(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116438811 A
(43)申请公布日 2023.07.14

(21)申请号 202180071762.9

(22)申请日 2021.09.08

(30)优先权数据
63/075,984 2020.09.09 US

(85)PCT国际申请进入国家阶段日
2023.04.20

(86)PCT国际申请的申请数据
PCT/CA2021/051238 2021.09.08

(87)PCT国际申请的公布数据
WO2022/051846 EN 2022.03.17

(71)申请人 沃伊斯亚吉公司
地址 加拿大魁北克

(72)发明人 V·马列诺夫斯基 T·维兰考特

(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 胡琪

(51)Int.Cl.
H04S 7/00(2006.01)

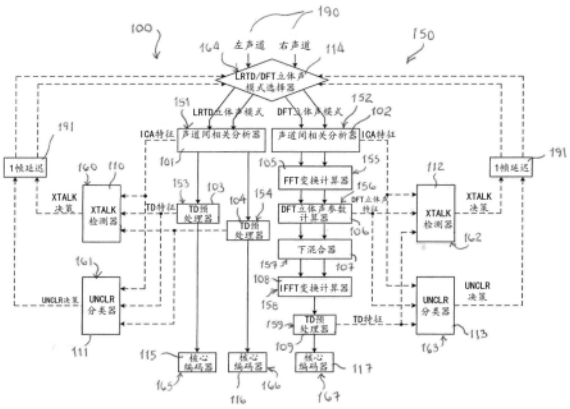
权利要求书17页 说明书40页 附图14页

(54)发明名称

用于声音编解码器中的非相关立体声内容的分类、串音检测和立体声模式选择的方法和设备

(57)摘要

本公开描述了输入立体声声音信号中的非相关立体声内容的分类(下文中称为“UNCLR分类”)和串音检测(下文中称为“XTALK检测”)。本公开还描述了立体声模式选择,例如自动LRTD/DFT立体声模式选择。另外,本公开使用所述分类以便选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码;响应于从包括左声道和右声道的立体声声音信号提取的特征来检测包括左声道和右声道的立体声声音信号中的串音;或者响应于从包括左声道和右声道的立体声声音信号提取的特征来对包括左声道和右声道的立体声声音信号中的非相关立体声内容进行分类。



CN 116438811 A

1.一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式以对包括左声道和右声道的立体声声音信号进行编码的设备,包括:

用于产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出的分类器;

用于产生指示在所述立体声声音信号中存在或不存在串音的第二输出的检测器;

用于计算辅助参数以用于选择用于对立体声声音信号进行编码的所述立体声模式的分析处理器;以及

用于响应于所述第一输出、所述第二输出和所述辅助参数来选择用于对立体声声音信号进行编码的所述立体声模式的立体声模式选择器。

2.如权利要求1所述的立体声模式选择设备,其中所述第一立体声模式是其中所述左声道和所述右声道被单独编码的时域立体声模式,并且所述第二立体声模式是频域立体声模式。

3.如权利要求1或2所述的立体声模式选择设备,其中在所述立体声声音信号的当前帧中,所述立体声模式选择器使用来自所述立体声声音信号的先前帧的所述第一输出和来自所述先前帧的所述第二输出。

4.如权利要求1至3中任一项所述的立体声模式选择设备,其中所述立体声模式选择器在所述第一立体声模式与所述第二立体声模式之间执行用于对所述立体声声音信号进行编码的所述立体声模式的初始选择。

5.如权利要求4所述的立体声模式选择设备,其中为了执行用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择,所述立体声模式选择器确定所述先前帧是否为语音帧。

6.如权利要求5所述的立体声模式选择设备,其中在用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择中,所述立体声模式选择器在所述立体声声音信号的每个帧中将用于对所述立体声声音信号进行编码的所述立体声模式初始化为在所述先前帧中选择的立体声模式。

7.如权利要求5或6所述的立体声模式选择设备,其中在所述立体声模式的所述初始选择中,如果(a)所述先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音,则所述立体声模式选择器选择所述第一立体声模式来用于对所述立体声声音信号进行编码。

8.如权利要求7所述的立体声模式选择设备,其中在用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择中,如果(i)条件(a)和(b)中的至少一个未被满足,并且(ii)在所述先前帧中选择的所述立体声模式是所述第二立体声模式,则所述立体声模式选择器选择所述第二立体声模式来用于对所述立体声声音信号进行编码。

9.如权利要求7或8所述的立体声模式选择设备,其中在所述立体声模式的所述初始选择中,如果(i)所述条件(a)和(b)中的至少一个未被满足,并且(ii)在所述先前帧中选择的所述立体声模式是所述第一立体声模式,则所述立体声模式选择器关于所述辅助参数中的一个辅助参数选择用于对所述立体声声音进行编码的所述立体声模式。

10.如权利要求9所述的立体声模式选择设备,其中所述一个辅助参数是辅助立体声模

式切换标志。

11.如权利要求4至7中任一项所述的立体声模式选择设备,其中在所述立体声模式的所述初始选择之后,如果数个给定条件被满足,则所述立体声模式选择器选择所述第二立体声模式来用于对所述立体声声音信号进行编码。

12.如权利要求11所述的立体声模式选择设备,其中所述给定条件包括以下条件中的至少一个:

-在所述立体声声音信号的所述先前帧中选择了所述第一立体声模式;

-在所述立体声声音信号的所述当前帧中初始选择了所述第一立体声模式;

-在所述当前帧中,所述检测器的所述第二输出指示在所述立体声声音信号中存在串音;

-(i)所述先前帧被确定为语音帧,并且(ii)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音;

-在所述先前帧中,使用所述第一立体声模式的连续帧的数量的计数值高于第一值;

-在所述先前帧中,使用所述第二立体声模式的连续帧的数量的计数值高于第二值;

-在所述先前帧中,所述立体声声音信号的类别在预定义类别集合内;以及

-(i)用于对所述立体声声音信号进行编码的总比特率等于或高于第三值,或者(ii)在所述先前帧中来自所述检测器的表示所述立体声声音信号中的串音的得分小于第四值。

13.如权利要求1至12中任一项所述的立体声模式选择设备,其中所述分析处理器计算辅助子模式标志以作为所述辅助参数中的一个辅助参数,所述辅助子模式标志指示所述第一立体声模式在子模式中操作,所述子模式应用于从所述第一立体声模式切换到所述第二立体声模式之前的短暂过渡。

14.如权利要求13所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的以下条件的帧中重置所述辅助子模式标志:(a)所述先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音。

15.如权利要求14所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的以下条件的帧中将所述辅助子模式标志重置为1:(1)由所述分析处理器计算的、作为辅助参数的辅助立体声模式切换标志等于1,(2)所述先前帧的所述立体声模式不是所述第一立体声模式,或者(3)使用所述第一立体声模式的帧的计数值小于给定值。

16.如权利要求15所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的其中条件(1)至(3)中均未被满足的帧中将所述辅助子模式标志重置为0。

17.如权利要求13至16中任一项所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的以下条件中的至少一个被满足的帧中不改变所述辅助子模式标志:(a)所述先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音。

18.如权利要求1至17中任一项所述的立体声模式选择设备,其中所述分析处理器包括使用所述第一立体声模式的连续帧的数量的计数值,以作为所述辅助参数中的一个辅助参

数。

19.如权利要求18所述的立体声模式选择设备,其中如果(a)所述先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音,则所述分析处理器递增使用所述第一立体声模式的连续帧的数量的所述计数值。

20.如权利要求18或19所述的立体声模式选择设备,其中如果在当前帧中所述立体声模式选择器选择所述第二立体声模式,则所述分析处理器将使用所述第一立体声模式的连续帧的数量的所述计数值重置为零。

21.如权利要求18至20中任一项所述的立体声模式选择设备,其中使用所述第一立体声模式的连续帧的数量的所述计数值被限制于上阈值。

22.如权利要求1至21中任一项所述的立体声模式选择设备,其中所述分析处理器包括使用所述第二立体声模式的连续帧的数量的计数值,以作为所述辅助参数中的一个辅助参数。

23.如权利要求22所述的立体声模式选择设备,其中如果在当前帧中所述第二立体声模式被选择,则所述分析处理器递增使用所述第二立体声模式的连续帧的数量的所述计数值。

24.如权利要求22或23所述的立体声模式选择设备,其中如果在当前帧中所述立体声模式选择器选择所述第一立体声模式,则所述分析处理器将使用所述第二立体声模式的连续帧的数量的所述计数值重置为零。

25.如权利要求22至24中任一项所述的立体声模式选择设备,其中使用所述第二立体声模式的连续帧的数量的所述计数值被限制于上阈值。

26.如权利要求1至25中任一项所述的立体声模式选择设备,其中所述分析处理器产生辅助立体声模式切换标志以作为所述辅助参数中的一个辅助参数。

27.如权利要求26所述的立体声模式选择设备,其中所述分析处理器(i)在(a)先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音的情况下,在当前帧中将所述辅助立体声模式切换标志初始化为1,并且(ii)在所述条件(a)和(b)中的至少一个未被满足时在当前帧中将所述辅助立体声模式切换标志初始化为0。

28.如权利要求26或27所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的所述左声道和所述右声道异相时将所述辅助立体声模式切换标志设置为0。

29.如权利要求10或15所述的立体声模式选择设备,其中所述分析处理器产生所述辅助立体声模式切换标志以作为所述辅助参数中的一个辅助参数。

30.如权利要求29所述的立体声模式选择设备,其中所述分析处理器(i)在(a)所述先前帧被确定为语音帧,并且(b)来自所述分类器的所述第一输出指示所述先前帧中存在非相关立体声内容,或者来自所述检测器的所述第二输出指示所述先前帧中的所述立体声声音信号中存在串音的情况下,在当前帧中将所述辅助立体声模式切换标志初始化为1,并且(ii)在所述条件(a)和(b)中的至少一个未被满足时在当前帧中将所述辅助立体声模式切换标志初始化为0。

31.如权利要求29或30所述的立体声模式选择设备,其中所述分析处理器在所述立体声声音信号的所述左声道和所述右声道异相时将所述辅助立体声模式切换标志设置为0。

32.如权利要求1至31中任一项所述的立体声模式选择设备,其中用于产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出的所述分类器包括如权利要求1至21中任一项所述的非相关立体声内容的所述分类器。

33.如权利要求1至32中任一项所述的立体声模式选择设备,其中用于产生指示在所述立体声声音信号中存在或不存在串音的第二输出的所述检测器包括如权利要求41至60中任一项所述的串音的所述检测器。

34.一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码的设备,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器实现:

用于产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出的分类器;

用于产生指示在所述立体声声音信号中存在或不存在串音的第二输出的检测器;

用于计算辅助参数以用于选择用于对立体声声音信号进行编码的所述立体声模式的分析处理器;以及

用于响应于所述第一输出、所述第二输出和所述辅助参数来选择用于对立体声声音信号进行编码的所述立体声模式的立体声模式选择器。

35.一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码的设备,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器:

产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出;

产生指示在所述立体声声音信号中存在或不存在串音的第二输出;

计算辅助参数以用于选择用于对立体声声音信号进行编码的所述立体声模式;以及

响应于所述第一输出、所述第二输出和所述辅助参数来选择用于对立体声声音信号进行编码的所述立体声模式。

36.一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码的方法,包括:

产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出;

产生指示在所述立体声声音信号中存在或不存在串音的第二输出;

计算辅助参数以用于选择用于对立体声声音信号进行编码的所述立体声模式;以及

响应于所述第一输出、所述第二输出和所述辅助参数来选择用于对立体声声音信号进行编码的所述立体声模式。

37.如权利要求36所述的立体声模式选择方法,其中所述第一立体声模式是其中所述左声道和所述右声道被单独编码的时域立体声模式,并且所述第二立体声模式是频域立体

声模式。

38.如权利要求36或37所述的立体声模式选择方法,其中在所述立体声声音信号的当前帧中,选择所述立体声模式包括使用来自所述立体声声音信号的先前帧的所述第一输出和来自所述先前帧的所述第二输出。

39.如权利要求36至38中任一项所述的立体声模式选择方法,其中选择所述立体声模式包括在所述第一立体声模式与所述第二立体声模式之间执行用于对所述立体声声音信号进行编码的所述立体声模式的初始选择。

40.如权利要求39所述的立体声模式选择方法,其中为了执行用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择,选择所述立体声模式包括确定所述先前帧是否为语音帧。

41.如权利要求40所述的立体声模式选择方法,其中在用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择中,选择所述立体声模式包括在所述立体声声音信号的每个帧中将用于对所述立体声声音信号进行编码的所述立体声模式初始化为在所述先前帧中选择的立体声模式。

42.如权利要求40或41所述的立体声模式选择方法,其中在所述立体声模式的所述初始选择中,选择所述立体声模式包括:如果(a)所述先前帧被确定为语音帧,并且(b)所述第一输出指示所述先前帧中存在非相关立体声内容,或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音,则选择所述第一立体声模式来用于对所述立体声声音信号进行编码。

43.如权利要求42所述的立体声模式选择方法,其中在用于对所述立体声声音信号进行编码的所述立体声模式的所述初始选择中,选择所述立体声模式包括:如果(i)条件(a)和(b)中的至少一个未被满足并且(ii)在所述先前帧中选择的所述立体声模式是所述第二立体声模式,则选择所述第二立体声模式来用于对所述立体声声音信号进行编码。

44.如权利要求42或43所述的立体声模式选择方法,其中在所述立体声模式的所述初始选择中,选择所述立体声模式包括:如果(i)所述条件(a)和(b)中的至少一个未被满足并且(ii)在所述先前帧中选择的所述立体声模式是所述第一立体声模式,则关于所述辅助参数中的一个辅助参数选择用于对所述立体声声音进行编码的所述立体声模式。

45.如权利要求44所述的立体声模式选择方法,其中所述一个辅助参数是辅助立体声模式切换标志。

46.如权利要求39至42中任一项所述的立体声模式选择方法,其中在所述立体声模式的所述初始选择之后,选择所述立体声模式包括:如果数个给定条件被满足,则选择所述第二立体声模式来用于对所述立体声声音信号进行编码。

47.如权利要求46所述的立体声模式选择方法,其中所述给定条件包括以下条件中的至少一个:

-在所述立体声声音信号的所述先前帧中选择了所述第一立体声模式;

-在所述立体声声音信号的所述当前帧中初始选择了所述第一立体声模式;

-在所述当前帧中,所述第二输出指示在所述立体声声音信号中存在串音;

-(i)所述先前帧被确定为语音帧,并且(ii)所述第一输出指示所述先前帧中存在非相关立体声内容,或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音;

-在所述先前帧中,使用所述第一立体声模式的连续帧的数量的计数值高于第一值;

-在所述先前帧中,使用所述第二立体声模式的连续帧的数量的计数值高于第二值;

-在所述先前帧中,所述立体声声音信号的类别在预定义类别集合内;以及

-(i)用于对所述立体声声音信号进行编码的总比特率等于或高于第三值,或者(ii)在所述先前帧中表示所述立体声声音信号中的串音的得分小于第四值。

48.如权利要求36至47中任一项所述的立体声模式选择方法,其中计算所述辅助参数包括计算辅助子模式标志以作为所述辅助参数中的一个辅助参数,所述辅助子模式标志指示所述第一立体声模式在子模式中操作,所述子模式应用于从所述第一立体声模式切换到所述第二立体声模式之前的短暂过渡。

49.如权利要求48所述的立体声模式选择方法,其中计算所述辅助参数包括在所述立体声声音信号的以下条件的帧中重置所述辅助子模式标志:(a)所述先前帧被确定为语音帧,并且(b)所述第一输出指示所述先前帧中存在非相关立体声内容,或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音。

50.如权利要求49所述的立体声模式选择方法,其中计算所述辅助参数包括在所述立体声声音信号的以下条件的帧中将所述辅助子模式标志重置为1:(1)作为辅助参数的辅助立体声模式切换标志等于1,(2)所述先前帧的所述立体声模式不是所述第一立体声模式,或者(3)使用所述第一立体声模式的帧的计数值小于给定值。

51.如权利要求50所述的立体声模式选择方法,其中计算所述辅助参数包括在所述立体声声音信号的其中条件(1)至(3)中均未被满足的帧中将所述辅助子模式标志重置为0。

52.如权利要求48至51中任一项所述的立体声模式选择方法,其中计算所述辅助参数包括在所述立体声声音信号的其中以下条件中的至少一个被满足的帧中不改变所述辅助子模式标志:(a)所述先前帧被确定为语音帧,并且(b)所述第一输出指示所述先前帧中存在非相关立体声内容,或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音。

53.如权利要求36至52中任一项所述的立体声模式选择方法,其中计算所述辅助参数包括计算使用所述第一立体声模式的连续帧的数量的计数值以作为所述辅助参数中的一个辅助参数。

54.如权利要求53所述的立体声模式选择方法,其中计算所述辅助参数包括:如果(a)所述先前帧被确定为语音帧,并且(b)所述第一输出指示所述先前帧中存在非相关立体声内容,或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音,则递增使用所述第一立体声模式的连续帧的数量的所述计数值。

55.如权利要求53或54所述的立体声模式选择方法,其中计算所述辅助参数包括:如果在当前帧中所述第二立体声模式被选择,则将使用所述第一立体声模式的连续帧的数量的所述计数值重置为零。

56.如权利要求53至55中任一项所述的立体声模式选择方法,包括将使用所述第一立体声模式的连续帧的数量的所述计数值限制于上阈值。

57.如权利要求36至56中任一项所述的立体声模式选择方法,其中计算所述辅助参数包括计算使用所述第二立体声模式的连续帧的数量的计数值以作为所述辅助参数中的一个辅助参数。

58.如权利要求57所述的立体声模式选择方法，其中计算所述辅助参数包括：如果在当前帧中所述第二立体声模式被选择，则递增使用所述第二立体声模式的连续帧的数量的所述计数值。

59.如权利要求57或58所述的立体声模式选择方法，其中计算所述辅助参数包括：如果在当前帧中所述立体声模式选择器选择所述第一立体声模式，则将使用所述第二立体声模式的连续帧的数量的所述计数值重置为零。

60.如权利要求57至59中任一项所述的立体声模式选择方法，包括将使用所述第二立体声模式的连续帧的数量的所述计数值限制于上阈值。

61.如权利要求36至60中任一项所述的立体声模式选择方法，其中计算所述辅助参数包括产生辅助立体声模式切换标志以作为所述辅助参数中的一个辅助参数。

62.如权利要求61所述的立体声模式选择方法，其中计算所述辅助参数包括：(i)在(a)先前帧被确定为语音帧，并且(b)所述第一输出指示所述先前帧中存在非相关立体声内容，或者所述第二输出指示所述先前帧中的所述立体声声音信号中存在串音的情况下，在当前帧中将所述辅助立体声模式切换标志初始化为1，并且(ii)在所述条件(a)和(b)中的至少一个未被满足时在当前帧中将所述辅助立体声模式切换标志初始化为0。

63.如权利要求61或62所述的立体声模式选择方法，其中计算所述辅助参数包括在所述立体声声音信号的所述左声道和所述右声道异相时将所述辅助立体声模式切换标志设置为0。

64.如权利要求45或50所述的立体声模式选择方法，其中计算所述辅助参数包括产生所述辅助立体声模式切换标志以作为所述辅助参数中的一个辅助参数。

65.如权利要求64所述的立体声模式选择方法，其中计算所述辅助参数包括：(i)在(a)所述先前帧被确定为语音帧，并且(b)所述第一输出指示在所述先前帧中存在非相关立体声内容，或者所述第二输出指示在所述先前帧中所述立体声声音信号中存在串音的情况下，在当前帧中将所述辅助立体声模式切换标志初始化为1，并且(ii)在所述条件(a)和(b)中的至少一个未被满足时在当前帧中将所述辅助立体声模式切换标志初始化为0。

66.如权利要求64或65所述的立体声模式选择方法，其中计算所述辅助参数包括在所述立体声声音信号的所述左声道和所述右声道异相时将所述辅助立体声模式切换标志设置为0。

67.如权利要求36至66中任一项所述的立体声模式选择方法，其中产生指示在所述立体声声音信号中存在或不存在非相关立体声内容的第一输出包括如权利要求22至40中任一项所述的用于对非相关立体声内容进行分类的方法。

68.如权利要求36至66中任一项所述的立体声模式选择方法，其中产生指示在所述立体声声音信号中存在或不存在串音的第二输出包括如权利要求61至78中任一项所述的用于检测串音的方法。

69.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的串音的检测器，包括：

响应于所述提取的特征的、表示所述立体声声音信号中的串音的得分的计算器；

用于检测所述立体声声音信号中的串音的辅助参数的计算器；以及

响应于所述串音的得分和所述辅助参数的、用于在指示在所述立体声声音信号中存在

串音的第一类别与指示在所述立体声声音信号中不存在串音的第二类别之间进行切换的类别切换机制。

70.如权利要求69所述的串音检测器,其中对串音的所述检测基于逻辑回归模型。

71.如权利要求69或70所述的串音检测器,其中在所述左声道和所述右声道被单独编码的时域立体声模式中,所述提取的特征包括以下特征中的至少一个:

-所述左声道中的FEC(帧擦除隐藏)类别与所述右声道中的FEC类别之间的差;

-所述左声道的最大自相关值与所述右声道的最大自相关值之间的差;

-所述左声道中的LSF(线谱频率)值的和与所述右声道中的LSF值的和之间的差;

-所述左声道与所述右声道之间的残差能量的差;

-所述左声道的相关图与所述右声道的相关图之间的差;

-所述左声道与所述右声道之间的噪声特性的差;

-所述左声道与所述右声道之间的非平稳性的差;

-所述左声道与所述右声道之间的频谱多样性的差;

-零滞后处的所述左声道和所述右声道的声道间相关函数的未归一化值;

-被计算为所述左声道和所述右声道的平均的单声道信号的能量与使用所述左声道与所述右声道之间的差计算的边信号的能量之间的比率;

-(a)所述左声道与所述单声道信号之间的点积以及所述右声道与所述单声道信号之间的点积的最大值与(b)所述左声道与所述单声道信号之间的所述点积以及所述右声道与所述单声道信号之间的所述点积中的最小值之间的差;

-零滞后处的所述左声道和所述右声道的声道间相关函数的值;

-所述声道间相关函数的演变;

-所述声道间相关函数的最大值的定位;

-所述声道间相关函数的最大值;

-所述左声道和所述单声道信号之间的所述点积与所述右声道和所述单声道信号之间的所述点积之间的差;以及

-所述边信号与所述单声道信号的所述能量之间的经平滑比率。

72.如权利要求69至71中任一项所述的串音检测器,包括每个提取的特征的归一化器,其中所述归一化器移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

73.如权利要求69至72中任一项所述的串音检测器,包括逻辑回归模型,在所逻辑回归模型述中输出被计算为所述提取的特征的线性组合。

74.如权利要求73所述的串音检测器,其中所述得分的计算器对所述逻辑回归模型的所述输出进行归一化。

75.如权利要求73或74所述的串音检测器,其中所述得分的计算器使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权,以产生表示所述立体声声音信号中的串音的所述得分。

76.如权利要求75所述的串音检测器,其中在对所述逻辑回归模型的所述输出进行加权之前,所述得分的计算器以反比例将所述当前帧的所述相对能量线性映射到给定区间。

77.如权利要求75或76所述的串音检测器,其中所述得分的计算器使用所述当前帧中

的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出，以产生表示所述立体声声音信号中的串音的经平滑得分。

78.如权利要求69或70所述的串音检测器，其中在频域立体声编码模式中，所述提取的特征包括以下特征中的至少一个：

-声道间电平差(ILD)增益；

-声道间相位差(IPD)增益；

-IPD旋转角度；

-表示所述左声道与所述右声道之间的相位差的预测增益；

-声道间相干性的均值能量；

-最大和最小声道内幅度积的比率；

-交叉声道频谱的总体幅值；

-具有相位差的广义交叉声道相关GCC-PHAT函数的最大值；

-所述GCC-PHAT函数的第一最高峰值和第二最高峰值的幅度之间的关系；

-所述GCC-PHAT函数的所述第二最高峰值的幅度；以及

-当前帧中的所述第二最高峰值的定位相对于先前帧中的所述第二最高峰值的所述定位的差。

79.如权利要求69、70和78中任一项所述的串音检测器，包括每个提取的特征的归一化器，其中所述归一化器移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

80.如权利要求69、70、78和79中任一项所述的串音检测器，包括逻辑回归模型，在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

81.如权利要求80所述的串音检测器，其中所述得分的计算器使用当前帧中的相对能量的上升沿来平滑所述逻辑回归模型的所述输出，以产生表示所述立体声声音信号中的串音的经平滑得分。

82.如权利要求69至81中任一项所述的串音检测器，其中所述类别切换机制产生二元状态输出，所述二元状态输出具有指示所述第一类别的第一值和指示所述第二类别的第二值。

83.如权利要求69至82中任一项所述的串音检测器，其中，所述类别切换机制将所述串音的得分和所述辅助参数与用于在所述第一类别与所述第二类别之间进行切换的给定值进行比较。

84.如权利要求69至83中任一项所述的串音检测器，其中，在所述左声道和所述右声道被单独编码的时域立体声编码模式中，所述辅助参数包括以下参数中的至少一个：

-所述立体声声音信号的所述左声道和所述右声道中的非相关立体声内容的分类器的输出；

-先前帧中的所述类别切换机制的输出，所述类别切换机制的输出是所述第一类别和所述第二类别中的一者；以及

-可以在立体声模式之间进行切换的帧的计数值。

85.如权利要求69至84中任一项所述的串音检测器，其中在频域立体声编码模式中，所述辅助参数包括以下参数中的至少一个：

-先前帧中的所述类别切换机制的输出,所述类别切换机制的输出是所述第一类别和所述第二类别中的一者;

-当前帧中的话音活动检测(VAD)标志;

-所述左声道和所述右声道的复交叉声道频谱的具有相位差的广义交叉声道相关GCC-PHAT函数的第一最高峰值和第二最高峰值的幅度;

-对应于所述GCC-PHAT函数的所述第一最高峰值和所述第二最高峰值的声道间时间差(ITD)定位;以及

-立体声信号静音标志。

86.如权利要求84所述的串音检测器,其中所述立体声模式包括时域立体声模式和频域立体声模式。

87.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的串音的检测器,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器实现:

响应于所述提取的特征的、表示所述立体声声音信号中的串音的得分的计算器;

用于检测所述立体声声音信号中的串音的辅助参数的计算器;以及

响应于所述串音的得分和所述辅助参数的、用于在指示在所述立体声声音信号中存在串音的第一类别与指示在所述立体声声音信号中不存在串音的第二类别之间进行切换的类别切换机制。

88.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的串音的检测器,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器:

响应于所述提取的特征来计算表示所述立体声声音信号中的串音的得分;

计算用于检测所述立体声声音信号中的串音的辅助参数;以及

响应于所述串音的得分和所述辅助参数而在指示在所述立体声声音信号中存在串音的第一类别与指示在所述立体声声音信号中不存在串音的第二类别之间进行切换。

89.一种用于响应于从包括左声道和右声道的立体声声音信号提取的特征来检测包括所述左声道和所述右声道的所述立体声声音信号中的串音的方法,包括:

响应于所述提取的特征来计算表示所述立体声声音信号中的串音的得分;

计算用于检测所述立体声声音信号中的串音的辅助参数;以及

响应于所述串音的得分和所述辅助参数而在指示在所述立体声声音信号中存在串音的第一类别与指示在所述立体声声音信号中不存在串音的第二类别之间进行切换。

90.如权利要求89所述的串音检测方法,其中串音的所述检测基于逻辑回归模型。

91.如权利要求89或90所述的串音检测方法,其中,在所述左声道和所述右声道被单独编码的时域立体声模式中,所述提取的特征包括以下特征中的至少一个:

-所述左声道中的FEC(帧擦除隐藏)类别与所述右声道中的FEC类别之间的差;

-所述左声道的最大自相关值与所述右声道的最大自相关值之间的差；

-所述左声道中的LSF(线谱频率)值的和与所述右声道中的LSF值的和之间的差；

-所述左声道与所述右声道之间的残差能量的差；

-所述左声道的相关图与所述右声道的相关图之间的差；

-所述左声道与所述右声道之间的噪声特性的差；

-所述左声道与所述右声道之间的非平稳性的差；

-所述左声道与所述右声道之间的频谱多样性的差；

-零滞后处的所述左声道和所述右声道的声道间相关函数的未归一化值；

-被计算为所述左声道和所述右声道的平均的单声道信号的能量与使用所述左声道与所述右声道之间的差计算的边信号的能量之间的比率；

-(a)所述左声道与所述单声道信号之间的点积以及所述右声道与所述单声道信号之间的点积中的最大值与(b)所述左声道与所述单声道信号之间的所述点积以及所述右声道与所述单声道信号之间的所述点积中的最小值之间的差；

-零滞后处的所述左声道和所述右声道的声道间相关函数的值；

-所述声道间相关函数的演变；

-所述声道间相关函数的最大值的定位；

-所述声道间相关函数的最大值；

-所述左声道和所述单声道信号之间的所述点积与所述右声道和所述单声道信号之间的所述点积之间的差;以及

-所述边信号与所述单声道信号的所述能量之间的经平滑比率。

92.如权利要求89至91中任一项所述的串音检测方法,包括对每个提取的特征进行归一化,其中对每个提取的特征进行归一化包括移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

93.如权利要求89至92中任一项所述的串音检测方法,包括使用逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

94.如权利要求93所述的串音检测方法,其中计算表示串音的得分包括对所述逻辑回归模型的所述输出进行归一化。

95.如权利要求93或94所述的串音检测方法,其中计算表示串音的得分包括使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权,以产生表示所述立体声声音信号中的串音的所述得分。

96.如权利要求95所述的串音检测方法,其中计算表示串音的所述得分包括:在对所述逻辑回归模型的所述输出进行加权之前,以反比例将所述当前帧的所述相对能量线性映射到给定区间。

97.如权利要求95或96所述的串音检测方法,其中计算表示串音的所述得分包括使用所述当前帧中的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出,以产生表示所述立体声声音信号中的串音的经平滑得分。

98.如权利要求89或90所述的串音检测方法,其中在频域立体声编码模式中,所述提取的特征包括以下特征中的至少一个:

-声道间电平差(ILD)增益；

-声道间相位差(IPD)增益；

-IPD旋转角度；

-表示所述左声道与所述右声道之间的相位差的预测增益；

-声道间相干性的均值能量；

-最大和最小声道内幅度积的比率；

-交叉声道频谱的总体幅值；

-具有相位差的广义交叉声道相关GCC-PHAT函数的最大值；

-所述GCC-PHAT函数的第一最高峰值和第二最高峰值的幅度之间的关系；

-所述GCC-PHAT函数的所述第二最高峰值的幅度;以及

-当前帧中的所述第二最高峰值的定位相对于先前帧中的所述第二最高峰值的所述定位的差。

99.如权利要求89、90和98中任一项所述的串音检测方法,包括对每个提取的特征进行归一化,其中对每个提取的特征进行归一化包括移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

100.如权利要求89、90、98和99中任一项所述的串音检测方法,包括使用逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

101.如权利要求100所述的串音检测方法,其中计算表示串音的所述得分包括使用当前帧中的相对能量的上升沿来平滑所述逻辑回归模型的所述输出,以产生表示所述立体声声音信号中的串音的经平滑得分。

102.如权利要求89至101中任一项所述的串音检测方法,其中在所述第一类别与所述第二类别之间进行切换包括产生二元状态输出,所述二元状态输出具有指示所述第一类别的第一值和指示所述第二类别的第二值。

103.如权利要求89至102中任一项所述的串音检测方法,其中在所述第一类别与所述第二类别之间进行切换包括将所述串音的得分和所述辅助参数与用于在所述第一类别与所述第二类别之间进行切换的给定值进行比较。

104.如权利要求89至103中任一项所述的串音检测方法,其中在所述左声道和所述右声道被单独编码的时域立体声编码模式中,所述辅助参数包括以下参数中的至少一个：

-所述立体声声音信号的所述左声道和所述右声道中的非相关立体声内容的分类器的输出；

-在所述第一类别与所述第二类别之间进行切换的输出,类别切换输出是所述第一类别和所述第二类别中的一者;以及

-可以在立体声模式之间进行切换的帧的计数值。

105.如权利要求89至104中任一项所述的串音检测方法,其中在频域立体声编码模式中,所述辅助参数包括以下参数中的至少一个：

-先前帧中的在所述第一类别与所述第二类别之间进行切换的输出,类别切换输出是所述第一类别和所述第二类别中的一者；

-当前帧中的话音活动检测(VAD)标志；

-所述左声道和所述右声道的复交叉声道频谱的具有相位差的广义交叉声道相关GCC-PHAT函数的第一最高峰值和第二最高峰值的幅度；

-对应于所述GCC-PHAT函数的所述第一最高峰值和所述第二最高峰值的声道间时间差(ITD)定位;以及

-立体声信号静音标志。

106.如权利要求104所述的串音检测方法,其中所述立体声模式包括时域立体声模式和频域立体声模式。

107.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的非相关立体声内容的分类器,包括:

响应于所述提取的特征的、表示所述立体声声音信号中的非相关立体声内容的得分的计算器;以及

响应于所述得分的、用于在指示所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二类别之间进行切换的类别切换机制。

108.如权利要求107所述的非相关立体声内容分类器,其中非相关立体声内容的所述分类基于逻辑回归模型。

109.如权利要求107或108所述的非相关立体声内容分类器,其中,在其中所述左声道和所述右声道被单独编码的时域立体声模式中,所述提取的特征包括以下特征中的至少一个:

-所述左声道和所述右声道的声道间互相关函数的最大值的定位;

-瞬时目标增益;

-零滞后处的声道间相关函数的绝对值的对数;

-对应于所述左声道与所述右声道之间的差的边信号与对应于所述左声道和所述右声道的平均的单声道信号之间的边比单声道能量比;

-(a)所述左声道与所述单声道信号之间的点积以及所述右声道与所述单声道信号之间的点积中的最大值与(b)所述左声道与所述单声道信号之间的所述点积以及所述右声道与所述单声道信号之间的所述点积中的最小值之间的差;

-所述左声道和所述单声道信号之间的所述点积与所述右声道和所述单声道信号之间的所述点积之间在对数域中的绝对差;

-交叉声道相关函数的零滞后值;以及

-所述声道间相关函数的演变。

110.如权利要求107至109中任一项所述的非相关立体声内容分类器,包括每个提取的特征的归一化器,其中所述归一化器移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

111.如权利要求107至110中任一项所述的非相关立体声内容分类器,包括逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

112.如权利要求111所述的非相关立体声内容分类器,其中所述得分的计算器使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权,以产生表示非相关立体声内容的所述得分。

113.如权利要求112所述的非相关立体声内容分类器,其中所述得分的计算器使用当前帧中的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出,以产生表示非相关

立体声内容的经平滑得分。

114.如权利要求107或108所述的非相关立体声内容分类器,其中在频域立体声编码模式中,所述提取的特征包括以下特征中的至少一个:

-声道间电平差(ILD)增益;

-声道间相位差(IPD)增益;

-以角度的形式表达声道间相位差(IPD)的IPD旋转角度;

-预测增益;

-声道间相干性的均值能量,所述均值能量表示未被所述声道间电平差(ILD)和所述声道间相位差(IPD)捕捉的所述左声道与所述右声道之间的差;

-最大和最小声道内幅度积的比率;

-交叉声道频谱幅值;以及

-具有相位差的广义交叉声道相关GCC-PHAT函数的最大值。

115.如权利要求114所述的非相关立体声内容分类器,包括每个提取的特征的归一化器,其中所述归一化器移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

116.如权利要求107、108、114和115中任一项所述的非相关立体声内容分类器,包括逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

117.如权利要求116所述的非相关立体声内容分类器,其中所述得分的计算器使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权,以产生表示非相关立体声内容的所述得分。

118.如权利要求117所述的非相关立体声内容分类器,其中所述得分的计算器使用所述当前帧中的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出,以产生表示非相关立体声内容的经平滑得分。

119.如权利要求107至118中任一项所述的非相关立体声内容分类器,其中所述类别切换机制产生二元状态输出,所述二元状态输出具有指示在所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一值以及指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二值。

120.如权利要求107至119中任一项所述的非相关立体声内容分类器,其中所述类别切换机制将所述得分与用于在所述第一类别与所述第二类别之间进行切换的给定值进行比较。

121.如权利要求107至120中任一项所述的非相关立体声内容分类器,包括可以在第一立体声模式与第二立体声模式之间进行切换的帧的计数值。

122.如权利要求121所述的非相关立体声内容分类器,其中所述第一立体声模式是所述左声道和所述右声道被单独编码的时域立体声模式,并且所述第二立体声模式是频域立体声模式。

123.如权利要求121或122所述的非相关立体声内容分类器,其中所述类别切换机制响应于所述得分和所述计数值两者,以用于在所述第一类别与所述第二类别之间进行切换。

124.如权利要求123所述的非相关立体内容分类器,其中所述得分来自当前帧,并且所述计数值来自先前帧。

125.如权利要求123或124所述的非相关立体声内容分类器,其中所述类别切换机制将所述得分和所述计数值两者与用于在所述第一类别与所述第二类别之间进行切换的给定值进行比较。

126.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的非相关立体声内容的分类器,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器实现:

响应于所述提取的特征的、表示所述立体声声音信号中的非相关立体声内容的得分的计算器;以及

响应于所述得分的、用于在指示所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二类别之间进行切换的类别切换机制。

127.一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括所述左声道和所述右声道的所述立体声声音信号中的非相关立体声内容的分类器,包括:

至少一个处理器;以及

存储器,所述存储器耦合到所述处理器并且包括非暂时性指令,所述非暂时性指令在被执行时使得所述处理器:

响应于所述提取的特征来计算表示所述立体声声音信号中的非相关立体声内容的得分;以及

响应于所述得分而在指示所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二类别之间进行切换。

128.一种用于响应于从包括左声道和右声道的立体声声音信号提取的特征来对包括所述左声道和所述右声道的所述立体声声音信号中的非相关立体声内容进行分类的方法,包括:

响应于所述提取的特征来计算表示所述立体声声音信号中的非相关立体声内容的得分;以及

响应于所述得分而在指示所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二类别之间进行切换。

129.如权利要求128所述的用于对非相关立体声内容进行分类的方法,其中对非相关立体声内容的所述分类基于逻辑回归模型。

130.如权利要求128或129所述的用于对非相关立体声内容进行分类的方法,其中在所述左声道和所述右声道被单独编码的时域立体声模式中,所述提取的特征包括以下特征中的至少一个:

-所述左声道和所述右声道的声道间互相关函数的最大值的定位;

-瞬时目标增益;

-零滞后处的声道间相关函数的绝对值的对数;

-对应于所述左声道与所述右声道之间的差的边信号与对应于所述左声道和所述右声道的平均的单声道信号之间的边比单声道能量比；

-(a)所述左声道与所述单声道信号之间的点积以及所述右声道与所述单声道信号之间的点积中的最大值与(b)所述左声道与所述单声道信号之间的所述点积以及所述右声道与所述单声道信号之间的所述点积中的最小值之间的差；

-所述左声道和所述单声道信号之间的所述点积与所述右声道和所述单声道信号之间的所述点积之间在对数域中的绝对差；

-交叉声道相关函数的零滞后值;以及

-所述声道间相关函数的演变。

131.如权利要求128至130中任一项所述的用于对非相关立体声内容进行分类的方法，包括对每个提取的特征进行归一化,其中对每个提取的特征进行归一化包括移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

132.如权利要求128至131中任一项所述的用于对非相关立体声内容进行分类的方法，包括使用逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

133.如权利要求132所述的用于对非相关立体声内容进行分类的方法,其中计算表示非相关立体声内容的所述得分包括使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权,以产生表示非相关立体声内容的所述得分。

134.如权利要求133所述的用于对非相关立体声内容进行分类的方法,其中计算表示非相关立体声内容的所述得分包括:使用所述当前帧中的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出,以产生表示非相关立体声内容的经平滑得分。

135.如权利要求128或129所述的用于对非相关立体声内容进行分类的方法,其中在频域立体声编码模式中,所述提取的特征包括以下特征中的至少一个：

-声道间电平差(ILD)增益；

-声道间相位差(IPD)增益；

-以角度的形式表达声道间相位差(IPD)的IPD旋转角度；

-预测增益；

-声道间相干性的均值能量,所述均值能量表示未被所述声道间电平差(ILD)和所述声道间相位差(IPD)捕捉的所述左声道与所述右声道之间的差；

-最大和最小声道内幅度积的比率；

-交叉声道频谱幅值;以及

-具有相位差的广义交叉声道相关GCC-PHAT函数的最大值。

136.如权利要求135所述的用于对非相关立体声内容进行分类的方法,包括对每个提取的特征进行归一化,其中对每个提取的特征进行归一化包括移除所述提取的特征的均值并且将所述提取的特征缩放至所述提取的特征的单位方差。

137.如权利要求128、129、135和136中任一项所述的用于对非相关立体声内容进行分类的方法,包括使用逻辑回归模型,在所述逻辑回归模型中输出被计算为所述提取的特征的线性组合。

138.如权利要求137所述的用于对非相关立体声内容进行分类的方法,其中计算表示

非相关立体声内容的所述得分包括使用当前帧的相对能量对所述逻辑回归模型的所述输出进行加权，以产生表示非相关立体声内容的所述得分。

139.如权利要求138所述的用于对非相关立体声内容进行分类的方法，其中计算表示非相关立体声内容的所述得分包括使用所述当前帧中的所述相对能量的上升沿来平滑所述逻辑回归模型的加权输出，以产生表示非相关立体声内容的经平滑得分。

140.如权利要求128至139中任一项所述的用于对非相关立体声内容进行分类的方法，其中在所述第一类别与所述第二类别之间进行切换包括产生二元状态输出，所述二元状态输出具有指示所述立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一值，以及指示所述非相关立体声内容和所述相关立体声内容中的另一者的第二值。

141.如权利要求128至140中任一项所述的用于对非相关立体声内容进行分类的方法，其中在所述第一类别与所述第二类别之间进行切换包括将所述得分与给定值进行比较。

142.如权利要求128至141中任一项所述的用于对非相关立体声内容进行分类的方法，包括可以在第一立体声模式与第二立体声模式之间进行切换的帧的计数值。

143.如权利要求142所述的用于对非相关立体声内容进行分类的方法，其中所述第一立体声模式是其中所述左声道和所述右声道被单独编码的时域立体声模式，并且所述第二立体声模式是频域立体声模式。

144.如权利要求142或143所述的用于对非相关立体声内容进行分类的方法，其中在所述第一类别与所述第二类别之间进行切换响应于所述得分和所述计数值两者。

145.如权利要求144所述的用于对非相关立体声内容进行分类的方法，其中所述得分来自当前帧，并且所述计数值来自先前帧。

146.如权利要求144或145所述的用于对非相关立体声内容进行分类的方法，其中在所述第一类别与所述第二类别之间进行切换包括将所述得分和所述计数值两者与用于在所述第一类别与所述第二类别之间进行切换的给定值进行比较。

用于声音编解码器中的非相关立体声内容的分类、串音检测和立体声模式选择的方法和设备

技术领域

[0001] 本公开涉及声音编码，具体地但不排他地涉及例如多声道声音编解码器中的非相关立体声内容的分类、串音检测和立体声模式选择，该多声道声音编解码器能够以低比特率和低延迟在复杂音频场景中产生良好的声音质量。

[0002] 在本公开和所附权利要求中：

[0003] -术语“声音”可以与语音、音频和任何其他声音相关；

[0004] -术语“立体声(stereo)”是“立体声(stereophonic)”的缩写；以及

[0005] -术语“单声道(mono)”是“单声道(monophonic)”的缩写。

背景技术

[0006] 历史上，会话式电话是利用仅具有一个换能器以仅向用户耳朵中的一者输出声音的手持机来实现的。在最近十年中，用户已开始结合头戴式耳机使用其便携式手持机以经由其两个耳朵接收声音，主要用于收听音乐，但有时也收听语音。然而，当便携式手持机被用于发送和接收会话式语音时，内容仍为单声道的，但在头戴式耳机被使用时被呈现给用户的两只耳朵。

[0007] 利用如参考文献[1](其全部内容通过引用并入本文)中描述的最新的3GPP语音编码标准EVS(增强型话音服务)，通过便携式手持机发送和接收的经编码声音(例如语音和/或音频)的质量已经被显著地改进。下一个自然步骤是发送立体声信息，使得接收器尽可能接近在通信链路的另一端捕捉的真实生活音频场景。

[0008] 在音频编解码器中，例如在参考文献[2](其全部内容通过引用并入本文)中描述的，立体声信息的发送常被使用。

[0009] 对于会话式语音编解码器，单声道信号是常态。当立体声声音信号被发送时，比特率常常会加倍，因为立体声声音信号的左声道和右声道均是使用单声道编解码器来编码的。这在大多数情况下工作良好，但是存在使比特率加倍并且不能利用两个声道(立体声声音信号的左声道和右声道)之间的任何潜在冗余的缺点。此外，为了将整体比特率保持在合理的水平，针对左声道和右声道中的每一个使用非常低的比特率，从而影响了总体音质。为了降低比特率，高效立体声编码技术已经被开发和使用。作为非限制性示例，在下面的段落中讨论可以在低比特率下高效使用的两种立体声编码技术。

[0010] 第一种立体声编码技术被称为参数立体声。参数立体声使用公共单声道编解码器将两个输入(左声道和右声道)编码为单声道信号加上表示立体声图像的一定量的立体声边信息(对应于立体声参数)。两个输入左声道和右声道被下混合(down-mixed)成单声道信号，并且立体声参数随后被计算。这通常是在频域(FD)中(例如在离散傅里叶变换(DFT)域中)执行的。立体声参数与所谓的双耳或声道间线索有关。双耳线索(参见例如参考文献[3]，其全部内容通过引用并入本文)包括耳间电平差(ILD)、耳间时间差(ITD)和耳间相关(IC)。取决于声音信号特性、立体声场景配置等，某些或全部双耳线索被编码并被发送到解

码器。关于哪些双耳线索被编码和发送的信息作为信令信息被发出，其通常是立体声边信息的部分。而且，给定的双耳线索可以使用不同的编码技术来进行量化，这导致可变数量的比特被使用。随后，除了经量化双耳线索之外，立体声边信息通常可以在中比特率和更高比特率处包含由下混合产生的经量化残差信号。残差信号可以是使用熵编码技术(例如，算术编码器)来编码的。在本公开的剩余部分中，参数立体声将被称为“DFT立体声”，因为参数立体声编码技术通常在频域中操作，并且本公开将描述使用DFT的非限制性实施例。

[0011] 另一立体声编码技术是在时域中操作的技术。此立体声编码技术将两个输入(左声道和右声道)混合为所谓的主声道和辅声道。例如，按照参考文献[4](其全部内容通过引用并入本文)中描述的方法，时域混合可以基于混合比率，该混合比率确定两个输入(左声道和右声道)在产生主声道和辅声道时的相应贡献。混合比率是从若干度量导出的，该度量例如两个输入(左声道和右声道)相对于单声道信号的归一化相关或两个输入(左声道和右声道)之间的长期相关差。主声道可以由公共单声道编解码器编码，而辅声道可以由较低比特率编解码器编码。辅声道的编码可以利用主声道与辅声道之间的相干性并且可以重用来自主声道的某些参数。在左声道和右声道表现出很小相关性的某些声音中，最好在时域中单独地或以最小的声道间参数化对立体声输入信号的左声道和右声道进行编码。编码器中的此类办法是时域TD立体声的特殊情况，并且贯穿本公开将被称为“LRTD立体声”。

[0012] 此外，在过去几年中，音频的生成、记录、表示、编码、发送和再现正朝着增强的、交互的和沉浸式的收听者体验发展。例如，沉浸式体验可以被描述为当声音来自所有方向时深度参与或陷入到声音场景中的状态。在沉浸式音频(也被称为3D(三维)音频)中，考虑到各种声音特性，如音色、方向性、混响、透明度和(听觉)空间感的准确性，声音图像被再现于在收听者周围的所有三个维度中。沉浸式音频被产生用于特定的声音回放或再现系统，诸如基于扬声器的系统、集成再现系统(条形音箱)或头戴式耳机。随后，声音再现系统的交互性可以包括，例如，调整声级、改变声音的定位或选择用于再现的不同语言的能力。

[0013] 存在三种实现沉浸式体验的基本办法。

[0014] 实现沉浸式体验的第一种办法是基于声道的音频办法，其使用多个间隔开的麦克风从不同方向捕捉声音，其中一个麦克风对应于特定扬声器布局中的一个音频声道。每个记录的声道随后被提供给给定位置的扬声器。基于声道的音频办法的示例是例如立体声、5.1环绕、5.1+4等。

[0015] 实现沉浸式体验的第二种办法是基于场景的音频办法，其通过维度分量的组合将本地空间上的期望声场表示为时间的函数。表示基于场景的音频的声音信号独立于音频源的定位，同时声场被变换成呈现器处的扬声器的所选布局。基于场景的音频的示例是环境立体声(“ambisonics”)。

[0016] 实现沉浸式体验的第三种办法是基于对象的音频办法，其将听觉场景表示为各个音频元素(例如，歌手、鼓、吉他等)的集合，该集合伴随有诸如它们的定位之类的信息，因此它们可以由声音再现系统在它们预期的位置处被呈现。这给予基于对象的音频办法很大的灵活性和交互性，因为每个对象保持离散并且可以被单独操纵。

[0017] 上面描述的用于实现沉浸式体验的音频办法中的每一个都呈现出优点和缺点。因此，通常不是只有一种音频办法，而是将若干音频办法组合在复杂的音频系统中以创建沉浸式的听觉场景。示例可以是将基于场景或基于声道的音频与基于对象的音频组合的音频

系统,例如具有若干离散音频对象的环境立体声。

[0018] 近年来,3GPP(第三代合作伙伴计划)开始致力于基于EVS编解码器(参见参考文献[5],其全部内容通过引用并入本文)开发被称为IVAS(沉浸式话音和音频服务)的、用于沉浸式服务的3D(三维)声音编解码器。

[0019] DFT立体声模式对于对单音("single-talk")发声进行编码是高效的。在两个或更多个讲话者的情况下,参数立体声技术难以完全描述场景的空间性质。当两个谈话者同时进行谈话时(串音("cross-talk")场景)以及当立体声输入信号的左声道和右声道中的信号弱相关或完全非相关时,该问题尤其明显。在那种情形中,最好是使用LRTD立体声模式在时域中单独地或以最小的声道间参数化对立体声输入信号的左声道和右声道进行编码。随着在立体声输入信号中捕捉的场景演变,期望的是基于立体声场景分类在DFT立体声模式和LRTD立体声模式之间进行切换。

发明内容

[0020] 根据第一方面,本公开涉及一种用于响应于从包括左声道和右声道的立体声声音信号提取的特征来对包括左声道和右声道的立体声声音信号中的非相关立体声内容进行分类的方法,包括:响应于提取的特征来计算表示立体声声音信号中的非相关立体声内容的得分:以及响应于得分,在指示立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示非相关立体声内容和相关立体声内容中的另一者的第二类别之间进行切换。

[0021] 根据第二方面,本公开提供一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括左声道和右声道的立体声声音信号中的非相关立体声内容的分类器,包括:响应于提取的特征的、表示立体声声音信号中的非相关立体声内容的得分的计算器;以及类别切换机制,其响应于得分,用于在指示立体声声音信号中的非相关立体声内容和相关立体声内容中的一者的第一类别与指示非相关立体声内容和相关立体声内容中的另一者的第二类别之间进行切换。

[0022] 本公开还涉及一种用于响应于从包括左声道和右声道的立体声声音信号提取的特征来检测包括左声道和右声道的立体声声音信号中的串音的方法,包括:响应于提取的特征来计算表示立体声声音信号中的串音的得分;计算辅助参数以用于检测立体声声音信号中的串音;以及响应于串音的得分和辅助参数而在指示在立体声声音信号中存在串音的第一类别与指示在立体声声音信号中不存在串音的第二类别之间进行切换。

[0023] 根据又一方面,本公开提供一种响应于从包括左声道和右声道的立体声声音信号提取的特征的、包括左声道和右声道的立体声声音信号中的串音的检测器,包括:响应于提取的特征的、表示立体声声音信号中的串音的得分的计算器;用于检测立体声声音信号中的串音的辅助参数的计算器;以及响应于串音的得分和辅助参数的、用于在指示立体声声音信号中存在串音的第一类别与指示立体声声音信号中不存在串音的第二类别之间进行切换的类别切换机制。

[0024] 本公开还涉及一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码的方法,包括:产生指示在立体声声音信号中存在或不存在非相关立体声内容的第一输出;产生指示在立体声声音信

号中存在或不存在串音的第二输出;计算辅助参数以用于选择用于对立体声声音信号进行编码的立体声模式;以及响应于第一输出、第二输出和辅助参数来选择用于对立体声声音信号进行编码的立体声模式。

[0025] 根据又一方面,本公开提供一种用于选择第一立体声模式和第二立体声模式中的一个立体声模式来用于对包括左声道和右声道的立体声声音信号进行编码的设备,包括:分类器,用于产生指示在立体声声音信号中存在或不存在非相关立体声内容的第一输出;检测器,用于产生指示在立体声声音信号中存在或不存在串音的第二输出;分析处理器,用于计算辅助参数以用于选择用于对立体声声音信号进行编码的立体声模式;以及立体声模式选择器,用于响应于第一输出、第二输出和辅助参数来选择用于对立体声声音信号进行编码的立体声模式。

[0026] 非相关立体声内容分类器和分类方法、串音检测器和检测方法、以及立体声模式选择设备和方法的前述和其他目的、优点和特征将在阅读其说明性实施例的以下非限制性描述后变得更加明显,这些非限制性描述仅仅是参考附图以示例的方式给出的。

附图说明

[0027] 在附图中:

[0028] 图1是同时图示出用于对立体声声音信号进行编码的设备和用于对立体声声音信号进行编码的对应方法的示意性框图;

[0029] 图2是示出串音场景的平面图的示意图,其中两个相对的讲话者由一对超心形麦克风捕捉;

[0030] 图3是示出GCC-PHAT函数中的峰值的位置的图;

[0031] 图4是用于真实记录的立体场景设置的俯视图;

[0032] 图5是图示出在LRTD立体声模式中的非相关立体声内容的分类中应用于LogReg模型的输出的归一化函数的图;

[0033] 图6是示出形成图1的用于对立体声声音信号进行编码的设备的部分的非相关立体声内容的分类器中的立体声内容类别之间的切换机制的状态机图;

[0034] 图7是具有AB麦克风设置的大型会议室的示意性平面图,其条件被模拟用于串音检测,其中AB麦克风由一对分开放置的心形或全向麦克风组成,被放置为使得它们覆盖空间而不产生彼此的相位问题;

[0035] 图8是图示出使用VAD(话音活动检测)对串音样本的自动标记的图;

[0036] 图9是表示在LRTD立体声模式中的串音检测中用于缩放LogReg模型的原始输出的函数的图;

[0037] 图10是图示出形成用于在LRTD立体声模式中对立体声声音信号进行编码的图1的设备的部分的串音检测器中的检测上升沿的机制的图;

[0038] 图11是图示出在LRTD立体声模式中的串音检测器的输出的状态之间进行切换的机制的逻辑图;

[0039] 图12是图示出在DFT立体声模式中的串音检测器的输出的状态之间进行切换的机制的逻辑图;

[0040] 图13是图示出在LRTD和DFT立体声模式之间进行选择的机制的示意性框图;以及

[0041] 图14是实现用于对立体声声音信号进行编码的方法和设备的硬件组件的示例配置的简化框图。

具体实施方式

[0042] 本公开描述了输入立体声声音信号中的非相关立体声内容的分类(下文中称为“UNCLR分类”)和串音检测(下文中称为“XTALK检测”)。本公开还描述了立体声模式选择,例如自动LRTD/DFT立体声模式选择。

[0043] 图1是同时图示出用于对立体声声音信号190进行编码的设备100和用于对立体声声音信号190进行编码的对应方法150的示意性框图。

[0044] 具体地,图1示出了UNCLR分类、XTALK检测和立体声模式选择如何被集成到立体声声音信号编码方法150和设备100中。

[0045] UNCLR分类和XTALK检测形成两个独立的技术。然而,它们基于相同的统计模型并且共享某些特征和参数。此外,UNCLR分类和XTALK检测两者是针对LRTD立体声模式和DFT立体声模式单独设计和训练的。在本公开中,给出LRTD立体声模式作为时域立体声模式的非限制性示例,并且给出DFT立体声模式作为频域立体声模式的非限制性示例。实现其他时域和频域立体声模式也在本公开的范围内。

[0046] UNCLR分类对从立体声声音信号190的左声道和右声道提取的特征进行分析,并且检测左声道与右声道之间的弱相关或零相关。另一方面,XTALK检测对在立体声场景中同时讲话的两个讲话者的存在进行检测。例如,UNCLR分类和XTALK检测两者都提供二元输出。这些二元输出在立体声模式选择逻辑中被组合在一起。作为非限制性的一般规则,当UNCLR分类和XTALK检测指示站在捕捉设备(例如麦克风)的相对侧的两个讲话者的存在时,立体声模式选择选择LRTD立体声模式。这种情形通常导致立体声声音信号190的左声道与右声道之间的弱相关。LRTD立体声模式或DFT立体声模式的选择是在逐帧的基础上执行的(如本领域所公知的,立体声声音信号190以给定的采样速率被采样,并且按照被划分为多个“子帧”的被称为“帧”的这些样本组来处理)。此外,立体声模式选择逻辑被设计成避免在LRTD与DFT立体声模式之间的频繁切换以及在感知上重要的信号片段内的立体声模式切换。

[0047] 在本公开中将参考被称为IVAS编解码器(或IVAS声音编解码器)的IVAS编码架构、仅通过示例来描述UNCLR分类、XTALK检测和立体声模式选择的非限制性、说明性实施例。然而,将此类分类、检测和选择并入任何其他声音编解码器中也在本公开的范围内。

[0048] 1.特征提取

[0049] UNCLR分类基于例如参考文献[9](其全部内容通过引用并入本文)中描述的逻辑回归(LogReg)模型。LogReg模型是针对LRTD立体声模式并且针对DFT立体声模式单独训练的。训练是使用从立体声声音信号编码设备100(立体声编解码器)提取的特征的大型数据库来完成的。类似地,XTALK检测基于LogReg模型,其是针对LRTD立体声模式并且针对DFT立体声模式单独训练的。在XTALK检测中使用的特征不同于在UNCLR分类中使用的特征。然而,某些特征由两个技术共享。

[0050] 在UNCLR分类中使用的特征和在XTALK检测中使用的特征是从以下操作中提取的:

[0051] -声道间相关分析;

[0052] -TD预处理;以及

[0053] -DFT立体声参数化。

[0054] 用于对立体声声音信号进行编码的方法150包括上述特征的提取的操作(未示出)。为了执行特征提取操作,用于对立体声声音信号进行编码的设备100包括特征提取器(未示出)。

[0055] 2.声道间相关分析

[0056] 特征提取的操作(未示出)包括用于LRTD立体声模式的声道间相关分析的操作151和用于DFT立体声模式的声道间相关分析的操作152。为了执行操作151和152,特征提取器(未示出)分别包括声道间相关的分析器101和声道间相关的分析器102。操作151和152以及分析器101和102是类似的并且将被同时描述。

[0057] 分析器101/102接收当前立体声声音信号帧的左声道和右声道作为输入。左声道和右声道首先被下采样到8kHz。例如,令经下采样的左声道和右声道被标示为:

[0058] $X_L(n), X_R(n)$, n=0,..,N-1 (1)

[0059] 其中n是当前帧中的样本索引,N=160是当前帧的长度(160个样本的长度)。经下采样的左声道和右声道被用于计算声道间相关函数。首先,使用例如以下关系来计算左声道和右声道的绝对能量:

[0060] $$E_L = \sum_{n=0}^{N-1} X_L^2(n)$$
$$E_R = \sum_{n=0}^{N-1} X_R^2(n) \quad (2)$$

[0061] 分析器101/102根据在滞后范围<-40,40>上左声道与右声道之间的点积来计算声道间相关函数的分子。对于负滞后,左声道与右声道之间的点积例如是使用以下关系来计算的:

[0062] $$C(k) = \sum_{n=0}^{N-1} X_L(n) X_R(n+k), \quad k = -40,\ldots,0 \quad (3)$$

[0063] 并且,对于正滞后,点积例如是通过以下关系给出的:

[0064] $$C(k) = \sum_{n=0}^{N-1} X_L(n-k) X_R(n), \quad k = 1,\ldots,40 \quad (4)$$

[0065] 分析器101/102随后使用例如以下关系来计算声道间相关函数:

[0066] $$R(k) = \frac{1}{2N} \frac{C(k) + C^{[-1]}(k)}{\sqrt{\left(E_L + E_L^{[-1]}\right)^2 \left(E_R + E_R^{[-1]}\right)^2}}, \quad k = -40,\ldots,40 \quad (5)$$

[0067] 其中上标[-1]标示对先前帧的参考。被动单声道信号是通过对左声道和右声道取平均来计算的:

[0068] $$X_M(n) = \frac{X_L(n) + X_R(n)}{2}, \quad n = 0,..,N-1 \quad (6)$$

[0069] 作为非限制性示例,使用以下关系,边信号被计算为左声道与右声道之间的差:

[0070] $$X_S(n) = \frac{X_L(n) - X_R(n)}{2}, \quad n = 0,..,N-1 \quad (7)$$

[0071] 最后,同样有用的是将左声道与右声道的每样本积定义为:

[0072] $X_P(n)=X_L(n)\cdot X_R(n),\ n=0,..,N-1$ (8)

[0073] 分析器101/102包括无限脉冲响应(IIR)滤波器(未示出),用于使用例如以下关系来平滑声道间相关函数:

[0074] $$R_{LT}^{[n]}(k)=\alpha_{ICA}R_{LT}^{[n-1]}(k)+(1-\alpha_{ICA})R^{[n]}(k),\quad k=-40,\ldots,40 \quad (9)$$

[0075] 其中上标[n]标示当前帧,上标[n-1]标示先前帧,并且α_{ICA}是平滑因子。

[0076] 平滑因子α_{ICA}是在立体声声音信号编码设备100(立体声编解码器)的声道间相关分析(ICA)模块(参考文献[1])内自适应地设置的。随后在预测峰值的区域中的位置处对声道间相关函数进行加权。用于峰值寻找和本地开窗的机制被实现于ICA模块内,并且将不在本文档中描述;关于ICA模块的附加信息见参考文献[1]。将ICA加权之后的声道间相关函数标示为$R_w(k)$,其中$k\in\langle -40,40\rangle$。

[0077] 声道间相关函数的最大值的定位是主导声音到达捕捉点的方向的重要指示符,并且被LRTD立体声模式中的UNCLR分类和XTALK检测用作特征。分析器101/102使用例如以下关系来计算也被LRTD立体声模式中的XTALK检测用作特征的声道间相关函数的最大值:

[0078] $$R_{\max}=\max_k\left[R_w(k)\right],\quad k=-40,..,40 \quad (10)$$

[0079] 并且作为非限制性实施例,使用以下关系来计算最大值的定位:

[0080] $$k_{\max}=\arg\max_k\left[R_w(k)\right],\quad k=-40,..,40 \quad (11)$$

[0081] 当声道间相关函数的最大值R_{max}为负时,其被设置为0。当前帧与先前帧中的最大值R_{max}之间的差例如被计算为:

[0082] $$d_{R\max}=R_{\max}-R_{\max}^{[-1]} \quad (12)$$

[0083] 其中上标[-1]标示对先前帧的参考。

[0084] 声道间相关函数的最大值的定位决定了哪个声道成为ICA模块中的"参考"声道(REF)和"目标"声道(TAR)。如果定位$k_{max}\geq 0$,则左声道(L)是参考声道(REF)并且右声道(R)是目标声道(TAR)。如果$k_{max}<0$,则右声道(R)是参考声道(REF),并且左声道(L)是目标声道(TAR)。目标声道(TAR)随后被移位以补偿其相对于参考声道(REF)的延迟。用于移位目标声道(TAR)的样本的数目可以例如被直接设置为$|k_{max}|$。然而,为了消除由连续帧之间的定位k_{max}的突然改变引起的伪影,可以利用ICA模块内的适当滤波器来平滑用于移位目标声道(TAR)的样本的数量。

[0085] 令用于移位目标声道(TAR)的样本的数量被标示为k_{shift},其中$k_{shift}>0$。令参考声道信号被标示为$X_{ref}(n)$,并且目标声道信号被标示为$X_{tar}(n)$。瞬时目标增益反映了参考声道(REF)与经移位目标声道(TAR)之间的能量比率。瞬时目标增益可以是例如使用以下关系来计算的:

[0086] $$g_t=\log 10\left(\frac{\sum_{n=0}^{N-k_{shift}}\left|X_{ref}(n)\right|}{\sum_{n=k_{shift}}^{N}\left|X_{tar}(n)\right|}+1.0\right) \quad (13)$$

[0087] 其中,N是帧长度。瞬时目标增益被LRTD立体声模式中的UNCLR分类用作特征。

[0088] 2.1声道间特征

[0089] 分析器101/102直接从声道间分析导出在UNCLR分类和XTALK检测中使用的第一特征系列。零滞后处的声道间相关函数的值R(0)其自身被LRTD立体声模式中的UNCLR分类和XTALK检测用作特征。通过计算C(0)的绝对值的对数，由LRTD立体声模式中的UNCLR分类和XTALK检测使用的另一特征被获得，如下所示：

[0090] $$p_{LR}(n)=\log_{10}\left|\sum_{n=0}^{N-1}X_L(n)\cdot X_R(n)\right|=\log_{10}|C(0)|,\quad n=0,..,N-1 \quad (14)$$

[0091] 边信号和单声道信号的能量比率也被LRTD立体声模式中的UNCLR分类和XTALK检测用作特征。该比率例如是使用以下关系来计算的：

[0092] $$r_{SM}(n)=\left|10\log_{10}\sqrt{\frac{\sum_{n=0}^{N-1}X_S^2(n)}{\sum_{n=0}^{N-1}X_M^2(n)}}\right| \quad (15)$$

[0093] 关系(15)的能量比率例如随时间被平滑，如下所示：

[0094] $$\overline{r_{SM}}(n)=\begin{cases}0.9\overline{r_{SM}}(n-1) & \text{如果 } c_{hang}\neq 0\\ 0.9\overline{r_{SM}}(n-1)+0.1r_{SM}(n) & \text{否则}\end{cases} \quad (16)$$

[0095] 其中，c_{hang}是作为立体声声音信号编码设备100(立体声编解码器)的VAD(话音活动检测)模块的部分而被计算的VAD拖尾(hangover)帧的计数值(参见例如参考文献[1])。关系(16)的经平滑比率被LRTD立体声模式中的XTALK检测用作特征。

[0096] 分析器101/102从左声道和单声道信号以及在右声道与单声道信号之间导出以下点积。首先，左声道与单声道信号之间的点积例如被表达为：

[0097] $$C_{LM}=\sum_{n=0}^{N-1}X_L(n)X_M(n) \quad (17)$$

[0098] 并且右声道与单声道信号之间的点积例如为：

[0099] $$C_{RM}=\sum_{n=0}^{N-1}X_R(n)X_M(n) \quad (18)$$

[0100] 两个点积都是正的，下界为0。基于这两个点积中的最大值与最小值之差的度量被LRTD立体声模式中的UNCLR分类和XTALK检测用作特征。其可以是使用以下关系来计算的：

[0101] $$d_{mmLR}=\max[C_{LM},C_{RM}]-\min[C_{LM},C_{RM}] \quad (19)$$

[0102] 被LRTD立体声模式中的UNCLR分类和XTALK检测用作独立特征的类似度量直接基于线性域和对数域中的两个点积之间的绝对差，该绝对差例如是使用以下关系来计算的：

[0103] $$\begin{aligned}\Delta_{LRM}&=C_{LM}-C_{RM}\\ d_{LRM}&=\log_{10}|C_{LM}-C_{RM}|\end{aligned} \quad (20)$$

[0104] 由LRTD立体声模式中的UNCLR分类和XTALK检测使用的最后特征是作为声道间相关分析操作151/152的部分来计算的，并且反映声道间相关函数的演变。其可以被计算如下：

[0105] $$RR=\frac{\sum_{k=-40}^{40}R(k)R^{[-2]}(k)}{\sqrt{\sum_{k=-40}^{40}R^2(k)\sum_{k=-40}^{40}\left(R^{[-2]}\right)^2(k)}} \quad (21)$$

[0106] 其中上标[-2]标示对当前帧之前的第二帧的参考。

[0107] 3.时域(TD)预处理

[0108] 在LRTD立体声模式中,不存在单声道下混合,并且输入立体声声音信号190的左声道和右声道两者在相应的时域预处理操作(即,用于立体声声音信号190的左声道的时域预处理的操作153和用于立体声声音信号190的右声道的时域预处理的操作154)中被分析以提取特征。为了执行操作153和154,特征提取器(未示出)包括如图1所示的相应的时域预处理器103和104。操作153和154以及对应的预处理器103和104是类似的并且将被同时描述。

[0109] 时域预处理操作153/154执行多个子操作以产生某些参数,这些参数被用作用于进行UNCLR分类和XTALK检测的所提取特征。此类子操作可以包括:

[0110] -频谱分析;

[0111] -线性预测分析;

[0112] -开环音调估计;

[0113] -话音活动检测(VAD);

[0114] -背景噪声估计;以及

[0115] -帧错误隐藏(FEC)分类。

[0116] 时域预处理器103/104使用Levinson-Durbin算法执行线性预测分析。Levinson-Durbin算法的输出是线性预测系数(LPC)的集合。Levinson-Durbin算法是迭代方法,并且Levinson-Durbin算法中的总迭代次数可以被标示为M。在每个第i次迭代中,其中i=1,..,M,残差能量$e_{LPC}^{[i-1]}$被计算。在本公开中,作为非限制性说明性实现,假设Levinson-Durbin算法以M=16的迭代运行。输入立体声声音信号190的左声道与右声道之间的残差能量的差被用作用于LRTD立体声模式中的XTALK检测的特征。残差能量的差可以被计算如下:

[0117] $$d_{LPC13}=\left|e_{LPC,L}^{[13]}-e_{LPC,R}^{[13]}\right| \quad (22)$$

[0118] 其中,下标L和R被添加以分别表示输入立体声声音信号190的左声道和右声道。在此非限制性实施例中,特征是使用来自第14次迭代而不是最后迭代的残差能量来计算的(差d_{LPC13}),因为实验发现此次迭代对于UNCLR分类具有最高的辨别潜力。关于Levinson-Durbin算法的更多信息和关于残差能量计算的细节可以例如在参考文献[1]中被找到。

[0119] 利用Levinson-Durbin算法估计的LPC系数被转换成线谱频率,LSF(i),i=0,..,M-1。LSF值的和可以用作输入立体声声音信号190的包络的重力点的估计。左声道和右声道中的LSF值的和之间的差包含关于两个声道的相似性的信息。因此,此差被用作LRTD立体声模式中的XTALK检测中的特征。左声道和右声道中的LSF值的和之间的差可以是使用以下关系来计算的:

[0120] $$d_{LSF}=\sum_{i=0}^{M-1}\left|LSF_L(i)-LSF_R(i)\right| \quad (23)$$

[0121] 关于上述LPC到LSF转换的附加信息可以在例如参考文献[1]中被找到。

[0122] 时域预处理器103/104执行开环音调估计并使用自相关函数，从中计算出左声道(L)/右声道(R)开环音调差。左声道(L)/右声道(R)开环音调差可以是使用以下关系来计算的：

[0123] $$d_T = \frac{1}{3}\sum_{k=1}^{3}\left|T_L^{[k]} - T_R^{[k]}\right| \quad (24)$$

[0124] 其中$T^{[k]}$是当前帧的第k片段中的开环音调估计。在本公开中，作为非限制性说明性示例，假定开环音调分析是在索引为k＝1,2,3的三个相邻半帧(片段)中执行的，其中两个片段位于当前帧中并且一个片段位于先前帧的第二半中。可以使用不同数量的片段以及不同的片段长度和重叠。关于开环音调估计的附加信息可以在例如参考文献[1]中被找到。

[0125] 输入立体声声音信号190的左声道和右声道的最大自相关值(发声)之间的差(由上述自相关函数确定)也被LRTD立体声模式中的XTALK检测用作特征。左声道和右声道的最大自相关值之间的差可以是使用以下关系来计算的：

[0126] $$d_v = \frac{1}{3}\sum_{k=1}^{3}\left|v_L^{[k]} - v_R^{[k]}\right| \quad (25)$$

[0127] 其中$v^{[k]}$表示第k半帧中的左(L)声道和右(R)声道的最大自相关值。

[0128] 背景噪声估计是话音活动检测(VAD)检测算法的部分(见参考文献[1])。具体来说，背景噪声估计使用依赖于特征集合的活动/非活动信号检测器(未显示)，该特征集合中的某些特征由UNCLR分类和XTALK检测使用。例如，活动/非活动信号检测器(未示出)产生左声道(L)和右声道(R)的非平稳性参数f_{sta}以作为频谱稳定性的测量。输入立体声声音信号190的左声道与右声道之间的非平稳性的差被LRTD立体声模式中的XTALK检测用作特征。左(L)声道与右(R)声道之间的非平稳性的差可以是使用以下关系来计算的：

[0129] $$d_{sta} = |f_{sta,L} - f_{sta,R}| \quad (26)$$

[0130] 活动/非活动信号检测器(未示出)依赖于包含相关图参数C_{map}的谐波分析。相关图是对输入立体声声音信号190的音调稳定性的测量，并且其被UNCLR分类和XTALK检测使用。左(L)声道和右(R)声道的相关图之间的差被LRTD立体声模式中的XTALK检测用作特征，并且是使用例如以下关系来计算的：

[0131] $$d_{cmap} = |C_{map,L} - C_{map,R}| \quad (27)$$

[0132] 最后，活动/非活动信号检测器(未示出)定期测量每个帧中的频谱多样性和噪声特性。这两个参数也被LRTD立体声模式中的UNCLR分类和XTALK检测用作特征。具体地，(a)左声道(L)与右声道(R)之间的频谱多样性的差可以被计算如下：

[0133] $$d_{sdiv} = |\log(S_{div,L}) - \log(S_{div,R})| \quad (28)$$

[0134] 其中S_{div}表示当前帧中的频谱多样性的测量，并且(b)左声道(L)与右声道(R)之间的噪声特性的差可以被计算如下：

[0135] $$d_{nchar} = |\log(n_{char,L}) - \log(n_{char,R})| \quad (29)$$

[0136] 其中n_{char}表示当前帧中的噪声特性的测量。关于相关图、非平稳性、频谱多样性和噪声特性参数的计算的细节，可以参考[1]。

[0137] 如参考文献[1]中所描述的，作为立体声声音信号编码设备100的部分的ACELP(代数码激励线性预测)核心编码器包括用于对清音(“unvoiced”)声音进行编码的特定设置。这些设置的使用是由多个因素来调节的，包括对当前帧内的短片段中的突然能量增加的测

量。用于ACELP核心编码器中的清音声音编码的设置仅在当前帧内没有突然的能量增加时被应用。通过比较左声道和右声道中的突然能量增加的测量，可以定位串音片段的开始定位。突然能量增加可以是类似于3GPP EVS编解码器(参考文献[1])中描述的E_d参数来计算的。左声道(L)与右声道(R)的突然能量增加的差可以是使用以下关系来计算的：

[0138] $$d_{dE}=|\log(E_{d,L})-\log(E_{d,R})| \quad (30)$$

[0139] 其中，下标L和R被添加以分别表示输入立体声声音信号190的左声道和右声道。

[0140] 时域预处理器103/104和预处理操作153/154使用包含用于FEC技术的状态机的FEC分类模块。每个帧中的FEC类别是基于评价函数在预定义的类别中选择的。在当前帧中针对左声道(L)与右声道(R)选择的FEC类别之间的差被LRTD立体声模式中的XTALK检测用作特征。然而，出于此类分类和检测的目的，FEC类别可能会受到以下限制：

[0141] $$t_{class}=\begin{cases}\text{VOICED} & \text{如果} t_{class}\geq\text{VOICED}\\ \text{UNVOICED} & \text{否则}\end{cases} \quad (31)$$

[0142] 其中t_{class}是在当前帧中选择的FEC类别。因此，FEC类别仅限于浊音(“VOICED”)和清音(“UNVOICED”)。左声道(L)与右声道(R)中的类之间的差可以被计算如下：

[0143] $$d_{class}=|t_{class,L}-t_{class,R}| \quad (32)$$

[0144] 对于关于FEC分类的附加细节，可以参考[1]。

[0145] 时域预处理器103/104和预处理操作153/154实现语音/音乐分类和对应的语音/音乐分类器。此语音/音乐分类根据功率谱散度和功率谱稳定性在每个帧中进行二元决策。左声道(L)与右声道(R)之间的功率谱散度的差例如是使用以下关系来计算的：

[0146] $$d_{Pdiff}=|P_{diff,L}-P_{diff,R}| \quad (33)$$

[0147] 其中，P_{diff}表示当前帧中的左声道(L)和右声道(R)中的功率谱散度，并且左声道(L)与右声道(R)之间的功率谱稳定性的差例如是使用以下关系来计算的：

[0148] $$d_{Psta}=|P_{sta,L}-P_{sta,R}| \quad (34)$$

[0149] 其中P_{sta}表示当前帧中的左声道(L)和右声道(R)的功率谱稳定性。

[0150] 参考文献[1]描述了关于在语音/音乐分类中计算的功率谱散度和功率谱稳定性的细节。

[0151] 4.DFT立体声参数

[0152] 用于对立体声声音信号190进行编码的方法150包括计算左声道(L)和右声道(R)的快速傅里叶变换(FFT)的操作155。为了执行操作155，用于对立体声声音信号190进行编码的设备100包括FFT变换计算器105。

[0153] 特征提取的操作(未示出)包括计算DFT立体声参数的操作156。为了执行操作156，特征提取器(未示出)包括DFT立体声参数的计算器106。

[0154] 在DFT立体声模式中，变换计算器105通过FFT变换将输入立体声声音信号190的左声道(L)和右声道(R)转换到频域。

[0155] 令左声道(L)的复频谱被标示为$\tilde{S}_L(k)$，并且右声道(R)的复频谱被标示为$\tilde{S}_R(k)$，其中$k=0,..,N_{FFT}-1$是频率元的索引，并且N_{FFT}是FFT变换的长度。例如，当输入立体声声音信号的采样速率是32kHz时，DFT立体声参数的计算器106在40ms的窗口上计算复频谱，得到$N_{FFT}=1280$个样本。随后，作为非限制性实施例，可以使用以下关系来计算复交叉声道频谱：

[0156] $$X_{LR}(k)=\tilde{S}_L(k)\tilde{S}_R^*(k),\quad k=0,..,N_{FFT}-1 \quad (35)$$

[0157] 其中星号上标指示复共轭。可以使用以下关系将复交叉声道频谱分解成实部和虚部：

[0158] $$\begin{aligned}\mathrm{Re}\left(X_{LR}(k)\right)&=\mathrm{Re}\left(\tilde{S}_L(k)\right)\cdot\mathrm{Re}\left(\tilde{S}_R(k)\right)+\mathrm{Im}\left(\tilde{S}_L(k)\right)\cdot\mathrm{Im}\left(\tilde{S}_R(k)\right),\quad k=0,..,N_{FFT}-1\\ \mathrm{Im}\left(X_{LR}(k)\right)&=\mathrm{Im}\left(\tilde{S}_L(k)\right)\cdot\mathrm{Re}\left(\tilde{S}_R(k)\right)-\mathrm{Re}\left(\tilde{S}_L(k)\right)\cdot\mathrm{Im}\left(\tilde{S}_R(k)\right),\quad k=0,..,N_{FFT}-1\end{aligned} \quad (36)$$

[0159] 使用实部和虚部分解，可以将复交叉声道频谱的绝对幅值表达为：

[0160] $$\left|X_{LR}(k)\right|=\sqrt{\mathrm{Re}\left(X_{LR}(k)\right)^2+\mathrm{Im}\left(X_{LR}(k)\right)^2},\quad k=0,..,N_{FFT}-1 \quad (37)$$

[0161] 通过使用以下关系对频率元上的复交叉声道频谱的绝对幅值求和，DFT立体声参数的计算器106获得复交叉声道频谱的总体绝对幅值：

[0162] $$\left|X_{LR}\right|=\sum_{k=0}^{N_{FFT}-1}\left|X_{LR}(k)\right| \quad (38)$$

[0163] 左声道(L)的能量谱和右声道(R)的能量谱可以被表达为：

[0164] $$\begin{aligned}E_L(k)&=\mathrm{Re}\left(\tilde{S}_L(k)\right)^2+\mathrm{Im}\left(\tilde{S}_L(k)\right)^2,\quad k=0,..,N_{FFT}-1\\ E_R(k)&=\mathrm{Re}\left(\tilde{S}_R(k)\right)^2+\mathrm{Im}\left(\tilde{S}_R(k)\right)^2,\quad k=0,..,N_{FFT}-1\end{aligned} \quad (39)$$

[0165] 通过使用以下关系在频率元上对左声道(L)的能量谱和右声道(R)的能量谱求和，可以获得左声道(L)和右声道(R)的总能量：

[0166] $$\begin{aligned}E_L&=\sum_{k=0}^{N_{FFT}-1}E_L(k)\\ E_R&=\sum_{k=0}^{N_{FFT}-1}E_R(k)\end{aligned} \quad (40)$$

[0167] DFT立体声模式中的UNCLR分类和XTALK检测使用复交叉声道频谱的总体绝对幅值作为它们的特征之一，但不是以如上定义的直接形式而是以能量归一化的形式，并且是在对数域中，如使用例如以下关系所表达的：

[0168] $$f_X=\sum_{k=0}^{N_{FFT}-1}\log\left(\frac{\left|X_{LR}(k)\right|}{E_L+E_R}\right) \quad (41)$$

[0169] DFT立体声参数计算器106可以使用例如以下关系来计算单声道下混合能量：

[0170] $E_M=E_L+E_R+2X_{LR}$ (42)

[0171] 声道间电平差(ILD)是由DFT立体声模式中的UNCLR分类和XTALK检测使用的特征，因为其包含关于主声音来自的角度的信息。出于UNCLR分类和XTALK检测的目的，声道间电平差(ILD)可以是以增益因子的形式来表达的。DFT立体声参数的计算器106使用例如以下关系来计算声道间电平差(ILD)增益：

[0172] $$g_{ILD}=\left|\frac{\sqrt{E_L/E_R}-1}{\sqrt{E_L/E_R}+1}\right| \quad (43)$$

[0173] 声道间相位差(IPD)包含收听者可以从其推断传入声音信号的方向的信息。DFT立体声参数的计算器106使用例如以下关系来计算声道间相位差(IPD)：

[0174] $$IPD = \arctan\left(\frac{\mathrm{Im}(X_{LR})}{\mathrm{Re}(X_{LR})}\right) \quad (44)$$

[0175] 其中：

[0176] $$\mathrm{Re}(X_{LR}) = \sum_{k=0}^{N_{FFT}-1} \mathrm{Re}\left(X_{LR}(k)\right)$$
$$\mathrm{Im}(X_{LR}) = \sum_{k=0}^{N_{FFT}-1} \mathrm{Im}\left(X_{LR}(k)\right) \quad (45)$$

[0177] 声道间相位差(IPD)相对于先前帧的差分值例如是使用以下关系来计算的：

[0178] $d_{IPD} = |IPD^{[n]} - IPD^{[n-1]}|$ (46)

[0179] 其中上标n被用于标示当前帧，并且上标n-1被用于标示先前帧。最后，计算器106可以将IPD增益计算为相位对齐的(IPD＝0)下混合能量(关系(47)的分子)与单声道下混合能量的能量E_M之间的比率：

[0180] $$g_{IPD_lin} = \frac{E_L + E_R + 2\,\mathrm{Re}(X_{LR})}{E_M} \quad (47)$$

[0181] IPD增益g_{IPD_lin}被限制在区间<0,1>。如果值超过上阈值1.0，则用来自先前帧的IPD增益的值代替。DFT立体声模式中的UNCLR分类和XTALK检测使用对数域中的IPD增益作为特征。计算器106使用例如以下关系来确定对数域中的IPD增益：

[0182] $g_{IPD} = \log(1 - g_{IPD_lin})$ (48)

[0183] 声道间相位差(IPD)也可以是以由DFT立体声模式中的UNCLR分类和XTALK检测用作特征的角度的形式来表达的，并且例如被计算如下：

[0184] $$\varphi_{rot} = \arctan\left(\frac{2\,\mathrm{Re}(X_{LR})}{E_L - E_R}\right) \quad (49)$$

[0185] 边声道可以被计算为左声道(L)与右声道(R)之间的差。可以使用以下关系、通过计算此差($E_L - E_R$)的能量绝对值相对于单声道下混合能量E_M的比率来表达边声道的增益：

[0186] $$g_{side} = \frac{|E_L - E_R|}{E_M} \quad (50)$$

[0187] 增益g_{side}越高，左声道(L)与右声道(R)的能量之间的差越大。边声道的增益g_{side}被限制为区间<0.01,0.99>。超出此范围的值是受限的。

[0188] 输入立体声声音信号190的左声道(L)与右声道(R)之间的相位差也可以是根据使用例如以下关系计算的预测增益来分析的：

[0189] $g_{pred_lin} = (1 - g_{side})E_L + (1 + g_{side})E_R - 2|X_{LR}|$ (51)

[0190] 其中预测增益g_{pred_lin}的值被限制在区间<0,∞>，即正值。g_{pred_lin}的上述表达捕捉交叉声道频谱(X_{LR})能量与单声道下混合能量$E_M = E_L + E_R + 2|X_{LR}|$之间的差。计算器106使用例如关系(52)将此增益g_{pred_lin}转换到对数域，以用于被DFT立体声模式中的UNCLR分类和XTALK检测用作特征：

[0191] $g_{pred}=\log(g_{pred_lin}+1)$ (52)

[0192] 计算器106还使用关系(39)的每仓(“per-bin”)声道能量来计算声道间相干性(ICC)的均值能量,该均值能量形成用于确定未被声道间时间差(ITD)(下文将描述)和声道间相位差(IPD)捕捉的左声道(L)与右声道(R)之间的差的线索。首先,计算器106使用例如以下关系来计算交叉声道频谱的总体能量:

[0193] $E_X=\mathrm{Re}(X_{LR})^2+\mathrm{Im}(X_{LR})^2$ (53)

[0194] 为了表达声道间相干性(ICC)的均值能量,计算以下参数是有用的:

[0195] $$\varphi_{tot}=\sqrt{(E_L-E_R)(E_L-E_R)+4E_X} \quad (54)$$

[0196] 随后,声道间相干性(ICC)的均值能量被DFT立体声模式中的UNCLR分类和XTALK检测用作特征,并且可以被表达为

[0197] $$E_{coh}=20\log_{10}\left(\frac{E_L+E_R+\varphi_{tot}}{E_L+E_R-\varphi_{tot}}\right) \quad (55)$$

[0198] 如果内项小于1.0,则均值能量E_{coh}的值被设置为0。声道间相干性(ICC)的另一可能的解释是被计算如下的边比单声道(“side-to-mono”)能量比率:

[0199] $$E_{S2M}=\left|\frac{E_L-E_R}{E_L+E_R}\right| \quad (56)$$

[0200] 最后,计算器106确定在UNCLR分类和XTALK检测中使用的最大与最小声道内幅度积的比率r_{pp}。被DFT立体声模式中的UNCLR分类和XTALK检测用作特征的此特征例如是使用以下关系来计算的:

[0201] $$r_{PP}=\log\left(1+\frac{\max(P_L,P_R)}{\min(P_L,P_R)}\right) \quad (57)$$

[0202] 其中,声道内幅度积被定义如下:

[0203] $$P_L=\prod_{k=0}^{N_{FFT}-1}\left|\tilde{S}_L(k)\right|$$
$$P_R=\prod_{k=0}^{N_{FFT}-1}\left|\tilde{S}_R(k)\right| \quad (58)$$

[0204] 在立体声声音信号再现中使用的一个参数是声道间时间差(ITD)。在DFT立体声模式中,DFT立体声参数的计算器106根据具有相位差的广义交叉声道相关(GCC-PHAT)函数来估计声道间时间差(ITD)。声道间时间差(ITD)对应于到达时间延迟(TDOA)估计。GCC-PHAT函数是用于估计混响信号上的声道间时间差(ITD)的稳健方法。GCC-PHAT例如是使用以下关系计算的:

[0205] $$X_{PHAT}(k)=IFFT\left[\frac{X_{LR}}{|X_{LR}|}\right],\quad k=0,..,N_{FFT}-1 \quad (59)$$

[0206] 其中IFFT表示逆快速傅里叶变换。

[0207] 声道间时间差(ITD)随后是使用例如以下关系根据GCC-PHAT函数来估计的:

[0208] $$d_{ITD}=\arg\max_d\left(X_{PHAT}(d)\right),\quad d=-200,..,200 \quad (60)$$

[0209] 其中d是对应于范围从-5ms到+5ms的时间延迟的样本中的时间滞后。对应于d_{ITD}的GCC-PHAT函数的最大值被DFT立体声模式中的UNCLR分类和XTALK检测用作特征，并且可以是使用以下关系来检索的：

[0210] $$G_{ITD}=\max_{d}\left(X_{PHAT}(d)\right),\quad d=-200,..,200 \tag{61}$$

[0211] 在单音场景中，在对应于声道间时间差(ITD)的GCC-PHAT函数中通常存在单个主导峰值。然而，在两个谈话者位于捕捉麦克风的相对侧上的串音情形中，通常存在彼此分开定位的两个主导峰值。图2图示出此类情形。具体地，根据非限制性说明性示例，图2是具有由一对超心形麦克风M1和M2捕捉的两个相对的谈话者S1和S2的串音场景的平面图，并且图3是示出GCC-PHAT函数中的两个主导峰值的位置的图。

[0212] 第一峰值的幅度G_{ITD}是使用关系(61)计算的，其定位d_{ITD}是使用关系(60)计算的。第二峰值的幅度是通过在相对于第一峰值相反的方向上搜索GCC-PHAT函数的第二最大值来定位的。更具体地，第二峰值的搜索的方向s_{ITD}是由第一峰值的定位d_{ITD}的符号确定的：

[0213] $$s_{ITD}=sgn(d_{ITD}) \tag{62}$$

[0214] 其中sgn(.)是符号函数。

[0215] DFT立体声参数的计算器106随后可以使用例如以下关系来检索在方向s_{ITD}上的GCC-PHAT函数的第二最大值(第二最高峰值)：

[0216] $$G_{ITD2}=\begin{cases}\max\limits_{d}\left(X_{PHAT}(d)\right) & d=thr_{xt},..,200 \quad \text{如果} s_{ITD}<0\\ \max\limits_{d}\left(X_{PHAT}(d)\right) & d=-200,..,-thr_{xt} \text{如果} s_{ITD}>0\end{cases} \tag{63}$$

[0217] 作为非限制性实施例，阈值$thr_{xt}=8$确保GCC-PHAT函数的第二峰值是在距开始($d_{ITD}=0$)至少8个样本的距离处搜索的。就串音(XTALK)的检测而言，这意味着场景中任何潜在的次要谈话者都必须存在于与第一“主导”谈话者和中间点(d=0)两者分开至少一定最小距离处。

[0218] GCC-PHAT函数的第二最高峰值的定位是通过以arg max(.)函数代替max(.)函数、使用关系(63)来计算的。GCC-PHAT函数的第二最高峰值的定位将被标示为d_{ITD2}。

[0219] GCC-PHAT函数的第一峰值与第二最高峰值的幅度之间的关系被DFT立体声模式中的XTALK检测用作特征，并且可以是使用以下比率来评估的：

[0220] $$r_{GITD12}=\frac{\left|G_{ITD}G_{ITD2}\right|}{\left|G_{ITD}+G_{ITD2}\right|} \tag{64}$$

[0221] 比率r_{GITD12}具有高辨别潜力，但是为了将其用作特征，XTALK检测消除了由于在DFT立体声模式中的频率变换期间应用的有限时间分辨率而导致的偶然假警报。这可以通过使用例如以下关系将当前帧中的比率r_{GITD12}的值与来自先前帧的相同比率的值相乘来完成：

[0222] $$r_{GITD12}\leftarrow r_{GITD12}(n)\cdot r_{GITD12}(n-1) \tag{65}$$

[0223] 其中添加索引n以标示当前帧，并且添加索引n-1以标示先前帧。为了简单起见，参数名称r_{GITD12}被重用于标识输出参数。

[0224] 第二最高峰值的幅度单独构成场景中的次要谈话者的强度的指示符。类似于比率r_{GITD12}，使用例如以下关系(66)来减少值G_{ITD2}的偶然随机“尖峰”，以获得由DFT立体声模式中的XTALK检测使用的另一特征：

[0225] $m_{ITD2}=G_{ITD2}(n)\cdot G_{ITD2}(n-1)$ (66)

[0226] 在DFT立体声模式中的XTALK检测中使用的另一特征是当前帧中的第二最高峰值的定位$d_{ITD2}(n)$相对于先前帧的差,其是使用例如以下关系来计算的:

[0227] $\Delta_{ITD2}=|d_{ITD2}(n)-d_{ITD2}(n-1)|$ (67)

[0228] 5.下混合和逆快速傅里叶逆变换(IFFT)

[0229] 在DFT立体声模式中,用于对立体声声音信号进行编码的方法150包括对立体声声音信号190的左声道(L)和右声道(R)进行下混合的操作157和计算下混合信号的IFFT变换的操作158。为了执行操作157和158,用于对立体声声音信号190进行编码的设备100包括下混合器107和IFFT变换计算器108。

[0230] 例如如参考文献[6](其全部内容通过引用并入本文)中所描述的,下混合器107将立体声声音信号的左声道(L)和右声道(R)下混合为单声道声道(M)和边声道(S)。

[0231] IFFT变换计算器108随后计算来自下混合器107的下混合单声道声道(M)的IFFT变换,以产生要在TD预处理器109中处理的时域单声道声道(M)。在计算器108中使用的IFFT变换是在计算器105中使用的FFT变换的逆。

[0232] 6.DFT立体声模式中的TD预处理

[0233] 在DFT立体声模式中,特征提取的操作(未示出)包括用于提取在UNCLR分类和XTALK检测中使用的特征的TD预处理操作159。为了执行操作159,特征提取器(未示出)包括响应于单声道声道(M)的TD预处理器109。

[0234] 6.1话音活动检测

[0235] UNCLR分类和XTALK检测使用话音活动检测(VAD)算法。在LRTD立体声模式中,VAD算法是在左声道(L)和右声道(R)上单独运行的。在DFT立体声模式中,VAD算法是在下混合单声道声道(M)上运行的。VAD算法的输出是二元标志f_{VAD}。VAD标志f_{VAD}不适用于UNCLR分类和XTALK检测,因为其太保守并且具有长的迟滞。这阻止了在LRTD立体声模式与DFT立体声模式之间的快速切换(例如,在谈话突发结束时或在发声中途的短暂停期间)。而且,VAD标志f_{VAD}对输入立体声声音信号190中的小改变敏感。这会导致串音检测中的假警报和立体声模式的错误选择。因此,UNCLR分类和XTALK检测使用基于相对帧能量的变化的、话音活动检测的替代测量。关于VAD算法的细节,参考[1]。

[0236] 6.1.1相对帧能量

[0237] UNCLR分类和XTALK检测使用利用关系(2)获得的左声道(L)的绝对能量E_L和右声道(R)的绝对能量E_R。输入立体声声音信号的最大平均能量可以是使用例如以下关系在对数域中计算的:

[0238] $$E_{ave}(n)=10\log_{10}\frac{\max\left(E_L(n),E_R(n)\right)}{N} \quad (68)$$

[0239] 其中添加索引n以标识当前帧,并且N=160是当前帧的长度(160个样本的长度)。对数域中的最大平均能量$E_{ave}(n)$的值被限制于区间$\langle 0;\infty\rangle$。

[0240] 随后可以通过使用例如以下关系将最大平均能量$E_{ave}(n)$线性映射到区间$\langle 0;0.9\rangle$中来计算输入立体声声音信号的相对帧能量:

[0241] $$E_{rl}=\left[E_{ave}(n)-E_{dn}(n)\right]\frac{0.9}{E_{up}(n)-E_{dn}(n)} \quad (69)$$

[0242] 其中$E_{up}(n)$标示相对帧能量$E_{rl}(n)$的上界，$E_{dn}(n)$标示相对帧能量$E_{rl}(n)$的下界，并且索引n标示当前帧。

[0243] 相对帧能量$E_{rl}(n)$的界限在每个帧中基于噪声更新计数值$a_{En}(n)$(其是TD预处理器103、104和109的噪声估计模块的部分)来被更新。有关此计数值的附加信息，参考[1]。计数值$a_{En}(n)$的目的是信令通知当前帧中的每个声道的背景噪声水平可以被更新。这种情形发生在计数值$a_{En}(n)$的值为零时。作为非限制性示例，每个声道中的计数值$a_{En}(n)$被初始化为6，并且在每个帧中递增或递减，其具有下阈值0和上阈值6。

[0244] 在LRTD立体声模式的情况下，噪声估计是在左声道(L)和右声道(R)上独立执行的。将两个噪声更新计数值分别表示为用于左声道(L)和右声道(R)的$a_{En,L}(n)$和$a_{En,R}(n)$。两个计数值随后可以利用以下关系被组合成单个二元参数：

[0245] $$f_{En}(n)=\begin{cases}1, & \text{如果} a_{En,L}(n)=6 \quad \text{或者} \quad a_{En,R}(n)=6\\ 0, & \text{否则}\end{cases} \tag{70a}$$

[0246] 在DFT立体声模式的情况下，噪声估计是在下混合单声道声道(M)上执行的。将单声道声道中的噪声更新计数值标示为$a_{En,M}(n)$。二元输出参数是利用以下关系来计算的：

[0247] $$f_{En}(n)=\begin{cases}1, & \text{如果} a_{En,M}(n)=6\\ 0, & \text{否则}\end{cases} \tag{70b}$$

[0248] UNCLR分类和XTALK检测使用二元参数$f_{En}(n)$来实现相对帧能量$E_{rl}(n)$的下界$E_{dn}(n)$或上界$E_{up}(n)$的更新。当参数$f_{En}(n)$等于零时，下界$E_{dn}(n)$被更新。当参数$f_{En}(n)$等于1时，上界$E_{up}(n)$被更新。

[0249] 相对帧能量$E_{rl}(n)$的上界$E_{up}(n)$是在其中参数$f_{En}(n)$等于1的帧中使用例如以下关系来更新的：

[0250] $$E_{up}(n)=\begin{cases}0.99E_{up}(n-1)+0.01E_{ave}(n) & \text{如果} E_{ave}(n)<E_{up}(n-1)\\ 0.95E_{up}(n-1)+0.05E_{ave}(n), & \text{否则}\end{cases} \tag{71}$$

[0251] 其中索引n表示当前帧，并且索引n-1表示先前帧。

[0252] 关系(71)中的第一行和第二行分别表示较慢更新和较快更新。因此，使用关系(71)，当能量增加时，上界$E_{up}(n)$被更快地更新。

[0253] 相对帧能量$E_{rl}(n)$的下界$E_{dn}(n)$是在其中参数$f_{En}(n)$等于0的帧中使用例如以下关系来更新的：

[0254] $E_{dn}(n)=0.9E_{dn}(n-1)+0.1E_{ave}(n)$ (72)

[0255] 其中下阈值为30.0。如果上界$E_{up}(n)$的值与下界$E_{dn}(n)$太接近，则其被修改，例如，如下所示：

[0256] $$E_{up}(n)=E_{dn}(n)+20.0, \text{如果} E_{up}(n)<E_{dn}(n)+20.0 \tag{73}$$

[0257] 6.1.2替代VAD标志估计

[0258] UNCLR分类和XTALK检测使用在关系(71)中计算的相对帧能量$E_{rl}(n)$的变化以作为用于计算替代VAD标志的基础。令当前帧中的替代VAD标志被标示为$f_{xVAD}(n)$。替代VAD标志$f_{xVAD}(n)$是通过将在LRTD立体声模式的情况下在TD预处理器103/104的噪声估计模块中生

成的VAD标志或者在DFT立体声模式的情况下在TD预处理器109中生成的VAD标志f_{VAD}与反映相对帧能量$E_{rl}(n)$的变化的辅助二元参数$f_{Erl}(n)$进行组合来计算的。

[0259] 首先，使用例如以下关系在10个先前帧的片段上对相对帧能量$E_{rl}(n)$求平均：

[0260] $$\overline{E_{rl}}(p)=\frac{1}{10-p}\sum_{k=p}^{9}E_{rl}(n-k),\quad p=0,..,2 \qquad (74)$$

[0261] 其中p是平均的索引。辅助二元参数例如是根据以下逻辑来设置的：

[0262] $$f_{Erl}(n)=\begin{cases}0, & \text{如果}\ \overline{E_{rl}}(0)<0.1\\ 0, & \text{如果}\ E_{rl}(n)<0.3\ \text{以及}\ \overline{E_{rl}}(1)<0.1\\ 0, & \text{如果}\ E_{rl}(n)<0.5\ \text{以及}\ \overline{E_{rl}}(2)<0.17\\ 1, & \text{否则}\end{cases} \qquad (75)$$

[0263] 在LRTD立体声模式中，替代VAD标志$f_{xVAD}(n)$是通过左声道(L)中的VAD标志$f_{VAD,L}(n)$、右声道(R)中的VAD标志$f_{VAD,R}(n)$、和辅助二元参数$f_{Erl}(n)$的逻辑组合，使用例如以下关系来计算的：

[0264] $f_{xVAD}(n)=(f_{VAD,L}(n)\ \mathrm{OR}\ f_{VAD,R}(n))\ \mathrm{AND}\ f_{Erl}(n)$ (76)

[0265] 在DFT立体声模式中，替代VAD标志$f_{xVAD}(n)$是通过下混合单声道声道(M)中的VAD标志$f_{VAD,M}(n)$、和辅助二元参数$f_{Erl}(n)$的逻辑组合，使用例如以下关系来计算的。

[0266] $f_{xVAD}(n)=f_{VAD,M}(n)\ \mathrm{AND}\ f_{Erl}(n)$ (77)

[0267] 6.2立体声静音标志

[0268] 在DFT立体声模式中，计算反映下混合单声道声道(M)的低电平的离散参数也是方便的。此类被称为立体声静音标志的参数可以是通过例如将活动信号的平均电平与特定预定义阈值进行比较来计算的。作为示例，在TD预处理器109的VAD算法内计算的长期活动语音电平$\bar{N}_{sp}(n)$可以被用作用于计算立体声静音标志的基础。关于VAD算法的细节，参考[1]。

[0269] 立体声静音标志随后可以是使用以下关系来计算的：

[0270] $$f_{sil}(n)=\begin{cases}2 & \text{如果}\ \bar{N}_{sp}(n)-E_M(n)>25\\ f_{sil}(n-1)-1 & \text{否则}\end{cases} \qquad (78)$$

[0271] 其中$E_M(n)$是当前帧中的下混合单声道声道(M)的绝对能量。立体声静音标志$f_{sil}(n)$被限制在区间<0;∞>。

[0272] 7.非相关立体声内容(UNCLR)的分类

[0273] LRTD立体声模式和DFT立体声模式中的UNCLR分类基于逻辑回归(LogReg)模型(见参考文献[9])。LogReg模型是在由相关和非相关立体声声音信号样本组成的大型标记数据库上针对LRTD立体声模式和DFT立体声模式单独训练的。非相关立体声训练样本是通过对随机选择的单声道样本进行组合来人工创建的。以下立体声场景可以利用这种单声道样本的人工混合来模拟：

[0274] -讲话者A在左声道中，讲话者B在右声道中(或反之亦然)；

[0275] -讲话者A在左声道中，音乐声在右声道中(或反之亦然)；

[0276] -讲话者A在左声道中，噪声在右声道中(或反之亦然)；

[0277] -讲话者A在左声道或右声道中,背景噪声在两个声道中;

[0278] -讲话者A在左声道或右声道中,背景音乐在两个声道中。

[0279] 在非限制性实施方式中,单声道样本是从以16kHz采样的AT&T单声道干净语音数据库中选择的。仅使用任何方便的VAD算法(例如,如参考文献[1]中描述的3GPP EVS编解码器的VAD算法)从单声道样本中提取活动片段。具有非相关内容的立体声训练数据库的总大小约为240MB。在将单声道信号组合以形成立体声声音信号之前,不对单声道信号应用电平调整。电平调整仅在此过程之后被应用。基于被动单声道下混合,每个立体声样本的电平被归一化为-26dBov。因此,声道间电平差没有改变,并且仍然是确定主导讲话者在立体声场景中的定位的主要因素。

[0280] 相关立体声训练样本是从立体声声音信号的各种真实记录中获得的。具有相关立体声内容的训练数据库的总大小约为220MB。在非限制性实施方式中,相关立体声训练样本包含来自图4所示的以下场景的样本,图4示出了用于真实记录的立体声场景设置的俯视图:

[0281] -讲话者S1在定位P1处,靠近麦克风M1,讲话者S2在定位P2处,靠近麦克风M6;

[0282] -讲话者S1在定位P4处,靠近麦克风M3,讲话者S2在定位P3处,靠近麦克风M4;

[0283] -讲话者S1在定位P6处,靠近麦克风M1,讲话者S2在定位P5处,靠近麦克风M2;

[0284] -仅讲话者S1在定位P4处,在M1-M2立体声记录中;

[0285] -仅讲话者S1在定位P4处,在M3-M4立体声记录中;

[0286] 令训练数据库的总大小被标示为:

[0287] $N_T = N_{UNC} + N_{CORR}$ (79)

[0288] 其中N_{UNC}是非相关立体声训练样本集合的大小,N_{CORR}是相关立体声训练样本集合的大小。标记是使用例如以下简单规则来手动指派的:

[0289] $$y(i) = \begin{cases} 1 & ,i \in \Omega_{UNC} \\ 0 & ,i \in \Omega_{CORR} \end{cases} \quad (80)$$

[0290] 其中Ω_{UNC}是非相关训练数据库的整个特征集合,并且Ω_{CORR}是相关训练数据库的整个特征集合。在此说明性、非限制性的实施方式中,从训练数据库中丢弃非活动帧(VAD=0)。

[0291] 非相关训练数据库中的每个帧都被标记为“1”,并且相关训练数据库中的每个帧都被标记为“0”。针对VAD=0的非活动帧在训练过程中被忽略。

[0292] 7.1LRTD立体声模式中的UNCLR分类

[0293] 在LRTD立体声模式中,用于对立体声声音信号190进行编码的方法150包括对非相关立体声内容(UNCLR)的分类的操作161。为了执行操作161,用于对立体声声音信号190进行编码的设备100包括UNCLR分类器111。

[0294] LRTD立体声模式中的UNCLR分类的操作161基于逻辑回归(LogReg)模型。通过在非相关立体声和相关立体声训练数据库两者上运行用于对立体声声音信号进行编码的设备100(立体声编解码器)提取的以下特征被用于UNCLR分类操作161:

[0295] -声道间互相关函数的最大值的定位,k_{max}(关系(11));

[0296] -瞬时目标增益,g_t(关系(13));

[0297] -零滞后处的声道间相关函数的绝对值的对数,p_{LR}(关系(14));

[0298] -边比单声道能量比率，r_{SM}(关系(15))；

[0299] -左/右声道与单声道信号之间的点积中的最大值与最小值之间的差，d_{mmLR}(关系(19))；

[0300] -左声道(L)和单声道信号(M)之间的点积与右声道和单声道信号(M)之间的点积之间在对数域中的绝对差，d_{LRM}(关系(20))；

[0301] -交叉声道相关函数的零滞后值，R_0(关系(5))；以及

[0302] -声道间相关函数的演变，RR(关系(21))。

[0303] UNCLR分类器111总共使用数量为F＝8的特征。

[0304] 在训练过程之前，UNCLR分类器111包括归一化器(未示出)，其执行通过移除特征的均值并将其缩放至单位方差来对特征集合进行归一化的子操作(未示出)。归一化器(未示出)为此目的使用例如以下关系：

[0305] $$f_i = \frac{f_{i,raw} - \bar{f}_i}{\sigma_{f_i}}, \quad i = 1,...,F \qquad (81)$$

[0306] 其中$f_{i,raw}$标示集合的第i特征，f_i标示归一化的第i特征，$\bar{f}_i$标示跨训练数据库的第i特征的全局均值，并且σ_{fi}是跨训练数据库的第i特征的全局方差。

[0307] 由UNCLR分类器111使用的LogReg模型将实值特征作为输入向量，并且对输入属于指示非相关立体声内容(UNCLR)的非相关类(类0)的概率进行预测。为此目的，UNCLR分类器111包括得分计算器(未示出)，其执行计算表示输入立体声声音信号190中的非相关立体声内容的得分的子操作(未示出)。得分计算器(未示出)以可以使用以下关系来表达、所提取特征的线性回归的形式计算LogReg模型的输出，该输出是实值的：

[0308] $y_p = b_0 + b_1 f_1 + ... + b_F f_F$ (82)

[0309] 其中，b_i标示LogReg模型的系数，并且f_i标示个体特征。随后使用例如以下逻辑函数将实值输出y_p变换成概率：

[0310] $$p(\text{class} = 0) = \frac{1}{1 + e^{-y_p}} \qquad (83)$$

[0311] 概率p(class＝0)取0与1之间的实值。直观地，更接近1的概率意味着当前帧是高度立体声非相关的，即，具有非相关立体声内容。

[0312] 学习过程的目标是基于训练数据找到用于系数b_i，i＝1,..,F的最佳值。系数是通过最小化训练数据库上的预测输出p(class＝0)与真实输出y之间的差来迭代地寻找的。LRTD立体声模式中的UNCLR分类器111是使用例如在参考文献[10](其全部内容通过引用并入本文)中描述的随机梯度下降(SGD)迭代方法来训练的。

[0313] 通过将概率输出p(class＝0)与固定阈值(例如0.5)进行比较，可以进行二元分类。然而，出于LRTD立体声模式中的UNCLR分类的目的，不使用概率输出p(class＝0)。替代地，LogReg模型的原始输出y_p被如下所示地进一步处理。

[0314] UNCLR分类器111的得分计算器(未示出)首先使用例如图5所示的函数对LogReg模型的原始输出y_p进行归一化。图5是图示出在LRTD立体声模式中的UNCLR分类中应用于LogReg模型的原始输出的归一化函数的图。

[0315] 图5的归一化函数可以在数学上描述如下：

[0316] $$y_{pn}(n)=\begin{cases}0.5 & \text{如果 } y_p(n)\geq 4.0\\ 0.125y_p(n) & \text{如果 } -4.0<y_p(n)<4.0\\ -0.5 & \text{如果 } y_p(n)\leq -4.0\end{cases} \quad (84)$$

[0317] 7.1.1基于相对帧能量的LogReg输出加权

[0318] UNCLR分类器111的得分计算器(未示出)随后使用例如以下关系用相对帧能量对LogReg模型的归一化输出$y_{pn}(n)$进行加权：

[0319] $$scr_{UNCLR}(n)=y_{pn}(n)\cdot E_{rl}(n) \quad (85)$$

[0320] 其中$E_{rl}(n)$是由关系(69)描述的相对帧能量。LogReg模型的归一化加权输出$scr_{UNCLR}(n)$被称为输入立体声声音信号190中的上述“得分”表示或非相关立体声内容。

[0321] 7.1.2上升沿检测

[0322] 对于UNCLR分类，得分$scr_{UNCLR}(n)$仍然不能被UNCLR分类器111直接使用，因为其包含由不完美的统计模型产生的偶然短期“峰值”。这些峰值可以由诸如一阶IIR滤波器之类的简单平均滤波器滤除。遗憾的是，此类平均滤波器的应用通常导致对表示输入立体声声音信号190中的立体声相关与非相关内容之间的过渡的上升沿的涂抹。为了保留上升沿，当在输入立体声声音信号190中检测到上升沿时，平滑过程(平均IIR滤波器的应用)被减少甚至被停止。对输入立体声声音信号190中的上升沿的检测是通过分析相对帧能量$E_{rl}(n)$的演变来完成的。

[0323] 相对帧能量$E_{rl}(n)$的上升沿是通过利用级联的P＝20个相同的一阶电阻-电容(RC)滤波器对相对帧能量进行滤波来得到的，每个滤波器具有例如以下形式：

[0324] $$F_p(z)=\frac{b_1z^{-1}}{a_0+a_1z^{-1}},\quad p=1,\ldots,20 \quad (86)$$

[0325] 常量a_0、a_1和b_1被选取为使得：

[0326] $$\frac{-a_1}{a_0}=\tau_{edge},\quad \frac{b_1}{a_0}=1-\tau_{edge} \quad (87)$$

[0327] 因此，单个参数τ_{edge}被用于控制每个RC滤波器的时间常量。在实验上，发现利用τ_{edge}＝0.3可以获得很好的结果。利用级联的P＝20个RC滤波器对相对帧能量$E_{rl}(n)$的滤波可以被执行如下执行：

[0328] $$\begin{aligned}E_f^{[0]}(n)&=t_{edge}\cdot E_f^{[0]}(n-1)+\left(1-t_{edge}\right)\cdot E_{rl}(n)\\ E_f^{[1]}(n)&=t_{edge}\cdot E_f^{[1]}(n-1)+\left(1-t_{edge}\right)\cdot E_f^{[0]}(n)\\ &\cdots\\ E_f^{[p]}(n)&=t_{edge}\cdot E_f^{[p]}(n-1)+\left(1-t_{edge}\right)\cdot E_f^{[p-1]}(n)\end{aligned} \quad (88)$$

[0329] 其中添加上标p＝0,1,...,P-1以表示RC滤波器级联中的级。级联的RC滤波器的输出等于来自最后一级的输出，即

[0330] $$E_f(n)=E_f^{[P-1]}(n)=E_f^{[19]}(n) \quad (89)$$

[0331] 使用级联的一阶RC滤波器而不是单个高阶RC滤波器的原因是为了降低计算复杂度。级联的多个一阶RC滤波器充当具有相对尖锐的阶跃函数的低通滤波器。当在相对帧能

量$E_{rl}(n)$上使用时，其趋向于抹掉偶然短期尖峰，同时保留较慢但重要的过渡，诸如开始和偏移。相对帧能量$E_{rl}(n)$的上升沿可以是通过使用例如以下关系计算相对帧能量与经滤波输出之间的差来量化的：

[0332] $f_{edge}(n)=0.95-0.05(E_{rl}(n)-E_f(n))$ (90)

[0333] 项$f_{edge}(n)$被限制于区间<0.9;0.95>。UNCLR分类器111的得分计算器(未示出)利用使用例如以下关系的、使用$f_{edge}(n)$作为遗忘因子的IIR滤波器来平滑LogReg模型的归一化加权输出$scr_{UNCLR}(n)$，以产生归一化、加权和平滑的得分(LogReg模型的输出)：

[0334] $wscr_{UNCLR}(n)=f_{edge}(n)\cdot wscr_{UNCLR}(n-1)+(1-f_{edge}(n))\cdot scr_{UNCLR}(n)$ (91)

[0335] 7.2DFT立体声模式中的UNCLR分类

[0336] 在DFT立体声模式中，用于对立体声声音信号190进行编码的方法150包括对非相关立体声内容(UNCLR)的分类的操作163。为了执行操作163，用于对立体声声音信号190进行编码的设备100包括UNCLR分类器113。

[0337] DFT立体声模式中的UNCLR分类与如上所述的LRTD立体声模式中的UNCLR分类类似地进行。具体地，DFT立体声模式中的UNCLR分类也基于逻辑回归(LogReg)模型。为了简单起见，标示来自LRTD立体声模式中的UNCLR分类的特定参数的符号/名称和相关联的数学符号也被用于DFT立体声模式。当同时引用来自多个部分的相同参数时，添加下标以避免模糊。

[0338] 通过在立体声非相关和立体声相关训练数据库两者上运行用于对立体声声音信号进行编码的设备100(立体声编解码器)而提取的以下特征被UNCLR分类器113用于DFT立体声模式中的UNCLR分类:

[0339] -ILD增益，g_{ILD}(关系(43))；

[0340] -IPD增益，g_{IPD}(关系(48))；

[0341] -IPD旋转角度，φ_{rot}(关系(49))；

[0342] -预测增益，g_{pred}(关系(52))；

[0343] -声道间相干性的均值能量，E_{coh}(关系(55))；

[0344] -最大和最小声道内幅度积的比率，r_{PP}(关系(57))；

[0345] -交叉声道频谱的总体幅值，f_X(关系(41))；以及

[0346] -GCC-PHAT函数的最大值，G_{ITD}(关系(61))。

[0347] UNCLR分类器113总共使用数量为F=8的特征。

[0348] 在训练过程之前，UNCLR分类器113包括归一化器(未示出)，其执行通过移除特征的均值并将其缩放至单位方差来对特征集合进行归一化的子操作(未示出)。归一化器(未示出)为此目的使用例如以下关系：

[0349] $$f_i=\frac{f_{i,raw}-\bar{f}_i}{\sigma_{f_i}},\quad i=1,\ldots,F \qquad (92)$$

[0350] 其中，$f_{i,raw}$标示集合的第i特征，$\bar{f}_i$标示跨整个训练数据库的第i特征的全局均值，并且σ_{fi}是再次跨整个训练数据库的第i特征的全局方差。应当注意，在关系(92)中使用的全局均值$\bar{f}_i$和全局方差σ_{fi}不同于在关系(81)中使用的相同参数。

[0351] 在DFT立体声模式中使用的LogReg模型类似于在LRTD立体声模式中使用的LogReg模型。LogReg模型的输出y_p由关系(82)描述，并且当前帧具有非相关立体声内容(类别=0)

的概率由关系(83)给出。上文描述了分类器训练过程和找到最佳决策阈值的过程。再次，为此目的，UNCLR分类器113包括得分计算器(未示出)，其执行计算表示输入立体声声音信号190中的非相关立体声内容的得分的子操作(未示出)。

[0352] 类似于在LRTD立体声模式中并且根据如图5所示的函数，UNCLR分类器113的得分计算器(未示出)首先将LogReg模型的原始输出y_p归一化。归一化在数学上可以被描述如下：

[0353] $$y_{pn}(n)=\begin{cases}0.5 & \text{如果 } y_p(n)\geq 4.0\\ 0.125y_p(n) & \text{如果 } -4.0<y_p(n)<4.0\\ -0.5 & \text{如果 } y_p(n)\leq -4.0\end{cases} \quad (93)$$

[0354] 7.2.1基于相对帧能量的LogReg输出加权

[0355] UNCLR分类器113的得分计算器(未示出)随后使用例如以下关系用相对帧能量$E_{rl}(n)$对LogReg模型的归一化输出$y_{pn}(n)$进行加权：

[0356] $scr_{UNCLR}(n)=y_{pn}(n)\cdot E_{rl}(n)$ (94)

[0357] 其中$E_{rl}(n)$是由关系(69)描述的相对帧能量。

[0358] LogReg模型的加权归一化输出被称为“得分”，并且其表示与上述LRTD立体声模式中相同的量。在DFT立体声模式中，当替代VAD标志$f_{xVAD}(n)$(关系(77)被设置为0时，得分$scr_{UNCLR}(n)$被重置为0。这由以下关系来表示：

[0359] $$scr_{UNCLR}(n)=0, \text{如果} f_{xVAD}(n)=0 \quad (95)$$

[0360] 7.2.2 DFT立体声模式中的上升沿检测

[0361] UNCLR分类器113的得分计算器(未示出)最终使用LRTD立体声模式中的UNCLR分类中的上述上升沿检测机制、利用IIR滤波器来平滑DFT立体声模式中的得分$scr_{UNCLR}(n)$。为此目的，UNCLR分类器113使用以下关系：

[0362] $wscr_{UNCLR}(n)=f_{edge}(n)\cdot wscr_{UNCLR}(n-1)+(1-f_{edge}(n))\cdot scr_{UNCLR}(n)$ (96)

[0363] 这与关系(91)相同。

[0364] 7.3二元UNCLR决策

[0365] UNCLR分类器111/113的最终输出是二元状态。令$c_{UNCLR}(n)$标示UNCLR分类器111/113的二元状态。二元状态$c_{UNCLR}(n)$具有值“1”以指示非相关立体声内容类别，或者具有值“0”以指示相关立体声内容类别。UNCLR分类器111/113的输出处的二元状态是可变的。其被初始化为“0”。UNCLR分类器111/113的状态在满足特定条件的帧中从当前类别改变到另一类别。

[0366] UNCLR分类器111/113中用于在立体声内容类别之间进行切换的机制在图6中以状态机的形式被描述。

[0367] 参考图6：

[0368] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“1”(601)，(b)当前帧的经平滑得分$wscr_{UNCLR}(n)$小于“-0.07”(602)，并且(c)先前帧的变量$cnt_{sw}(n-1)$大于“0”(603)，则当前帧的二元状态$c_{UNCLR}(n)$被切换为“0”(604)；

[0369] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“1”(601)，并且(b)当前帧的经平滑得分

$wscr_{UNCLR}(n)$不小于“-0.07”(602)，则在当前帧中不存在二元状态$c_{UNCLR}(n)$的切换；

[0370] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“1”(601)，(b)当前帧的经平滑得分$wscr_{UNCLR}(n)$小于“-0.07”(602)，并且(c)先前帧的变量$cnt_{sw}(n-1)$不大于“0”(603)，则在当前帧中不存在二元状态$c_{UNCLR}(n)$的切换。

[0371] 以相同的方式，参考图6：

[0372] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“0”(601)，(b)当前帧的经平滑得分$wscr_{UNCLR}(n)$大于“0.1”(605)，并且(c)先前帧的变量$cnt_{sw}(n-1)$大于“0”(606)，则当前帧的二元状态$c_{UNCLR}(n)$被切换为“1”(607)；

[0373] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“0”(601)，并且(b)当前帧的经平滑得分$wscr_{UNCLR}(n)$不大于“0.1”(605)，则在当前帧中不存在二元状态$c_{UNCLR}(n)$的切换；

[0374] -如果(a)先前帧的二元状态$c_{UNCLR}(n-1)$为“0”(601)，(b)当前帧的经平滑得分$wscr_{UNCLR}(n)$大于“0.1”(605)，并且(c)先前帧的变量$cnt_{sw}(n-1)$不大于“0”(606)，则在当前帧中不存在二元状态$c_{UNCLR}(n)$的切换。

[0375] 最后，当前帧中的变量$cnt_{sw}(n)$被更新(608)，并且过程针对下一帧被重复(609)。

[0376] 变量$cnt_{sw}(n)$是UNCLR分类器111/113的其中可以在LRTD与DFT立体声模式之间进行切换的帧的计数值。此计数值被初始化为零，并且是在每个帧中使用例如以下逻辑来更新(608)的：

[0377]
$$cnt_{sw}(n)=\begin{cases} cnt_{sw}(n)+1, \text{如果} c_{type}\in(\text{GENERIC, UNVOICED, INACTIVE}) \\ cnt_{sw}(n)+1, \text{如果} \text{VAD}=0 \\ 0, \quad \text{否则} \end{cases} \tag{97}$$

[0378] 计数值$cnt_{sw}(n)$的上限为100。变量c_{type}指示用于对立体声声音信号进行编码的设备100中的当前帧的类型。帧类型通常是在用于对立体声声音信号进行编码的设备100(立体声编解码器)的预处理操作中确定的，具体地是在预处理器103/104/109中确定的。当前帧的类型通常是基于输入立体声声音信号190的以下特性来选择的：

[0379] -音调时段

[0380] -发声(“voicing”)

[0381] -频谱倾斜

[0382] -过零率

[0383] -帧能量差(短期、长期)

[0384] 作为非限制性实例，来自如参考文献[1]中描述的3GPP EVS编解码器的帧类型可以在UNCLR分类器111/113中用作关系(97)的参数c_{type}。3GPP EVS编解码器中的帧类型选自以下类别集合：

[0385] $c_{type}\in$(INACTIVE,UNVOICED,VOICED,GENERIC,TRANSITION,AUDIO)

[0386] 关系(97)中的参数VAD0是没有任何拖尾添加的VAD标志。没有拖尾添加的VAD标志通常是在用于对立体声声音信号进行编码的设备100(立体声编解码器)的预处理操作中计算的，具体地是在TD预处理器103/104/109中计算的。作为非限制性实例，来自如参考文献[1]中描述的3GPP EVS编解码器的没有拖尾添加的VAD标志可以在UNCLR分类器111/113中被用作参数VAD0。

[0387] 如果当前帧的类型是通用(“GENERIC”)、清音(“UNVOICED”)或非活动(“INACTIVE”),或者如果没有拖尾添加的VAD标志指示输入立体声声音信号中的非活动(VAD0＝0),则UNCLR分类器111/113的输出二元状态$c_{UNCLR}(n)$可以被更改。此类帧通常适合于在LRTD与DFT立体声模式之间进行切换,因为它们位于稳定的片段中或者位于在感知上对质量具有低影响的片段中。目标是最小化切换伪影的风险。

[0388] 8.串音(XTALK)的检测

[0389] XTALK检测基于LogReg模型,其是针对LRTD立体声模式并且针对DFT立体声模式单独训练的。两个统计模型都是在从真实立体声记录和人工准备的立体声样本的大型数据库收集的特征上训练的。在训练数据库中,每个帧被标记为单音或串音。标记在真实立体声记录的情况下被手动进行,或者在人工准备的样本的情况下被半自动地进行。手动标记是通过标识具有串音特性的短紧凑片段来进行的。半自动标记是在将单声道信号混合为立体声声音信号之前使用来自单声道信号的VAD输出来进行的。在本部分8的末尾处提供了细节。

[0390] 在本公开中描述的实现的非限制性示例中,真实立体声记录是以32kHz来被采样的。这些真实立体声记录的总大小约为263MB,对应于约30分钟。人工准备的立体声样本是通过使用ITU-T G.191混响工具混合来自单声道干净语音数据库的随机选择的讲话者来创建的。人工准备的立体声样本是通过模拟具有如图7中所示的AB麦克风设置的大型会议室中的条件来准备的。图7是具有AB麦克风设置的大型会议室的示意性平面图,其条件被模拟用于XTALK检测。

[0391] 考虑两种类型的房间,回声(LEAB)和无回声(LAAB)。参考图7,对于每种类型的房间,第一讲话者S1可以出现在定位P4、P5或P6处,并且第二讲话者S2可以出现在定位P10、P11和P12处。每个讲话者S1和S2的定位是在准备训练样本期间随机选择的。因此,讲话者S1总是靠近第一模拟麦克风M1,而讲话者2总是靠近第二模拟麦克风M2。麦克风M1和M2在图7中所示的非限制性实现中是全向的。麦克风对M1和M2构成模拟AB麦克风设置。在进一步处理之前,单声道样本从训练数据库中被随机选择,被下采样到32kHz,并且被归一化到-26dBov(dB(过载)——与在削波发生之前设备可以处理的最大值相比的音频信号的幅度)。ITU-T G.191混响工具包含用于每个讲话者/麦克风对的房间脉冲响应(RIR)的真实测量的数据库。

[0392] 针对讲话者S1和S2的随机选择的单声道样本随后与对应于给定讲话者/麦克风定位的房间脉冲响应(RIR)进行卷积,从而模拟真实AB麦克风捕捉。每个麦克风M1和M2中的来自两个讲话者S1和S2的贡献被加在一起。在卷积之前,在4-4.5秒范围内随机选择的偏移被添加到讲话者样本中的一者。这确保了在所有训练语句中总是存在某一时段的单音语音,接着是短时段的串音语音和另一时段的单音语音。在RIR卷积和混合之后,样本被再次归一化为-26dBov,这次被应用于被动单声道下混合。

[0393] 标记是使用常规VAD算法(例如,如参考文献[1]中描述的3GPP EVS编解码器的VAD算法)来半自动地创建的。VAD算法单独地应用于第一讲话者(S1)文件和第二讲话者(S2)文件。二元VAD决策随后通过逻辑“与”来被组合。这产生标记文件。组合输出等于“1”的片段确定串音片段。这在图8中被示出,图8示出了图示出使用VAD对串音样本的自动标记的图。在图8中,第一行示出来自讲话者S1的语音样本,第二行示出对来自讲话者S1的语音样本的二元VAD决策,第三行示出来自讲话者S2的语音样本,第四行示出对来自讲话者S2的语音样本

的二元VAD决策，并且第五行示出串音片段的位置。

[0394] 训练集是不平衡的。串音帧与单音帧的比例大约为1比5，即，只有大约21％的训练数据属于串音类别。这在LogReg训练过程期间通过应用如参考文献[6](其全部内容通过引用并入本文)中描述的类别权重来补偿。

[0395] 训练样本被串接并且被用作用于对立体声声音信号进行编码的设备100(立体声编解码器)的输入。特征针对每个20ms的帧、在编码过程期间、在单独的文件中被单独地收集。这构成了训练特征集合。令训练特征集合中的帧的总数被标示为例如：

[0396] $N_T = N_{XTALK} + N_{NORMAL}$ (98)

[0397] 其中N_{XTALK}是串音帧的总数，而N_{NORMAL}是单音帧的总数。

[0398] 此外，令对应的二元标记被标示为例如：

[0399] $$y(i) = \begin{cases} 1 & ,i \in \Omega_{XTALK} \\ 0 & ,i \in \Omega_{NORMAL} \end{cases} \quad (99)$$

[0400] 其中Ω_{XTALK}是所有串音帧的超集，并且Ω_{NORMAL}是所有单音帧的超集。非活动帧(VAD＝0)从训练数据库中被移除。

[0401] 8.1LRTD立体声模式中的XTALK检测

[0402] 在LRTD立体声模式中，用于对立体声声音信号进行编码的方法150包括检测串音(XTALK)的操作160。为了执行操作160，用于对立体声声音信号进行编码的设备100包括XTALK检测器110。

[0403] 在LRTD立体声模式中检测串音(XTALK)的操作160与如上所述的LRTD立体声模式中的UNCLR分类类似地进行。XTALK检测器110基于逻辑回归(LogReg)模型。为了简单起见，在本部分中也使用来自UNCLR分类的参数名称和相关联的数学符号。当指代来自不同部分的相同参数名称时，将下标添加到符号以避免模糊。

[0404] 以下特征由XTALK检测器110使用：

[0405] -L/R类别差，d_{class}(关系(32))；

[0406] -最大自相关的L/R差，d_v(关系(25))；

[0407] -LSF之和的L/R差，d_{LSF}(关系(23))；

[0408] -残差能量的L/R差，d_{LPC13}(关系(22))；

[0409] -相关图的L/R差，d_{cmap}(关系(27))；

[0410] -噪声特性的L/R差，d_{nchar}(关系(29))；

[0411] -非平稳性的L/R差，d_{sta}(关系(26))；

[0412] -频谱多样性的L/R差，d_{sdiv}(关系(28))；

[0413] -滞后0处的声道间相关函数的未归一化值，p_{LR}(关系(14))；

[0414] -边比单声道能量比率，r_{SM}(关系(15))；

[0415] -左声道与单声道声道之间以及右声道与单声道信号之间的点积的最大值与最小值之间的差，d_{mmLR}(关系(19))；

[0416] -交叉声道相关函数的零滞后值，R_0(关系(5))；

[0417] -声道间互相关函数的演变，RR(关系(21))；

[0418] -声道间互相关函数的最大值的定位，k_{max}(关系(11))；

[0419] -声道间相关函数的最大值，R_{max}(关系(10))；

[0420] -L/M与R/M点积之间的差，Δ_{LRM}(关系(20))；以及

[0421] -边信号和单声道信号的经平滑的能量比率，$\overline{r_{SM}(n)}$(关系(16))。

[0422] 相应地，XTALK检测器110使用总数为F＝17的特征。

[0423] 在训练过程之前，XTALK检测器110包括归一化器(未示出)，其执行通过移除特征的均值并将其缩放至单位方差来对17个特征f_i的集合进行归一化的子操作(未示出)。归一化器(未示出)使用例如以下关系：

[0424] $$f_i = \frac{f_{i,raw} - \overline{f_i}}{\sigma_{f_i}}, \quad i = 1,...,F \qquad (100)$$

[0425] 其中，$f_{i,raw}$标示集合的第i特征，$\overline{f_i}$标示跨训练数据库的第i特征的全局均值，并且σ_{fi}是跨训练数据库的第i特征的全局方差。此处，在关系(100)中使用的参数$\overline{f_i}$和σ_{fi}不同于在关系(81)中使用的相同参数。

[0426] LogReg模型的输出y_p由关系(82)描述，并且当前帧属于串音片段类别(类别0)的概率p(类别＝0)由关系(83)给出。训练过程和找到最优决策阈值的过程的细节在上文中在LRTD立体声模式中的UNCLR分类的描述中被提供。如上所述，为此目的，XTALK检测器110包括得分计算器(未示出)，其执行计算表示输入立体声声音信号190中的非相关立体声内容的得分的子操作(未示出)。

[0427] XTALK检测器110的得分计算器(未示出)利用例如图9所示并且经进一步处理的函数对LogReg模型的原始输出y_p进行归一化。图9是表示在LRTD立体声模式中的XTALK检测中用于缩放LogReg模型的原始输出的函数的图。此类归一化在数学上可以被描述如下：

[0428] $$y_{pn}(n) = \begin{cases} 1.0 & \text{如果 } y_p(n) \geq 3.0 \\ 0.333 y_p(n) & \text{如果 } -3.0 < y_p(n) < 3.0 \\ -1.0 & \text{如果 } y_p(n) \leq -3.0 \end{cases} \qquad (101)$$

[0429] 如果先前帧是以DFT立体声模式编码的并且当前帧是以LRTD立体声模式编码的，则LogReg模型的归一化输出$y_{pn}(n)$被设置为0。此类过程防止了切换伪影。

[0430] 8.1.1基于相对帧能量的LogReg输出加权

[0431] XTALK检测器110的得分计算器(未示出)基于相对帧能量$E_{rl}(n)$对LogReg模型的归一化输出$y_{pn}(n)$进行加权。在LRTD立体声模式中的XTALK检测器110中应用的加权方案类似于LRTD立体声模式中的UNCLR分类器111中应用的加权方案(如上所述)。主要区别在于相对帧能量$E_{rl}(n)$不直接用作如关系(85)中的乘法因子。替代地，XTALK检测器110的得分计算器(未示出)以反比例将相对帧能量$E_{rl}(n)$线性映射到区间<0;0.95>中。此映射可以是例如使用以下关系来完成的：

[0432] $$w_{relE}(n) = -2.375 E_{rl}(n) + 2.1375 \qquad (102)$$

[0433] 因此，在具有较高相对能量的帧中，权重将接近0，而在具有低能量的帧中，权重将接近0.95。XTALK检测器110的得分计算器(未示出)随后例如使用以下关系、使用权重$w_{relE}(n)$来对LogReg模型的归一化输出$y_{pn}(n)$进行滤波：

[0434] $$scr_{XTALK}(n) = w_{relE} scr_{XTALK}(n-1) + (1 - w_{relE}) y_{pn}(n) \qquad (103)$$

[0435] 其中，索引n标示当前帧，并且n-1标示先前帧。

[0436] 来自XTALK检测器110的归一化加权输出$scr_{XTALK}(n)$被称为表示输入立体声声音信号190中的串音的“XTALK得分”。

[0437] 8.1.2上升沿检测

[0438] 以与LRTD立体声模式中的UNCLR分类类似的方式，XTALK检测器110的得分计算器(未示出)平滑LogReg模型的归一化加权输出$scr_{XTALK}(n)$。原因是为了抹掉偶然短期“峰值”和“下降”，否则其将导致假警报或错误。平滑被设计成保留LogReg输出的上升沿，因为这些上升沿可能表示输入立体声声音信号190中的串音与单音片段之间的重要过渡。LRTD立体声模式中的XTALK检测器中用于上升沿的检测的机制不同于上文关于LRTD立体声模式中的UNCLR分类描述的上升沿的检测的机制。

[0439] 在XTALK检测器110中，上升沿检测算法分析来自先前帧的LogReg输出值，且将其与具有不同斜率的预计算的“理想”上升沿进行比较。“理想”上升沿被表示为帧索引n的线性函数。图10是图示出LRTD立体声模式中的XTALK检测器110中的检测上升沿的机制的图。参考图10，x轴包含当前帧0之前的帧的索引n。小灰色矩形是在当前帧之前的六个帧的时段上XTALK得分$scr_{XTALK}(n)$的示例性输出。从图10可以看出，在XTALK得分$scr_{XTALK}(n)$中，从当前帧之前三个帧处开始存在上升沿。虚线表示不同长度的片段上的四个“理想”上升沿的集合。

[0440] 对于每个“理想”上升沿，上升沿检测算法计算虚线与XTALK得分$scr_{XTALK}(n)$之间的均方误差。上升沿检测算法的输出是所测试的“理想”上升沿之中的最小均方误差。由虚线表示的线性函数是基于分别针对最小值scr_{min}和最大值scr_{max}的预定义阈值来预计算的。这在图10中由大的浅灰色矩形示出。每个“理想”上升沿线性函数的斜率取决于最小和最大阈值以及片段的长度。

[0441] 上升沿检测仅在满足以下准则的帧中由XTALK检测器110执行：

[0442]
$$\min_{k=0,..,K}(scr_{XTALK}(n-k))<0 \quad \text{以及}$$
$$\max_{k=0,..,K}(scr_{XTALK}(n-k))>0 \quad \text{以及} \qquad (104)$$
$$\max_{k=0,..,K}(scr_{XTALK}(n-k))-\min_{k=0,..,K}(scr_{XTALK}(n-k))>0.2$$

[0443] 其中K＝4是所测试的上升沿的最大长度。

[0444] 令上升沿检测算法的输出值被标示为ε_{0_1}。下标“0_1”的使用强调了上升沿检测的输出值被限制在区间<0;1>中的事实。对于不满足关系(104)中的准则的帧，上升沿检测的输出值被直接设置为0，即

[0445] $\varepsilon_{0_1}=0$ (105)

[0446] 表示所测试的“理想”上升沿的线性函数的集合在数学上可以是利用以下关系来表达的：

[0447]
$$t(l,n-k)=scr_{\max}-k\frac{scr_{\max}-scr_{\min}}{l},\quad l=1,..,K,k=1,..,l \qquad (106)$$

[0448] 其中索引l标示所测试上升沿的长度，并且n-k是帧索引。每个线性函数的斜率由三个参数确定：所测试上升沿的长度l、最小阈值scr_{min}、和最大阈值scr_{max}。出于LRTD立体声模式中的XTALK检测器110的目的，阈值被设置为$scr_{max}=1.0$和$scr_{min}=-0.2$。这些阈值的值是通过实验找到的。

[0449] 针对每个所测试上升沿的长度，上升沿检测算法使用例如以下关系来计算线性函数t(关系(106))与XTALK得分scr_{XTALK}之间的均方误差：

[0450] $$\varepsilon(l)=\frac{1}{l}\varepsilon_0+\frac{1}{l}\sum_{k=1}^{l}\left[scr_{XTALK}(n-k)-t(l,n-k)\right]^2,\quad l=1,..,K \qquad (107)$$

[0451] 其中ε_0是由下式给出的初始误差：

[0452] $$\varepsilon_0=[scr_{XTALK}(n)-scr_{max}]^2 \qquad (108)$$

[0453] 最小均方误差由XTALK检测器110使用下式计算：

[0454] $$\varepsilon_{\min}=\min_{l=1,..,K}(\varepsilon(l)) \qquad (109)$$

[0455] 最小均方误差越低，检测的上升沿越强。在非限制性实现中，如果最小均方误差高于0.3，则上升沿检测的输出被设置为0，即：

[0456] $$\varepsilon_{0_1}=0 \text{ 如果 } \varepsilon_{\min}>0.3 \qquad (110)$$

[0457] 并且上升沿检测算法退出。在所有其他情况下，可以使用例如以下关系将最小均方误差线性映射在区间<0;1>中：

[0458] $$\varepsilon_{0_1}=1-2.5\varepsilon_{min} \qquad (111)$$

[0459] 在上述示例中，上升沿检测的输出与最小均方误差之间的关系是成反比的。

[0460] XTALK检测器110将上升沿检测的输出归一化在区间<0.5;0.9>中以产生使用例如以下关系计算的边缘锐度参数：

[0461] $$f_{edge}(n)=0.9-0.4\varepsilon_{0_1} \qquad (112)$$

[0462] 其中0.5和0.9分别用作下限和上限。

[0463] 最后，XTALK检测器110的得分计算器(未示出)通过XTALK检测器110的IIR滤波器来平滑LogReg模型的归一化加权输出$scr_{XTALK}(n)$，其中$f_{edge}(n)$被用于代替遗忘因子。此类平滑使用例如以下关系：

[0464] $$wscr_{XTALK}(n)=f_{edge}(n)\cdot wscr_{XTALK}(n-1)+(1-f_{edge}(n))\cdot scr_{XTALK}(n) \qquad (113)$$

[0465] 在其中关系(77)中计算的替代VAD标志为零的帧中，经平滑输出$wscr_{XTALK}(n)$(XTALK得分)被重置为0。即：

[0466] $$wscr_{XTALK}(n)=0, \text{ 如果 } f_{xVAD}(n)=0 \qquad (114)$$

[0467] 8.2DFT立体声模式中的串音检测

[0468] 在DFT立体声模式中，用于对立体声声音信号190进行编码的方法150包括检测串音(XTALK)的操作162。为了执行操作162，用于对立体声声音信号190进行编码的设备100包括XTALK检测器112。

[0469] DFT立体声模式中的XTALK检测与LRTD立体声模式中的XTALK检测类似地进行。逻辑回归(LogReg)模型被用于输入特征向量的二元分类。为了简单起见，在本部分中也使用来自LRTD立体声模式中的XTALK检测的特定参数的名称和其相关联的数学符号。当同时参考来自两个部分的相同参数时，添加下标以避免模糊。

[0470] 通过在单音和串音训练数据库两者上运行DFT立体声模式，从用于对立体声声音信号190进行编码的设备100提取以下特征：

[0471] -ILD增益，g_{ILD}(关系(43))；

[0472] -IPD增益，g_{IPD}(关系(48))；

[0473] -IPD旋转角度，φ_{rot}(关系(49))；

[0474] -预测增益，g_{pred}(关系(52))；

[0475] -声道间相干性的均值能量，E_{coh}(关系(55))；

[0476] -最大和最小声道内幅度积的比率，r_{PP}(关系(57))；

[0477] -交叉声道频谱的总体幅值，f_X(关系(41))；

[0478] -GCC-PHAT函数的最大值，G_{ITD}(关系(61))；

[0479] -GCC-PHAT函数的第一和第二最高峰值的幅度之间的关系，r_{GITD12}(关系(64))；

[0480] -GCC-PHAT的第二最高峰值的幅度，m_{ITD2}(关系(66))；以及

[0481] -当前帧中的第二最高峰值的定位相对于先前帧中的第二最高峰值的定位的差，Δ_{ITD2}(关系(67))。

[0482] XTALK检测器112总共使用数量为F＝11的特征。

[0483] 在训练过程之前，XTALK检测器112包括归一化器(未示出)，其使用例如以下关系来执行通过移除特征的全局均值并将其缩放至单位方差来对所提取特征的集合进行归一化的子操作(未示出)。

[0484] $$f_i = \frac{f_{i,raw} - \bar{f}_i}{\sigma_{f_i}}, \quad i = 1,...,F \qquad (115)$$

[0485] 其中$f_{i,raw}$标示集合的第i特征，f_i标示归一化的第i特征，$\bar{f}_i$是跨训练数据库的第i特征的全局均值，并且σ_{fi}是跨训练数据库的第i特征的全局方差。关系(115)中使用的参数$\bar{f}_i$和σ_{fi}不同于关系(81)中使用的参数。

[0486] LogReg模型的输出完全由关系(82)描述，并且当前帧属于串音片段类别(类别0)的概率由关系(83)给出。训练过程和找到最优决策阈值的过程的细节在上文中在LRTD立体声模式中的UNCLR分类的部分中被提供。再次，为此目的，XTALK检测器112包括得分计算器(未示出)，其执行计算表示输入立体声声音信号190中的XTALK检测的得分的子操作(未示出)。

[0487] XTALK检测器112的得分计算器(未示出)使用图5所示并且经进一步处理的函数对LogReg模型的原始输出y_p进行归一化。LogReg模型的归一化输出被标示为y_{pn}。在DFT立体声模式中，不使用基于相对帧能量的加权。因此，LogReg模型的归一化加权输出(具体地，XTALK得分$scr_{XTALK}(n)$)由下式给出：

[0488] $$scr_{XTALK}(n) = y_{pn} \qquad (116)$$

[0489] 当替代VAD标志$f_{xVAD}(n)$被设置为0时，XTALK得分$scr_{XTALK}(n)$被重置为0。这可以被表达如下：

[0490] $$scr_{XTALK}(n) = 0, \text{如果} f_{xVAD}(n) = 0 \qquad (117)$$

[0491] 8.2.1上升沿检测

[0492] 如同在LRTD立体声模式中的XTALK检测的情况下，XTALK检测器112的得分计算器(未示出)平滑XTALK得分$scr_{XTALK}(n)$以移除短期峰值。此类平滑是使用如关于LRTD立体声模式中的XTALK检测器110描述的上升沿检测机制、通过IIR滤波来执行的。XTALK得分scr_{XTALK}

(n)是利用使用例如以下关系的IIR滤波器来平滑的：

[0493] $wscr_{XTALK}(n)=f_{edge}(n)\cdot wscr_{XTALK}(n-1)+(1-f_{edge}(n))\cdot scr_{XTALK}(n)$ (118)

[0494] 其中$f_{edge}(n)$是在关系(112)中计算的边缘锐度参数。

[0495] 8.3二元XTALK决策

[0496] XTALK检测器110/112的最终输出是二元的。令$c_{XTALK}(n)$标示XTALK检测器110/112的输出，其中“1”表示串音类别，而“0”表示单音类别。输出$c_{XTALK}(n)$也可以被看作是状态变量。其被初始化为0。状态变量仅在满足特定条件的帧中从当前类改变为另一类。用于串音类别切换的机制类似于上文在7.3部分中已详细描述的非相关立体声内容的类别切换的机制。然而，对于LRTD立体声模式和DFT立体声模式两者，差异都是存在的。这些差异将在下文讨论。

[0497] 在LRTD立体声模式中，XTALK检测器110使用如图11所示的串音切换机制。参考图11：

[0498] -如果当前帧n中的UNCLR分类器111的输出$c_{UNCLR}(n)$等于“1”(1101)，则在当前帧n中不存在XTALK检测器110的输出$c_{XTALK}(n)$的切换。

[0499] -如果(a)当前帧n中的UNCLR分类器111的输出$c_{UNCLR}(n)$等于“0”(1101)，并且(b)先前帧n-1中的XTALK检测器110的输出$c_{XTALK}(n-1)$等于“1”(1102)，则在当前帧n中不存在XTALK检测器110的输出$c_{XTALK}(n)$的切换。

[0500] -如果(a)当前帧n中的UNCLR分类器111的输出$c_{UNCLR}(n)$等于“0”(1101)，(b)先前帧n-1中的XTALK检测器110的输出$c_{XTALK}(n-1)$等于“0”(1102)，并且(c)当前帧n中的经平滑的XTALK得分$wscr_{XTALK}(n)$不大于0.03(1104)，则在当前帧n中不存在XTALK检测器110的输出$c_{XTALK}(n)$的切换。

[0501] -如果(a)当前帧n中的UNCLR分类器111的输出$c_{UNCLR}(n)$等于“0”(1101)，(b)先前帧n-1中的XTALK检测器110的输出$c_{XTALK}(n-1)$等于“0”(1102)，(c)当前帧n中的经平滑的XTALK得分$wscr_{XTALK}(n)$大于0.03(1104)，并且(d)先前帧n-1中的计数值$cnt_{sw}(n-1)$不大于“0”(1105)，则在当前帧n中不存在XTALK检测器110的输出$c_{XTALK}(n)$的切换。

[0502] -如果(a)当前帧n中的UNCLR分类器111的输出$c_{UNCLR}(n)$等于“0”(1101)，(b)先前帧n-1中的XTALK检测器110的输出$c_{XTALK}(n-1)$等于“0”(1102)，(c)当前帧n中的经平滑的XTALK得分$wscr_{XTALK}(n)$大于0.03(1104)，并且(d)先前帧n-1中的计数值$cnt_{sw}(n-1)$大于“0”(1105)，则当前帧n中的XTALK检测器110的输出$c_{XTALK}(n)$被切换为“1”(1106)。

[0503] 最后，当前帧n中的计数值$cnt_{sw}(n)$被更新(1107)，并且过程针对下一帧被重复(1108)。

[0504] 计数值$cnt_{sw}(n)$对于UNCLR分类器111和XTALK检测器110是公共的，并且是在关系(97)中定义的。计数值$cnt_{sw}(n)$的正值指示状态变量$c_{XTALK}(n)$(XTALK检测器110的输出$c_{XTALK}(n)$)的切换是被允许的。如图11中可见，切换逻辑使用当前帧中的UNCLR分类器111的输出$c_{UNCLR}(n)$(1101)。因此假设UNCLR分类器111在XTALK检测器110之前运行，因为它使用其输出。而且，图11的状态切换逻辑是单向的，意思是XTALK检测器110的输出$c_{XTALK}(n)$只能从“0”(单音)改变为“1”(串音)。用于相反方向的状态切换逻辑(即从“1”(串音)到“0”(单音)是DFT/LRTD立体声模式切换逻辑的部分，稍后将在本公开中对其进行描述。

[0505] 在DFT立体声模式中，XTALK检测器112包括辅助参数计算器(未示出)，其执行计算

以下辅助参数的子操作(未示出)。具体地,串音切换机制使用XTALK检测器112的输出$wscr_{XTALK}(n)$,以及以下辅助参数:

[0506] -当前帧中的话音活动检测(VAD)标志(f_{VAD});

[0507] -GCC-PHAT函数的第一和第二最高峰值的幅度,G_{ITD},m_{ITD2}(分别为关系(61)和(66));

[0508] -对应于GCC-PHAT函数的第一和第二最高峰值的定位(ITD值),d_{ITD},d_{ITD2}(分别为关系(60)和(段落[00111]));以及

[0509] -DFT立体声静音标志,f_{sil}关系(78)。

[0510] 在DFT立体声模式中,XTALK检测器112使用如图12所示的串音切换机制。参考图12:

[0511] -如果$d_{ITD}(n)$等于“0”(1201),则$c_{XTALK}(n)$被切换为“0”(1217);

[0512] -如果(a)$d_{ITD}(n)$不等于“0”(1201),并且(b)$c_{XTALK}(n-1)$不等于“0”(1202),

[0513] ■如果(c)$c_{XTALK}(n-1)$不等于“1”(1215),则不存在$c_{XTALK}(n)$的切换;

[0515] ■如果(c)$c_{XTALK}(n-1)$等于“1”(1215),并且(d)$wscr_{XTALK}(n)$不小于“0.0”(1216),则不存在$c_{XTALK}(n)$的切换;

[0516] ■如果(c)$c_{XTALK}(n-1)$等于“1”(1215),并且(d)$wscr_{XTALK}(n)$小于“0.0”(1216),则$c_{XTALK}(n)$被切换为“0”(1219);

[0517] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),并且(c)f_{VAD}不等于“1”(1203),

[0518] ■如果(d)$c_{XTALK}(n-1)$不等于“1”(1215),则不存在$c_{XTALK}(n)$的切换;

[0519] ■如果(d)$c_{XTALK}(n-1)$等于“1”(1215),并且(e)$wscr_{XTALK}(n)$不小于“0.0”(1216),则不存在$c_{XTALK}(n)$的切换;

[0520] ■如果(d)$c_{XTALK}(n-1)$等于“1”(1215),并且(e)$wscr_{XTALK}(n)$小于“0.0”(1216),则$c_{XTALK}(n)$被切换为“0”(1219);

[0521] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),(d)$0.8G_{ITD}(n)$小于$m_{ITD2}(n)$(1204),(e)$0.8G_{ITD}(n-1)$小于$m_{ITD2}(n-1)$(1205),(f)$d_{ITD2}(n)-d_{ITD2}(n-1)$小于“4.0”(1206),(g)$G_{ITD}(n)$大于“0.15”(1207),(h)$G_{ITD}(n-1)$大于“0.15”(1208),则$c_{XTALK}(n)$被切换为“1”(1218);

[0522] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),并且(d)测试1204到1208中的任一个是否定的,

[0523] ■如果(e)$wscr_{XTALK}(n)$大于“0.8”(1209),则$c_{XTALK}(n)$被切换为“1”(1218);

[0524] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),(d)测试1204到1208中的任一个是否定的,(e)$wscr_{XTALK}(n)$不大于“0.8”(1209),并且(f)$f_{sil}(n)$不等于“1”(1210),

[0525] ■如果(g)$c_{XTALK}(n-1)$不等于“1”(1215),则不存在$c_{XTALK}(n)$的切换;

[0526] ■如果(g)$c_{XTALK}(n-1)$等于“1”(1215),并且(h)$wscr_{XTALK}(n)$不小于“0.0”(1216),则不存在$c_{XTALK}(n)$的切换;

[0527] ■如果(g)$c_{XTALK}(n-1)$等于“1”(1215),并且(h)$wscr_{XTALK}(n)$小于“0.0”(1216),则$c_{XTALK}(n)$被切换为“0”(1219);

[0528] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),(d)测试1204到1208中的任一个是否定的,(e)$wscr_{XTALK}(n)$不大于“0.8”(1209),(f)$f_{sil}(n)$等于“1”(1210),(g)$d_{ITD}(n)$大于“8.0”(1211),并且(h)$d_{ITD}(n-1)$小于“-8.0”,则$c_{XTALK}(n)$被切换为“1”(1218);

[0529] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),(d)测试1204到1208中的任一个是否定的,(e)$wscr_{XTALK}(n)$不大于“0.8”(1209),(f)$f_{sil}(n)$等于“1”(1210),(g)测试1211和1212中的任一个是否定的,(h)$d_{ITD}(n-1)$大于“8.0”(1213),并且(i)$d_{ITD}(n)$小于“-8.0”(1214),则$c_{XTALK}(n)$被切换为“1”(1218);

[0530] -如果(a)$d_{ITD}(n)$不等于“0”(1201),(b)$c_{XTALK}(n-1)$等于“0”(1202),(c)f_{VAD}等于“1”(1203),(d)测试1204到1208中的任一个是否定的,(e)$wscr_{XTALK}(n)$不大于“0.8”(1209),(f)$f_{sil}(n)$等于“1”(1210),(g)测试1211和1212中的任一个是否定的,(h)测试1213和1214中的任一个是否定的,

[0531] ■如果(i)$c_{XTALK}(n-1)$不等于“1”(1215),则不存在$c_{XTALK}(n)$的切换;

[0532] ■如果(i)$c_{XTALK}(n-1)$等于“1”(1215),并且(j)$wscr_{XTALK}(n)$不小于“0.0”(1216),则不存在$c_{XTALK}(n)$的切换;

[0533] ■如果(i)$c_{XTALK}(n-1)$等于“1”(1215),并且(j)$wscr_{XTALK}(n)$小于“0.0”(1216),则$c_{XTALK}(n)$被切换为“0”(1219);

[0534] 最后,当前帧n中的计数值$cnt_{sw}(n)$被更新(1220),并且过程针对下一帧被重复(1221)。

[0535] 变量$cnt_{sw}(n)$是其中可以在LRTD与DFT立体声模式之间进行切换的帧的计数值。此计数值$cnt_{sw}(n)$对于UNCLR分类器113和XTALK检测器112是公共的。计数值$cnt_{sw}(n)$被初始化为零,并且是在每个帧中根据关系(97)来更新的。

[0536] 9.DFT/LRTD立体声模式选择

[0537] 用于对立体声声音信号190进行编码的方法150包括选择LRTD或DFT立体声模式的操作164。为了执行操作164,用于对立体声声音信号190进行编码的设备100包括LRTD/DFT立体声模式选择器114,其接收延迟一帧(191)的来自XTALK检测器110的XTALK决策、来自UNCLR分类器的UNCLR决策111、来自XTALK检测器112的XTALK决策、以及来自UNCLR分类器113的UNCLR决策。

[0538] LRTD/DFT立体声模式选择器114基于UNCLR分类器111/113的二元输出$c_{UNCLR}(n)$和XTALK检测器110/112的二元输出$c_{XTALK}(n)$来选择LRTD或DFT立体声模式。LRTD/DFT立体声模式选择器114还考虑某些辅助参数。这些参数主要用于防止感知上敏感片段中的立体声模式切换,或者防止其中UNCLR分类器111/113和XTALK检测器110/112都不提供准确输出的片段中的频繁切换。

[0539] 选择LRTD或DFT立体声模式的操作164是在输入立体声声音信号190的下混合和编码之前执行的。因此,如图1中的191所示,操作164使用来自先前帧的UNCLR分类器111/113和XTALK检测器110/112的输出。选择LRTD或DFT立体声模式的操作164在图13的示意性框图中被进一步描述。

[0540] 如将在以下描述中描述的,在操作164中使用的DFT/LRTD立体声模式选择机制包括以下子操作:

[0541] -初始DFT/LRTD立体声模式选择;以及

[0542] -检测到串音内容时的LRTD到DFT立体声模式切换。

[0543] 9.1初始DFT/LRTD立体声模式选择

[0544] DFT立体声模式是用于对在输入立体声声音信号190的左(L)声道与右(R)声道之间具有高声道间相关的单音语音进行编码的优选模式。

[0545] LRTD/DFT立体声模式选择器114通过确定先前处理的帧是否“可能是语音帧”来开始立体声模式的初始选择。例如,这可以是通过检查“语音”类别与“音乐”类别之间的对数似然比来完成的。对数似然比被定义为由“音乐”源生成的输入立体声声音信号帧的对数似然与由“语音”源生成的输入立体声声音信号帧的对数似然之间的绝对差。以下关系可以被用于计算对数似然比:

[0546] $dL_{SM}(n)=L_{M}(n)-L_{S}(n)$ (119)

[0547] 其中$L_{S}(n)$是“语音”类别的对数似然,而$L_{M}(n)$是“音乐”类别的对数似然。

[0548] 例如,来自如参考文献[7](其全部内容通过引用并入本文)中描述的3GPP EVS编解码器的高斯混合模型(GMM)可以被用于估计“语音”类别的对数似然$L_{S}(n)$,以及“音乐”类别的对数似然$L_{M}(n)$。语音/音乐分类的其他方法也可以被用于计算对数似然比(差分得分)$dL_{SM}(n)$。

[0549] 利用例如使用以下关系的、具有不同遗忘因子的两个IIR滤波器来平滑对数似然比$dL_{SM}(n)$:

[0550] $$\begin{aligned} wdL_{SM}^{(1)}(n)&=0.97\cdot wdL_{SM}^{(1)}(n-1)+0.03\cdot dL_{SM}(n-1)\\ wdL_{SM}^{(2)}(n)&=0.995\cdot wdL_{SM}^{(2)}(n-1)+0.005\cdot dL_{SM}(n-1)\end{aligned} \quad (120)$$

[0551] 其中相应地,上标(1)指示第一IIR滤波器,而上标(2)指示第二IIR滤波器。

[0552] 随后将经平滑值$wdL_{SM}^{(1)}(n)$和$wdL_{SM}^{(2)}(n)$与预定阈值进行比较,并且如果例如满足以下组合条件,则将新的二元标志$f_{SM}(n)$设置为1:

[0553] $$f_{SM}(n)=\begin{cases}1 & \text{如果 } wdL_{SM}^{(1)}(n)<1.0 \text{ 以及 } wdL_{SM}^{(2)}(n)<0.0\\ 0 & \text{否则}\end{cases} \quad (121)$$

[0554] 标志$f_{SM}(n)=1$是先前帧可能是语音帧的指示符。阈值1.0是通过实验找到的。

[0555] 如果先前帧n-1中的UNCLR分类器111/113的二元输出$c_{UNCLR}(n-1)$或XTALK检测器110/112的二元输出$c_{XTALK}(n-1)$被设置为1,并且如果先前帧可能是语音帧,则初始DFT/LRTD立体声模式选择机制将新的二元标志$f_{UX}(n)$设置为1。这可以由以下关系来表达:

[0556] $$f_{UX}(n)=\begin{cases}1 & \text{如果 } f_{SM}(n)=1 \text{ 以及 } (c_{UNCLR}(n-1)=1 \text{ 或者 } c_{XTALK}(n-1)=1)\\ 0 & \text{否则}\end{cases} \quad (122)$$

[0557] 令$M_{SMODE}(n)\in(LRTD,DFT)$是标示当前帧n中所选择的立体声模式的离散变量。立体声模式在每个帧中是利用来自先前帧n-1的值来初始化的,即:

[0558] $M_{SMODE}(n)=M_{SMODE}(n-1)$ (123)

[0559] 如果标志$f_{UX}(n)$被设置为1,则LRTD立体声模式被选择用于在当前帧n中进行编码。这可以被表达如下:

[0560] $$M_{SMODE}(n) \leftarrow \text{LRTD} \text{ 如果 } f_{UX} = 1 \qquad (124)$$

[0561] 如果在当前帧n中标志$f_{UX}(n)$被设置为0，并且先前帧n-1中的立体声模式是LRTD立体声模式，则来自LRTD/DFT立体声模式选择器114的LRTD能量分析处理器1301的辅助立体声模式切换标志$f_{TDM}(n-1)$(将在下文描述)被分析以使用例如以下关系来选择当前帧n中的立体声模式：

[0562] $$M_{SMODE}(n) \leftarrow \begin{cases} \text{LRTD} & \text{如果 } f_{SM}(n) = 1 \text{ 以及 } f_{TDM}(n-1) = 1 \\ \text{DFT} & \text{否则} \end{cases} \qquad (125)$$

[0563] 仅在处于LRTD模式中的每个帧中更新辅助立体声模式切换标志$f_{TDM}(n)$。参数$f_{TDM}(n)$的更新在下文的描述中被描述。

[0564] 如图13中所示，LRTD/DFT立体声模式选择器114包括LRTD能量分析处理器1301以产生稍后在本公开中更详细描述的辅助参数$f_{TDM}(n)$、$c_{LRTD}(n)$、$c_{DFT}(n)$和$m_{TD}(n)$。

[0565] 如果在当前帧n中标志$f_{UX}(n)$被设置为0，并且上一先前帧n-1中的立体声模式是DFT立体声模式，则不执行立体声模式切换，并且在当前帧n中同样选择DFT立体声模式。

[0566] 9.2XTALK检测时的LRTD到DFT立体声模式切换

[0567] 已在前面的描述中描述了LRTD模式中的XTALK检测器110。从图11可以看出，XTALK检测器110的二元输出$c_{XTALK}(n)$只能在串音内容在当前帧中被检测到时被设置为1。结果，如上所述的初始立体声模式选择逻辑在XTALK检测器110指示单音内容时无法选择DFT立体声模式。在串音立体声声音信号片段后面跟着单音立体声声音信号片段的情形中，这可能导致LRTD立体声模式的不希望的扩展。因此，已经实现了用于在检测到单音内容时从LRTD立体声模式切换回DFT立体声模式的附加机制。该机制在以下描述中被描述。

[0568] 如果LRTD/DFT立体声模式选择器114在先前帧n-1中选择LRTD立体声模式并且初始立体声模式选择在当前帧n中选择LRTD模式，并且如果同时XTALK检测器110的二元输出$c_{XTALK}(n-1)$为1，则立体声模式可以从LRTD改变为DFT立体声模式。后者的改变例如当满足下列条件时被允许：

[0569] $$M_{SMODE}(n) \leftarrow \text{DFT} \text{ 如果 } \begin{bmatrix} M_{SMODE}(n) = \text{LRTD} \text{ 以及} \\ M_{SMODE}(n-1) = \text{LRTD} \text{ 以及} \\ c_{XTALK}(n) = 1 \text{ 以及} \\ f_{UX}(n-1) = 1 \text{ 以及} \\ c_{LRTD}(n-1) > 15 \text{ 以及} \\ c_{DFT}(n-1) > 3 \text{ 以及} \\ clas(n-1) \in \begin{pmatrix} \text{UNVOICED_CLAS} \\ \text{UNVOICED_TRANSITION} \\ \text{VOICED_TRANSITION} \end{pmatrix} \text{ 以及} \\ (brate \geq 16400 \text{ 或者 } wscr_{XTALK}(n-1) < 0.01) \end{bmatrix} \qquad (126)$$

[0570] 上面定义的条件集合包含对clas和brate参数的引用。brate参数是高级常量，包含用于对立体声声音信号进行编码的设备100(立体声编解码器)使用的总比特率。其是在立体声编解码器的初始化期间设置的，并在编码过程期间保持不变。

[0571] clas参数是一个离散变量，包含有关帧类型的信息。对clas参数的估计通常为立体声编解码器的信号预处理的部分。作为非限制性示例，来自如参考文献[1]中描述的3GPP EVS编解码器的帧擦除隐藏(FEC)模块的clas参数可以被用于DFT/LRTD立体声模式选择机制。来自3GPP EVS编解码器的FEC模块的clas参数是在考虑帧擦除隐藏和解码器恢复策略的情况下选择的。clas参数是从以下预定义的类别集合中选择的：

[0572] $$clas \in \begin{pmatrix} \text{UNVOICED_CLAS, UNVOICED_TRANSITION, VOICED_TRANSITION,} \\ \text{VOICED_CLAS, ONSET_CLAS, AUDIO_CLAS} \end{pmatrix}$$

[0573] 利用帧类型分类的其他手段实现的DFT/LRTD立体声模式选择机制在本公开的范围内。

[0574] 在上面定义的条件集合(126)中，条件

[0575] $$clas(n-1) \in \begin{pmatrix} \text{UNVOICED_CLAS} \\ \text{UNVOICED_TRANSITION} \\ \text{VOICED_TRANSITION} \end{pmatrix}$$

[0576] 是指在用于对立体声声音信号进行编码的设备100在DFT立体声模式下运行时、在下混合单声道(M)声道的预处理期间计算的clas参数。

[0577] 在用于对立体声声音信号进行编码的设备100处于LRTD立体声模式的情况下，条件应当被替换为：

[0578] $$clas_L(n-1) \in \begin{pmatrix} \text{UNVOICED_CLAS} \\ \text{UNVOICED_TRANSITION} \\ \text{VOICED_TRANSITION} \end{pmatrix} \text{以及} \; clas_R(n-1) \in \begin{pmatrix} \text{UNVOICED_CLAS} \\ \text{UNVOICED_TRANSITION} \\ \text{VOICED_TRANSITION} \end{pmatrix}$$

[0579] 其中索引“L”和“R”分别指在左(L)声道和右(R)声道的预处理模块中计算的clas参数。

[0580] 参数$c_{LRTD}(n)$和$c_{DFT}(n)$分别是LRTD和DFT帧的计数值。作为LRTD能量分析处理器1301的部分，这些计数值在每个帧中被更新。两个计数值$c_{LRTD}(n)$和$c_{DFT}(n)$的更新在下一部分中被详细描述。

[0581] 9.3在LRTD能量分析模块中计算的辅助参数

[0582] 当用于对立体声声音信号进行编码的设备100在LRTD立体声模式下运行时，LRTD/DFT立体声模式选择器114计算或更新若干辅助参数以改进DFT/LRTD立体声模式选择机制的稳定性。

[0583] 对于某些特殊类型的帧，LRTD立体声模式在所谓的“TD子模式”中运行。TD子模式通常被应用于从LRTD立体声模式切换到DFT立体声模式之前的短暂过渡时段。LRTD立体声模式是否将在TD子模式中运行由二元子模式标志$m_{TD}(n)$指示。二元标志$m_{TD}(n)$是辅助参数之一，并且可以在每个帧中被初始化如下：

[0584] $$m_{TD}(n) = f_{TDM}(n-1) \qquad (127)$$

[0585] 其中$f_{TDM}(n)$为本部分稍后描述的上述辅助切换标志。

[0586] 在其中$f_{UX}(n) = 1$的帧中，二元子模式标志$m_{TD}(n)$被重置为0或1。用于重置$m_{TD}(n)$的条件例如被定义如下：

[0587] $$m_{TD}(n) \leftarrow \begin{cases} 1 & \text{如果 } f_{UX}(n)=1 \text{ 则} \\ & \text{如果 } f_{TDM}(n-1)=1 \text{ 或者 } M_{SMODE}(n-1) \neq \text{LRTD 或者 } c_{LRTD}(n-1)<5 \\ 0 & \text{否则} \end{cases} \quad (128)$$

[0588] 如果$f_{UX}(n)=0$,则二元子模式标志$m_{TD}(n)$不变。

[0589] LRTD能量分析处理器1301包括上述两个计数值$c_{LRTD}(n)$和$c_{DFT}(n)$。计数值$c_{LRTD}(n)$是辅助参数之一,并且对连续的LRTD帧的数量进行计数。此计数值在DFT立体声模式在用于对立体声声音信号进行编码的设备100中被选择的每个帧中被设置为0,并且在LRTD立体声模式被选择的每个帧中递增1。这可以被表达如下:

[0590] $$c_{LRTD}(n)=\begin{cases} c_{LRTD}(n-1)+1 & \text{如果 } f_{UX}(n)=1 \\ 0 & \text{否则} \end{cases} \quad (129)$$

[0591] 实质上,计数值$c_{LRTD}(n)$包含自从最后的DFT->LRTD切换点以来的帧的数量。计数值$c_{LRTD}(n)$受阈值100的限制。计数值$c_{DFT}(n)$对连续DFT帧的数量进行计数。计数值$c_{DFT}(n)$是辅助参数之一,并且在LRTD立体声模式在用于对立体声声音信号进行编码的设备100中被选择的每个帧中被设置为0,并且在DFT立体声模式被选择的每个帧中递增1。这可以被表达如下:

[0592] $$c_{DFT}(n)=\begin{cases} c_{DFT}(n-1)+1 & \text{如果 } f_{TDM}(n)=0 \\ 0 & \text{否则} \end{cases} \quad (130)$$

[0593] 实质上,计数值$c_{DFT}(n)$包含自从最后的LRTD->DFT切换点以来的帧的数量。计数值$c_{DFT}(n)$受阈值100的限制。

[0594] 在LRTD能量分析处理器1301中计算的最后辅助参数是辅助立体声模式切换标志$f_{TDM}(n)$。此参数在每个帧中利用二元标志$f_{UX}(n)$来初始化,如下所示:

[0595] $$f_{TDM}(n)=f_{UX}(n) \quad (131)$$

[0596] 当输入立体声声音信号190的左(L)声道和右(R)声道异相(OOP)时,辅助立体声模式切换标志$f_{TDM}(n)$被设置为0。用于OOP检测的示例性方法可以例如在参考文献[8](其全部内容通过引用并入本文)中找到。当检测到OOP情形时,二元标志s2m在当前帧n中被设置为1,否则其被设置为零。当二元标志s2m被设置为1时,LRTD立体声模式中的辅助立体声模式切换标志$f_{TDM}(n)$被设置为0。这可以利用关系(132)来表达:

[0597] $$f_{TDM}(n) \leftarrow 0 \text{ 如果 } s2m(n)=1 \quad (132)$$

[0598] 如果二元标志s2m(n)被设置为零,则辅助切换标志$f_{TDM}(n)$可以例如基于以下条件集合被重置为零:

[0599] $$f_{TDM}(n) \leftarrow 0 \text{ 如果} \left[\begin{array}{l} m_{TD}(n)=1 \text{ 以及} \\ c_{LRTD}(n)>10 \text{ 以及} \\ \left[\begin{array}{l} \text{VAD}=0 \text{ 或} \\ c_{UNCLR}(n)=0 \text{ 以及 } \left(scr_{XTALK}(n)<-0.8 \text{ 或 } wscr_{XTALK}(n)<-0.13\right) \text{ 或} \\ c_{UNCLR}(n)=1 \text{ 以及 } clas(n-1)=\text{UNVOICED_CLAS 以及 } \left|wscr_{UNCLR}(n)\right|<0.005 \end{array}\right] \end{array}\right]$$

[0600] (133)

[0601] 当然,DFT/LRTD立体声模式切换机制可以是利用用于OOP检测的其他方法来实现的。

[0602] 辅助立体声模式切换标志$f_{TDM}(n)$也可以基于以下条件集合被重置为0:

[0603] $$f_{TDM}(n)\leftarrow 0 \text{ 如果 } \left[\begin{array}{l} m_{TD}(n)=0 \text{ 以及} \\ \left[\begin{array}{l} \mathrm{VAD}=0 \text{ 或} \\ c_{UNCLR}(n)=0 \text{ 以及 } \left(scr_{XTALK}(n)\leq 0 \text{ 或 } wscr_{XTALK}(n)\leq 0.1\right) \text{ 或} \\ c_{UNCLR}(n)=1 \text{ 以及 } clas(n-1)=\mathrm{UNVOICED_CLAS} \text{ 以及 } \left|wscr_{UNCLR}(n)\right|<0.025 \end{array}\right] \end{array}\right] \quad (134)$$

[0604] 在如上面定义的两个条件集合中,条件

[0605] clas(n-1)=UNVOICED_CLAS

[0606] 是指在用于对立体声声音信号进行编码的设备100在DFT立体声模式下运行时、在下混合单声道(M)声道的预处理期间计算的clas参数。

[0607] 在用于对立体声声音信号进行编码的设备100处于LRTD立体声模式的情况下,条件应当被替换为:

[0608] $$clas_L(n-1)=\mathrm{UNVOICED_CLAS} \text{ 以及 } clas_R(n-1)=\mathrm{UNVOICED_CLAS}$$

[0609] 其中索引“L”和“R”分别指在左(L)声道和右(R)声道的预处理期间计算的clas参数。

[0610] 10.核心编码器

[0611] 用于对立体声声音信号进行编码的方法150包括在LRTD立体声模式中对立体声声音信号190的左声道(L)进行核心编码的操作115、在LRTD立体声模式中对立体声声音信号190的右声道(R)进行核心编码的操作116、以及在DFT立体声模式中对立体声声音信号190的下混合单声道(M)进行核心编码的操作117。

[0612] 为了执行操作115,用于对立体声声音信号进行编码的设备100包括核心编码器115,例如单声道核心编码器。为了执行操作116,设备100包括核心编码器116,例如单声道核心编码器。最后,为了执行操作167,用于对立体声声音信号进行编码的设备100包括核心编码器117,其能够在DFT立体声模式中操作以对立体声声音信号190的下混合单声道(M)声道进行编码。

[0613] 选择合适的核心编码器115、116和117被认为是在本领域普通技术人员的知识内。相应地,在本公开中将不进一步描述这些编码器。

[0614] 11.硬件实现

[0615] 图14是形成上述用于对立体声声音信号进行编码的设备100和方法150的硬件组件的示例配置的简化框图。

[0616] 用于对立体声声音信号进行编码的设备100可以被实现为移动终端的部分、便携式媒体播放器的部分或在任何类似设备中。设备100(在图14中被标识为1400)包括输入1402、输出1404、处理器1406和存储器1408。

[0617] 输入1402被配置为接收数字或模拟形式的图1的输入立体声声音信号190。输出1404被配置为供应输出的、经编码的立体声声音信号。输入1402和输出1404可以被实现于

公共模块中，例如串行输入/输出设备。

[0618] 处理器1406可操作地连接到输入1402、输出1404和存储器1408。处理器1406被实现为用于执行代码指令的一个或多个处理器，该代码指令支持如图1所示的用于对立体声声音信号进行编码的设备100的各种组件的功能。

[0619] 存储器1408可以包括用于存储可由处理器1406执行的代码指令的非暂态存储器，具体地，包括/存储非暂时性指令的处理器可读存储器，指令在被执行时导致处理器实现如本公开中描述的用于对立体声声音信号进行编码的方法150和设备100的操作和组件。存储器1408还可以包括随机存取存储器或缓冲器，以存储来自由处理器1406执行的各种功能的中间处理数据。

[0620] 本领域普通技术人员将认识到，对用于对立体声声音信号进行编码的设备100和方法150的描述仅是说明性的，并且不旨在以任何方式进行限制。受益于本公开的本领域普通技术人员将很容易想到其他实施例。此外，所公开的用于对立体声声音信号进行编码的设备100和方法150可以被定制，以提供针对对声音进行编码和解码的现有需求和问题的有价值的解决方案。

[0621] 为了清楚起见，并未示出和描述用于对立体声声音信号进行编码的设备100和方法150的实现的所有常规特征。当然，将理解，在用于对立体声声音信号进行编码的设备100和方法150的任何此类实际实现的开发中，可能需要做出许多特定于实现的决定以便实现开发者的特定目标，诸如符合与应用、系统、网络和商业相关的约束，并且这些特定目标将因实现而异，并且因开发人员而异。此外，应当理解，虽然开发工作可能是复杂且耗时的，但是对于受益于本公开的声音处理领域的普通技术人员来说仍然是常规的工程任务。

[0622] 根据本公开，本文描述的组件/处理器/模块、处理操作和/或数据结构可以是使用各种类型的操作系统、计算平台、网络设备、计算机程序和/或通用机器来实现的。另外，本领域普通技术人员将认识到，也可以使用通用性质较差的设备，诸如硬连线设备、现场可编程门阵列(FPGA)、专用集成电路(ASIC)等。在包括一系列操作和子操作的方法由处理器、计算机或机器实现并且那些操作和子操作可以被存储为处理器、计算机或机器可读的一系列非暂态代码指令的情况下，它们可以被存储在有形和/或非暂态介质上。

[0623] 如本文所述的用于对立体声声音信号进行编码的设备100和方法150可以使用适用于本文描述的目的的软件、固件、硬件或者软件、固件或硬件的任何组合。

[0624] 在如本文所述的用于对立体声声音信号进行编码的设备100和方法150中，各种操作和子操作可以是以各种次序执行的，并且操作和子操作中的某些可以是任选的。

[0625] 尽管上文已经通过本公开的非限制性、说明性实施例描述了本公开，但是这些实施例可以在所附权利要求的范围内被任意修改，而不脱离本公开的精神和本质。

[0626] 12.参考文献

[0627] 本公开提及以下参考文献，其全部内容通过引用并入本文：

[0628] [1]3GPP TS 26.445,v.12.0.0,“Codec for Enhanced Voice Services(EVS);Detailed Algorithmic Description(用于增强型话音服务(EVS)的编解码器;详细的算法描述)”,2014年9月。

[0629] [2]M.Neuendorf,M.Multrus,N.Rettelbach,G.Fuchs,J.Robillard,J.Lecompte,S.Wilde,S.Bayer,S.Disch,C.Helmrich,R.Lefevbre,P.Gournay等人的“The ISO/MPEG

Unified Speech and Audio Coding Standard-Consistent High Quality for All Content Types and at All Bit Rates(ISO/MPEG统一语音和音频编码标准——针对所有内容类型和所有比特率的一致高质量)”,音频工程学会杂志,第61卷,第12期,第956-977页,2013年12月。

[0630] [3]F.Baumgarte,C.Faller,“Binaural cue coding-Part I:Psychoacoustic fundamentals and design principles(双耳线索编码——部分I:心理声学基础和设计原则)”,IEEE语音音频处理会刊,第11卷,第509-519页,2003年11月。

[0631] [4]Tommy Vaillancourt的“Method and system using a long-termcorrelation difference between left and right channels for time domain down mixing a stereo sound signal into primary and secondary channels(用于将立体声声音信号时域下混合到主声道和辅声道中的、使用左声道与右声道之间的长期相关差的方法和系统)”,美国专利10,325,606B2。

[0632] [5]3GPP SA4文稿S4-170749“New WID on EVS Codec Extension for Immersive Voice and Audio Services(用于沉浸式话音和音频服务的EVS编解码器扩展上的新WID)”,SA4会议#94,2017年6月26-30日,http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/TSGS4_94/Docs/S4-170749.zip

[0633] [6]I.Mani,J.Zhang在关于从不平衡数据集学习的研讨会记录中的“kNN approach to unbalanced data distributions:A case study involving information extraction(不平衡数据分布的kNN办法:涉及信息提取的案例研究)”,第1-7页,2003.KNN

[0634] [7]V.Malenovsky,T.Vaillancourt,W.Zhe,K.Choo和V.Atti的“Two-stage speech/music classifier with decision smoothing and sharpening in the EVS codec(EVS编解码器中具有决策平滑和锐化的两级语音/音乐分类器)”,2015IEEE国际声学会议,语音和信号处理(ICASSP),布里斯班,昆士兰,2015,第5718-5722页。

[0635] [8]Vaillancourt,T.的“Method and system for time-domain down mixing astereo sound signal into primary and secondary channels using detecting an out-of-phase condition on the left and right channels(用于使用检测左右声道上的异相条件将立体声声音信号时域下混合到主声道和辅声道的方法和系统)”,美国专利US 10,522,157。

[0636] [9]Maalouf,Maher.的“Logistic regression in data analysis:An overview(数据分析中的逻辑回归:概论)”,2011国际数据分析技术与策略杂志.3.281-299.10.1504/IJDATS.2011.041335。

[0637] [10]Ruder,S.的“An overview of gradient descent optimization algorithms(梯度下降优化算法概论)”.2016.ArXiv预印本ArXiv:1609.04747。

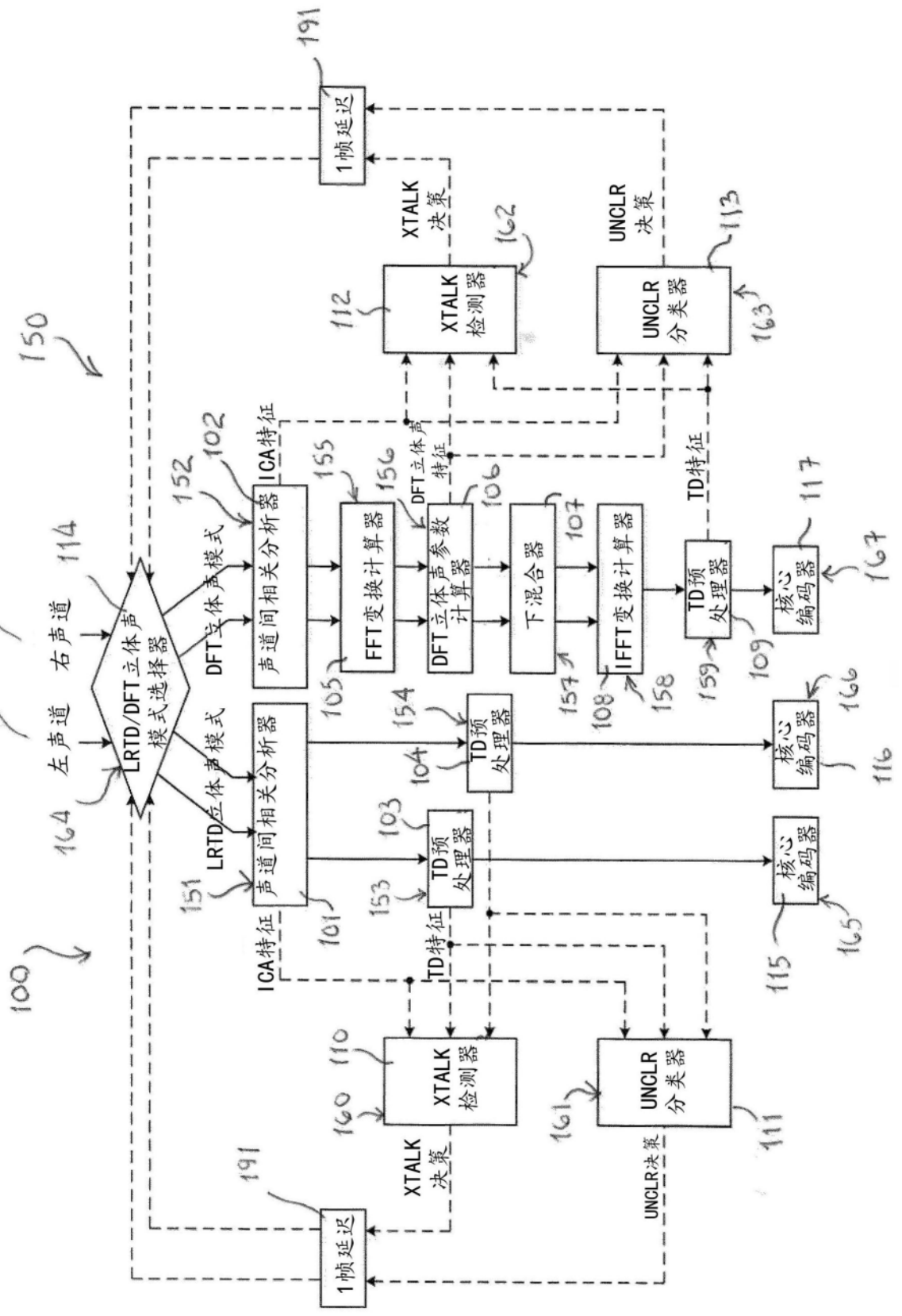

100
150
190
191
左声道
右声道
LRTD/DFT立体声模式选择器
LRTD立体声模式
DFT立体声模式
声道间相关分析器
ICA特征
声道间相关分析器
ICA特征
FFT变换计算器
DFT立体声参数计算器
DFT立体声特征
下混合器
IFFT变换计算器
TD特征
TD预处理器
TD预处理器
TD预处理器
核心编码器
核心编码器
核心编码器
XTALK检测器
XTALK检测器
XTALK决策
UNCLR分类器
UNCLR分类器
UNCLR决策
1帧延迟
1帧延迟

图1

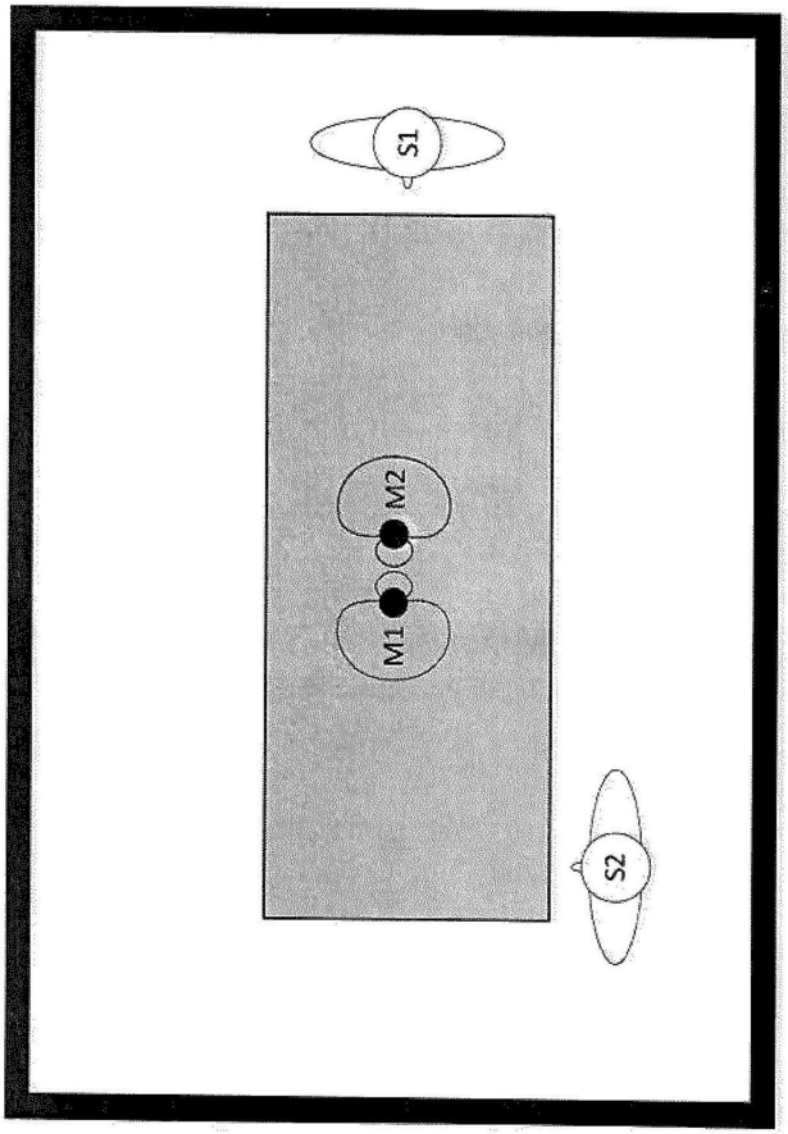


图2

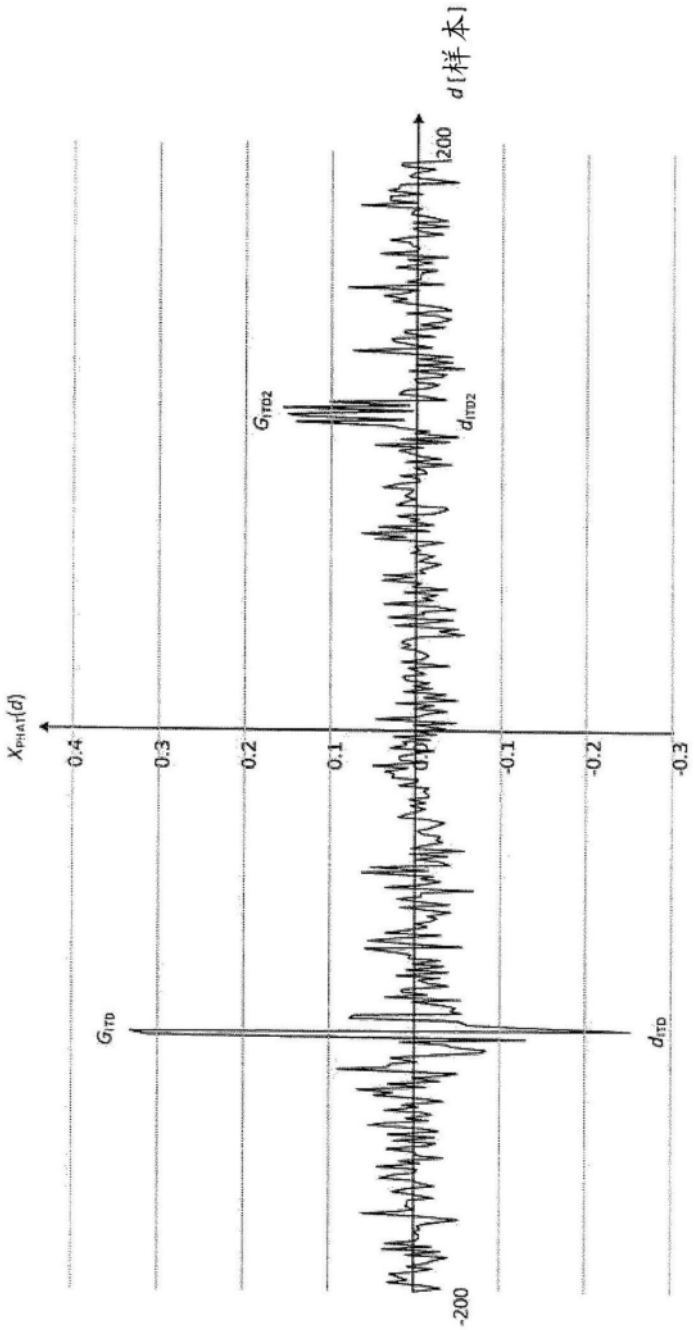


图3

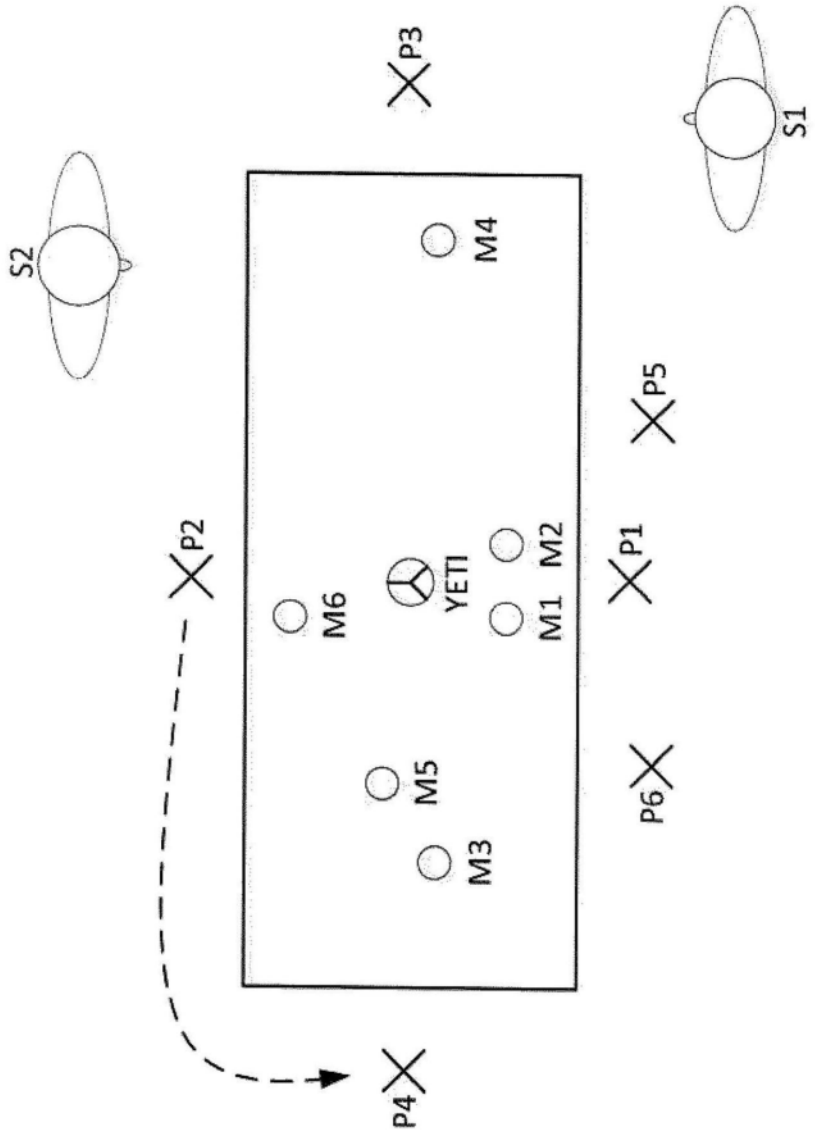


图4

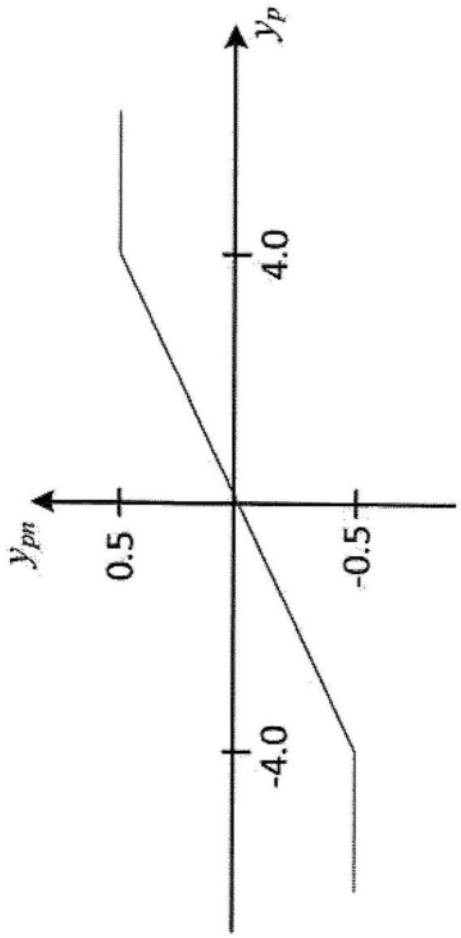


图5

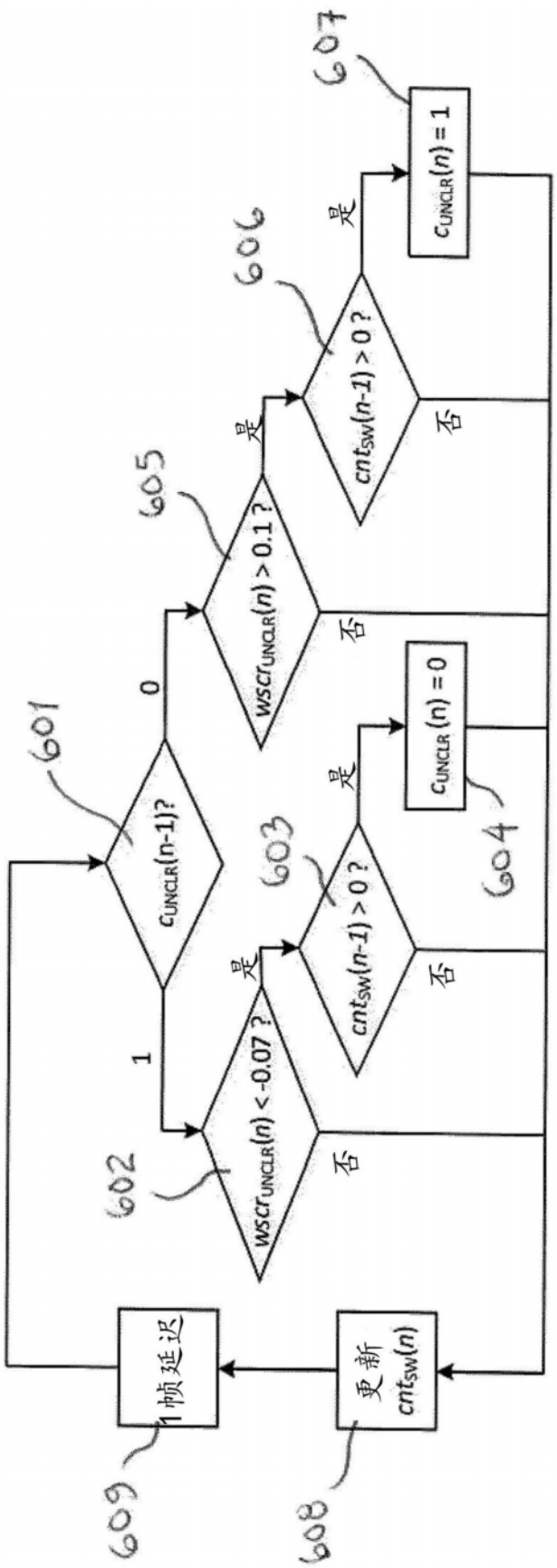

601
$c_{UNCLR}(n-1)$?
0
1
602
$wscr_{UNCLR}(n) < -0.07$?
是
否
603
$cnt_{SW}(n-1) > 0$?
是
否
604
$c_{UNCLR}(n) = 0$
605
$wscr_{UNCLR}(n) > 0.1$?
是
否
606
$cnt_{SW}(n-1) > 0$?
是
否
607
$c_{UNCLR}(n) = 1$
608
更新 $cnt_{SW}(n)$
609
1帧延迟

图6

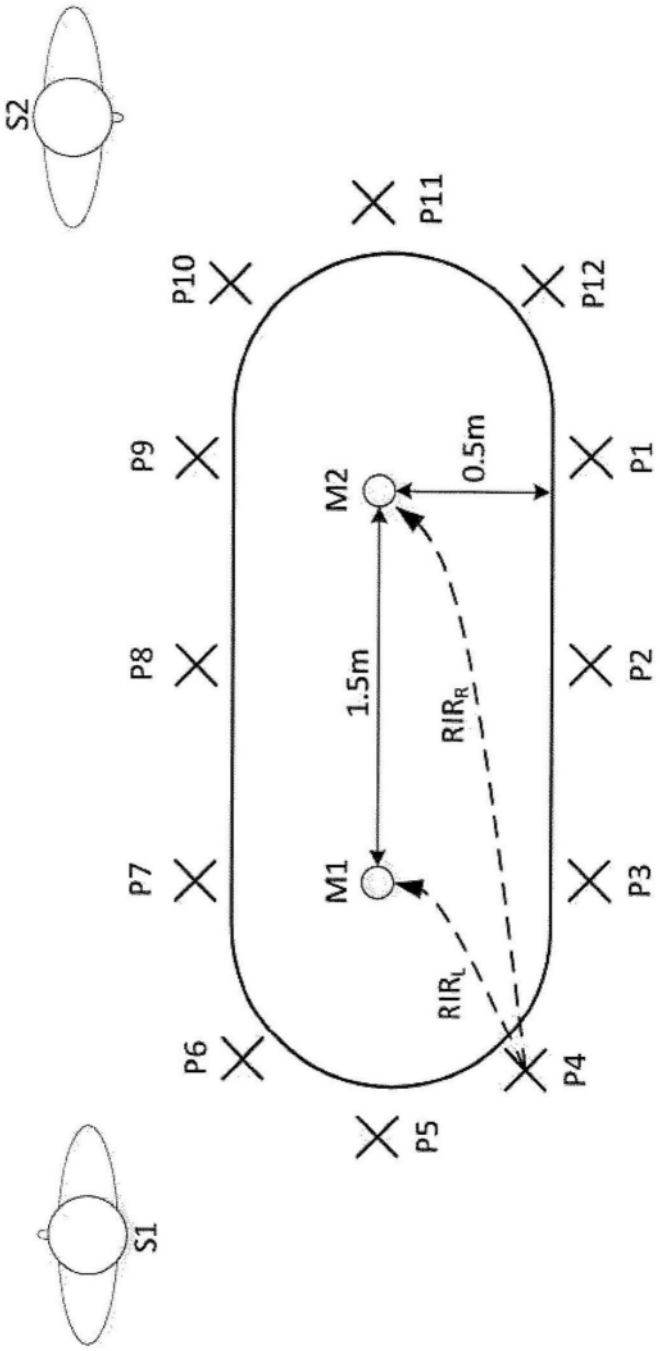


图7

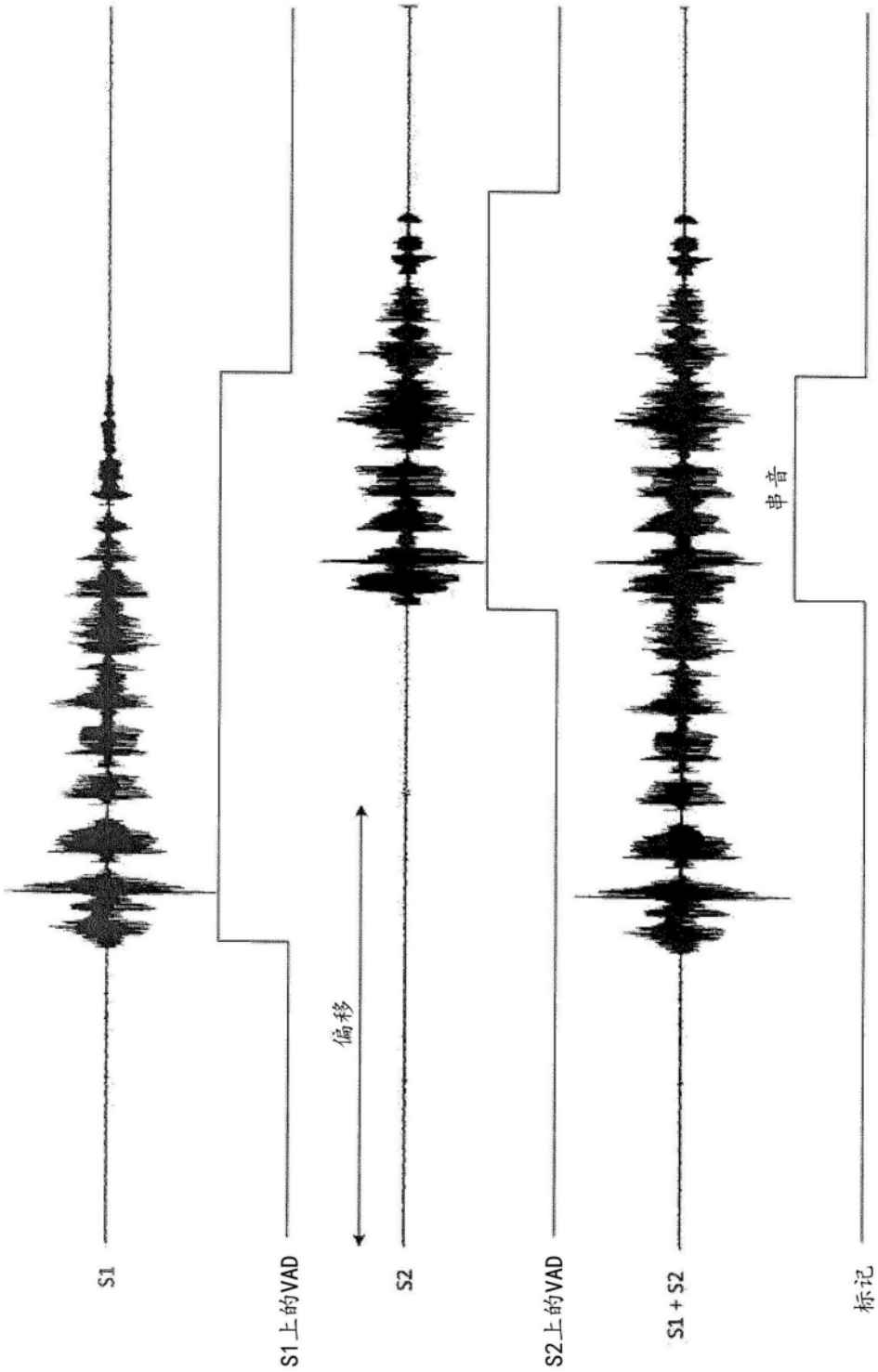


图8

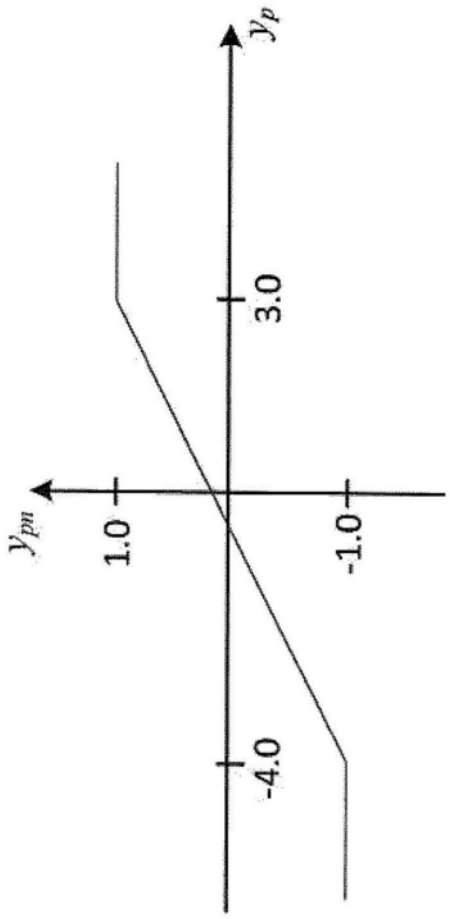


图9

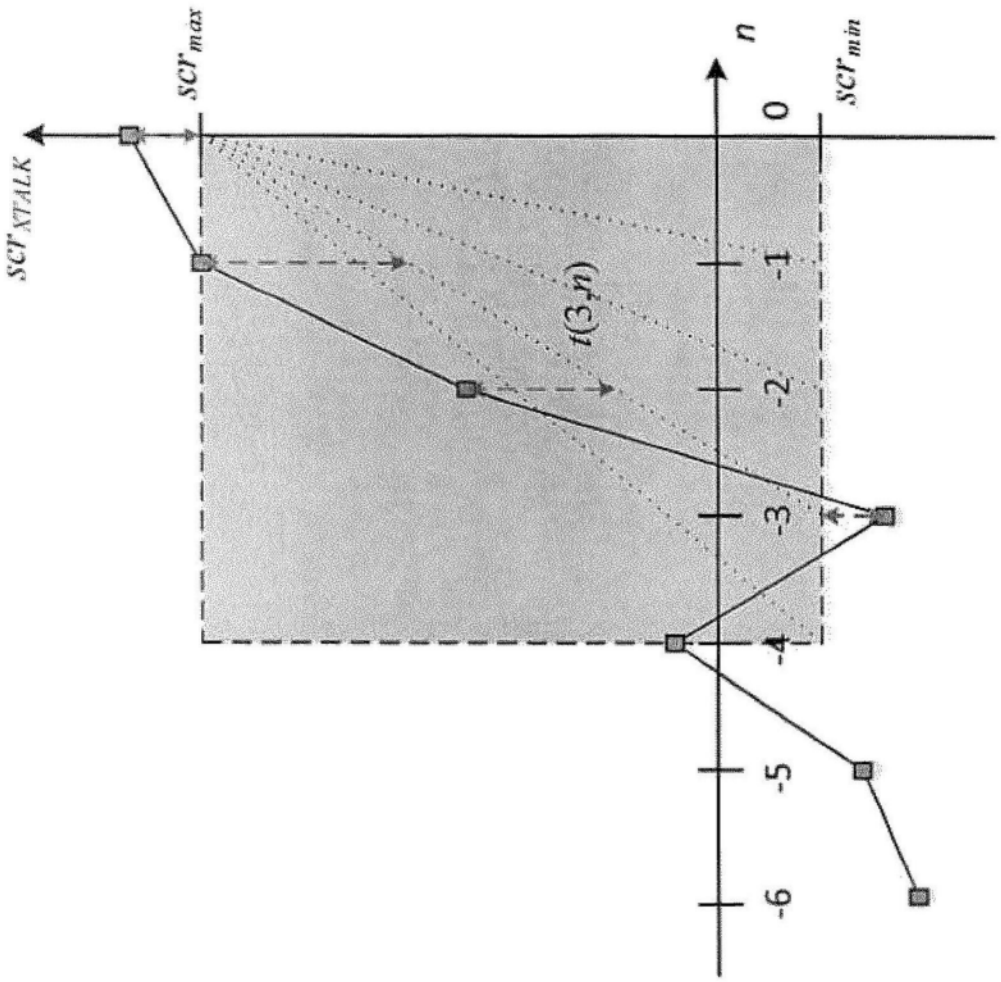


图10

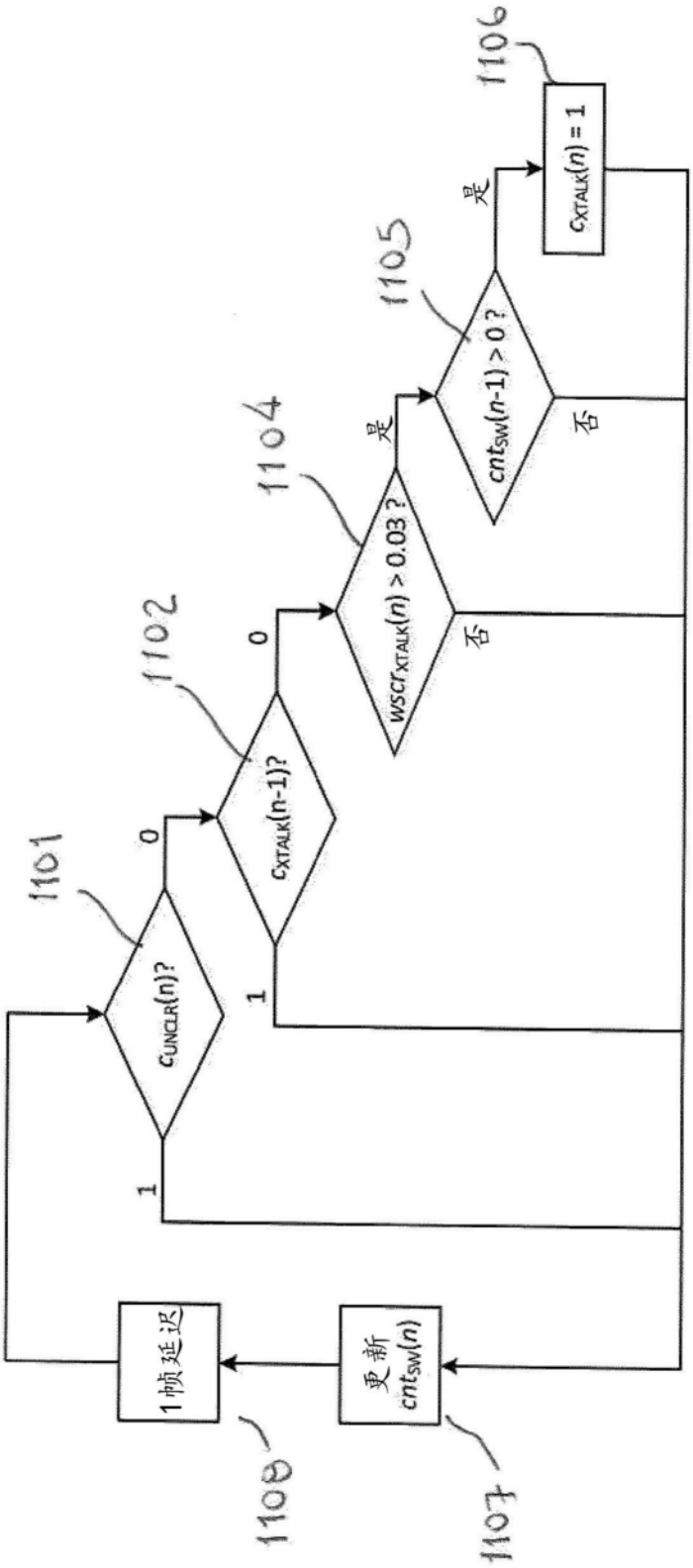

1101
$c_{UNCLA}(n)$?
0
1
1102
$c_{XTALK}(n-1)$?
0
1
1104
$WSCF_{XTALK}(n) > 0.03$?
是
否
1105
$cnt_{SW}(n-1) > 0$?
是
否
1106
$c_{XTALK}(n) = 1$
1107
更新 $cnt_{SW}(n)$
1108
1帧延迟

图11

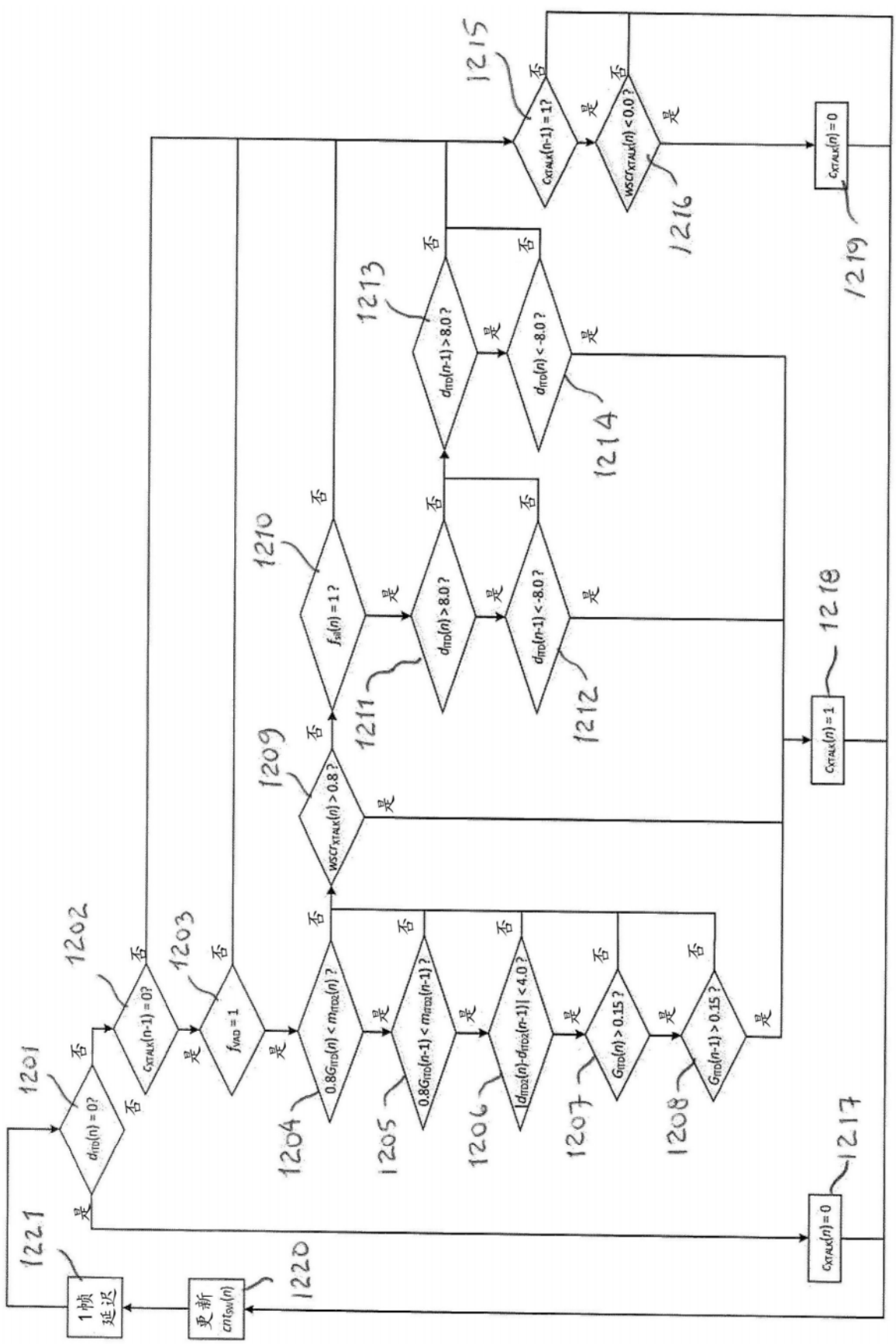

1201
$d_{ITD}(n) = 0$?
1202
$c_{XTALK}(n-1) = 0$?
1203
$f_{VAD} = 1$
1204
$0.8G_{ITD}(n) < m_{ITD2}(n)$?
1205
$0.8G_{ITD}(n-1) < m_{ITD2}(n-1)$?
1206
$|d_{ITD2}(n)-d_{ITD2}(n-1)| < 4.0$?
1207
$G_{ITD}(n) > 0.15$?
1208
$G_{ITD}(n-1) > 0.15$?
1209
$wSCF_{XTALK}(n) > 0.8$?
1210
$f_{all}(n) = 1$?
1211
$d_{ITD}(n) > 8.0$?
1212
$d_{ITD}(n-1) < -8.0$?
1213
$d_{ITD}(n-1) > 8.0$?
1214
$d_{ITD}(n) < -8.0$?
1215
$c_{XTALK}(n-1) = 1$?
1216
$wSCF_{XTALK}(n) < 0.0$?
1217
$c_{XTALK}(n) = 0$
1218
$c_{XTALK}(n) = 1$
1219
$c_{XTALK}(n) = 0$
1220
更新 $cnt_{SW}(n)$
1221
1帧延迟
否
是

图12

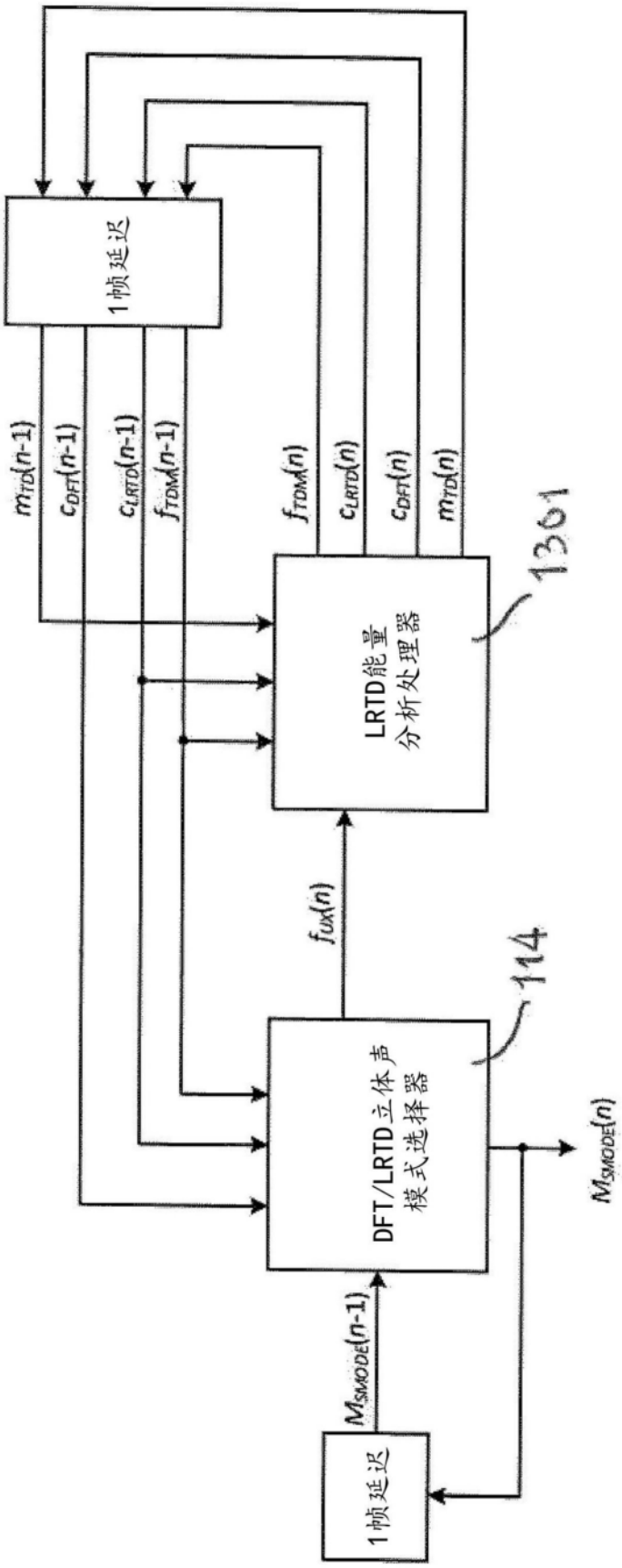

1帧延迟
$m_{TD}(n-1)$
$c_{DFT}(n-1)$
$c_{LRTD}(n-1)$
$f_{TDM}(n-1)$
$f_{TDM}(n)$
$c_{LRTD}(n)$
$c_{DFT}(n)$
$m_{TD}(n)$
LRTD能量分析处理器
1301
$f_{UX}(n)$
DFT/LRTD立体声模式选择器
114
$M_{SMODE}(n)$
$M_{SMODE}(n-1)$
1帧延迟

图13

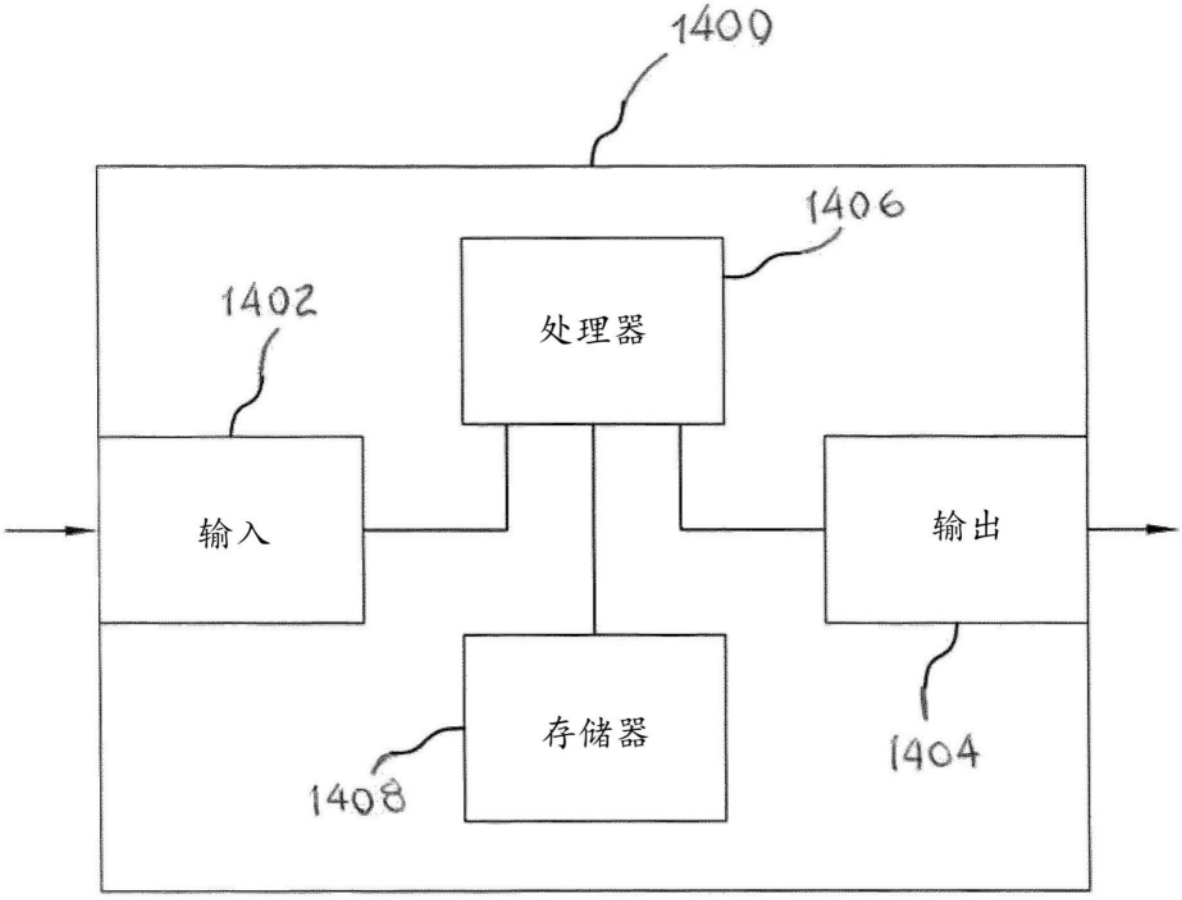


图14