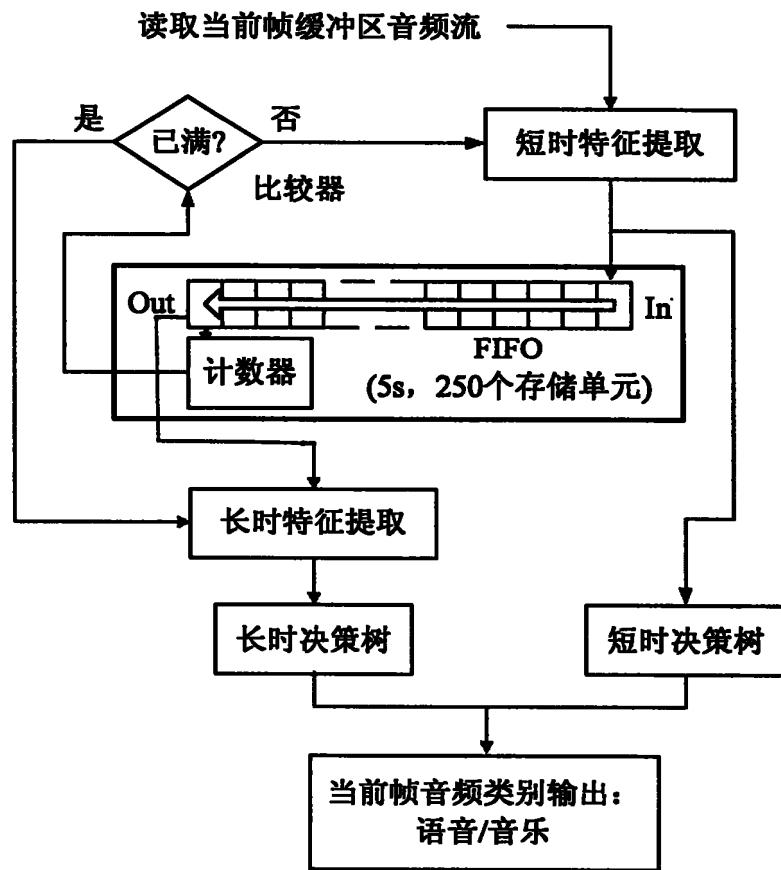


图 1

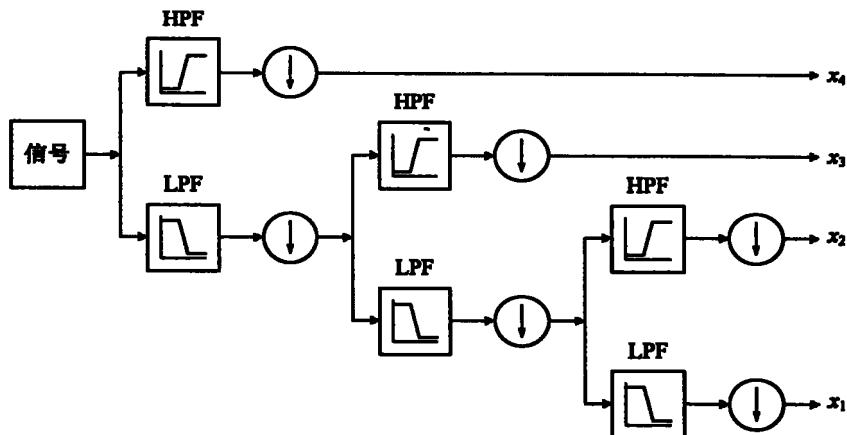


图 2

短时音频特征输入

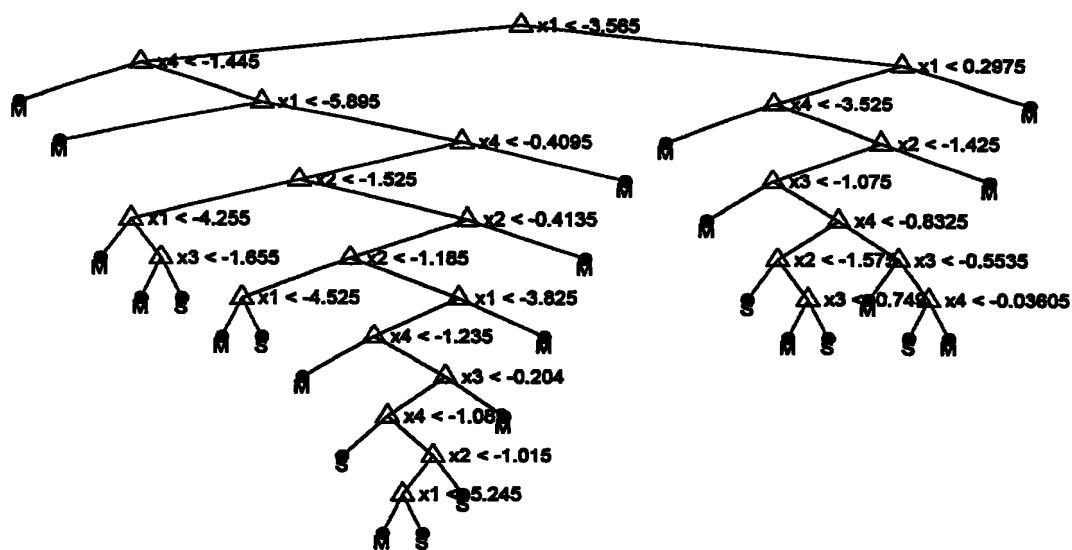


图 3

长时音频特征输入

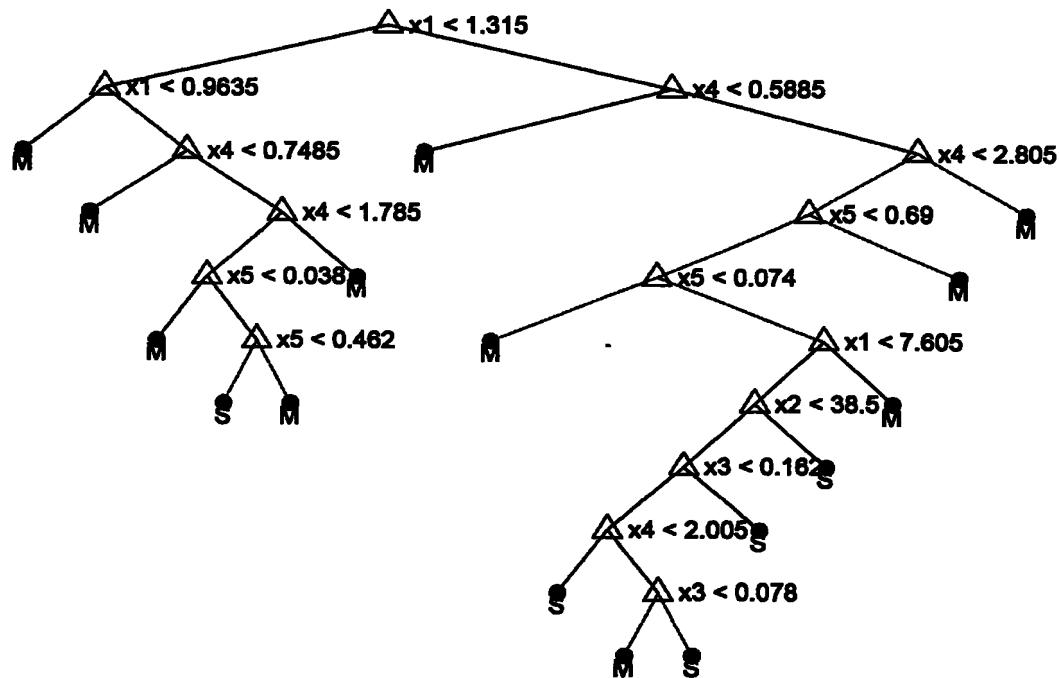


图 4