



[12] 发明专利申请公开说明书

[21] 申请号 03819918.1

[43] 公开日 2005 年 10 月 5 日

[11] 公开号 CN 1679082A

[22] 申请日 2003.8.15 [21] 申请号 03819918.1

[30] 优先权

[32] 2002.8.30 [33] US [31] 10/233,073

[86] 国际申请 PCT/US2003/025627 2003.8.15

[87] 国际公布 WO2004/021332 英 2004.3.11

[85] 进入国家阶段日期 2005.2.22

[71] 申请人 杜比实验室特许公司

地址 美国加利福尼亚

[72] 发明人 马克·S·温登

查尔斯·Q·鲁宾逊

肯尼思·J·刚德瑞

史蒂文·J·维尼兹尔

杰弗里·C·瑞德米勒尔

[74] 专利代理机构 中国国际贸易促进委员会专利商
标事务所

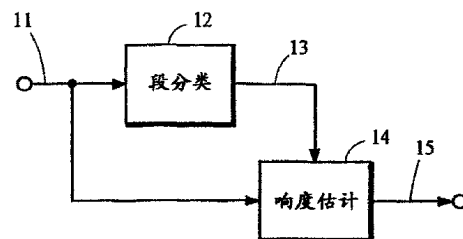
代理人 李德山

权利要求书 5 页 说明书 21 页 附图 2 页

[54] 发明名称 控制包含语音和其它类型音频素材
的信号中的语音响度

[57] 摘要

通过将音频信息各段分类为语音或非语音来获得包含语音和其它类型音频素材的音频信号的响度指示。估计语音段的响度并且该估计用于导出响度指示。响度指示可以用于控制音频信号电平以减少不同节目之间的语音响度的变化。公开了分类语音段的一种优选方法。



1. 一种信号处理的方法，包括：

接收输入信号并从输入信号中获得音频信息，其中该音频信息代表音频信号的一个片段；

检查该音频信息以将该音频信息的各段分类为代表归类为语音的音频信号部分的语音段或代表不归类为语音的音频信号部分的非语音段，其中由一段代表的该音频信号的每部分具有相应响度，并且所述语音段的响度低于一个或多个音量大的非语音段的响度；

检查该音频信息以获得所述语音段的估计响度；以及

通过生成比对于所述非语音段代表的音频信号部分的响度更响应于所述语音段估计响度的控制信息，提供该音频信号片段的响度指示。

2. 权利要求 1 的方法，该方法包括：

响应于控制信息而控制所述音频信号片段的响度，以减少该语音段响度的变化，其中当该语音段代表的音频信号部分的响度增加时，由所述一个或多个音量大的非语音段代表的音频信号部分的响度增加。

3. 权利要求 1 的方法，该方法包括：

将该音频信息的表示和控制信息组合到一个输出信号中并发送该输出信号。

4. 权利要求 1 或 2 的方法，该方法通过计算语音段代表的音频信号的频率加权形式的平均功率来获得该语音段的估计响度。

5. 权利要求 1 或 2 的方法，该方法通过对音频信息应用响度心理声学模型来获得该语音段的估计响度。

6. 权利要求 1 或 2 的方法，该方法通过从该音频信息导出音频信号的多个特征，用各自重要性的度量加权每个特征，并按照各加权特征的组合分类各段来对各段分类。

7. 权利要求 1 或 2 的方法，该方法通过仅在音频能量的度量小

于阈值的音频信号片段期间调整响度来控制该音频信号片段的响度。

8. 权利要求 1 或 2 的方法，其中音频信号片段的响度指示仅响应于语音段的估计响度。

9. 权利要求 1 或 2 的方法，该方法包括估计所述一个或多个非语音段的响度，其中该音频信号片段的响度指示比对于所述一个或多个非语音段的估计响度更响应于所述语音段的估计响度。

10. 权利要求 1 或 2 的方法，该方法包括：

提供指示相应段代表的音频信号具有语音特征的程度的语音度量；以及

提供该响度指示以使其根据相应段的语音度量响应于该相应段的估计响度。

11. 权利要求 1 或 2 的方法，该方法包括提供该响度指示以使其根据各段的时间次序响应于相应段的估计响度。

12. 权利要求 1 或 2 的方法，该方法包括响应于音频信息的特征而适配音频信息各段的长度。

13. 一种媒体，该媒体对于某一装置是可读的并且传送该装置可执行的指令程序以执行信号处理方法，其中该方法包括执行下面操作的步骤：

接收输入信号并从输入信号中获得音频信息，其中该音频信息代表音频信号的一个片段；

检查该音频信息以将该音频信息的各段分类为代表归类为语音的音频信号部分的语音段或代表不归类为语音的音频信号部分的非语音段，其中由一段代表的该音频信号的每部分具有相应响度，并且所述语音段的响度低于一个或多个音量大的非语音段的响度；

检查该音频信息以获得所述语音段的估计响度；以及

通过生成比对于所述非语音段代表的音频信号部分的响度更响应于所述语音段估计响度的控制信息，提供该音频信号片段的响度指示。

14. 权利要求 13 的媒体，其中的方法包括：

响应于控制信息控制音频信号片段的响度以减少该语音段响度的变化，其中当该语音段代表的音频信号部分的响度增加时由一个或多个音量大的非语音段代表的音频信号部分的响度增加。

15. 权利要求 13 的媒体，其中的方法包括：

将该音频信息和控制信息合成为一个输出信号并发送该输出信号。

16. 权利要求 13 或 14 的媒体，其中的方法通过计算语音段代表的音频信号频率加权复本的平均功率来获得该语音段的估计响度。

17. 权利要求 13 或 14 的媒体，其中的方法通过对音频信息应用响度心理声学模型来获得该语音段的估计响度。

18. 权利要求 13 或 14 的媒体，其中的方法通过从该音频信息导出多个音频信号特征，用各自重要性的度量加权每个特征，并按照该加权特征的组合分类来划分段的类型。

19. 权利要求 13 或 14 的媒体，其中的方法通过仅对音频能量的度量小于阈值的音频信号片段调整响度来控制该音频信号片段的响度。

20. 权利要求 13 或 14 的媒体，其中音频信号片段的响度指示仅对应于语音段的估计响度。

21. 权利要求 13 或 14 的媒体，其中的方法包括估计一个或多个非语音段的响度，其中该音频信号片段的响度指示比一个或多个非语音段的估计响度更反映语音段的估计响度。

22. 权利要求 13 或 14 的媒体，其中的方法包括：

提供指示相应段代表的音频信号具有语音特征的程度的语音度量；

提供响度指示以使其根据相应段的语音度量反映该相应段的估计响度。

23. 权利要求 13 或 14 的媒体，其中的方法包括提供响度指示使其根据该段的时间次序反映相应段的估计响度。

24. 权利要求 13 或 14 的媒体，其中的方法包括针对音频信息的

特征调整音频信息段的长度。

25. 一种信号处理的装置，其中该装置包括：

一个接收输入信号的输入终端；

存储器；以及

与输入终端和存储器相连的处理电路；其中该处理电路适合于：

接收输入信号并从输入信号中获得音频信息，其中该音频信息代表音频信号的一个片段；

检查该音频信息以将该音频信息的各段分类为代表归类为语音的音频信号部分的语音段或代表不归类为语音的音频信号部分的非语音段，其中由一段代表的该音频信号的每部分具有相应响度，并且所述语音段的响度低于一个或多个音量大的非语音段的响度；

检查该音频信息以获得所述语音段的估计响度；以及

通过生成比对于所述非语音段代表的音频信号部分的响度更响应于所述语音段估计响度的控制信息，提供该音频信号片段的响度指示。

26. 权利要求 25 的装置，其中处理电路适合于根据控制信息来控制音频信号片段的响度以减少该语音段响度的变化，其中当该语音段代表的音频信号部分的响度增加时，由所述一个或多个音量大的非语音段代表的音频信号部分的响度随之增加。

27. 权利要求 25 的装置，其中处理电路适合于将该音频信息的表示和控制信息组合到一个输出信号并发送该输出信号。

28. 权利要求 25 或 26 的装置，其中处理电路适合于通过计算语音段代表的音频信号的频率加权形式的平均功率来获得该语音段的估计响度。

29. 权利要求 25 或 26 的装置，其中处理电路适合于通过对音频信息应用响度心理声学模型来获得该语音段的估计响度。

30. 权利要求 25 或 26 的装置，其中处理电路适合于通过从该音频信息导出音频信号的多个特征，用各自重要性的度量加权每个特征，并按照各加权特征的组合分类来对各段分类。

31. 权利要求 25 或 26 的装置，其中处理电路适合于通过仅在音频能量的度量小于阈值的音频信号片段期间调整响度来控制该音频信号片段的响度。

32. 权利要求 25 或 26 的装置，其中音频信号片段的响度指示仅响应于语音段的估计响度。

33. 权利要求 25 或 26 的装置，其中处理电路适合于估计一个或多个非语音段的响度，其中该音频信号片段的响度指示比对于所述一个或多个非语音段的估计响度更响应于所述语音段的估计响度。

34. 权利要求 25 或 26 的装置，其中处理电路适合于：

提供指示相应段代表的音频信号具有语音特征的程度的语音度量；以及

提供该响度指示以使其根据相应段的语音度量响应于该相应段的估计响度。

35. 权利要求 25 或 26 的装置，其中处理电路适合于提供该响度指示以使其根据各段的时间次序响应于相应段的估计响度。

36. 权利要求 25 或 26 的装置，其中处理电路适合于检测音频信息的特征以及根据检测的特征适配音频信息各段的长度。

控制包含语音和其它类型 音频素材的信号中的语音响度

技术领域

本发明涉及与度量和控制包含语音和其它类型音频素材的音频信号中的语音响度有关的音频系统和方法。

背景技术

当听收音机或电视广播时，听者经常选择音量控制设置以获得满意的语音响度。所需的音量控制设置受到多种因素的影响，如收听环境中的背景噪声，再现系统的频率响应以及个人喜好等。在选择了音量控制设置之后，听者通常希望不管其它节目素材如音乐或声音效果存在与否，语音的响度都保持相对恒定。

当节目发生变化或者选择了其它频道时，新节目中的语音响度通常是不同的，需要调整音量控制设置以恢复所需的响度。若有的话，在通过模拟广播技术传送的节目中该设置通常仅仅需要适度的变化以调整语音响度，这是因为大多数模拟广播传送节目的语音都接近通过模拟广播系统可能传送的最大允许电平。这通常通过压缩音频节目素材的动态范围来提高相对于广播系统各个组件引入的噪声的语音信号电平。即便如此，对于不同频道接收的节目以及对于相同频道接收的不同类型的节目如商业通告或“广告”和它们打断的节目，在语音响度方面仍然存在着不希望有的差别。

因为数字广播可能以足够的信噪比传送信号而不压缩动态范围，不设置语音电平为接近最大的允许电平，所以数字广播技术的引入可能使该问题恶化。结果，在相同频道的不同节目之间和不同频道节目之间的语音响度很有可能存在大得多的差别。例如，已经发现在由模拟和数字电视频道接收的节目之间语音电平的差别有时会超过 20dB。

可以减少响度差别的一个方法是所有的数字广播将语音电平设置为一个正好低于最大电平的标准响度，这将为宽动态范围素材提供足够的峰值储备以避免压缩或限幅的需要。不幸的是，该解决方法需要改变广播实际，是不可能发生的。

通过在美国为数字电视广播采用的 AC-3 音频编码技术提供了另一种解决方法。遵守 AC-3 标准的数字广播将元数据与编码的音频数据一起传送。元数据包括已知为“调节规格 (dialnorm)”的控制信息，该信息可以在接收机端用于调整信号电平以提供一致的或规格化的语音响度。换句话说，调节规格信息允许接收机自动完成否则是听者必须做的工作，适当为每个节目或频道调整音量。听者为一个特定的节目调整音量控制设置以获得所需的语音响度水平，而尽管在不同节目或频道之间存在其他方面的差别，接收机使用调节规格信息以保证维持所需的水平。可以从高级电视系统委员会 (ATSC) 2001 年 8 月 20 日出版的标题为“Revision A to Digital Audio Compression (AC-3) Standard”的 A/52A 文件和 ATSC 1995 年 10 月 4 日出版的标题为“Guide to Use of the ATSC Digital Television Standard” A/54 文件获得描述调节规格信息使用的附加信息。

调节规格的适当值对于生成 AC-3 适应编码信号的编码系统部分必须是可用的。编码处理需要一个衡量或估计特定节目的语音响度的方法，以确定可以用于保持来自接收机的该节目语音响度的调节规格的值。

估计语音响度的方法有许多种。国际电工技术委员会 (IEC) 出版的标题为“Integrating-averaging sound level meters”的标准 IEC 60804 (2000-10) 描述了一个基于频率加权和时间平均声压级的度量。国际标准化组织出版的标题为“Method for calculating loudness level”的 ISO 标准 532: 1975 描述了从计算的频率子带功率水平的组合中获得响度度量的方法。可以用于估计响度的心理声学模型的示例参见 Moore, Glasberg 和 Baer 的文章“A model for prediction of thresholds, loudness and partial loudness”, J. Audio Eng. Soc., vol. 45, no. 4,

April 1997, 和 Glasberg 与 Moore 的文章“A model of loudness applicable to time-varying sounds”, J. Audio Eng. Soc., vol. 50, no. 5, May 2002.

不幸的是, 没有应用这些及其它已知技术的便利的途径。例如, 在广播应用中广播者必须选择音频素材的一个片段, 衡量或估计该选择片段的语音响度并把该度量传递给将调节规格信息插入 AC-3 适应数字数据流的设备。选择的片段必须包含代表性的语音而不包含会使响度度量失真的其他类型的音频素材。因为节目中包含特意比语音更响或更低音的其他成分, 所以度量一个音频节目的整个响度一般是不可接受的。通常需要比推荐的语音电平响得多的音乐和声音效果。而对于背景声音效果如风、遥远的交通或静静流淌的水就不需要和语音一样的响度, 这也是显而易见的。

发明人已经意识到确定音频信号是否包含语音的技术可以用于改进调节规格信息的适当值的创建。可以采用多种语音检测技术的任何一种。下面几种技术的说明参见引用的参考文献。

1981 年 7 月 28 日发布的美国专利 4281218 说明的技术通过提取一个或多个语音特征如短时功率来划分信号是语音还是非语音。该划分用于为语音或非语音选择适当的信号处理方法。

1992 年 3 月 17 日发布的美国专利 5097510 说明的技术分析输入信号幅度包络的变化。快速的变化被认为是语音并从信号中滤掉, 将剩余部分划分为四种噪声类型的一种, 该划分用于选择对输入信号消噪滤波的不同类型。

1995 年 10 月 10 日发布的美国专利 5457769 说明的技术检测语音以操作语音控制开关。通过辨别信号中的频率分量相互之间的间隔大约为 150Hz 来检测语音, 该条件指示信号可能传送语音共振峰。

1999 年 10 月 14 日出版的 EP 专利申请公报 0737011 和 1999 年 3 月 2 日发布的美国专利 5878391 说明的技术生成表示音频信号是语音信号的的概率的信号。该概率是通过提取一个或多个信号特征来导出的, 如不同频谱部分之间功率比值的变化。这些参考文献指出如果推导中

采用更多的特征可以提高导出概率的可靠性。

2000年5月9日发布的美国专利 6061647 说明了一种检测语音的技术，该技术通过存储没有语音的噪声模型，将输入信号与该模型相比较来确定是否存在语音，并使用辅助探测器来决定什么时候输入信号可以用于噪声模型的更新。

1998年6月25日出版的国际专利申请公报 WO 98/27543 说明了一种从音乐中辨别语音的技术。该技术通过从输入信号中提取一系列特征并对每个特征使用几种分类技术中的一个。最佳的特征序列及适用于每个特征的分类技术是通过经验确定的。

这些参考文献中说明的技术和所有其它已知的语音检测技术都试图检测语音或将音频信号分类，以便可以用一种不同于用来处理或控制非语音信号的方法处理或控制语音。

1998年10月6日发布的美国专利 5819247 说明的技术构造一个用于分类装置的假设，如光学字符识别设备。从实例中构造弱的假设然后进行估计，通过一个迭代的过程为弱的假设构造一个更强的假设。虽然没有提及语音检测，但发明人已经意识到该技术可以用于改进已知的语音检测技术。

发明内容

提供包含语音和其它类型音频素材的信号中的语音响度控制是本发明的一个目的。

根据本发明，信号的处理是通过接收输入信号并从输入信号中获得代表音频信号片段的音频信息，检查该音频信息以划分音频信息段是属于语音段还是非语音段，检查该音频信息以获得语音段的估计响度，并通过生成比对于非语音段代表的音频信号部分的响度更响应于语音段估计响度的控制信息来提供该音频信号片段的响度指示。

响度指示可以用于控制音频信号的响度以减少语音段响度的变化。当该语音段代表的音频信号部分的响度增加时，由非语音段代表的音频信号部分的响度随之增加。

通过参考下面的说明和附图，可以更好的理解本发明的各个特征及其最佳实施例，附图的几个图中相同的参考编号是指相同的单元。下面说明的内容仅仅作为实施例列出且不能理解为对本发明范畴的限制。

附图简述

图 1 是可包含本发明各个方面的音频系统的示意框图。

图 2 是可以用于控制包含语音和其它类型音频素材的音频信号响度的装置示意框图。

图 3 是可以用于生成和发送代表音频信号的音频信息和代表语音响度的控制信息的装置示意框图。

图 4 是可以用于提供包含语音和其它类型音频素材的音频信号中语音响度指示的装置示意框图。

图 5 是可以用于分类音频信息段的装置示意框图。

图 6 是可以用于实施本发明的各个方面的装置示意框图。

具体实施方式

A. 系统概述

图 1 是一个音频系统的示意框图，其中发射机 2 接收来自路径 1 的音频信号，处理该音频信号以生成代表音频信号的音频信息，然后沿路径 3 发送音频信息。路径 3 可以代表一个传送即时应用的音频信息的通信路径，或者可以代表一个与存储媒体相匹配的信号路径，该媒体存储音频信息以用于后续的检索和应用。接收机 4 接收来自路径 3 的音频信息，处理该音频信息以生成音频信号，并将该音频信号沿路径 5 发送给收听者。

图 1 所示的系统包含一个单独的发射机和接收机，但本发明可以应用于包含多个发射机和/或多个接收机的系统。本发明的各个方面可以仅仅在发射机 2，仅仅在接收机 4，或者在发射机 2 和接收机 4 实施。

在一个实施例中，发射机 2 执行的处理是将音频信号编码为比音

频信号信息容量需求更低的编码音频信息，使得音频信息可以在更低带宽的信道中传输或存储于更少空间的媒体中。解码器 4 执行的处理是将编码音频信息解码为可以用于生成感知上与输入音频信号类似或相同的更适宜的音频信号的形式。例如，发射机 2 和接收机 4 可以按照 AC-3 编码标准或运动图像专家组（MPEG）发布的几种标准中的任何一个编码和解码数字比特流。本发明可以应用于且有益于使用编码和解码处理的系统，但是这些处理对于本发明的实施并不是必需的。

虽然本发明可以用模拟信号处理技术实施，但用数字信号处理技术实施通常更方便。下面的实施例尤其参考数字信号处理。

B. 语音响度

本发明针对在包含语音和其它类型音频素材的信号中控制语音的响度。表 I 和 III 的条目表示在不同节目中不同类型音频素材的声音水平。

表 I 包括如同那些可能传送给电视接收者的三个节目中语音相对响度的信息。在新闻广播 1 中，两个人以不同的音量说话。在新闻广播 2 中，一个人以低的音量说话而其所在的位置其他声音的音量有时会超过语音。音乐有时会以低音量出现。在广告中，一个人以非常高的音量说话而音乐的响度有时候甚至更高。

表 I

新闻广播 1	新闻广播 2	广告
语音 1 -24 dB	其他声音 -33dB	音乐 -17dB
语音 2 -27dB	语音 -37dB	语音 -20dB
	音乐 -38dB	

本发明允许音频系统自动控制这三个节目中音频素材的响度以自动减少语音响度的变化。也可以控制新闻广播 1 中音频素材的响度以减少两个语音音量之间的差别。例如，如果所有语音的所需音量是 -24 dB，那么表 I 中所示音频素材的响度可以调整为如表 II 所示的水

平。

表 II

新闻广播 1	新闻广播 2 (+13dB)	广告 (-4dB)
语音 1 -24 dB	其他声音 -20dB	音乐 -21dB
语音 2(+3 dB) -24dB	语音 -24dB	语音 -24dB
	音乐 -25dB	

表 III 包括一个或多个运动画面的三个不同场景中不同声音的相对响度的信息。在场景 1 中，人们在一艘船的甲板上说话。背景声音包括音量大大低于语音音量的波浪拍击和远处汽笛声。该场景还包括一阵比语音响得多的该船的汽笛声。在场景 2 中，人们在低声说话而背景中有钟的嘀嗒声。该场景的语音比正常语音响度小而钟的嘀嗒声就更小了。在场景 3 中，人们在一个响声非常大的机器旁边高声说话。高声说话的音量高于正常语音。

表 III

场景 1	场景 2	场景 3
本船的汽笛声 -12 dB		机器声 -18dB
正常语音 -27dB	轻声细语 -37dB	高声说话 -20dB
远处汽笛 -33dB	钟的嘀嗒声 -43dB	
波浪 -40dB		

本发明允许音频系统自动控制这三个场景中音频素材的响度以自动减少语音响度的变化。例如，可以调整音频素材的响度以使所有场景的语音响度相同或基本相同。

另一方面，可以调整音频素材的响度以使语音的响度位于一个特定的范围内。例如，如果语音响度的特定范围是-24 dB 到-30 dB，那么表 III 中所示音频素材的响度可以调整为如表 IV 所示的水平。

表 IV

场景 1 (不变)	场景 2 (+7 dB)	场景 3 (-4 dB)
本船的汽笛声 -12 dB		机器声 -22dB
正常语音 -27dB	轻声细语 -30dB	高声说话 -24dB
远处汽笛 -33dB	钟的嘀嗒声 -36dB	
波浪 -40dB		

在另一个实施例中，控制音频信号的音量以使估计响度的某个平均值保持在所需的水平。该平均可以得自一个特定的时间间隔如 10 分钟，或得自所有节目或其特定的一部分。仍然参考表 III 所示的响度信息，假设这三个场景位于同一个运动画面，则整个运动画面的平均语音响度的估计值为-25 dB，而所需的语音响度是-27 dB。控制这三个场景的信号音量以使每个场景的估计响度得到如表 V 所示的修改。在该实施例中，保留了节目或运动画面中的语音响度的变化，却减少了与其他节目或运动画面中的平均响度的变化。换句话说，节目或部分节目之间语音响度的变化可以在不需要压缩那些节目或部分节目的动态范围的情况下获得。

表 V

场景 1 (-2 dB)	场景 2 (-2 dB)	场景 3 (-2 dB)
本船的汽笛声 -14dB		机器声 -20dB
正常语音 -29dB	轻声细语 -39dB	高声说话 -22dB
远处汽笛 -35dB	钟的嘀嗒声 -45dB	
波浪 -42dB		

动态范围的压缩也可能是需要的，但是该特征是可选的并可以在需要的时候提供。

C. 控制语音响度

可以通过在发射机或接收机内执行的独立过程，或者通过在发射机和接收机内联合执行的协作过程来实施本发明。

1. 独立过程

图2是可以用于在发射机或接收机内实施独立过程的装置示意框图。该装置接收来自路径11代表音频信号片段的音频信息。分类器12检查音频信息并将音频信息段分类，即将代表属于语音的音频信号部分归类为“语音段”，或者将代表不属于语音的音频信号部分归类为“非语音段”。分类器12也可以将非语音段划分为若干类型。前面提及了可以用于划分音频信息类型的技术。下面说明一种优选的技术。

由音频信息段表示的音频信号的每一部分都具有相应的响度。响度估计器检查语音段并得到该语音段响度的估计，该估计响度的指示沿路径15传递。在另一个实施例中，响度估计器14至少还检查部分非语音段并得到这些段的响度估计。前面提及了可以估计响度的一些方法。

控制器16接收来自路径15的响度指示和来自路径11的音频信息，然后对音频信息进行必要的修改以减少语音段所代表的部分音频信号响度的变化。如果控制器16增加了语音段的响度，那么它也会增加包括那些响度甚至高于语音段的所有非语音段的响度。修改的音频信息沿路径17传递以备后续的处理。例如，在发射机中，可以编码或者准备传送或存储修改的音频信息。在接收机中，可以处理修改的音频信息并将其提供给收听者。

分类器12、响度估计器14和控制器16按照这样的方式排列，即语音段的估计响度用于控制非语音段以及语音段的响度。这可以用多种方式进行，在一个实施例中，响度估计器14为每个语音段提供估计响度。控制器16利用该估计响度对被估计响度的语音段的响度进行任何需要的调整，该控制器还利用相同的估计对后续的非语音段的响度进行任何需要的调整，直到接收到下一个语音段的新估计。当不能预先检查音频信号而信号的电平必须实时调整时，该实施例是适合的。然而在另一个实施例中，估计的电平对应于音频信息语音和非语音段

的一个或多个特征进行调整,该信息可以由分类器 12 通过图中虚线所示的路径提供。

在优选的实施例中,控制器 16 也接收所有段的响度或信号能量指示并仅仅在响度或能量水平低于某个阈值的段内进行调整。另一方面,分类器 12 或响度估计器 14 可以提供给控制器 16 一个该段可以进行调整的段指示。

2. 协作过程

图 3 是可以用于在发射机中实施部分协作过程的装置示意框图。发射机接收来自路径 11 代表音频信号片段的音频信息。分类器 12 和响度估计器 14 的工作基本上与前面描述的相同。响度估计器 14 提供的估计响度的指示沿路径 15 传递。在如图所示的实施例中,编码器 18 沿路径 19 生成从路径 11 接收的音频信息的编码表示。编码器 18 基本上可以采用可能需要的任何类型的编码,包括所谓的感知编码。例如,图 3 所示的装置可以与音频编码器相结合以提供集成到 AC-3 适应编码数据流中的调节规格信息。编码器 18 对本发明并不是不可缺少的。在另一个实施例中忽略了编码器 18,音频信息本身沿路径 19 传递。格式化器 20 将从路径 19 接收的音频信息表示和从路径 15 接收的估计响度指示和成为输出信号,并沿路径 21 传递以用于发送或存储。

在没有图示出的互补接收机中,接收沿路径 21 生成的信号并进行处理以提取音频信息表示和估计响度指示。估计响度指示用于控制由音频信息表示生成的音频信号的信号电平。

3. 响度仪

图 4 是可以用于提供包含语音和其他类型音频素材的音频信号中语音的响度指示的装置示意图。该装置接收来自路径 11 代表音频信号片段的音频信息。分类器 12 和响度估计器 14 的工作基本上与前面描述的相同。响度估计器 14 提供的估计响度指示沿路径 15 传递。该指示可以以任何需要的形式显示,或者也可以提供给其他装置作后续的处理。

D. 段分类

本分发明基本上可以采用任何能够将音频信息段划分为包括语音类型在内的两类或更多类型的技术。前面提及了几种合适的分类技术的实例。在优选的实施例中，音频信息段的分类采用下面所描述技术的几种形式。

图5是可以用于按照优选分类技术划分音频信息段的装置示意框图。采样频率转换器30接收来自路径11的音频信息数字样本，并对音频信息进行必要的重采样以获得特定采样率的数字样本。在下面说明的实施例中，采样率是每秒16k个样本。采样率转换对本发明的实施不是必需的，但是当输入采样率高于划分语音信息所需的频率且较低的采样率允许更有效执行分类过程时，转换音频信息采样率通常是值得的。此外，如果将每个组件设计为仅仅工作在一个采样频率，那么提取特征组件的实施通常可以得到简化。

在所示的实施例中，通过提取组件31、32和33提取音频信息的三个特征。在另一个实施例中，可以提取少至一个特征多至有效处理资源可以处理的许多特征。语音检测器35接收提取的特征并利用它们确定是否应该将音频信息段归类为语音。下面说明特征提取和语音检测。

1. 特征

为了便于说明，如图5所示特定的实施例中仅仅示出了从音频信息中提取三个特征的组件。但是在优选的实施例中，段分类是基于下面描述的7个特征。每个特征提取组件通过对按帧排列的样本块进行计算来提取音频信息的一个特征。用于七个特征每一个的块大小以及每帧块的数目如表VI所示。

表 VI

特征	块大小 (样本数)	块长度 (毫秒)	每帧 的块数
加权频谱通量的平均平方 l_2 范数	1024	64	32
功率谱密度最佳拟合回归线的斜度	512	32	64
停顿数	256	16	128
过零率的倾斜系数	256	16	128
过零率的均值与中值比	256	16	128
短节奏度量	256	16	128
长节奏度量	256	16	128

在该实施例中，每帧是 32768 个样本或大约 2.057 秒的长度。下面说明该表中七个特征的每一个。在下面的说明中，块中的样本数目用符号 N 表示，每帧中块的数目用符号 M 表示。

a) 加权频谱通量的平均平方 l_2 范数

加权频谱通量的平均平方 l_2 范数利用了语音通常具有快速变化的频谱这个事实。语音信号通常有两种形式之一：类似乐音的浊音，或类似噪声的清音。这两种形式的过渡引起了频谱的突变。此外，在语音的浊音期间，大多数说话人改变基频以表示强调，语言的因袭，或者因为该变化是语言的本质组分。非语音信号如音乐可能也会有快速的频谱变化，但这些变化通常不经常发生。即使音乐的歌唱部分也没有频繁的变化，这是因为歌手通常在某个可以感知的时间内以相同的频率歌唱。

计算加权频谱通量平均平方 l_2 范数的过程中第一步是对音频信息样本块应用变换，如离散傅里叶变换(DFT)，得到变换系数的幅值。在应用变换之前最好用窗函数 $w[n]$ ，如汉明窗函数对样本块加权。DFT 系数的幅值可以用下面所示的公式计算。

$$|X_m[k]| = \left| \sum_{n=0}^{N-1} x[mN+n] \cdot w[n] \cdot e^{-j2\pi kn/N} \right| \quad 0 \leq k \leq N/2 \quad (1)$$

其中， N 表示块中的样本数；

$x[n]$ 表示第 m 块的第 n 个样本；以及

$X_m[k]$ 表示第 m 块中样本的第 k 个变换系数。

下一步由当前和前面块的平均功率计算当前块的权值 W 。如果样本 $x[n]$ 是实数而不是复数或虚数，那么利用 PARSEVAL 定理可以由变换系数按照下面所示的公式计算平均功率。

$$W_m = \frac{\sum_{k=0}^{N-1} (X_{m-1}[k]^2 + |X_m[k]|^2)}{N} \quad (2)$$

其中 W_m 表示当前块 m 的权值。

下一步计算当前和前面块的频谱分量差的平方，并将结果除以按照公式 2 计算的当前块的块权值 W_m 以产生加权频谱通量，然后计算 l_2 范数或欧几里得距离。加权频谱通量和 l_2 范数的计算公式如下所示。

$$\|l_m\| = \sqrt{\sum_{k=0}^{N-1} \frac{(X_{m-1}[k] - X_m[k])^2}{W_m}} \quad (3)$$

其中 $\|l_m\|$ 表示块 m 加权频谱通量的 l_2 范数。

通过计算该帧内每个块的 l_2 范数平方和得到一帧的特征。这个和如下面的公式所示。

$$F_1(t) = \sum_{m=0}^{M-1} (\|l_m\|)^2 \quad (4)$$

其中 M 表示一帧中块的数目；以及

$F_1(t)$ 表示第 t 帧加权频谱通量平均平方 l_2 范数的特征。

b) 功率谱密度最佳拟合回归线的斜度

对数功率谱密度最佳拟合回归线的梯度或斜度给出了信号谱倾斜或谱强调的一个估计。如果信号强调较低频率，则信号倾斜的谱形状的近似直线向下指向较高频率，该直线的斜率是负的。如果信号强调较高频率，则信号倾斜的谱形状的近似直线向上指向较高频率，该直线的斜率是正的。

语音在浊音期间强调低频而在清音期间强调高频。浊音频谱形状近似直线的斜率是负的而清音频谱形状近似直线的斜率是正的。因为与清音相比语音中浊音占大多数，所以语音频谱形状近似直线的斜率在大多数时间都是负的，仅在正和负斜率中间快速变化。结果，直线

斜率或梯度的分布必然强烈倾向于负值。对于音乐和其他类型的音频素材斜率的分布更加对称。

可以通过计算信号的对数功率谱密度最佳拟合回归线来获得信号频谱形状的近似直线。通过计算变换系数可以获得信号的功率谱密度,该系数采用的变换如前面公式 1 所示。功率谱密度的计算如下面公式所示。

$$|X_m[k]|^2 = \left| \sum_{n=0}^{N-1} x[mN+n] \cdot w[n] \cdot e^{\frac{-j2\pi kn}{N}} \right|^2 \quad 0 \leq k \leq N/2 \quad (5)$$

然后将公式 5 中计算的功率谱密度转换到对数域,如下式所示。

$$X_m^{dB}[k] = 10 \cdot \log_{10}(|X_m[k]|^2) \quad 0 \leq k \leq N/2 \quad (6)$$

然后计算最佳拟合回归线的梯度,如下式所示,这是用最小二乘法推导出来的。

$$G_m = \frac{\frac{N}{2} \sum_{k=0}^{\frac{N-1}{2}} k X_m^{dB} - \sum_{k=0}^{\frac{N-1}{2}} k \cdot \sum_{k=0}^{\frac{N-1}{2}} X_m^{dB}[K]}{\frac{N}{2} \sum_{k=0}^{\frac{N-1}{2}} k^2 - \left(\sum_{k=0}^{\frac{N-1}{2}} k \right)^2} \quad (7)$$

其中 G_m 表示块 m 的回归系数。

第 t 帧的特征是对整个帧斜度的估计,如下面给出的公式。

$$F_2(t) = \sum_{m=0}^{M-1} \left(G_m - \sum_{m=0}^{M-1} \frac{G_m}{M} \right)^3 \quad (8)$$

其中 $F_2(t)$ 表示第 t 帧对数功率谱密度最佳拟合回归线的梯度特征。

c) 停顿数

停顿数特征利用的事实是信号中具有很小或没有音频功率的停顿或短时间间隔通常出现在语音中而其他类型的音频素材通常没有这种停顿。

特征提取的第一步是计算一帧内各个块 m 音频信息的功率 $P[m]$ 。这可以用下面的公式来表示。

$$P[m] = \sum_{n=0}^{N-1} \frac{x[n]^2}{N} \quad (9)$$

其中 $P[m]$ 表示块 m 的计算功率。

第二步是计算该帧内的音频信息的功率 P_F 。第 t 帧的停顿数特征 $F_3(t)$ 等于在该帧内相应功率小于等于 $\frac{1}{4}P_F$ 的块的数目。该四分之一的值是由经验导出的。

d) 过零率的倾斜系数

过零率是音频信息所代表的音频信号在一段时间内穿过零值的次数。过零率可以通过统计一短时音频信息样本穿过零的次数来估计。在此说明的实施例中，样本块的持续时间是 16 毫秒 256 个样本。

虽然概念很简单，但是过零率导出的信息可以提供当前音频信号是否是语音的一个相当有效的指示。语音的浊音部分具有相对低的过零率，而语音的清音部分具有相对高的过零率。而且因为与清音部分相比，语音通常包含更多的浊音部分和停顿，过零率的分布一般向更低的比值倾斜。一个可以提供帧 t 内倾斜指示的特征是过零率的倾斜系数，可以用下面的公式计算

$$F_4(t) = \frac{\sum_{m=0}^{M-1} \left(Z_m - \sum_{m=0}^{M-1} \frac{Z_m}{M} \right)^3}{\left(\sum_{m=0}^{M-1} \left(Z_m - \sum_{m=0}^{M-1} \frac{Z_m}{M} \right)^2 \right)^{3/2}} \quad (10)$$

其中 Z_m 表示块 m 的过零数；以及

$F_4(t)$ 表示帧 t 的过零率的倾斜系数特征。

e) 过零率的均值与中值比

可以提供帧 t 内过零率倾斜分布指示的另一个特征是过零率的均值与中值比，这可以用下面的公式得出

$$F_5(t) = \frac{Z_{median}}{\sum_{m=0}^{M-1} \frac{Z_m}{M}} \quad (11)$$

其中 Z_{median} 表示帧 t 内所有块的块过零数的中值；以及

$F_5(t)$ 表示帧 t 的过零率的均值与中值比特征。

f) 短节奏度量

使用前面所描述特征的技术可以检测许多类型音频素材中的语音，但是在高节奏音频素材中，如所谓的“rap”和许多 pop 音乐的场合这些技术就会产生检测错误。通过检测高节奏素材并将这些素材从

分类中除去或者提高将该素材归类为语音所需要的置信级可以更可靠地将音频信息段归类为语音。

按照如下所示的公式先计算每个块中样本的方差可以计算一帧的短节奏度量。

$$\sigma_x^2[m] = \sum_{n=0}^{N-1} \frac{(x[n] - \bar{x}_m)^2}{N} \quad (12)$$

其中 $\sigma_x^2[m]$ 表示块 m 内样本 x 的方差；以及

$\bar{x}_m$ 表示块 m 内样本 x 的均值。

帧内所有块的方差导出一个零均值序列，如下式所示。

$$\delta[m] = \sigma_x^2[m] - \bar{\sigma}_x^2 \quad 0 \leq m < M \quad (13)$$

其中 $\delta[m]$ 表示零均值序列中对应于块 m 的元素；以及

$\bar{\sigma}_x^2$ 表示对于帧内所有块的方差的均值。

按照如下所示的公式可以得到零均值序列的自相关。

$$A_t[l] = \frac{1}{M} \sum_{m=0}^{M-1-l} \delta[m] \cdot \delta[m+l] \quad 0 \leq l < M \quad (14)$$

其中 $A_t[l]$ 表示第 t 帧与块滞后 l 的自相关的值。

短节奏度量特征由这些自相关值的最大值导出。该最大值不包括针对块滞后 $l = 0$ 的值，所以最大值是得自块滞后 $l \geq L$ 的一组值。 L 值代表所期望的最快节奏的周期。在一个实施例中 L 的值为 10，表示最小周期是 160 毫秒。如下式所示，通过用最大值除以块滞后 $l = 0$ 的自相关值来计算该特征。

$$F_6(t) = \frac{\max_{L \leq n < M} (A_t[n])}{A_t[0]} \quad (15)$$

其中 $F_6(t)$ 表示第 t 帧的短节奏度量特征。

g) 长节奏度量

除了用频谱权值代替零均值序列，长节奏度量按照与前面短节奏度量类似的方法导出。通过先得到对数功率谱密度来计算这些频谱权值，如前面公式 5 和 6 所示，其描述与对数功率谱密度最佳拟合回归线的梯度的斜度有关。有必要指出的是，在此描述的实施例中，计算长节奏度量的块长度并不等于用于梯度斜度计算的块长度。

下一步按照下面的公式所示得到对数域功率谱值的最大值。

$$O_m = \max_{0 \leq k < \frac{N}{2}} (X_m^{dB}[k]) \quad (16)$$

其中 O_m 表示块 m 中最大对数功率谱值。

由超过阈值 ($O_m \cdot \alpha$) 的对数域功率谱峰值的数目来确定每个块的谱权值。这可以用下面的公式来表示。

$$W[m] = \sum_{k=0}^{\frac{N}{2}-1} \frac{\text{sign}(X_m^{dB}[k] - O_m \cdot \alpha) + 1}{2} \quad (17)$$

其中 $W[m]$ 表示块 m 的谱权值；

如果 $n \geq 0$ ，则 $\text{sign}(n) = +1$ ；如果 $n < 0$ 则 $\text{sign}(n) = -1$ ；以及 α 表示经验导出的常数，等于 0.1。

在每一帧的末尾，将前面帧的 M 个谱权值的序列和当前帧的 M 个谱权值的序列串联起来构造一个 $2M$ 个谱权值的序列。然后按照下面的公式计算该长序列的自相关。

$$AL_t[l] = \frac{1}{2M} \sum_{m=-M+1}^{M-1-l} W[m] \cdot W[m+l] \quad 0 \leq l < 2M \quad (18)$$

其中 $AL_t[l]$ 表示第 t 帧的自相关值。

长节奏度量特征由这些自相关值的最大值导出。该最大值不包括针对块滞后 $l = 0$ 的值，所以最大值是得自块滞后 $l \geq LL$ 的一组值。 LL 值代表所期望的最快节奏的周期。在此说明的一个实施例中 LL 的值为 10。如下式所示，通过用最大值除以块滞后 $l = 0$ 的自相关值来计算该特征。

$$F_A(t) = \frac{\max_{LL \leq n < M} (AL_t[n])}{AL_t[0]} \quad (19)$$

其中 $F_A(t)$ 表示第 t 帧的长节奏度量特征。

2. 语音检测

语音检测器 35 合并针对各个帧提取的特征以确定是否应该将音频信息段归类为语音。可以采用的一种合并特征的方法是实现一组简单的或中间分类器。一个中间分类器通过将前面讨论的特征之一与一个阈值相比较来计算一个二元值。然后将该二元值用系数加权。每个中间分类器基于一个特征进行中间分类。特定特征可以由超过一个的中间分类器使用。中间分类器的实现可以按照下面的公式计算。

$$C_j = c_j \cdot \text{sign}(F_i - Th_j)$$

其中 C_j 表示中间分类器 j 提供的二元值分类;

c_j 表示中间分类器 j 的系数;

F_i 表示从音频信息提取的特征 i ; 以及

Th_j 表示中间分类器 j 的阈值。

在该特定实施例中, 中间分类值 $C_j = 1$ 说明中间分类器 j 倾向于支持的结论是音频信息的特定帧应该归类为语音。中间分类值 $C_j = -1$ 说明中间分类器 j 倾向于支持的结论是音频信息的特定帧不应该归类为语音。

表 VII 的条目示出的系数和阈值以及几个中间分类器的适当特征可以用于一个对音频信息进行分类的实施例。

表 VII

中间分类器序号 j	系数 c_j	阈值 Th_j	特征序号 i
1	1.175688	5.721547	1
2	-.672672	0.833154	5
3	0.631083	5.826363	1
4	-0.29152	0.232458	6
5	0.5.2359	1.474436	4
6	-0.310641	0.269663	7
7	0.266078	5.806366	1
8	-0.101095	0.218851	6
9	0.097274	1.474855	4
10	0.058117	5.810558	1
11	-0.042538	0.264982	7
12	0.034076	5.811342	1
13	-0.044324	0.850407	5
14	-0.066890	5.902452	3
15	-0.029350	0.263540	7
16	0.035183	5.812901	1
17	0.033141	1.497580	4
18	-0.015365	0.849056	5
19	0.016036	5.813189	1
20	-0.016559	0.263945	7

最终的分类基于中间分类的组合。可以按照下面的公式进行。

$$C_{final} = \text{sign} \left(\sum_{j=1}^J C_j \right) \quad (21)$$

其中 C_{final} 表示音频信息帧的最终分类；以及
 J 表示用于分类的中间分类器的数目。

通过优化中间分类器的选择，优化中间分类器的系数和阈值可以提高语音检测器的可靠性。这种优化可以用多种方式实施，包括前面引用的美国 5819247 号专利以及 Schapire 于 1999 年发表在 Proc. of the 16th Int. Joint Conf. on Artificial Intelligence 的文章“A Brief Introduceion to Boosting”中的技术。

在另一个实施例中，语音检测不是通过二元判决来指示的，而是用分级度量的分类来表示。该度量可以表示语音分类中语音的估计概率或者置信级。这可以以多种方式进行，例如从中间分类的和得到最终的分类结果而不是得到一个如公式 21 所示的二值结果。

3. 样本块

前面描述的实施例是对相连的、不重叠的固定长度块提取特征。作为选择，分类技术可以应用于相连的不重叠的可变长度块，应用于固定或可变长度的重叠块，或者应用于固定或可变长度的非相连块。例如，块长度可以针对瞬变，停顿或极小或无能量段而进行适应以使得每个块的音频信息更加稳定。帧长度也可以通过变换每一帧块的数目和/或变换一帧中块的长度来适应。

E. 响度估计

响度估计器 14 检查音频信息段以获得该语音段的响度估计。在一个实施例中，对归类为语音段的每一帧估计响度。基本上可以对所需任意时长估计响度。

在另一个实施例中，估计过程开始于启动处理的一个请求并且持续至接收到一个停止处理的请求。例如在接收机 4 中，这些请求可以通过路径 3 接收信号中的特殊编码来传递。另一方面，这些请求可以通过用于估计响度的装置上的开关或控制来提供。可以提供附加的控

制来促使响度估计器 14 暂停处理并保持当前的估计。

在一个实施例中，对归类为语音段的所有音频信息段估计响度。但是原则上可以仅对选择的语音段估计响度，例如仅仅针对音频能量超过阈值的那些段。通过一个分类器 12 把低能量段归类为非语音段然后对所有的语音段估计响度也可以获得类似的效果。可能有其它的变化，例如在估计响度计算中较过时的段可以赋予较低的权值。

在另一个实施例中，响度估计器 14 至少估计一些非语音段的响度。非语音段的估计响度可以用于音频信息段的响度计算；这些计算应该更加对应于语音段的估计。非语音段的估计也可以用于提供段类型分级度量的实施例中。音频信息段的响度计算可以用考虑分类分级度量的方式反应语音和非语音段的估计响度。例如，分级度量可以代表音频信息段包含语音的置信级指示。通过在响度估计计算中赋予具有较高置信级的段以较高的权值可以使响度的估计更加针对这些段。

可以用包括前面讨论的那些方法在内的多种方法估计响度。对于本发明特定的估计技术是不重要的；但是，在实际的实施例中通常优先选择需要更少计算资源的更简单的技术。

F. 实施

本发明的各个方面可以用非常多的方法实施，包括通用计算机系统中或其它一些装置中的软件，这些装置包括更专门化的元件如与元件相连的数字信号处理器（DSP）电路，其元件类似于通用计算机中的那些元件。图 6 是可以用于在音频编码发射机或音频解码接收机上实施本发明各个方面的装置 70 的框图。DSP 72 提供计算资源。RAM 73 是系统的随机存储器（RAM），被 DSP 72 用于信号处理。ROM 74 代表永久存储的某种形式以存储需要操作装置 70 的程序，如只读存储器（ROM）。I/O 控制器 75 代表通过信道 76、77 接收和发送信号的接口电路。当需要接收和/或发送模拟音频信号时，I/O 控制器 75 可以包含模-数转换器和数-模转换器。在所示装置中，所有主要的系统组件都与总线 71 相连，该总线代表的物理总线可不止一种；但总线结构对于本发明的实施不是必需的。

在通用计算机系统的实施例中，可以包含附加的元件，用于连接键盘或鼠标以及显示器等装置，以及用于控制具有如磁带或磁盘或光学介质等存储媒体的存储装置。存储介质可以用于记录操作系统、实用程序及应用的指令程序，可以包含实施本发明各个方面的程序本身。

实践本发明各个方面所需的功能可以用通过很多种方法实现的专用元件来执行，包括离散逻辑元件，一个或多个 ASIC 和/或程控处理器。这些元件的实现方式对于本发明都是不重要的。

本发明的软件实施可以通过多种机器可读媒体传输，如覆盖从超声到紫外频段频率范围的基带或调制通信线路，或通过包含使用基本上任何磁性或光学记录技术传送信息的存储介质传送，包括磁带、磁盘和光盘。也可以在计算机系统 70 的各个组件中采用处理电路实施各个方面，如 ASIC，通用集成电路，通过包含在各种 ROM 或 RAM 中的程序控制的微处理器以及其它技术。

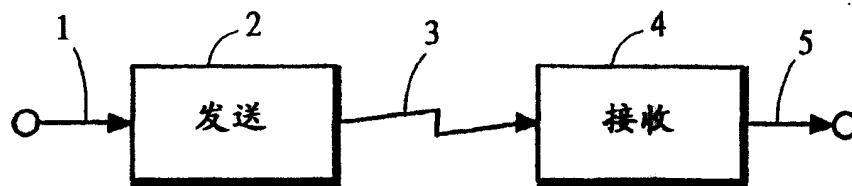


图1

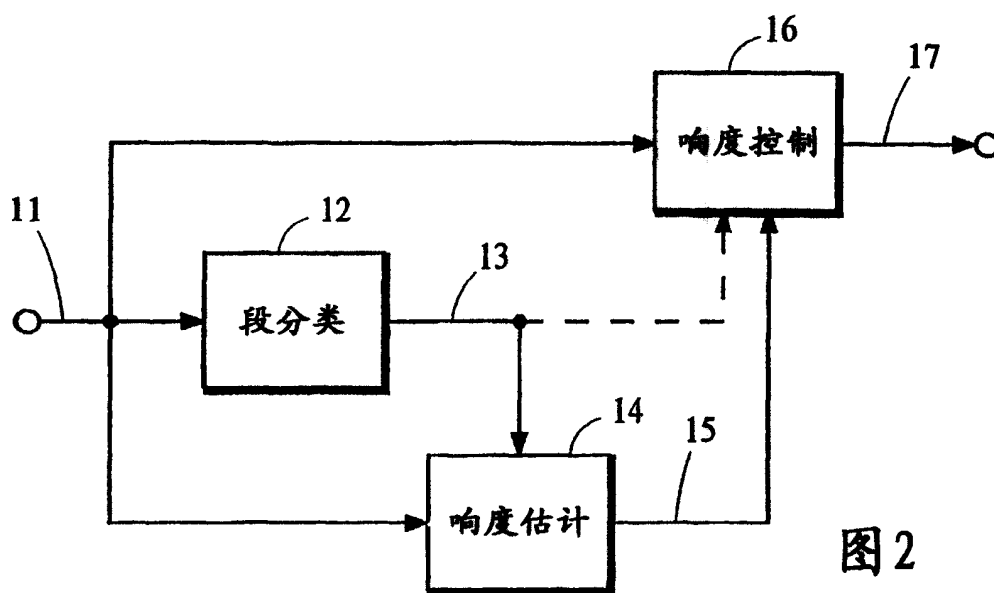


图2

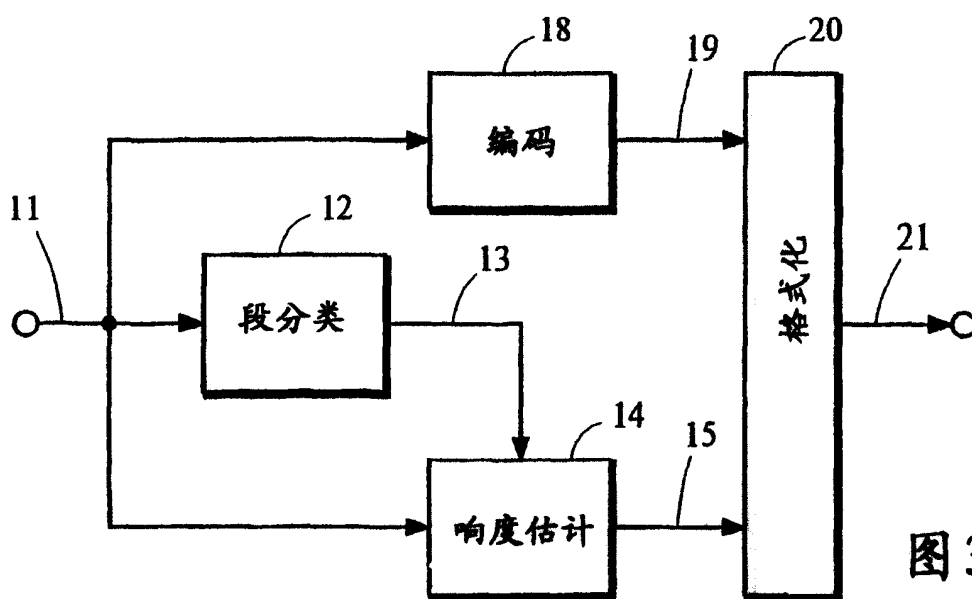


图3

