[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 21/02 (2006.01)
G10L 15/20 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200510067777.0

[43] 公开日 2006 年 2 月 22 日

[11] 公开号 CN 1737906A

[22] 申请日 2005.3.22
[21] 申请号 200510067777.0
[30] 优先权
[32] 2004. 3.23 [33] US [31] 60/555,582
[71] 申请人 哈曼贝克自动系统 - 威美科公司
地址 加拿大英属哥伦比亚
[72] 发明人 P·赫瑟林顿 P·扎卡拉乌斯卡斯
S·帕尔文

[74] 专利代理机构 北京纪凯知识产权代理有限公司
代理人 沙 捷 刘 颖

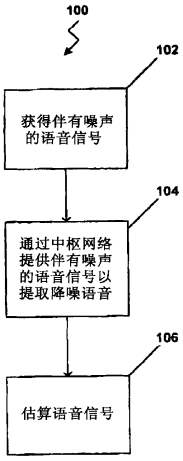
权利要求书 4 页 说明书 12 页 附图 14 页

[54] 发明名称

利用中枢网络分离语音信号

[57] 摘要

一种语音信号分离系统，它能够分离和重建在语音信号的频率成分被背景噪声掩盖的环境中传输的语音信号。 该语音分离系统从一个音频源获得一个伴有噪声的语音信号。 然后噪声语音信号可通过一个已训练为能够从背景噪声中隔离和重建纯净语音信号的中枢网络进行馈送。 一旦噪声语音信号通过中枢网络进行馈送，该语音信号分离系统就产生一个充分降噪的估值语音信号。



1、一种从音频信号中背景噪声内提取语音信号的语音信号分离系统，包括：

用于估算多个频率上的音频信号的背景噪声强度的背景噪声估值部分；

用于从背景噪声中提取语音估值信号的中枢网络部分，以及

用于从音频信号和基于背景噪声估值提取的语音中产生重建的语音信号的混合部分。

2、如权利要求1所述的系统，还包括用于将所述音频信号从时间序列信号转换到频域信号的频率转换部分。

3、如权利要求2所述的系统，还包括用于产生已减少频率子带数目的压缩音频信号的压缩部分。

4、如权利要求3所述的系统，其中：中枢网络具有第一组输入节点，其数目等于压缩音频信号中的频率子带数目，用来接收所述压缩音频信号。

5、如权利要求4所述的系统，其中：中枢网络包括第二组输入节点，其数目等于频率子带数目，用于接收所述背景噪声估值。

6、如权利要求4所述的系统，其中：中枢网络包括第二组输入节点，其数目等于压缩音频信号中的频率子带数目，用于接收先前时间步骤中的压缩音频信号。

7、如权利要求4所述的系统，其中：中枢网络包括第二组输入节点，其数目等于压缩音频信号中的频率子带数目，用于接收先前时间步骤中的中枢网络的输出。

8、如权利要求 4 所述的系统，其中：中枢网络包括第二组输入节点，用于接收先前时间步骤的中间结果。

9、如权利要求 1 所述的系统，其中：混合部分用于将音频信号中强度高于背景噪声估值的部分和对应音频信号中强度低于背景噪声估值的部分所提取的语音部分相混合。

10、一种从包含语音部分和背景噪声的音频信号中分离语音信号的方法，所述方法包括：

将时序音频信号变换为频域；

估算多个频带上音频信号中的背景噪声；

从音频信号中提取语音信号估值；

将部分语音信号估值和基于背景噪声估值的部分音频信号混合，以提供降低背景噪声的重建语音信号。

11、如权利要求 10 所述的方法，其中：从音频信号中提取语音信号估值包括将音频信号作为输入分配给中枢网络。

12、如权利要求 10 所述的方法，其中：混合语音信号估值和音频信号包括建立一个高于背景噪声估值的强度上限，并将音频信号中强度高于该强度上限的部分与语音信号估值混合。

13、如权利要求 10 所述的方法，其中语音信号估值和音频信号的混合包括建立一个处于或接近背景噪声估值的强度下限，并混合对应于音频信号中强度低于该强度下限部分的语音信号估值部分。

14、如权利要求 10 所述的方法，其中混合语音信号估值和音频信号包括建立上下限强度值，并将部分音频信号和对应于音频信号中强度值位于上下限强度值之间部分的语音信号估值相混合。

15、如权利要求 14 所述的方法，其中混合部分音频信号和部分语

音信号估值包括加权音频信号和语音信号估值，使得在部分音频信号有较接近强度下限的强度值时，语音信号估值比音频信号有更大的权重；在部分音频信号有较接近强度上限的强度值时，音频信号比语音信号估值有更大的权重。

16、如权利要求 11 所述的方法，还包括在中枢网络中应用背景噪声估值。

17、如权利要求 11 中所述的方法，还包括在中枢网络中应用先前时间步骤中的语音信号估值。

18、如权利要求 11 中所述的方法，还包括在中枢网络中应用先前时间步骤中的语音信号估值的中间结果。

19、如权利要求 11 中所述的方法，还包括在中枢网络中应用先前时间步骤中的音频信号。

20、一种语音信号增强系统，包括：

提供时间序列音频信号的音频信号源，所述音频信号包括语音部分和背景噪声；

提供频率变换功能的信号处理器，其用于将音频信号从时域变换到频域；

背景噪声估值器；

中枢网络；以及

信号混合器，

所述背景噪声估值器产生所述音频信号中的背景噪声估值，所述中枢网络从所述音频信号中提取语音信号估值，所述信号混合器混合语音信号估值和基于背景噪声估值的音频信号来产生充分降低背景噪声后的重建语音信号。

21、如权利要求 20 所述的系统，其中：中枢网络包括第一组输入

节点，用于接收音频信号。

22、如权利要求21所述的系统，其中：中枢网络包括第二组输入节点，用于接收来自先前时间步骤的音频信号。

23、如权利要求21所述的系统，其中：中枢网络包括第二组输入节点，用于接收背景噪声估值。

24、如权利要求21所述的系统，其中：中枢网络包括第二组输入节点，用于接收来自先前时间步骤的语音信号估值。

25、如权利要求21所述的系统，其中：中枢网络包括第二组输入节点，用于接收来自先前时间步骤的中间结果。

26、一种从背景噪声中分离语音信号的方法，包括：

接收音频信号；

识别音频信号中信号的准确性已知为具有较高可信度的部分；以及

训练中枢网络，来估计一个重建信号，该重建信号已经显著降低了音频信号中音频信号的准确性被怀疑的部分的背景噪声。

利用中枢网络分离语音信号

相关申请

本申请要求美国临时专利申请 No. 60 / 555，582 的优先权，其申请日为 2004 年 3 月 23 日.

技术领域

本发明涉及语音处理系统，特别是涉及在噪声环境下对语音信号的检测和分离。

背景技术

声音是通过任何弹性材料，固体，液体，或气体传播的振动。一种普通声音的类型就是人的语音。当信号在一个噪声环境下传递语音信号时，它经常被背景噪声所掩盖。声音可以通过频率来表征。频率是指产生在单元时间内一个周期性过程的完整循环数量。信号可以根据代表时间的 x-轴和代表振幅的 y-轴来绘制。典型的信号可从其原点产生到一个正的峰值，然后回落到负的峰值。然后信号回到它初始值，从而完成第一个周期。正弦信号的周期是信号重复的周期间隔。

频率通常以赫兹（Hz）为单位进行计量。人耳通常能察觉到的声音频率范围是 20-20,000Hz。声音可能包含多种频率。多频声音的振幅是组成频率在每个时间取样点振幅的总和。由于谐波关系两个或两个以上的频率可能相关。如果第一个频率是第二个频率的整数倍，那么第一个频率就是第二个频率的谐波。

多频声音根据包含它们的频率模式而进行表征。通常，噪音以某一角度在频率特性曲线上衰减。这种频率模式被称为“粉红噪声”。粉红噪声包含高强度低频信号。随着频率增高，声音的强度就减小。“褐色噪声”与“粉红噪声”相似，但是表现为更快的衰减。褐色噪声可以是汽车的声音，例如从车身面板中发出的低频隆隆声，在所有频率下表现为相等能量的声音被称为“白噪声”。

声音也可以它的强度来表征，通常以分贝（dB）为计量单位。分贝是声音强度的对数单位，或者是声音强度和其它参考强度的比值对数的 10 倍。对人耳而言，分贝的范围是从平均可察觉声音 0(dB)到平均疼痛程度 130（dB）。

人的声音由喉门产生。喉门是位于咽喉上部的声带间的开口。人的声音是由呼气时气体经过声带振动时产生。声音可表征为喉门的振动频率。大多数声音范围是 70-400Hz。男人一般说话的声音频率范围是 80-150Hz。女人一般说话的范围是 125-400Hz。

人的语音包括辅音和元音。辅音，例如“TH”和“F”特点是白噪声。这种声音的频谱类似与一把台扇。辅音“S”可表征为宽带噪声，通常由 3000Hz 左右开始扩展到大约 10,000Hz。元音，“T”，“B”，和“P”，被称为爆破音，也可表征为宽带段噪声，但是不同于“S”的突然上升。元音还产生独特的频谱。元音的频谱特点是共振峰频率。共振峰包括元音所特有的几个共鸣带中的任何一个。

语音检测和记录的主要问题是从背景噪声中分离出语音信号。背景噪声能干扰并且降低语音信号。在噪声的环境下，语音信号中的频率成分可能被背景噪声的频率部分，或者甚至是全部地掩盖。这样，就存在一种需要在目前的背景噪声下可以分离和重建语音信号的语音信号分离系统。

发明内容

本发明公开了一种语音信号分离系统，它能够分离和重建在语音信号的频率成分被背景噪声掩盖的环境中传输的语音信号。本发明的一个实施例中，通过中枢网络分析伴有噪声的语音信号，能够从伴有噪声的语音信号中生成纯净的语音信号。中枢网络被训练成能够从背景噪声中分离出语音信号。

本发明其它的系统，方法，特征和优点对于本领域的技术人员来说将在以下的附图和详细描述中变得显而易见。以下描述中所包括的所有相似的附加系统、方法、特征和优点，都在本发明范围之内，并得到后面权利要求的保护。

附图说明

参考以下的附图和说明能够更好的理解本发明。图中各部件并不需要按比例绘制，而是重点要求能说明本发明的原理。此外，附图中，相同的参考标号在不同的附图中表示相对应的部件。

图 1 是语音信号分离系统的方框图。

图 2 是典型元音的频谱图。

图 3 是典型元音部分被噪声掩盖的频谱图。

图 4 是中枢网络图。

图 5 是语音信号分离系统的语音信号处理方法的方框图。

图 6 是典型元音部分被噪声掩盖和其平滑包络的示意图。

图 7 是压缩后的语音信号图。

图 8 是语音信号分离系统使用的中枢网络结构示意图。

图 9 是根据本发明的另一中枢网络结构示意图。

图 10 是另一中枢网络结构示意图。

图 11 是另一包含反馈的中枢网络结构示意图。

图 12 是另一包含反馈的中枢网络结构示意图。

图 13 是另一包含反馈以及附加隐藏层的中枢网络结构示意图。

图 14 是语音信号分离系统的方框图。

具体实施方式

本发明涉及一种从背景噪声中分离信号的系统和方法。这种系统和方法特别适于从噪声环境里产生的音频信号中恢复语音信号。然而，本发明决不限于声音信号，可以用于任何被噪声掩盖的信号。

图 1 描述了一种将语音信号从背景噪声中分离出来的方法 100。该方法 100 能够分离和重建在语音信号的频率成分被背景噪声掩盖的环境中传输的语音信号。在以下的描述中，将会举出许多具体细节来对语音信号分离方法 100 及其相关的实现该方法的相对应系统 10 进行更为完整的说明。但是，对于本领域的技术人员来说，显而易见，实施本发明并不局限于这些具体细节。在其它情况下，没有详细说明众所周知的特征是为了避免混淆发明。从背景噪声中分离语音信号的方法 10 包括获得或接收伴有噪声的语音信号的步骤 102。第二个步骤 104

是通过中枢网络提供语音信号，该中枢网络用于从噪声输入信号中提取出降噪的语音。最后的步骤 106 用来估算语音。

图 14 是语音信号分离系统 10。该语音信号分离系统包括音频信号装置例如麦克风 12，或者其它任何配置成能够提供音频信号的音频源。设置一 A/D 转换器 14 以能将麦克风 12 的模拟语音信号转换为数字语音信号，并将数字语音信号作为输入提供给信号处理单元 16。如果音频信号装置提供的是数字音频信号，A/D 转换器可以省略。数字处理单元 16 可以是数字信号处理器，计算机，或者其它类型能够处理音频信号的电路或系统。信号处理单元包括一个中枢网络部件 18，一个背景噪声估算部件 20，和一个信号混合部件 22。噪声估算部件在一系列频率子带上所接收的信号中估算噪声级别。中枢网络部件 18 用来接收音频信号和从音频信号的背景噪声部分中分离出音频信号中的语音部分。信号混合部件 22 重建完整的降噪声语音信号作为所分离语音部分和音频信号的函数。因此，语音信号分离系统 10 能够从背景噪声分离语音信号，显著地减小或消除背景噪声，然后通过估算如果在原始信号中没有背景噪声时真实语音信号看起来和听起来像什么来重建完整的语音信号。

图 2 是典型元音声的频谱图，并且作为表示语音信号如何进行表征的例子。元音声是非常有趣的，因为它们通常是语音信号强度最高的部分，并且在干涉语音信号的噪声上方的上升具有最高的相似性。虽然图 2 中表示的是元音声，但是该语音信号分离系统 10 和方法 100 能处理作为输入而接收到的任何类型的语音信号。

元音或语音信号 200 表征为它的构成频率及每个频带的强度。语音信号 200 通过频率（Hz）轴 202 和强度（dB）轴 204 绘制。频率图通常包括任意数量离散的段（bin）或带。频率组 206 表示从语音信号 200 获得 256 个频率带（256 个频段）。信号频带数量的选择对于本领域的技术人员来说是一种公知的方法，并且这里使用的 256 频带长度仅仅用来说明，因为也可以使用其它频带长度。接近水平的线 208 代表获得语音信号 200 的环境中背景噪声的强度。通常来说，语音信号 200 必须对比背景环境噪声来检测。强度范围高于噪声 208 的语音信号 200 很容易被检测出来。但是，必须在低于噪声水平的强度水平处从背

景噪声中提取语音信号 200。此外，在强度水平等于或接近噪声水平 208 时，从噪声 208 中分辨语音将变得十分困难。

再参考图 1 和 14，在步骤 102 中，语音信号分离系统 100 能从外部设备中例如麦克风等获得语音信号。通常情况下，语音信号 200 可能包含背景噪声，例如音乐会时人群的噪声或者汽车噪声或者来自其它噪声源的噪声。如图 2 中线 208 所示，背景噪声掩盖了部分语音信号 200。语音信号 200 的峰值在一个或多个位置处高出线 208，但是语音信号 200 低于分辨线 208 的部分由于背景噪声的存在很难或者不可能被分辨出来。在方框 104 中，语音信号 200 通过中枢网络由语音信号分离系统 10 进行馈送，中枢网络被训练为能够在噪声环境下分离和重建语音信号。在步骤 106 中，语音信号 200 通过中枢网络从背景噪声中分离，以用来产生明显减小或消除背景噪声后的估值语音信号。

语音检测的主要问题是从背景噪声中分离语音信号 200。在噪声环境下，语音信号 200 许多频率部分或完全被噪声频率掩盖。图 3 清楚地说明了这种现象。噪声 302 干扰语音信号 300 以致于语音信号 300 的部分 304 被噪声 302 掩盖，并且只有高出噪声 302 的部分 306 容易被检测。因为区域 306 仅仅包含语音信号 300 的一部分，所以噪声的存在使某些语音信号 300 被丢失或被掩盖。

正如这里所谈及，中枢网络是一种不严格模拟人脑神经互联系统的计算机结构。中枢网络模仿大脑辨别模式的能力。在使用中，中枢网络提取出输入到网络中的数据的关联性。中枢网络被训练能够识别这些关联性，就像教会孩童或动物一项任务一样。中枢网络通过反复试验方法学进行学习。随着课程的重复，中枢网络性能就会得到提高。

图 4 表示语音信号分离系统 10 使用的典型中枢网络 400。中枢网络 400 包含 3 个计算层。输入层 402 包括输入神经元 404。隐藏层 406 包括隐藏神经元 408。输出层 410 包括输出神经元 412。如图所示，每层 402，406，和 410 中的每个神经元 404，408 和 412 与随后的 402，406 和 410 层中的每个神经元 404，408 和 412 完全互联。因此，每个输入神经元 404 通过连接 414 和每个隐藏神经元 408 相连。此外，每个隐藏神经元 408 通过连接 416 和每个输出神经元 412 相连。每个连接 414 和 416 通过加权系数相关。

每个神经元可能在一定值范围内触发。这个范围例如可能是从 0 到 1。输入神经元 404 的输入可能根据应用确定，或根据网络环境设定。隐藏神经元 408 的输入可能是输入神经元 404 状态乘以连接 414 的加权系数或用该系数进行调整。输出神经元 412 的输入可能是输入神经元 408 状态乘以连接 416 的加权系数或用该系数进行调整。隐藏或输出神经元 412 各自的启动是对所述节点输入总和使用 Squashing 或 S 函数的结果。Squashing 函数可能是非线性函数，它将输入总和限制在一定范围的值中。范围还是从 0 到 1。

当范例（具有已知结果）演示给中枢网络时，它就进行“学习”。重复调整加权系数以使输出逼近正确的结果。训练之后，在实践中，每个输入神经元 404 状态根据应用分配或根据网络环境设定，输入神经元 404 的输入通过加权连接 414 传到每个隐藏神经元 408。然后每个隐藏神经元 408 的合成结果的状态传到每个输出神经元 412。每个输出神经元 412 合成结果的状态是网络呈现给输入层 402 的模式的答案。

图 5 进一步表示语音信号分离系统 10 执行语音信号处理过程的方框图。在步骤 500 中，从外部语音信号装置，例如麦克风中获取语音信号。语音信号被用大约 46 微秒（ms）的时间序列取样，但也可使用其他时间序列。本领域的技术人员应该认识到可从几种不同类型的声源获得语音信号。例如，语音信号可从希望去除背景噪声而净化的录音中获得，或者从嘈杂的汽车里面的一个或几个麦克风中获得。

在步骤 502 中，执行了一个从时域到频域的变换。这种变换可以是快速傅立叶变换（FFT），也可以是 DFT、DCT、滤波器组，或者其他能在频率上估算语音信号功率的方法。FFT 是一种表示作为正弦余弦加权和的波形的技术。FFT 是计算一系列离散的数据值的傅立叶变换的算法。给定一系列有限的数据点，例如是从声音信号中提取的周期性采样，FFT 按照这些数据的组成频率来表示上述数据。正如下文所述，它也解决了从频率数据重建时域信号所必须的相同的反转问题。

进一步来讲，在步骤 504 中，包含在语音信号中的背景噪声被估算。背景噪声可以通过任何已知方法进行估算。例如从一段安静，或没有检测到语音的周期计算出平均值。平均值可根据在各个频率处的信号与噪声估算的比值进行连续调整，该平均值在信噪比值较低的频

率被较快地更新。或者可用中枢网络自身来估算噪声。

在步骤 502 产生的语音信号和在 504 产生的噪声估值在步骤 506 进行压缩。在一个实例中，使用“Mel 频率标度”算法压缩语音信号。语音往往在低频比高频具有更复杂的结构，因此非线性压缩往往在压缩频段上平均分布频率信息。

语音中的信息以对数方式衰减。在较高频中，只能发现“S”或”T”的音，因此几乎没有信息需要保持。Mel 频率标度优化压缩来保存声音信息：在低频区为线性；在高频区为对数方式。Mel 频率标度通过下列公式对应于实际的频率（f）：

$$mel(f)=2595\log(1+f/700)$$

其中 f 的单位是赫兹（Hz）。信号压缩的结果值被存储在“Mel 频率组”中。Mel 频率组是将中心频率设置为使 Mel 值均匀分开而生成的滤波器组。压缩的结果是一个突出语音信号信息内容的平滑信号，以及一个压缩的噪声信号。

Mel 标度代表音调的心理声学比率标度。也可以使用其它压缩标度，例如以 2 为底的频率对数标度，或 Bark 标度或 ERB（等效矩形带宽）标度。后两个是基于临界频带的心理声学现象的经验标度。

在压缩前，502 的语音信号也被平滑。这种平滑减少了高音谐波的易变性对压缩信号平滑的影响。平滑可使用 LPC、或频谱平均、或内插法完成。

在步骤 508 中，通过分配压缩信号作为信号处理神经元 16 的中枢网络部分 18 的输入，来从背景噪声中提取语音信号。提取的语音信号表示的是在没有任何背景噪声下的初始语音信号的估值。在步骤 510 中，步骤 508 产生的提取信号和在步骤 506 中产生的压缩信号混合。混合过程尽可能的保存更多的原始压缩语音信号（来自步骤 506），只在需要时才依赖提取的语音估值。回到图 3，例如 306 中，部分远远高出背景噪声 302 水平的原始语音信号很容易被检测。因此，语音信号的这些部分将被保存在混合信号中，用来尽可能保存更多的初始语音信号特征。在完全被背景噪声掩盖的信号的原始信号部分中，没有任何选择，只能依赖在步骤 508 通过中枢网络提取的语音信号估值，假设提取的信号没有超出背景噪声或初始信号的强度。在信号强度位于

或接近背景噪声相同水平的区域，将压缩的初始信号和在步骤 508 提取的信号组合用来获得尽可能与原始信号尽可能接近的估值。混合过程产生压缩的重建语音信号，该语音信号尽可能多地具有初始语音信号特征，但是具有显著减小的背景噪声。

其余框图描绘了执行压缩重建语音信号的步骤。执行实时重建语音信号的步骤根据语音信号的应用而变化。例如，重建的语音信号可能直接转换为与一种汽车语音识别系统兼容的形式。步骤 520 表示 Mel 倒频谱系数（Frequency Cepstral Coefficient（MFCC））变换。步骤 520 的输出直接输入到语音识别系统中。可选择的是，通过在步骤 516 对压缩重建信号执行一个相反的频域-时序变换，步骤 510 产生的压缩重建语音信号直接转换回时序或可听见的语音信号。这样就会产生一个极大地减小或完全消除了背景噪声的时序信号。另一种选择是，压缩的重建语音信号可在步骤 512 解压缩。在步骤 514 谐波被加回到信号上，信号再次被混合。这时由于原始未被压缩的语音信号和混合信号转换回时序，语音信号或者信号在加上谐波之后不做附加混合立即被转换回时序信号。无论如何，其结果是改善的时序语音信号，其中的全部或大部分背景噪声被消除。

无论是从第一混合步骤 510，第二混合步骤 522，或在步骤 514 增添附加谐波之后输出，都可使用 502 中所用的时域到频域变换的反变换在 516 中将语音信号变换回时域。

图 6 描述了图 5 中的步骤 506 表示的语音信号压缩过程的第一阶段。语音信号 600 被表征为它的组成频率和每个频带的强度。语音信号 600 以频率（Hz）轴 602 和强度（dB）轴 604 绘制。频率图通常包含任意数量的离散频带。频率组 606 表示包括了语音信号 600 的 256 个频带。信号带数量的选择对于本领域的技术人员来说是一种公知的方法，这里的 256 频带长度只是用来说明。分辨线 608 表示背景噪声的强度。

语音信号 600 包含许多频率尖峰 610。这些频率尖峰 610 由语音信号 600 所带的谐波产生。这些频率尖峰 610 的存在掩盖了真实的语音信号，也使语音分离过程变得复杂化。平滑过程可消除这些频率尖峰 610。平滑过程可包括在语音信号 600 的谐波之间内插一个信号。在语

音信号 600 谐波信息稀疏的区域，内插算法平均了在其余信号上的插入值。插入信号 612 就是平滑处理的结果。

图 7 是被压缩语音信号 700 的示意图。被压缩语音信号 700 是以 Mel 频带轴 702 和强度（dB）轴 704 来绘制的。被压缩噪声估值 706 也被显示。信号压缩的结果是由较少数量的频带表示的信号，在本例中是在 20 和 36 个频带之间。代表较低频率的频带通常代表未压缩信号的 4 到 5 个频带。中间频率的频带代表大约 20 个压缩前的带。高频的频带通常代表大约 100 个压缩前的频带。

图 7 也表示步骤 508 预期的结果。被压缩噪声语音信号 700（实线）输入到信号处理单元 15（图 14）的中枢网络部分 18。中枢网络的输出是被压缩语音信号 708（虚线）。信号 708 表示语音信号中所有噪声的影响都被消除或抵消的理想情况。被压缩语音信号 708 被称为重建语音信号。

图 7 也示出了在步骤 510 混合处理所使用的强度门限值。强度上限值 710 限定强度等级高出背景噪声很多的强度。初始语音信号高于此门限的部分在没有去除背景噪声的情况下就很容易被检测。因此，对于原始语音信号强度等级高于强度上限 710 的部分，进行混合处理时仅使用原始信号。强度下限值 712 限定刚刚低于背景噪声平均强度的强度等级。原始信号强度等级低于强度下限值 712 的部分不能从背景噪声中区分出来。因此，对于原始语音信号强度水平低于强度下限值 712 的部分，假设提取的信号不超出背景噪声或原始信号的强度，则仅对通过步骤 508 产生的重建语音信号进行混合处理。对原始语音信号强度等级在强度下限值 712 和强度上限值 710 范围之间的部分，初始语音信号包括的内容仍然是有价值的，它们提供的信息能够增加语音信号的清晰度和质量，但是因为它们接近背景噪声的平均值并且可能包含噪声，所以其可靠性不强。因此，对初始信号强度在强度下限值 712 和强度上限值 710 范围之间的部分，步骤 510 对初始语音压缩信号和在步骤 508 的重建压缩信号部分都进行混合处理。对重建信号强度值在强度上和下限值之间的部分，步骤 510 的混合处理使用滑动标度（sliding scale）方法。原始信号中的信息靠近强度上限值离噪声门限较远，因此比接近强度下限值 712 的信息更具有可靠性。考虑

到这点，当信号强度靠近强度上限值时，混合处理给初始语音信号更高的权值，靠近强度下限值 712 时，给初始语音信号较低的权值。以相反的方式，对步骤 508 的压缩重建信号初始信号强度水平靠近强度下限值 712 的部分，混合处理给其较高的权值，对压缩重建信号初始信号强度水平靠近强度上限值 710 的部分，给其较低的权值。

图 8 是另一典型语音分离中枢网络的示意图。中枢网络 800 包括 3 个处理层：输入层 802，隐藏层 804 和输出层 806。输入层 802 包括输入神经元 808。隐藏层 804 包括隐藏神经元 810。输出层 806 包括输出神经元 812。输入层 802 的每个输入神经元 808 经过一个或更多连接 814 与隐藏层 804 的每个隐藏神经元 810 完全互连。隐藏层 804 的每个隐藏神经元 810 经过一个或更多连接 816 和输出层 806 的每个输出神经元 812 完全互连。

尽管没有特别说明，输入层 802 中输入神经元 808 的数量与频率组 702 的频带数目相一致。输出神经元 812 的数量也等于频率组 702 的频带数目。隐藏层 804 的隐藏神经元 810 的数目介于 10 和 80 之间。输入神经元 808 的状态由频率组 702 的强度值决定。实际上，中枢网络 800 接收噪声语音信号例如 700 作为其输入，并且产生作为输出的纯净的语音信号例如 708。

图 9 是另一典型语音分离中枢网络 900 的示意图。中枢网络 900 包括三个处理层：输入层 902，隐藏层 904，和输出层 906。输入层 902 包括两组输入神经元，语音信号输入层 908 和掩蔽输入层 910。语音信号输入层 908 包括输入神经元 912。掩蔽输入层 910 包括输入神经元 914。隐藏层 904 包括隐藏神经元 916。输出层 906 包括输出神经元 918。语音信号输入 908 的每个语音信号输入神经元 912 和噪声信号输入层 910 的每个输入神经元 914 经过一个或更多连接 920 和隐藏层 904 的每个隐藏神经元 916 完全互连。隐藏层 904 的每个隐藏神经元 916 经过一个或更多连接 922 和输出层 916 的每个输出神经元 918 完全互连。

语音信号输入层 908 的神经元 912 的数量与频率组 702 的频带数目一致。相同的，掩蔽信号输入层 910 的神经元 914 的数量与频率组 702 的频带数目一致。输出神经元 918 的数目也等于频率组 702 的频带数。隐藏层 904 的隐藏神经元 916 的数目介于 10 和 80 之间。输出神

经元 912 和输入神经元 904 的状态由频率组 702 的强度值决定。

实际上，中枢网络 900 接收噪声语音信号例如 700 作为输入，并且产生降噪语音信号例如 708 作为输出。掩蔽输入层 910 或直接或间接提供来自 506 或者用 700 表示的语音信号质量信息。换句话说，本发明的一个实施例中，掩蔽输入层 910 作为输入压缩噪声估值 706。

在本发明的另一实施例中，二元掩码可通过噪声估值 706 和压缩噪声信号 700 比较来计算。在 702 的每个压缩频带，当 700 和 706 之间的强度差超过门限值时，例如 3dB，掩码将被设为 1，否则设为 0。掩码是表示频率带是否携带有指示语音的可靠或有用信息的指示。506 的功能是仅重建掩码为 0，或被噪声 706 掩盖的 700 的那些部分。

在本发明的另一个实施例中，掩码不是二元的，而是 700 和 706 之间的差。因此，这种“失真（fuzzy）”的掩码为中枢网络指示了可靠性信任度。在 700 和 706 相交的区域设为 0，和二元掩码一样，在 700 十分接近 706 的区域具有较小的值，表示较低的可靠性或可信度，而 700 远远超过 706 的区域说明很好的语音信号质量。

中枢网络可学习时间以及在频率上的关联性。这对语音十分重要，因为嘴、喉、声道的物理构成限制了在发出一个声音之后多快形成另一个声音。因此，声音从一个时间帧到另一个往往是相关的，并且能够学习这种相关性的中枢网络可以优于其它不能学习这种相关性的中枢网络。

图 10 是另一个语音分离神经元网络 1000 的示意图。为简要起见，没有被列出单独的神经元。中枢网络 1000 包括 3 个处理层：输入层 1002-1008，隐藏层 1010，和输出层 1012。网络 1000 可等同于 900，只是输入层 1002 到 1006 中的神经元触发值可以从以前时间步骤压缩语音信号的值中指定。例如，在时间 t，1002 被指定在 t-2 中的压缩噪声信号 700， 1004 被指定给在 t-1 的 700， 1006 被指定给在时间 t 的 700，并且，如上所述，1008 可被指定掩码。因此，1010 能够学习压缩语音信号之间暂时关联性。

图 11 是另一语音分离中枢网络 1100 的示意图。中枢网络 1100 包括三个处理层：输入层 1102-1106，隐藏层 1108，和输出层 1110。网络 1100 可等同 900，除了神经元输入层 1106 的触发值可从以前时间步

骤中从 1110 提取的语音信号值中指派之外。例如，在时间 t，1102 被指派 t-1 的压缩噪声信号 700，1104 被指派给掩码，并且 1106 被指派给在时间 t-1 时 1110 的状态。上述网络在文献中被大家熟知为 Jordan 网络，它可以学会根据当前输入和先前输出改变其输出。

图 12 是另一语音分离中枢网络 1200 的示意图。中枢网络 1200 包括 3 个处理层：输入层 1202－1206，隐藏层 1208，和输出层 1210。网络 1200 可等同于 1100，除了输入层 1206 的神经元触发值可从以前时间步骤从 1208 提取的语音信号中的值指派以外。例如，在时间 t，1202 被指派 t-1 的压缩噪声信号 700，1204 被指派给掩码，而 1206 被指派给在时间 t-1 时 1206 的状态。上述网络在文献中为大家熟知为 Elman 网络，并且可以学会根据当前输入和先前内部的或隐藏的活动改变它的输出。

图 13 是另一语音分离中枢系统 1300 的示意图。中枢网络 1300 等同于 1200，除了它包含另一隐藏单元层 1310 之外。上述附加层允许更高等级相关性的学习，这样可以更好的提取语音。

隐藏或输出单元的强度值可由与它相连的每个输入神经元的强度和它们之间连接的权数的乘积的总和决定。使用非线性函数减少隐藏或输出神经元的活动范围，这种非线性函数可以是任何 S 函数，对数（logistic）或双曲线函数，或具有绝对极限的直线（line with absolute limits）。这些函数对于本领域的技术人员是熟知的。

中枢网络可在加入了真实或模拟噪声的干净的复合语音信号中来进行训练。

虽然本发明已以各种实施例披露如上，显而易见本领域的技术人员在本发明的范围内能够实施更多的实施例或应用。因此，本发明除了根据所附权利要求和其等效范围不应该受到限制。

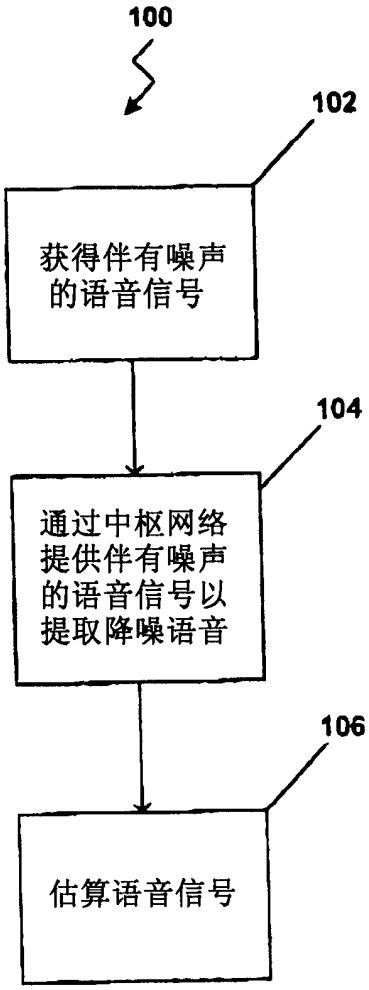

100
102
获得伴有噪声的语音信号
104
通过中枢网络提供伴有噪声的语音信号以提取降噪语音
106
估算语音信号

图1

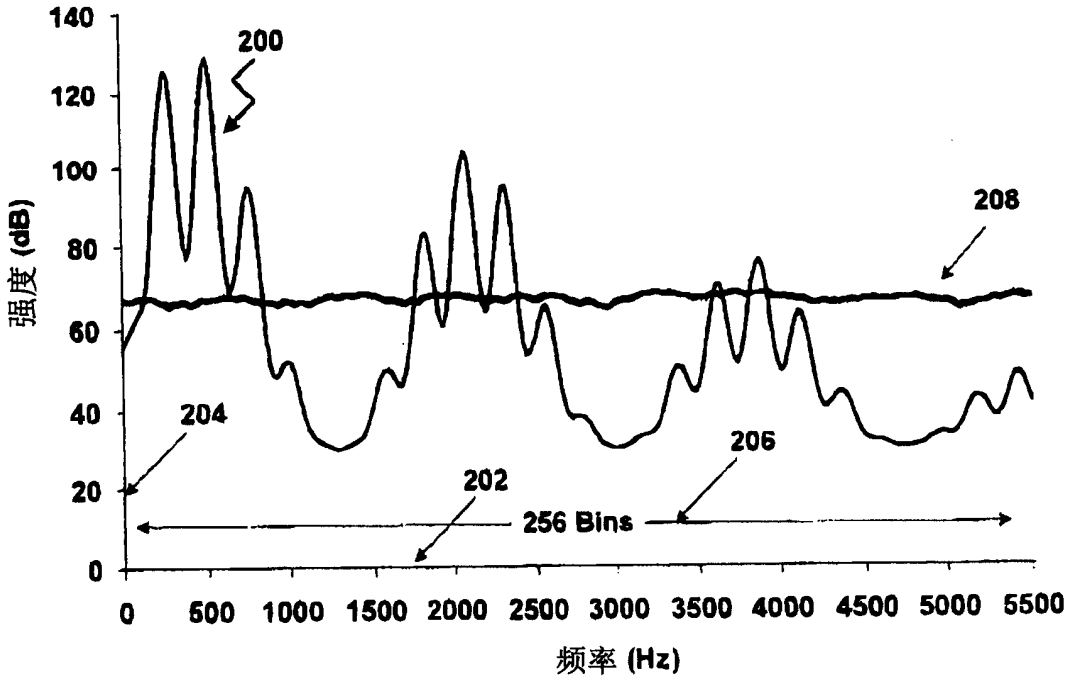

140
120
100
80
60
40
20
0
强度 (dB)
200
208
204
202
206
256 Bins
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
频率 (Hz)

图2

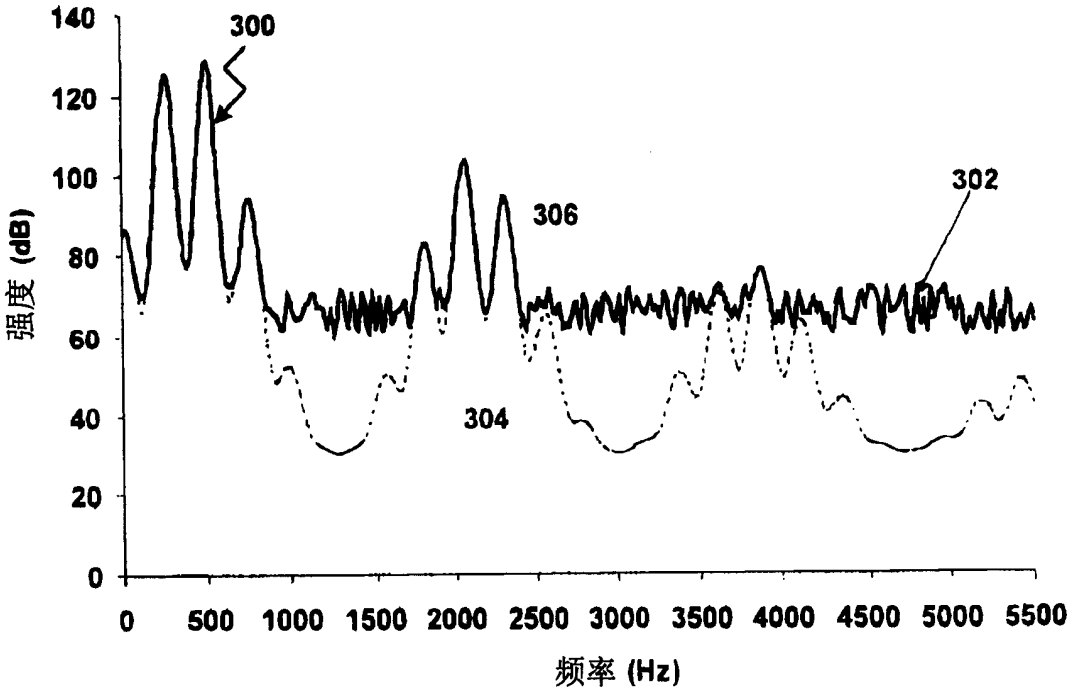

140
120
100
80
60
40
20
0
强度 (dB)
300
306
302
304
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
频率 (Hz)

图3

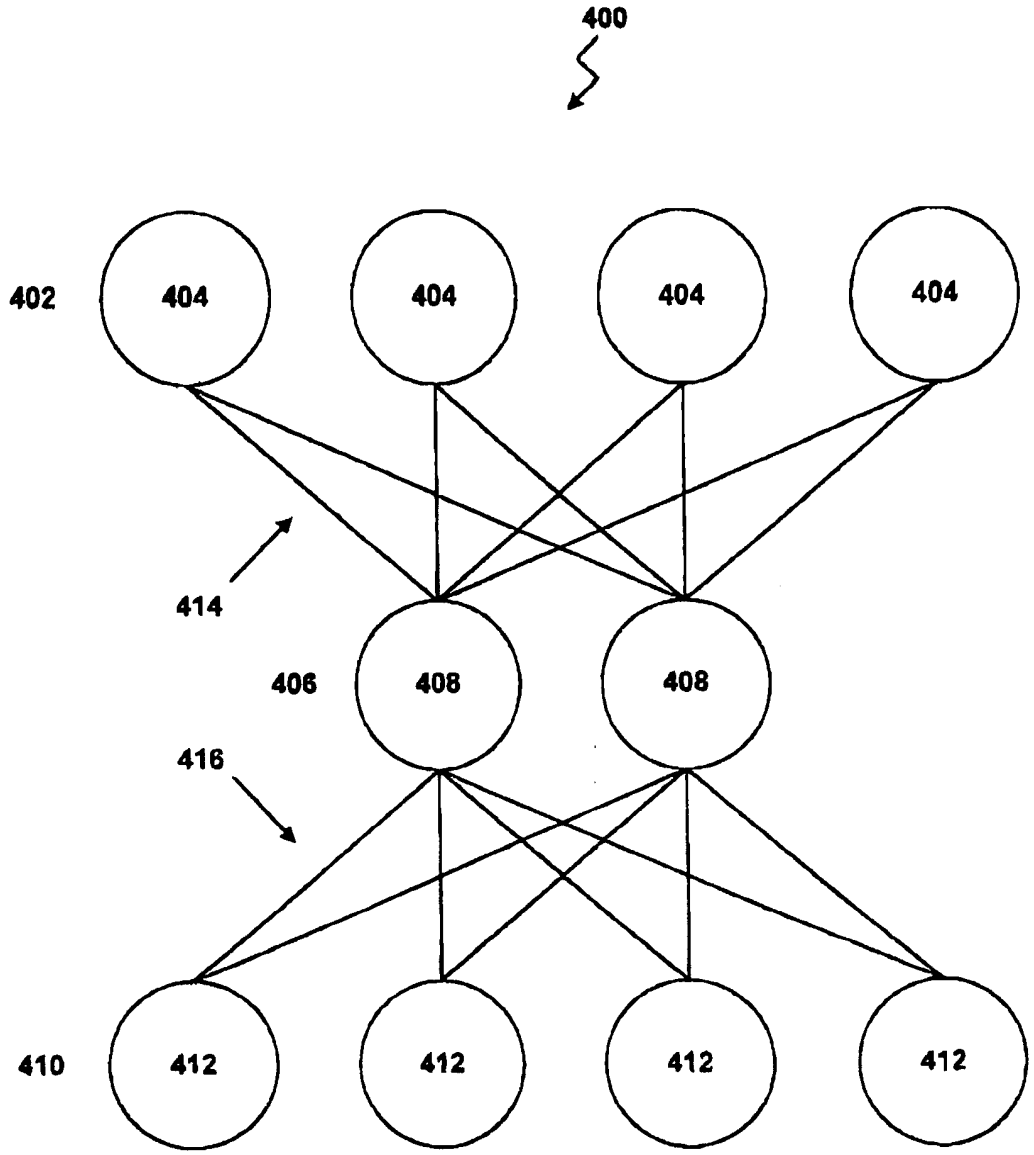


图4

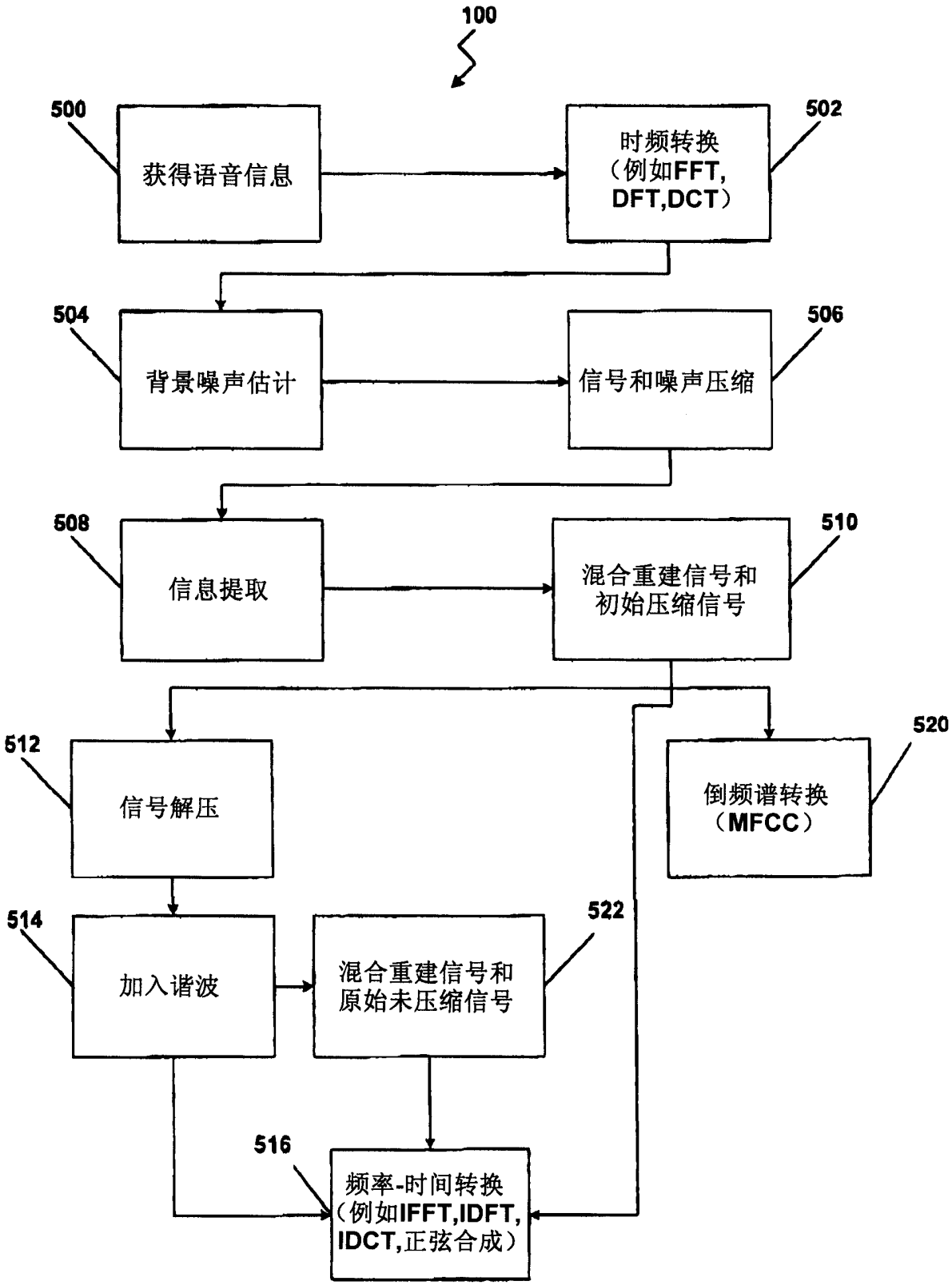

100
500
获得语音信息
502
时频转换
（例如FFT,
DFT,DCT）
504
背景噪声估计
506
信号和噪声压缩
508
信息提取
510
混合重建信号和
初始压缩信号
512
信号解压
520
倒频谱转换
（MFCC）
514
加入谐波
522
混合重建信号和
原始未压缩信号
516
频率-时间转换
（例如IFFT,IDFT,
IDCT,正弦合成）

图5

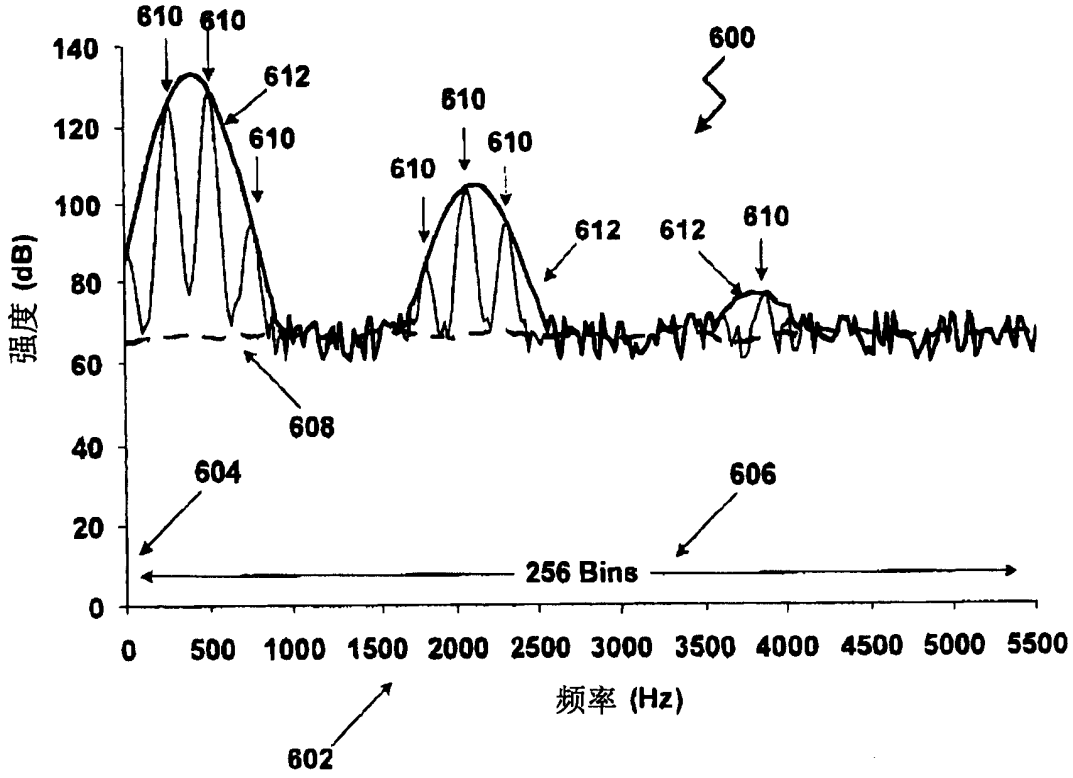

610
610
612
610
610
610
610
600
612
612
610
608
604
606
256 Bins
602
强度 (dB)
频率 (Hz)
140
120
100
80
60
40
20
0
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500

图6

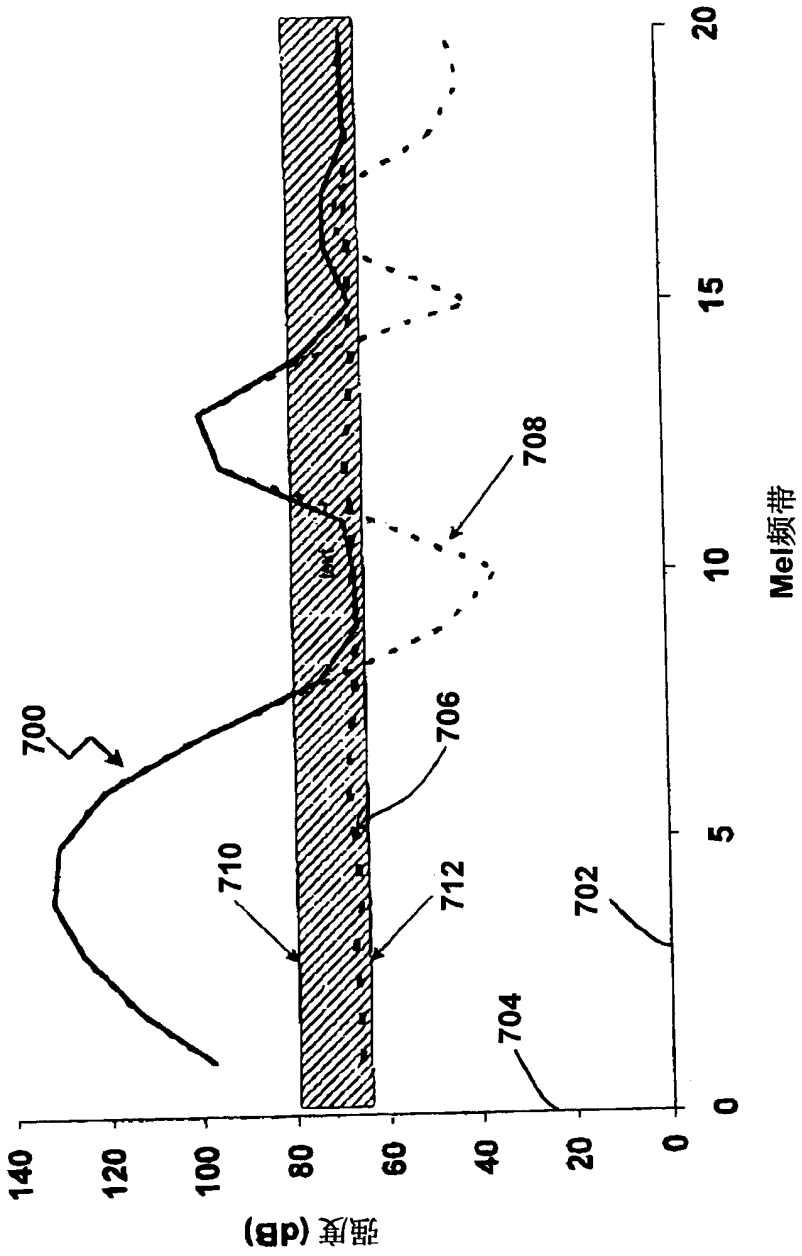

强度(dB)
140
120
100
80
60
40
20
0
0
5
10
15
20
Mel频带
700
702
704
706
708
710
712

图7

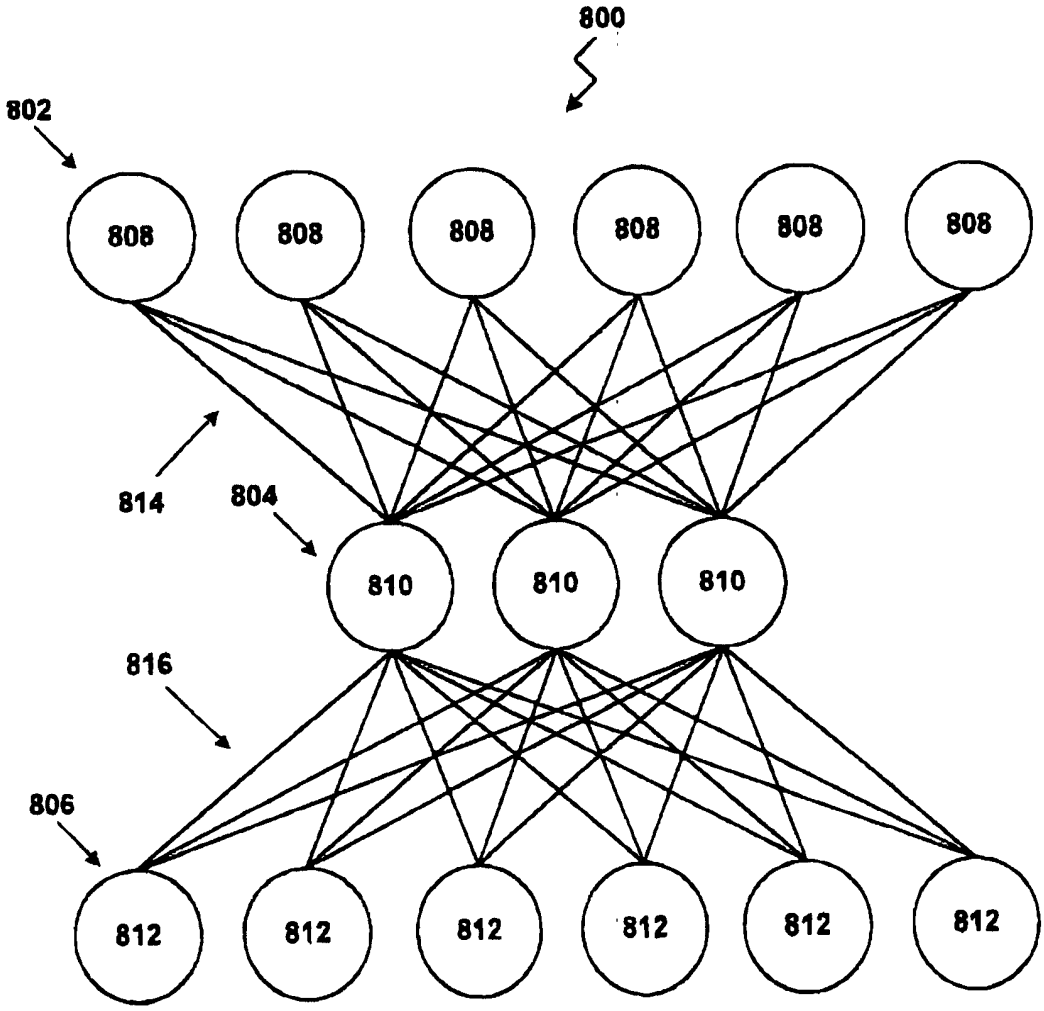

800
802
808
808
808
808
808
808
814
804
810
810
810
816
806
812
812
812
812
812
812

图8

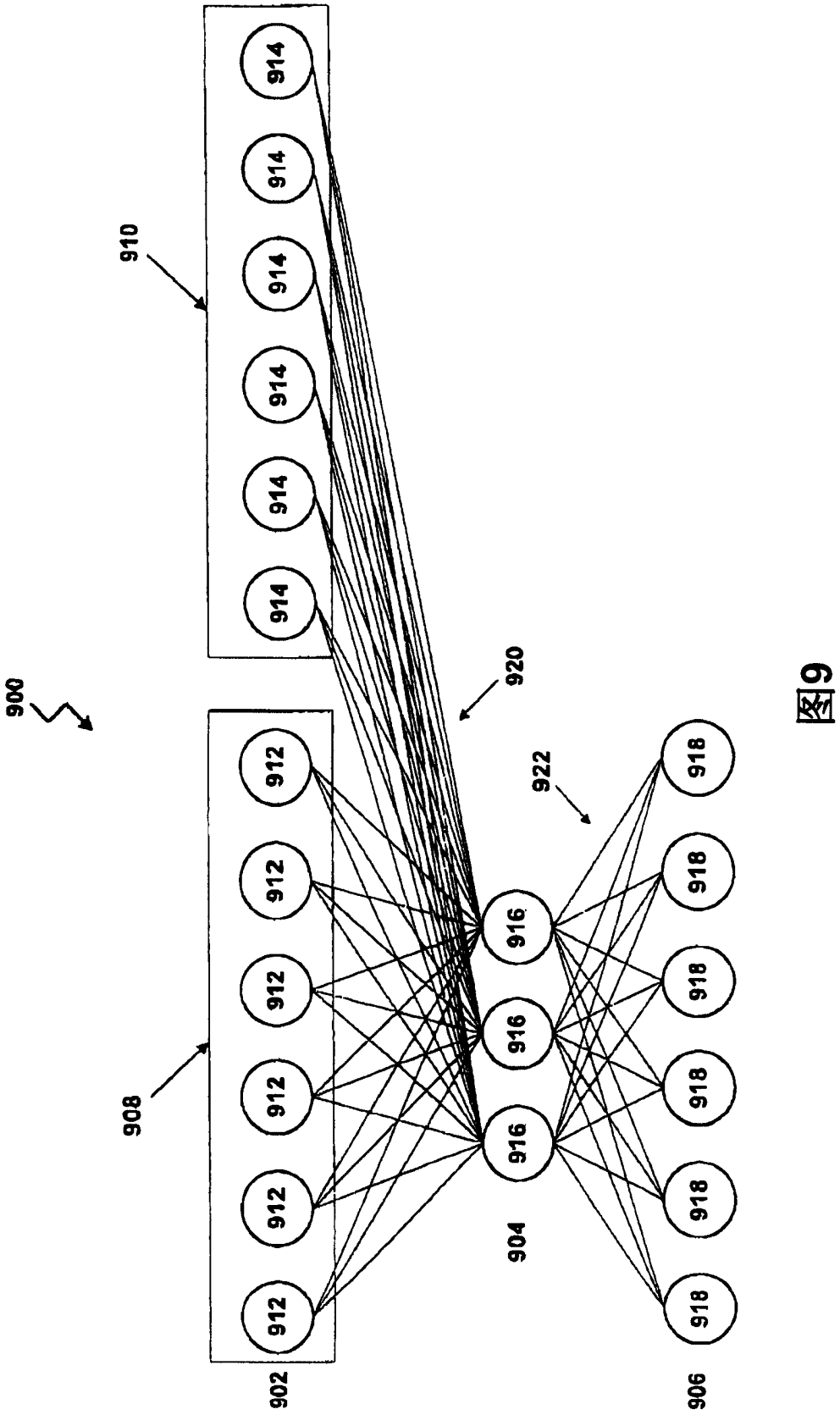

900
910
914
914
914
914
914
914
908
912
912
912
912
912
912
902
920
922
916
916
916
904
918
918
918
918
918
918
906

图9

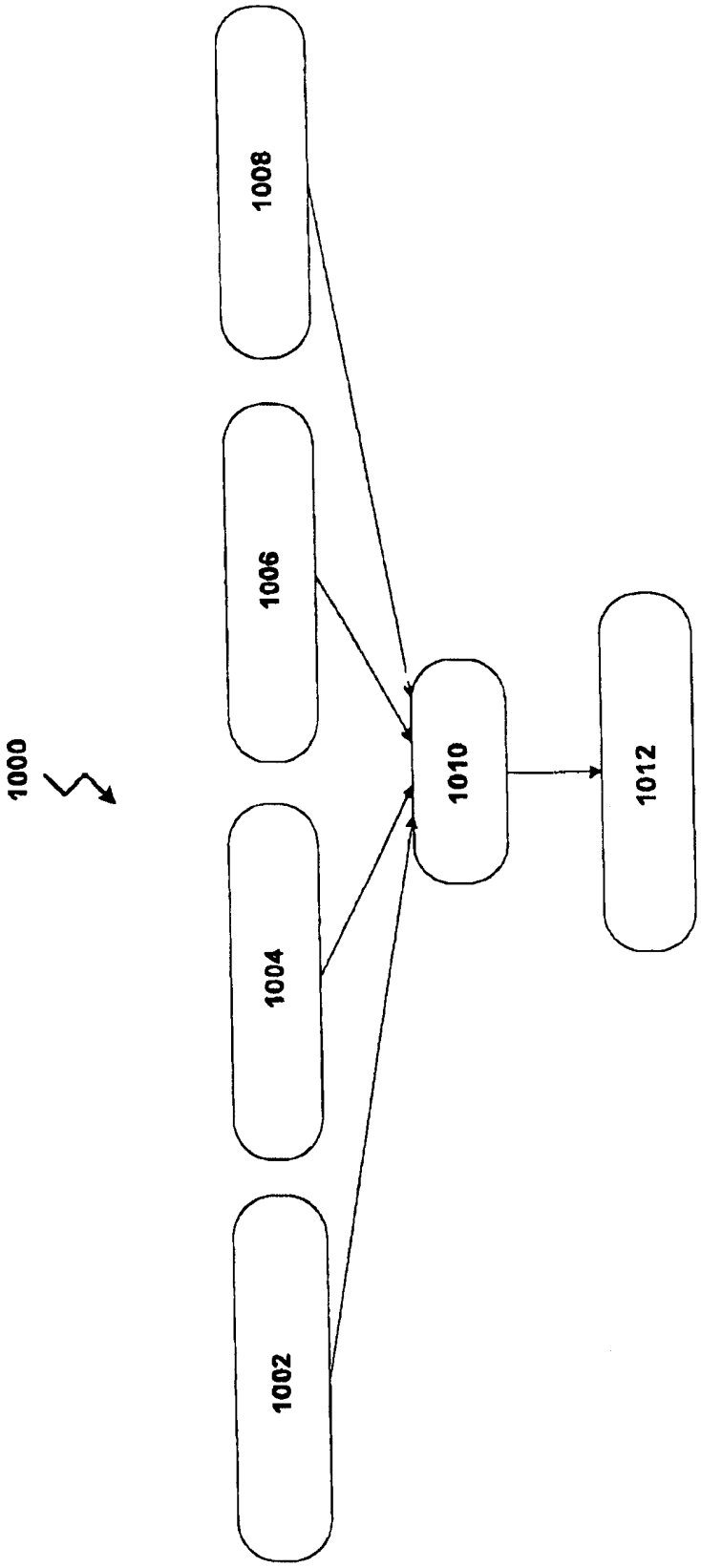


图10

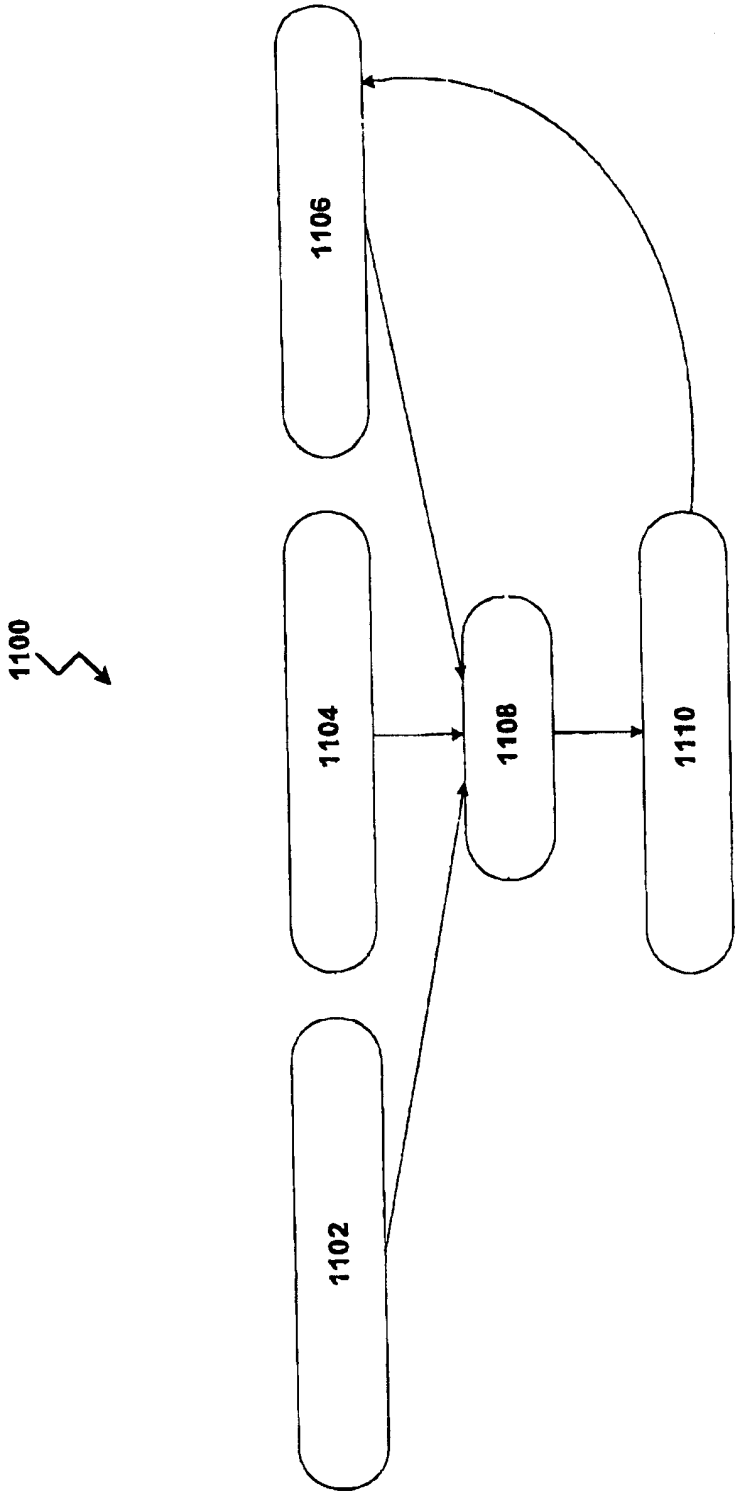


图11

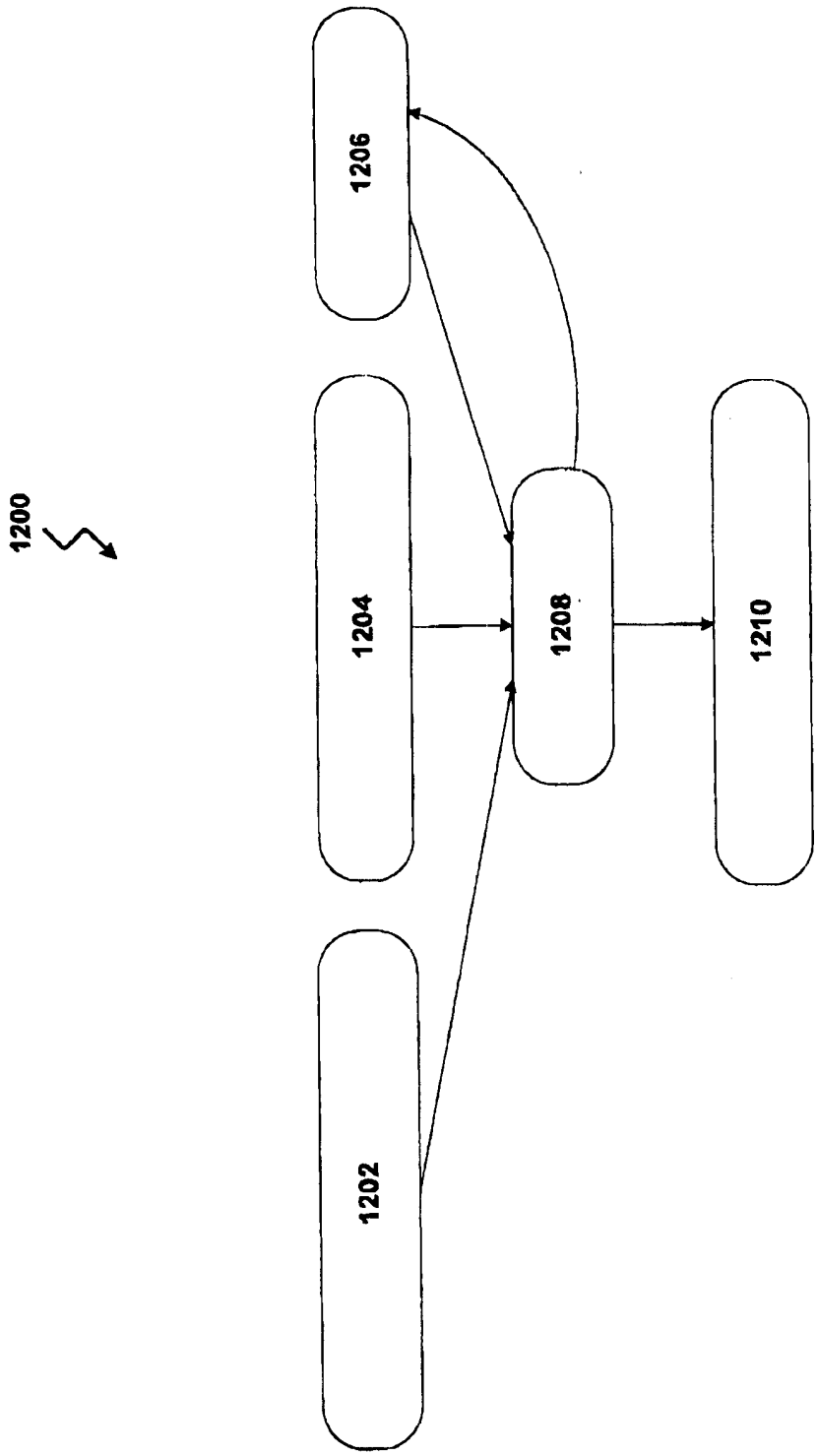


图12

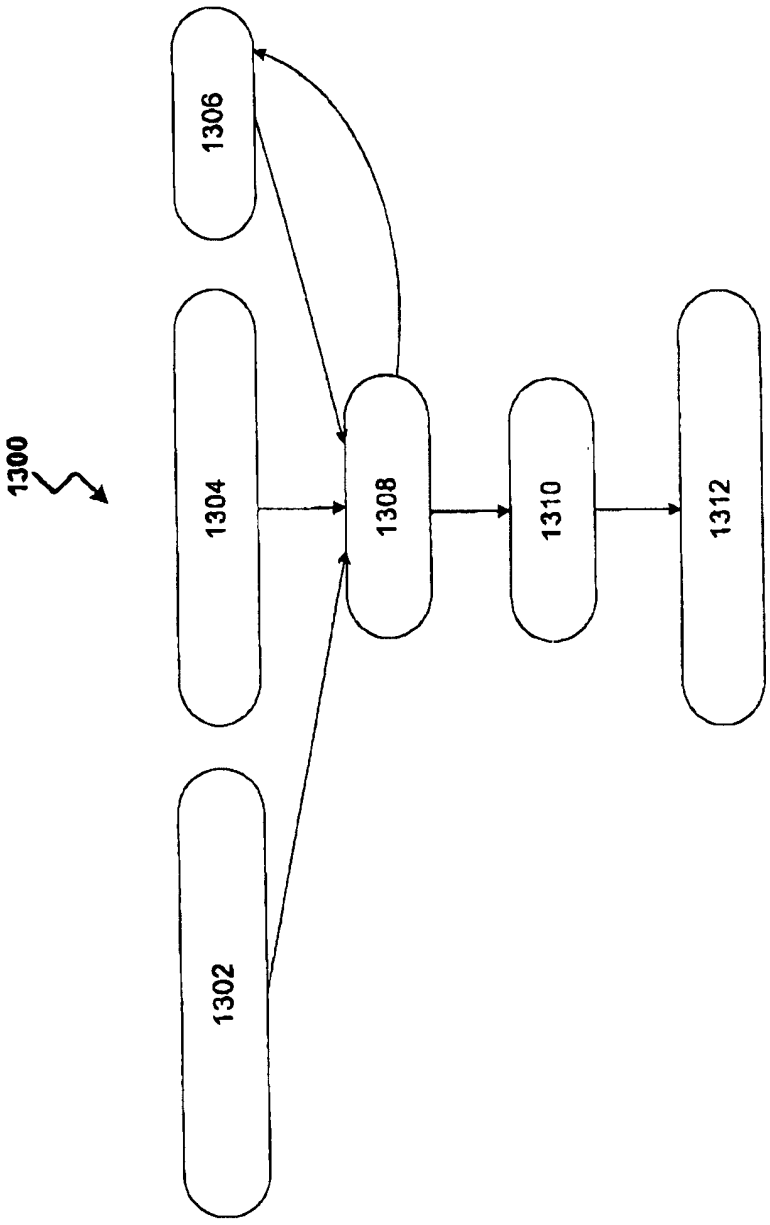


图13

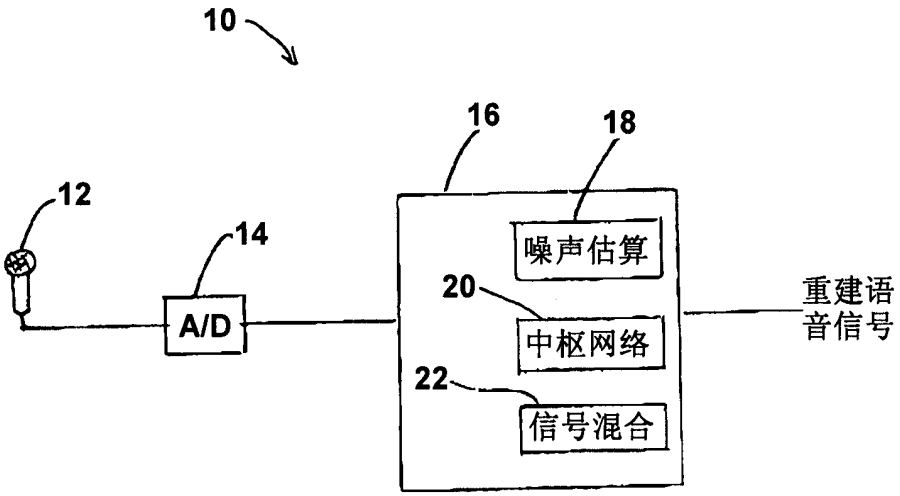


图14