(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102843540 A
(43)申请公布日 2012.12.26

(21)申请号 201210195470.9

(22)申请日 2012.05.17

(30)优先权数据
13/163,837 2011.06.20 US

(71)申请人 宝利通公司
地址 美国加利福尼亚

(72)发明人 P·L·楚 冯津伟 K·萨伊

(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 11038
代理人 陈新

(51)Int.Cl.
H04N 7/15(2006.01)
H04N 5/232(2006.01)

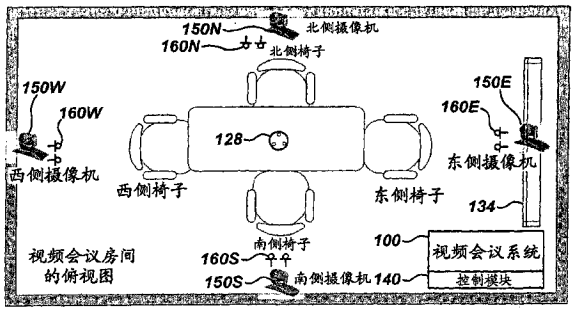
权利要求书 2 页 说明书 11 页 附图 8 页

(54)发明名称

用于视频会议的自动摄像机选择

(57)摘要

本公开涉及用于视频会议的自动摄像机选择。在视频会议摄像选择中，与用于视频会议的摄像机相关联的音频输入可以分别被处理成分别对应于第一和第二频率范围的第一和第二音频能量。该选择随后确定哪个音频输入具有第一音频能量与第二音频能量的最大比率，并选择相关联的摄像机画面以输出用于视频会议的视频。该选择也可以单独对来自摄像机的视频输入进行处理，或者结合音频处理对来自摄像机的视频输入进行处理。无论哪种处理方法，该选择对每个视频输入进行针对至少一个面部特征的处理，并确定哪个视频输入具有拍摄到人脸的最大可能性。最后，所述选择至少部分基于该视频的确定来选择相关联的摄像机画面，以输出用于视频会议的视频。



CN 102843540 A

1. 一种视频会议摄像机选择方法，包括：

获取用于视频会议的多个音频输入，每个音频输入与多个摄像机画面中的一个相关联；

将每个音频输入处理为第一频率范围的第一音频能量和第二频率范围的第二音频能量，第一频率范围高于第二频率范围；

确定哪个音频输入具有第一音频能量相对于第二音频能量的最大比率；

选择与具有最大比率的音频输入相关联的摄像机画面；以及

输出所选择的摄像机画面的视频用于视频会议。

2. 根据权利要求 1 的方法，其中第一频率大于约为 2500Hz 的阈值，而第二频率范围小于该阈值。

3. 根据权利要求 1 的方法，其中将每个音频输入处理为第一频率范围的第一音频能量包括：使用约 4000Hz 到约 7000Hz 的第一频率范围。

4. 根据权利要求 3 的方法，其中将每个音频输入处理为第二频率范围的第二音频能量包括：使用约 500Hz 到约 1000Hz 的第二频率范围。

5. 根据权利要求 1 的方法，进一步包括检测指示语音的音频，作为处理每个音频输入的先决条件。

6. 根据权利要求 1 的方法，选择相关联的摄像机画面包括：从一个摄像机画面切换到另一个摄像机画面以进行输出。

7. 根据权利要求 1 的方法，进一步包括调整所选择的摄像机画面的平移、倾斜和缩放中的一项或多项。

8. 根据权利要求 1 的方法，进一步包括：

获取多个视频输入，每个视频输入与一个摄像机画面相关联；

对每个视频输入进行针对至少一个面部特征的处理；以及

基于该处理确定哪个视频输入具有拍摄到人脸的最大可能性，

其中选择摄像机画面至少部分基于与具有最大可能性的视频输入相关联的摄像机画面。

9. 根据权利要求 8 的方法，其中至少一个面部特征是从由以下项组成的组中选择的：人脸的容貌特征、表示人皮肤的色调、表示人的运动、以及它们的组合。

10. 根据权利要求 1 的方法，其中摄像机画面与多个任意布置的摄像机相关联。

11. 根据权利要求 10 的方法，其中每个音频输入包括一个或多个通过与至少一个摄像机接近而与之相关联的麦克风。

12. 根据权利要求 1 的方法，其中摄像机画面与至少一个可控制的摄像机相关联。

13. 一种程序存储装置，其上存储有用于使可编程控制设备执行根据权利要求 1 的方法的程序指令。

14. 一种视频会议设备，包括：

视频接口，接收多个摄像机画面的视频输入；

音频接口，接收音频输入，每个音频输入与所述多个摄像机画面中的一个相关联；

网络接口，通信地耦接到网络；以及

处理单元，操作性地耦接到视频接口、音频接口和网络接口，该处理单元被编程以：

将每个音频输入处理为第一频率范围的第一音频能量和第二频率范围的第二音频能量，第一频率范围高于第二频率范围；

确定哪个音频输入具有第一音频能量相对于第二音频能量的最大比率；

选择与具有最大比率的音频输入相关联的摄像机画面；以及

通过网络接口输出所选择的摄像机画面的视频输入。

15. 根据权利要求 14 的设备，

其中处理单元被编程以：

对每个视频输入进行针对至少一个面部特征的处理；以及

基于该处理确定哪个视频输入具有拍摄到人脸的最大可能性，并且

其中摄像机画面的选择至少部分基于与具有最大可能性的视频输入相关联的摄像机画面。

16. 根据权利要求 14 的设备，进一步包括通信地耦接到视频接口的多个任意布置的摄像机，每个摄像机与所述多个摄像机画面中的一个相关联。

17. 根据权利要求 16 的设备，还包括通信地耦接到音频接口的多个麦克风，每个麦克风通过与至少一个所述任意布置的摄像机接近而与之相关联。

18. 一种视频会议摄像机选择方法，包括：

获取用于视频会议的多个视频输入，每个视频输入与多个摄像机画面中的一个相关联；

对每个视频输入进行针对至少一个面部特征的处理；以及

基于该处理确定哪个视频输入具有拍摄到人脸的最大可能性，

选择与具有最大可能性的视频输入相关联的摄像机画面；

输出所选择的摄像机画面的视频输入用于视频会议。

19. 根据权利要求 18 的方法，进一步包括：

获取用于视频会议的多个音频输入，每个音频输入与多个摄像机画面中的一个相关联；

将每个音频输入处理为第一频率范围的第一音频能量和第二频率范围的第二音频能量，第一频率范围高于第二频率范围；以及

确定哪个音频输入具有第一音频能量相对于第二音频能量的最大比率；

其中选择摄像机画面至少部分基于与具有最大比率的音频输入相关联的摄像机画面。

20. 一种视频会议设备，包括：

至少一个可控制的摄像机，用于获取具有定向画面的视频；

至少一个麦克风，用于获取音频；以及

处理单元，操作性地耦接到所述至少一个可控制的摄像机和所述至少一个麦克风，该处理单元被编程以：

将来自所述至少一个麦克风的音频处理为第一频率范围的第一音频能量和第二频率范围的第二音频能量，第一频率范围高于第二频率范围；

确定第一音频能量与第二音频能量的比率，以及

基于所确定的比率控制所述至少一个可控制的摄像机的定向画面。

用于视频会议的自动摄像机选择

背景技术

[0001] 对于在房间或其他环境中使用的大多数视频会议系统来说，在视频会议中讲话的参与者的画面受限制仍然是一个持续的问题。例如，图 1A 示出了具有典型配置的视频会议房间的平面图。单个摄像机 14 安装在显示器 12 顶上以用于视频会议系统 10。当从该摄像机 14 捕捉的视频被发送到远端时，在该远端的画面受该摄像机画面（比如东侧摄像机）的限制。如果坐在南侧椅子的参与者正在对房间中的其他人讲话，则远端的观看者将看到讲话者的侧面而不是更理想的正面画面。即使可以控制摄像机 14 的平移、倾斜和缩放，这种情况也不会改变。最终，得到的受限制的参与者画面可能并不是远端的观看者所想要的。

[0002] 为解决这些问题，视频会议系统 10 可以如图 1B 所示的在视频会议环境中使用多个摄像机 14。在这里，多个摄像机 14(N, S, E, W) 安放在房间四周以获得参与者的更多画面。使用在桌上的麦克风箱 16 里的麦克风处接收的全频带能量，系统 100 可以发现当前正在讲话的参与者的方向。为此，在箱 16 中拾取最强能量的麦克风可以指示当前讲话者的方向。在此基础上，系统 10 接着选择具有与该方向相关联的画面的摄像机 14(N, S, E, W)。

[0003] 不幸的是，在人们讲话时，单纯的能量并不能可靠地指示讲话的人的头部是如何转动的。例如，坐在南侧椅子的参与者可能正在讲话，从麦克风箱 16 确定的具有最大音频能量的方向将指示北侧的摄像机 150N 是获得讲话的参与者的画面的最佳摄像机。在此基础上，视频会议系统 10 将选择北侧的摄像机 150N 用于输出视频。

[0004] 然而，南侧椅子的参与者在讲话时，实际上可能将他的头转向东侧椅子或显示器 12 处的参与者，从而使他的会话指向东侧。依赖于桌子处的最强麦克风能量的视频会议系统 10 将不能确定参与者的头是如何转动的。结果在他讲话时，尽管他面朝东侧（面向东侧椅子或面向显示器 12），视频会议系统 10 将发送来自北侧摄像机 140N 的参与者侧面画面。远端观看者获得的将是一幅不太满意的参与者讲话的画面。

[0005] 本公开的主题旨在克服，或至少减轻上面所提出的一个或多个问题的影响。

发明内容

[0006] 公开了用于执行自动化视频会议技术的方法、可编程存储装置以及视频会议设备。所述视频会议设备利用与各个麦克风耦接的音频接口来获取用于视频会议的音频输入。所述音频输入中的每一个与多个摄像机中的一个相关联。例如，可以任意地在视频会议环境周围布置各个摄像机，并且每个摄像机可以就近关联有一个或多个麦克风。

[0007] 作为阈值确定，该设备可以在处理每个音频输入前，首先检测指示语音的音频。可以使用检测与人类语音相关联的预期水平和预期频率范围内的声学能量的语音检测技术来实现这一点。

[0008] 在任何情况下，该设备都将每个音频输入处理成分别对应于第一和第二频率范围的第一和第二音频能量。通常，第一频率范围大于第二频率范围。更特别地，第一频率范围可以是约 4000Hz 到约 7000Hz，而第二频率范围可以是约 500Hz 到 1000Hz。

[0009] 在确定能量后，该设备接着确定哪一个音频输入具有这些不同能量的最大比率。

使用该比率结果，该设备选择与最大比率关联的摄像机来输出视频以用于视频会议。因为在更高频率处，可以更好地辨别出正在说话的人的头部的方向性，所以所选择的摄像机更可能具有能够指向当前正在说话的视频会议参与者的面部的画面。

[0010] 随着视频会议的进行，该设备可以根据哪个参与者正在说话以及确定他们面对的是哪个摄像机，而在各个摄像机之间切换以用于输出。基于音频的摄像机确定可以单独使用或是与下面描述的基于视频的确定结合使用。同样地，基于视频的确定也可以单独使用。

[0011] 在基于视频的方案中，该设备获取与分别与多个摄像机之一相关联的输入视频。由此，该设备针对至少一个面部特征处理每个视频输入，并确定哪个视频输入具有拍摄到人脸的最大可能性。至少部分基于该视频确定，该设备选择相关联的摄像机来输出用于视频会议的视频。通常，面部特征可以包括摄像机画面中的人脸的容貌特征、表示人皮肤的色调、表示人的运动、以及它们的组合。当与基于音频的确定一起使用时，该视频确定可以进一步改进最终的摄像机选择。

[0012] 以上概述并不是意图概括本公开的每个可能的实施例或是每个方面。

附图说明

[0013] 图 1A 示出了具有现有技术视频会议系统的配置的视频会议房间的平面图。

[0014] 图 1B 示出了具有现有技术视频会议系统的另一配置的视频会议房间的平面图。

[0015] 图 2 示出了具有根据本公开的视频会议系统的视频会议房间的平面图。

[0016] 图 3A 图解说明了根据本公开的特定教导的视频会议系统。

[0017] 图 3B 图解说明了图 3A 所示的视频会议系统的部件。

[0018] 图 4 原理性示出了所公开的视频会议系统的麦克风、摄像机和控制模块的布置。

[0019] 图 5A 原理性示出了通过所公开的系统的控制模块执行的基于音频的摄像机选择处理。

[0020] 图 5B 图解说明了声音的音频频率是如何与频率和声调谱的方向性信息相关的。

[0021] 图 6 原理性示出了通过所公开系统的控制模块执行的基于视频的摄像机选择处理。

[0022] 图 7 示出了摄像机和麦克风相对于参与者和麦克风箱的另一原理性布置，以帮助更详细地描述摄像机选择技术。

[0023] 图 8A-8B 示出了图 7 布置中的参与者的示例性摄像机画面。

[0024] 图 9A 示出了具有摄像机、麦克风及摄像机操控和处理部件的摄像机模块。

[0025] 图 9B 示出了图 9A 的摄像机模块所捕捉的画面。

具体实施方式

[0026] A. 具有多个摄像机和麦克风的视频会议系统概述

[0027] 图 2 原理性示出了具有根据本公开的视频会议系统 100 的视频会议房间的平面图。系统 100 具有布置在该房间四周的多台摄像机 150(N, S, E, W)。尽管只示出了四台摄像机 150(N, S, E, W)，然而根据执行情况可以使用更多或更少的摄像机。例如，一个可控制的摄像机 150 可以负责多个摄像机画面。

[0028] 进一步地，图 2 的摄像机 150 对称布置或有组织地布置在该房间四周。然而并不

总是按这种情况布置摄像机，因为为了得到不同画面，可以任意将系统 100 中的摄像机 150 设置在多个位置，而且所述布置可以从一个视频会议到下一个视频会议改变，或者甚至在视频会议过程中也可以进行改变。因此，可以明了本公开的教导可以应用到多种任意布置中的任一种，且不仅限于应用于所示的视频会议系统 100 的有序的、预先配置的布置。

[0029] 通常，摄像机 150 可以是适用于视频会议的任何适合的摄像机，其具有固定的画面，或者可以包括可操控的平移 - 倾斜 - 缩放 (PTZ) 或电子式平移 - 倾斜 - 缩放 (PTZ) 摄像机。因此，视频会议系统 100 可以包括用于根据现有技术中的自动化技术来引导各个摄像机 150 的平移、倾斜和缩放的特征。例如，给定的摄像机 150 可能能够检测和定位音频源并自动执行必要的平移、倾斜和缩放来拍摄到该音频源。

[0030] 每台摄像机 150 具有与之关联的一个或多个麦克风 160(N, S, E, W)。这些麦克风 160 可以相对应摄像机 150 单独安装，或是需要时可以集成于摄像机 150 中。如果集成于摄像机 150 中，相关的麦克风 160 可以用来检测音频源的方向，从而摄像机 150 可以应用现有技术向该音频源自动平移、倾斜和缩放。然而总的来说，相关联的麦克风 160 可以任意布置在房间内。由此那些紧邻给定摄像机 150 的麦克风 160 与该给定摄像机 150 相关联。这样的关联可以针对房间预先配置，或通过系统 100 的用户人为配置，或是由系统 100 使用自动检测技术（比如侦测 (ping) 音频信号、红外信号等等）而自动检测。

[0031] 另外，可以在桌上使用麦克风箱 128，以获取视频会议的主音频。这样，与摄像机关联的麦克风 160 可以用于这里所公开的摄像机引导和选择。当然，如这里所公开的，系统 100 可以仅仅使用与摄像机关联的麦克风 160 来同时用于会议音频和摄像机选择。

[0032] 在视频会议中，摄像机和麦克风信号传送给视频会议系统 100。视频会议系统 100 处理这些信号，接着选择要通过通信网络输出给远端（未示出）的摄像机画面。如接下来所详细描述的，摄像机的选择取决于哪个参与者正在讲话以及他们如何朝向。然而，在探究视频会议系统 100 如何确定选择哪个摄像机画面之前，下文中讨论的图 3A-3B 将示出视频会议系统 100 的更多细节。

[0033] B. 视频会议系统的详细描述

[0034] 如图 3A 所示，视频会议系统 100 通过网络 22 与一个或多个远程端点 20 进行通信。在一些通常的部件中，系统 100 包括具有音频编解码器 122 的音频模块 120 和具有视频编解码器 132 的视频模块 130。这些模块 120/130 可操作地耦接到控制模块 140 和网络模块 160。控制模块 140 可以是单独的部件或是集成于系统 100 内。

[0035] 在视频会议中，摄像机 150 捕捉视频，并将捕捉的视频提供给视频模块 130 和编解码器 132 进行处理。另外，麦克风 128/160 捕捉音频，并将音频提供给音频模块 120 和编解码器 122 进行处理。如之前所述，麦克风箱 128 可以放置在桌上（或者可以使用天花板麦克风），系统 100 可以使用麦克风箱 128 捕捉的音频主要用于会议音频。

[0036] 另外，麦克风 160 与摄像机 150 相关联，以捕捉音频，并如上所述地将该音频提供给音频模块 122 进行处理。例如，每台摄像机 150 可以具有一个或多个麦克风 160，并且相关联的麦克风 160 可以布置成正交阵列，以在视频会议中确定音频源的位置。可替换的是，麦克风 160 可以是与摄像机 150 分离的部件，但是就近与其进行关联。通常，系统 100 可以使用来自这些麦克风 160 的音频主要用于摄像机追踪和选择的目的，而并不是将其用于会议音频，尽管这些音频也可以用于会议。

[0037] 对于摄像机追踪，例如，视频会议系统 100 可以包括一个或多个摄像机控制器 152，该摄像机控制器 152 能够处理麦克风 128/160 拾取的音频信号和摄像机 150 产生的视频信号，以确定在视频会议中正在讲话的参与者的位置。（一个或多个控制器 152 可以与摄像机 150 分离，或并入摄像机的单元，或是视频会议系统 100 的一部分）。这样，使用自动摄像机控制的视频会议系统 100 可以确定参与者面部的精确位置，并可以对该位置自动“放大”。实现该功能的示例性系统在美国专利号 5778082 名称为“Method and Apparatus for Localication of an Acoustic Source”，美国专利号 6593956 名称为“Locating an Audio Source”以及美国专利号 6980485 名称为“Automatic Camera Tracking using Beamforming”的专利文件中被公开，它们通过参考全部结合于此。这些和其他已知的技术可以应用于下面所讨论的本公开的摄像机选择技术。

[0038] 当捕捉到音频和视频时，系统 100 使用任何通常编码标准，比如 MPEG-1、MPEG-2、MPEG-4、H.261、H.263 和 H.264，来编码这些信号。接着，网络模块 160 使用任何合适的协议，通过网络 22 输出编码的音频和视频到远程端点 20。类似地，网络模块 160 通过网络 22 从远程端点 20 接收会议音频和视频，并发送这些到其各自的编解码器 122/132 来处理。最后，扬声器 126 输出会议音频，并且显示器 134 输出会议视频。这些模块和其他部件中的许多可以以本领域熟知的传统方式进行操作，因此这里没有提供进一步的详细描述。

[0039] 与一些传统特征相对比，系统 100 使用控制模块 140，以自动化且协调的方式来从各个摄像机 150 选择视频输出。通常，在视频会议中的任何特定时间，系统 100 仅输出来自多个摄像机 150 之一的视频，并且优选地，输出的视频捕捉到的是正在说话的视频会议参与者。随着视频会议的继续，来自系统 100 的输出视频可以相据哪个参与者正在说话而在各个摄像机 150 的画面之间不时地切换。

[0040] 为了选择由摄像机 150 捕捉的画面以用于输出，系统 100 使用基于音频的定位器 142 和基于视频的定位器 144，来确定参与者的位置以及拍摄房间和参与者的画面。接着，操作性地耦接到音频和视频模块 120/130 的控制模块 140 使用来自这些定位器 142/144 的音频和 / 或视频信息，来选择摄像机 150 的画面以用于输出和 / 或发送摄像机命令以引导摄像机 150 改变其朝向和它们捕捉的画面。

[0041] 例如以下所更详细描述的，控制模块 140 使用来自远程麦克风 160 的、由基于音频的定位器 142 处理的音频信息。为避免关注非话音相关的音频，基于音频的定位器 142 可以使用语音检测器 143 在来自麦克风 160 的捕捉音频中检测语音。语音检测技术可以检测与人类语音相关联的预期水平和预期频率范围内的声学能量，从而可忽略视频会议期间的噪音和无关声音。

[0042] 控制模块 140 使用来自基于音频的定位器 142 的、所确定的当前讲话者的位置，来切换到最佳画面摄像机 150，和 / 或操控最佳画面摄像机 150 朝向当前讲话者。也如下文中详细描述的，控制模块 140 可以使用由基于视频的定位器 144 利用来自摄像机 150 的视频而处理的视频信息，来确定参与者的位置，确定画面的取景，以及操控摄像机 150 朝向参与者的面部。

[0043] 图 3B 示出了图 3A 的视频会议系统的一些示例性部件。如上文中所示和所讨论的，系统 100 具有两个或更多个摄像机 150 和多个麦克风 128/160。除这些之外，系统 100 具有处理单元 110、网络接口 112、存储器 114 以及通用输入 / 输出 (I/O) 接口 118，它们都通过

总线 111 耦接。

[0044] 存储器 114 可以是任意的常规存储器，如 SDRAM，并可以以软件和硬件的形式存储用于控制系统 100 的模块 116。除前文讨论的视频和音频编解码器以及其他模块以外，模块 116 还可以包括操作系统、使用户能够控制系统 100 图形用户界面 (GUI)、以及接下来所讨论的用于处理音频 / 视频信号和控制摄像机 150 的算法。

[0045] 网络接口 112 提供系统 100 和远程端点（未示出）之间的通信。作为对照，通用 I/O 接口 118 提供与本地设备的数据传输，本地设备诸如 :键盘，鼠标，打印机，头顶投影仪，显示器，外部扬声器，附加的摄像机，麦克风箱等等。系统 100 可以也包括内部扬声器 126。

[0046] 摄像机 150 和麦克风 160 在视频会议环境中分别捕捉视频和音频，并产生通过总线 111 传输到处理单元 110 的视频和音频信号。这里，处理单元 110 使用模块 116 中的算法处理视频和音频。例如，如这里所公开的，处理单元 110 处理麦克风 128/160 捕捉的音频以及摄像机 150 捕捉的视频，以确定参与者的位置和引导摄像机的画面。最后，处理过的音频和视频可以发送到与接口 112/118 耦接的本地和远程设备。

[0047] 通过以上所提供的对视频会议系统 100 的理解，讨论现在转向视频会议系统 100 如何使用基于音频和视频的技术来选择最佳的摄像机画面以捕捉视频会议中当前正在说话的参与者。

[0048] C. 视频会议系统的摄像机选择

[0049] 如前文所提到的，视频会议期间的摄像机和麦克风信号传送给视频会议系统 100 的控制模块 140。基于图 2 的系统 100 的示例性布置，上述内容在图 4 中进行了原理性的描述。控制模块 140 将用于视频会议的所捕捉视频来源切换到指向当前说话的参与者的面部的摄像机 150。作为该选择的结果，控制模块 140 将来自所选择的摄像机 150 的所捕捉视频引导到系统的视频编码器 170，以作为输出传输到远端。

[0050] 不同于像现有技术中通常所做的那样仅基于那些麦克风具有最强能量来选择摄像机画面，控制模块 140 确定正在说话的参与者实际上可能如何朝向。这样，如果坐在图 2 的南侧椅子上的参与者正在说话并且他的头面向东侧椅子或显示器 (134)，则控制模块 140 可以更适当地选择东摄像机 (160E) 来捕捉参与者的视频，即使麦克风箱 (128) 检测到的麦克风能量可能指示的是其他方向。为实现上述操作，控制模块 140 使用图 4 的摄像机选择处理 146 来确定参与者在说话时面向哪个摄像机 (150)。

[0051] 在一种布置中，摄像机选择处理 146 可以使用基于音频的处理 200，其基于房间中的麦克风 (128/160) 捕捉的声学特性来确定讲话者面向哪个摄像机 (150)。图 5A-5B 示出了基于音频的处理 200 的特征。作为基于音频的处理 200 的可替换选择或与之结合，摄像机选择处理 146 可以使用基于视频的处理 300，该处理 300 可以使用基于各个摄像机 (150) 所捕捉的视频。图 6 示出了基于视频的处理 300 的特征。

[0052] 1. 基于音频的选择处理

[0053] 图 5A 原理性示出了通过控制模块 (140) 执行的基于音频的摄像机选择处理 200 的示例。（引用之前的部件和数字，以有助于解释。）来自与各个摄像机 (150) 相关联的麦克风 (160) 的音频信号 206(N,S,E,W) 作为输入到达系统 (100)，并且系统的音频部件的滤波器组 210 将每个进来的音频信号 206 滤波到合适的频带，任何合适频带都可以被用于滤波音频信号 206，优选地，频带及其范围有助于这里所公开的语音处理和检测目的。

[0054] 各个频带中的音频以适当的间隔（比如通常使用 20ms）被采样，并且样本的能量水平使用本领域已知的技术来计算。从该采样和计算所得到的能量水平信号 212 用在后面的处理中。对于每一个相关联的麦克风 (160) 及其信号 206 进行采样和能量水平计算，在该示例中包括北、南、东、西，但是也可以应用任意的布置。

[0055] 这样，基于音频的选择处理 200 执行每一个能量水平信号 212(N, S, E, W) 的比较处理 220。这里，处理 220 在两个不同的频率范围 222/224 中比较每个能量水平信号 212。在该比较中，特定信号 212 在高频率 222 范围内的能量与该特定信号 212 在低频率 224 范围内的能量进行比较。如下文详细描述的，挑选的频率范围被选择为最佳地确定正在说话的参与者面对哪个方向。

[0056] 如图 5B 所示，人声音的方向（从而，人所面对的方向）最佳地在较高频率处而不是在较低频率处被确定。在图 5B 中，例如，极坐标图示出在不同频率处且具有不同发音的人声音在水平平面中的方向。在该示例中，声音以不同的元音唱出，这些图从 Katz, Brian F.G.&d'Alessandro, Chirstophe, "Directivity Measurements of the Singing Voice", Proceedings of the 19th International Congress on Acoustics(ICA'2007), Madrid, 2-7September 2007 中获得。

[0057] 如这些图一般性示出的，当发声时人的头部的方向性可以最佳地在较高频率处确定。例如，在频率 160Hz，极坐标图显示声音大致是全向的。相对的是，在频率 8000Hz 的极坐标图明显更具有方向性。

[0058] 基于这种相关性，图 5A 所示的基于音频的选择处理 200 将由麦克风 (160) 捕捉的人声音的音频频率与参与者头部的方向信息相关联。为实现上述操作，之前所提到的比较处理 220 使用频率谱并在每一个麦克风 (160) 处比较高频率声学能量 222 与低频率声学能量 224 的比率。通常，较高频率范围可以大于约 2500Hz 的阈值，而较低频率范围可以低于该阈值。

[0059] 然而实验中发现，对于与每个摄像机 (150) 相关联的麦克风 (160)，通过采用约 4000Hz 到 7000Hz 的较高频率能量 222 除以约 500Hz 到 1000Hz 的较低频率能量 224 所得到的能量比率，具有最大比率的摄像机 (150) 很可能是正在说话的参与者所面向的摄像机 (150)。发现这些频率范围特别适合于视频会议环境中的语音。然而，特定的频率范围可以根据实施而改变，并可以基于会议区域有多大、出席的参与者有多少、使用多少个麦克风、以及其他考虑因素而改变。

[0060] 因此，比较处理 220 获取北侧摄像机麦克风的约 4000Hz 到 7000Hz 的较高频率的第一声学能量 222，并将其除以相同麦克风 (160) 的约 500Hz 到 1000Hz 的较低频率的第二声学能量 224。在结果 230 中，存储所得到的比率值 R1 并将之与对应的摄像机位置（即，北侧摄像机）相关联。对每个摄像机 (150) 及其相关联的麦克风 (160) 重复上述操作。最后，处理 200 在结果 230 中选择 240 具有最高比率 R 的摄像机 (150)，因为该摄像机 (150) 最有可能是当前面向正在讲话的参与者的摄像机。

[0061] 返回图 2，现在已经选择了摄像机 150，控制模块 140 引导系统 100 的操作，以使得来自该选中的摄像机 150 的所捕捉视频成为到远端的输出的一部分。当然，控制模块 140 可以在切换之前提供延迟，可以验证结果，以及执行其他的通常功能，以避免摄像机画面之间的伪切换或频繁切换。同样，之前描述的自动摄像机控制处理可以控制所选择的摄像机

(150)的平移、倾斜和缩放，以最佳地拍摄到参与者。总之，意图很好地控制摄像机画面之间的切换时机并使切换自然，以及避免画面间的频繁改变和类似的问题。

[0062] 2. 基于视频的选择处理

[0063] 作为上文所述的声学配置的替换或与之结合，控制模块 140 可以使用基于视频的选择处理 300，其基于各个摄像机所捕捉的视频。现在转向图 6，其示出了该基于视频的选择处理 300 的详细描述。

[0064] 在处理 300 中，分别来自各个摄像机 (150)(如 N,S,E,W) 的视频信号 305 作为到视频接口 310 的输入。接着，接口 310 对输入的视频信号 305 执行用于面部识别 312 的视频处理。该视频处理可以使用运动检测、皮肤色调检测、面部检测以及其他算法来处理视频信号 305，以指示所捕捉的视频是否包括人脸。

[0065] 例如，视频处理可以使用面部识别技术在摄像机画面中检测和定位面部。为实现这一点，所述视频处理可以搜寻可能包括人皮肤色调的区域，然后从这些区域中寻找指示面部在画面中的位置的区域，从而找到面部。与皮肤色调和面部检测（以及音频定位）的相关的详细描述被公开在名称为“Locating an Audio Source”的美国专利 No.6593956 中，其通过引用结合在本文中。

[0066] 另外，用于面部识别 312 的视频处理可以使用面部检测或是面部搜索算法，通过确定当前摄像机画面确实拍摄到了具有人脸的画面，来增加追踪准确度。该面部检测可以检测正面面部画面，也可能能够检测面部的左侧面和右侧面画面。视频面部检测的一个可用算法是 OpenCV（开源计算机视觉），其是用于实时计算机视觉的编程函数库。也可以使用本领域可用的许多其他算法，尽管对于面部检测优选 OpenCV。

[0067] 使用该面部检测处理，说话的参与者所面对的摄像机 (150) 在正面面部检测器中的置信度得分将会是最高的。面部检测可以非常有用，因为当确定了正面面部的捕捉画面时，其可以具有最佳效果。面部的其他方向可以以较低的可靠度被检测，因此面部检测可以用于被处理的所捕捉摄像机画面的阈值测量。

[0068] 进一步，用于面部识别 312 的视频处理，可以使用本领域已知的技术在视频信号 305 中检测皮肤的色调，并可以对摄像机 (150) 捕捉的视频执行运动检测算法以检查当前画面中的运动。作为这些技术的一个示例，可以通过在摄像机 (150) 捕捉的视频帧中识别出具有皮肤色调频色的区域，而检测出面部的图像。简言之，视频处理可以获得一帧或是一帧的一部分中的色度值的平均值。如果该平均值在与皮肤色调相关联的范围内，则认为该帧或其一部分具有皮肤色调特征。摄像机 (150) 捕捉的该视频也可以表示移动的面部，这例如是通过将视频的该帧与视频的前一帧比较以确定由于视频中的运动所致的改变（大概是由参与者的移动而产生的）来确定的。

[0069] 使用用于面部识别 312 的这些视频处理技术中的一项或多项，处理 300 为相关联的摄像机 (150) 计算面部加权 314。面部加权 314 可以以任意方法计算出来，这取决于处理结果是如何获得的。因此，阈值、置信度水平、比较、平均、或是类似的处理可以用来确定加权 314。

[0070] 处理 300 接着使用这些面部加权 314 来执行比较 330，以在比较的结果 335 中找到具有拍摄到当前说话者的面部的最大可能性的摄像机 (150)。最后，摄像机选择 340 可以至少部分基于这些比较结果而做出。

[0071] 尽管基于视频的选择处理 300 可以完全基于视频，但是多于一个摄像机 (150) 可能拍摄到参与者的面部，并且单独的面部识别 312 可能无法确定此人是否当前正在讲话。因此，基于音频的处理 320 可以被包括在图 6 所示的处理 300 中，作为基于视频的选择的补充。这样，可以将音频 322 输入到音频接口 324 中，并且根据所公开的技术的处理可以确定输入音频的音频能量 326，其然后可用在比较 330 中。基于音频的处理 320 可以使用本文描述的任何各种技术来确定会议中哪个参与者正在说话以及参与者面朝哪个方向。所述技术可以包括传统的波束形成技术或图 5A 所示的频率能量比率比较 (220)。

[0072] 作为一个示例，在麦克风箱 (128) 处（诸如在桌上）捕捉的能量可以指示哪个人正在说话，并且该信息可以与视频处理一起结合到比较 330 中以改进摄像机选择 340 的结果 335。因此，用于面部识别 312 的视频处理可以在图 2 中的北侧摄像机 150N 和东侧摄像机 150E 有可能拍摄到在西侧椅子和南侧椅子上的参与者时，给它们高的面部加权。在这种情况下，使用来自麦克风箱 128 的音频进行的音频处理可以指示在南侧椅子上的参与者当前正在讲话。因此，系统 100 可以选择北侧摄像机 150N。

[0073] 可替换的是，如果使用来自麦克风箱 128 的音频进行的音频处理指示在南侧椅子上的参与者当前正在讲话，摄像机 (150) 之间的面部识别加权 314 可能指示在南侧椅子上的参与者正在面对着东侧摄像机 (150E) 而不是其他摄像机。这是有可能的，因为来自东侧摄像机 (150E) 的视频的面部加权可能表明，参与者正指向该摄像机 (150E) 而不是其他摄像机这种情况具有更高可能性。

[0074] 如之前所提到的，在麦克风箱 (128) 处所捕捉的能量和波束形成可能并不能可靠地指示当前讲话者面对着哪个方向。因此，图 6 所示的基于音频的处理 320 可以使用这里提到的与摄像机相关联的麦克风 (160)，并可以根据图 5A-5B 所讨论的基于音频的处理，着眼于高频与低频的比率。继而，图 6 所示的音频处理 320 可随后增加到比较 330 中的视频处理（举例来说，使用如本文所讨论的面部检测或类似的方法）中，以提供用于找到讲话的人当前面对着哪个摄像机的另一层面。

[0075] 最后，在与摄像机相关联的麦克风 (160) 处所捕捉的能量可以指示哪个人正在说话，并且该信息可以与视频处理一起结合到比较 330 中以改进摄像机选择 340 的结果 335。例如，使用来自与摄像机相关联的麦克风 (160) 的音频进行的音频处理可能指示在图 2 的南侧椅子上的参与者当前正面对着北侧摄像机 (150N)、或东侧摄像机 (150E)、或是北侧和东侧摄像机之间的某个点，因为基于音频的确定不是很精确。在这种情况下，用于面部识别 312 的视频处理可能赋予北侧摄像机 (150N) 比东侧摄像机 (150E) 更高的面部加权 314。如果这样的话，则系统 100 选择北侧摄像机 (150N) 来输出在南侧椅子上的当前讲话者的视频，因为该参与者最可能面对着北侧摄像机 (150N)。

[0076] D. 具有多个摄像机和麦克风的视频会议系统的另外的布置

[0077] 在上面参照图 2 的讨论中，摄像机 150 和麦克风 160 的布置是相当有序和对称的。如前文中所述，情况并非必须如此，而是可以在会议环境中任意布置视频会议系统 100。不过，在这些任意的布置中，本文公开的摄像机选择技术可能特别有益。

[0078] 例如，图 7 示出了使摄像机 150 和麦克风 160 相对于参与者和麦克风箱 128 布置的的系统 100 的另一种示意性布置。该示例布置将帮助更详细地描述所公开的摄像机选择技术，因为各个摄像机 150 并不是绕着房间、桌子和参与者对称布置的。因此，前方摄像机

150(R, C, L) 中的一个或甚至侧方摄像机 150S 可能最佳地捕捉和拍摄到正在说话的参与者，即使另外一个摄像机 150 可能位于该参与者的前面、接近该参与者、或更一般而言面向该参与者。然而，使用这里所公开的基于音频和 / 或视频的处理，视频会议系统 100 和控制模块 140 在上述布置和其他类似的布置中可以提供更好的摄像机选择。

[0079] 举一个简单的例子，参与者 P3 可能正在说话。使用麦克风箱 128 的常规波束形成技术可能通过指示说话的参与者 P3 相对于箱 128 的的方向来表明情况是这样的。（例如，使用波束形成的自动摄像机追踪技术的相关细节可以在美国专利 No. 6980485 中找到，其通过引用被全文结合于此。）这样的摄像机追踪技术可以被使用，以便利用来自侧方摄像机 150S 的画面来拍摄到参与者 P3。然而，参与者 P3 可能正在转过头说话，这样将使得由可能更合适于捕捉该参与者的面部的另一个摄像机进行拍摄。

[0080] 使用相关联的麦克风 160，控制模块 140 可以基于这里所公开的音频处理技术进行摄像机选择，以确定说话的参与者 P3 当前面向哪个摄像机 150。在本例中，该特定的摄像机可以是摄像机 150L，其随后可用于输出用于视频会议的视频。

[0081] 例如，图 8A 示出了来自摄像机 150S/150L 的画面。来自摄像机 150S 的第一画面示出了在说话时稍微转头的参与者 P3，而来自摄像机 150L 的第二画面示出了在说话时的参与者面部的更正面的画面。所公开的基于音频的处理技术可以帮助指示参与者 P3 很可能正在面向着左摄像机 150L，因此系统 100 可以选择其画面来输出视频会议的视频。一旦选择，可以使用已知技术使得摄像机 150L 的画面进一步被引导以指向参与者 P3 并拍摄到参与者 P3。当然，如本文所讨论的，也可以使用基于视频的选择技术。

[0082] 在另一个简短的例子中，参与者 P1 可能正在说话。使用麦克风箱 128 的常规波束形成技术可以通过指示说话的参与者 P1 相对于箱 128 的方向来表明情况是这样的。这可以用于利用来自前方摄像机 150(R, C, L) 中的任一个的画面来拍摄到参与者 P1。然而，参与者 P1 可能正在转过头说话，这样将使得使用某个特定摄像机来拍摄可以更合适地捕捉参与者的面部。

[0083] 使用相关联的麦克风 160，控制模块 140 因此可基于这里所公开的音频处理技术来进行摄像机选择，以确定说话的参与者 P1 当前正在面对着哪个摄像机 150。在本例中，该特定摄像机可以是摄像机 150L，其随后可用于输出用于视频会议的视频。

[0084] 例如，图 8B 示出了来自摄像机 150L/150C/150R 的画面。来自左摄像机 150L 的第一画面显示了参与者 P1 的头部面向摄像机 150L，来自中间摄像机 150C 的第二画面显示了参与者 P1 的面部稍微偏转，来自右摄像机 150R 的第三画面显示了参与者 P1 的面部偏转得更多。基于音频的处理技术可以帮助指示参与者 P1 很可能正在面向着摄像机 150L，从而系统 100 可以选择它来输出视频会议的视频。一旦选择，可以使用已技术，使得摄像机 150L 的画面进一步被引导以指向并拍摄到参与者 P1。

[0085] 然而，如果参与者 P1 的头部的方向性不能明确地被分辨，以明确选择这些摄像机 150(R, C, L) 中的一个，则可以使用这里所公开的基于视频的处理技术，来进一步改进该确定。特别是，这里所公开的面部识别技术可以指示一个摄像机（即，150L）相比于其他摄像机（即，150R, C），更可能捕捉参与者面部的正面画面，而其他摄像机捕捉的是参与者面部的部分侧面。最后，该附加的确定可以更好地指示要使用哪个摄像机 150L 来输出说话的参与者 P1 的被捕捉视频。

[0086] E. 具有音频 / 视频处理能力的摄像机模块

[0087] 在之前的布置中，视频会议系统 100 具有多个摄像机和麦克风，并使用了这里所公开的基于音频和视频的处理技术来选择哪个摄像机用于输出。尽管可以在针对多个摄像机和麦克风的宽尺度上使用，所公开的技术可以用单独的摄像机模块以更离散的方式使用，这些模块可以是较大视频会议系统的一部分。例如，图 9A 示出了具有可控制的摄像机 450 和麦克风阵列 460 的摄像机模块 450。模块 450 具有摄像机操控部件 454 和音频 / 视频处理部件 456，它们可集成到模块 450 中或是与之分离。这样的部件 454/456 可以由较大视频会议系统的处理单元来操纵，或可以由模块 450 在本地操纵。

[0088] 在操作中，摄像机操控部件 454 可以使用从麦克风阵列 460 输入的音频，来操控摄像机 452 朝向正在说话的参与者。其可以使用旨在利用麦克风阵列进行波束操控的很多技术，例如在并入的美国专利 No. 6593956 中公开的。除这样的技术之外，模块 450 还使用这里所公开的音频 - 视频处理技术来改进对摄像机 452 的操控以朝向说话的参与者。

[0089] 特别是，该基于 A/V 的处理部件 456 使用上文中公开的基于音频的处理来确定参与者头部的方向性，其中该基于音频的处理测量不同频率范围的能量的比率。例如，可以对来自模块中的麦克风 460 之一的音频进行滤波和采样。可以确定高频率范围（如，4000Hz 到 7000Hz）的第一能量水平，并将其与正在说话的参与者头部的方向性相关。

[0090] 可替换的是，来自麦克风 460 中的多个麦克风的音频可被处理以得到第一能量水平，以及得到较低频率范围（如，500Hz 到 1000Hz）的第二能量水平。第一能量水平除以第二能量水平的比率最大的麦克风 460 则可以表示正在说话的参与者的头部最可能指向的麦克风 460。

[0091] 最后，不同于阵列中固定的麦克风 460，摄像机模块 450 可以包括随着摄像机 452 移动的麦克风 462。当摄像机操控部件 454 操控摄像机 452 朝向说话的参与者时，可以滤波和采样来自该移动麦克风 462 的音频，以确定高频和低频范围的第一和第二能量水平的比率。随着摄像机 452 和麦克风 462 的移动，处理部件 456 可以确定在下降之前，该比率何时达到最大或最高水平。在这一时刻的摄像机 452 位置将表明，摄像机 452 与正在说话的参与者头部的方向对齐得最好。

[0092] 为帮助图解说明该精细的操控，图 9B 示出了在摄像机操控和处理期间，图 9A 的摄像机模块 450 捕捉的画面。如画面 470 所示出的，摄像机模块 450 可能正在捕捉正在说话的参与者的画面。然而，参与者的头部可能从摄像机的视角方向转开或稍微偏离。使用上文所描述的基于音频的处理，摄像机模块 450 可以进一步操控摄像机 452，并确定参与者头部的方向性，以使摄像机 452 的视角更好地与参与者头部对准。这可得到如图 9B 所示的更精确的画面 475。

[0093] 除基于音频的技术之外，也可以应用这里公开的基于视频的技术，以便在参与者讲话时，更好地使摄像机 452 指向参与者的面部并且将其“放大”。举例来说，这可以使用在引入的美国专利 No. 6593956 中公开的多种技术，也可以使用前面公开的面部识别技术。

[0094] F. 总结

[0095] 如前文所见，这里所公开的基于音频和基于视频的摄像机选择技术，可以单独使用或彼此结合使用，以确定使用哪个摄像机来来输出说话的参与者的视频，从而最佳地捕捉参与者的面部。受益于本公开，本领域技术人员可以明白在摄像机 150、麦克风 160/128、

参与者、座位布置等的各自可能的布置中可以如何使用这些技术。

[0096] 不脱离于权利要求的范围，可以对所图解说明的操作方法的细节进行各种改变。例如，说明性的流程图步骤或是处理步骤可以以与本文公开不同的次序来执行相同的步骤。可替换的是，一些实施例可以组合在这里作为分开的步骤描述的动作。类似地，基于实施该方法的特定操作环境，可以省略所描述的步骤中的一个或多个。

[0097] 另外，根据流程图或处理步骤的动作可以由运行指令的可编程控制设备来执行，这些指令被组织为在非暂态可编程存储设备上的一个或多个程序模块。可编程控制设备可以是单个程序处理器、专用处理器（如，数字信号处理器，“DSP”）、由通信链路耦接的多个处理器、或定制设计的状态机。定制设计的状态机可以体现在硬件设备中，比如集成电路，包括但不限于专用集成电路（“ASIC”）或现场可编程门阵列（“FPGA”）。适于有形地表达程序指令的非暂态可编程存储设备（有时称为计算机可读介质）包括但不限于：磁盘（固定的、软盘和可移动的）和磁带；光媒体，如 CD-ROM 和数字视频盘 (DVD)；以及半导体存储设备，如电可编程只读存储器（“EPROM”）、电可擦除可编程只读存储器（“EEPROM”）、可编程门阵列、和闪存设备。

[0098] 之前的优选和其他实施例的描述并不意图限制或约束申请人所构想的发明概念的范围或应用。作为公开这里所包含的发明概念的交换，申请人希望获得所附权利要求所提供的所有专利权。因此，意图的是，所附权利要求包括位于权利要求或其等同的最大程度范围内的所有的修改和变动。

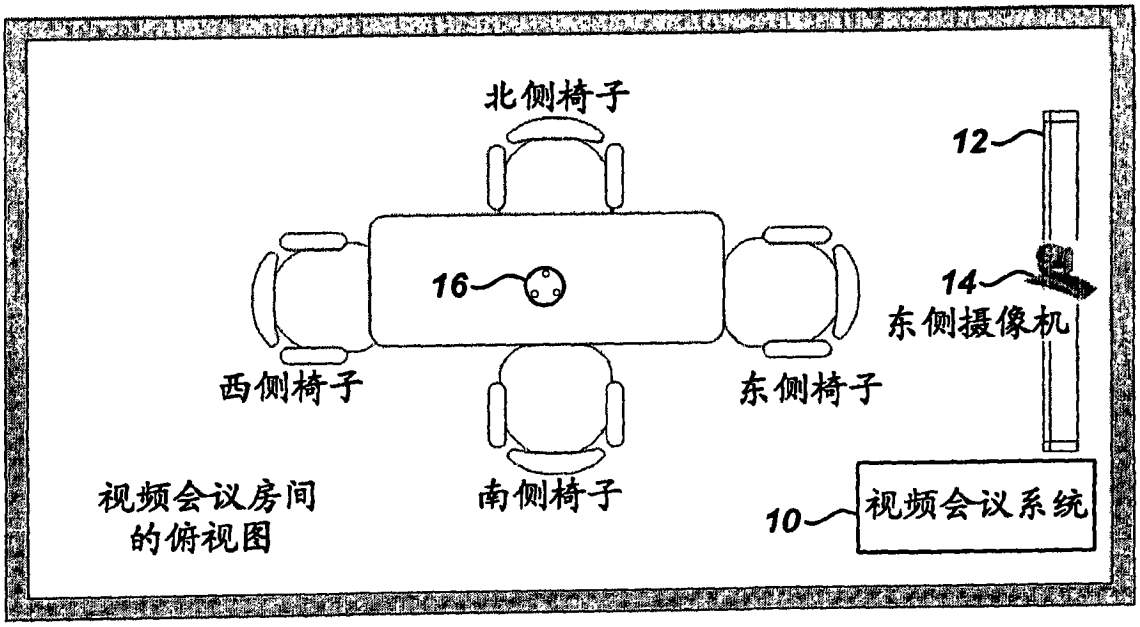


图 1A（现有技术）

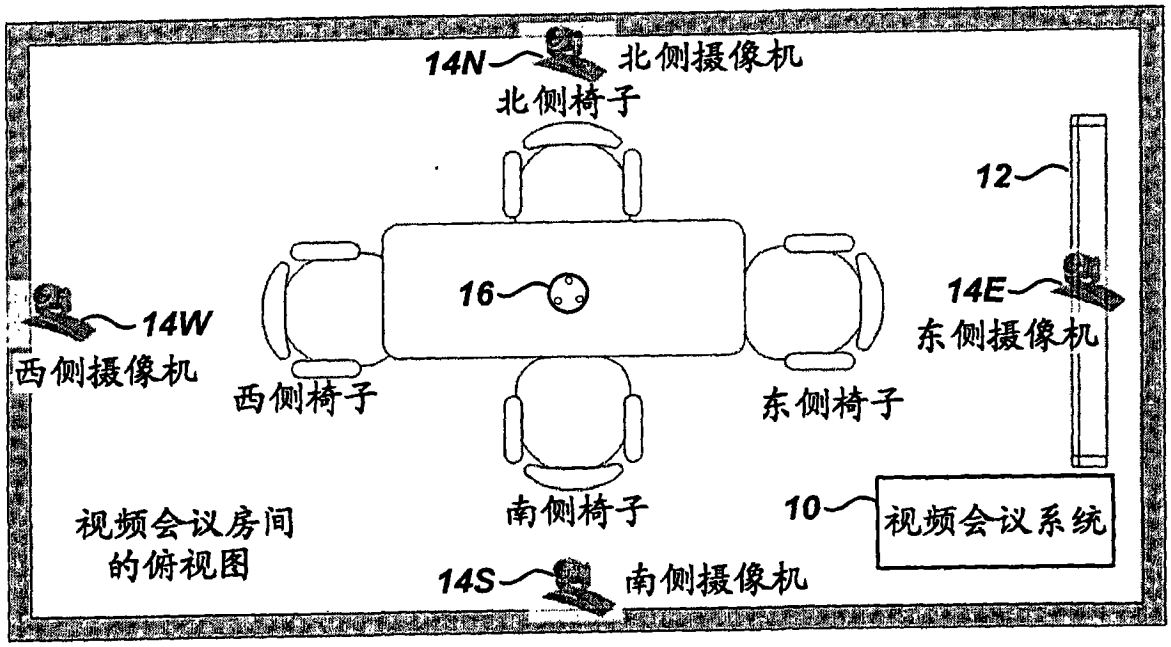


图 1B（现有技术）

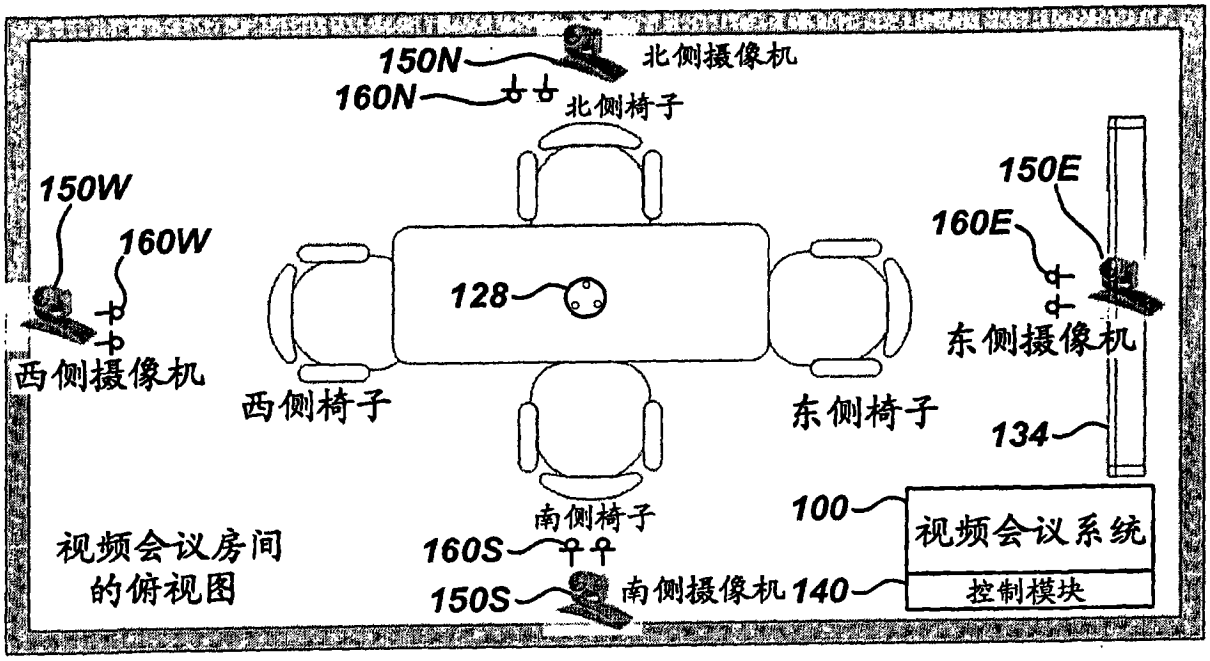

150N
北侧摄像机
160N
北侧椅子
150W
160W
西侧摄像机
128
西侧椅子
150E
160E
东侧摄像机
东侧椅子
134
南侧椅子
视频会议房间
的俯视图
160S
150S
南侧摄像机
100
视频会议系统
140
控制模块

图 2

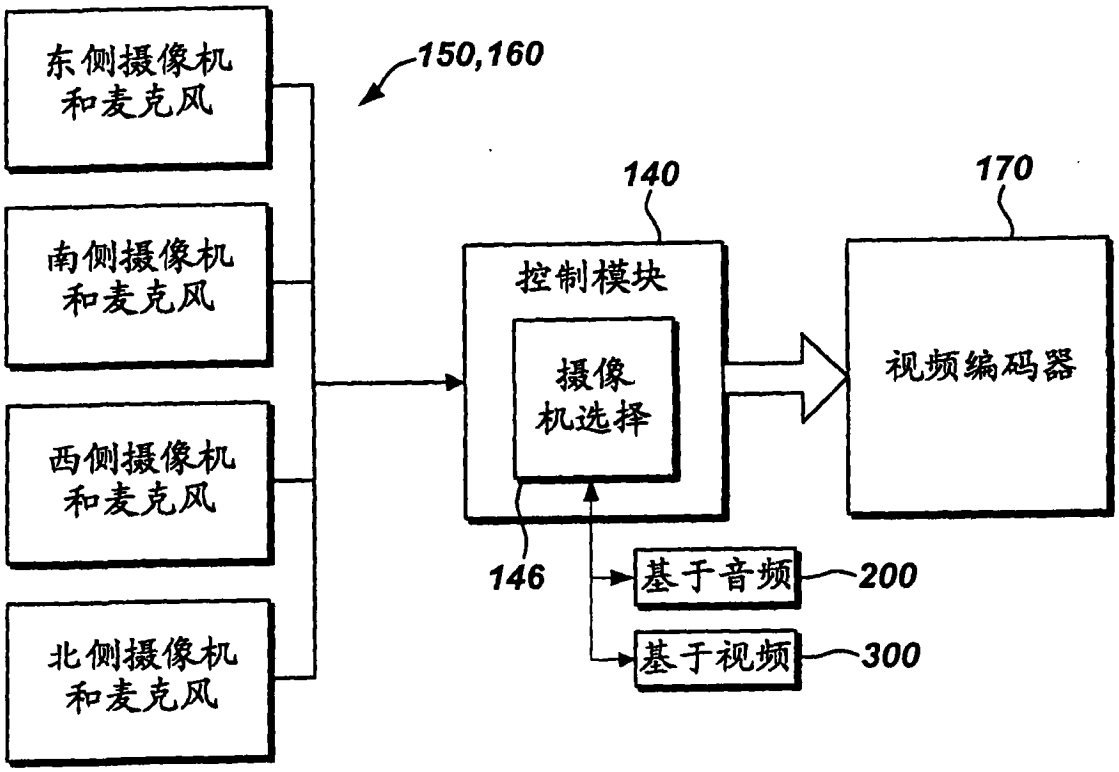

东侧摄像机
和麦克风
南侧摄像机
和麦克风
西侧摄像机
和麦克风
北侧摄像机
和麦克风
150,160
140
控制模块
摄像
机选择
170
视频编码器
146
基于音频
200
基于视频
300

图 4

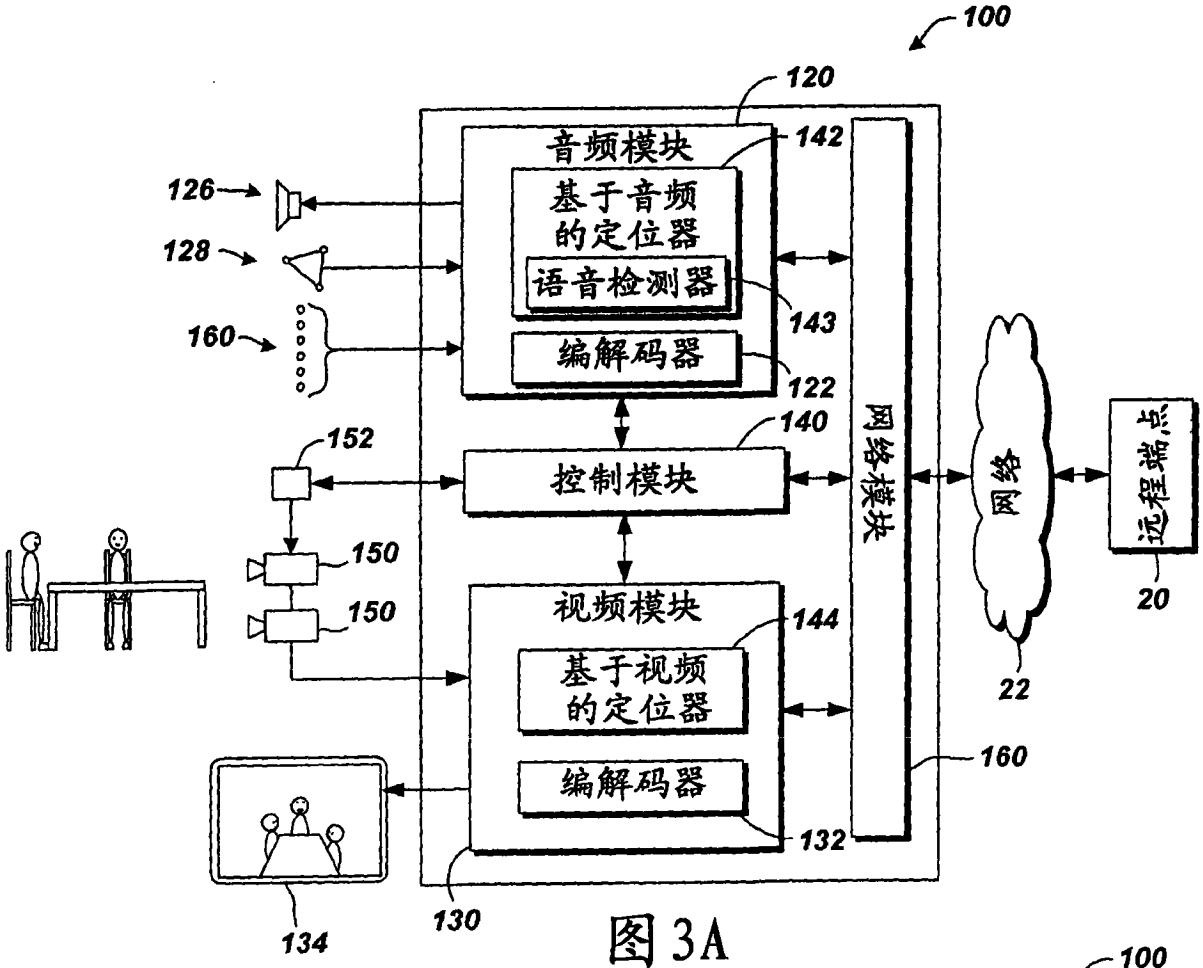

100
120
音频模块
142
基于音频的定位器
语音检测器
143
编解码器
122
126
128
160
152
140
控制模块
网络模块
网络
远程端点
20
22
150
150
视频模块
144
基于视频的定位器
编解码器
132
160
130
134

图 3A

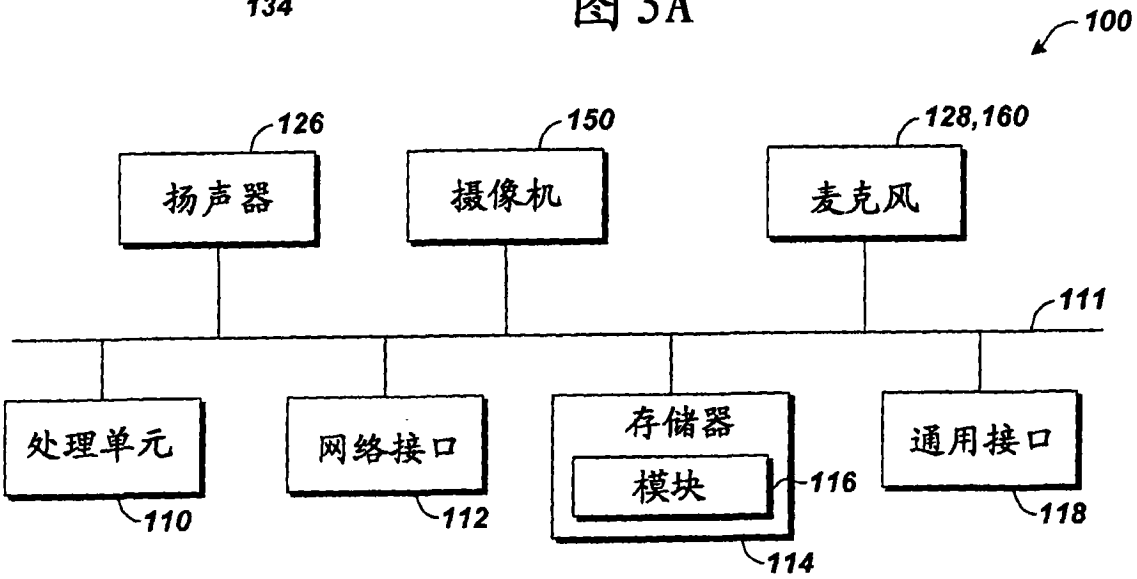

100
126
扬声器
150
摄像机
128,160
麦克风
111
处理单元
110
网络接口
112
存储器
模块
116
114
通用接口
118

图 3B

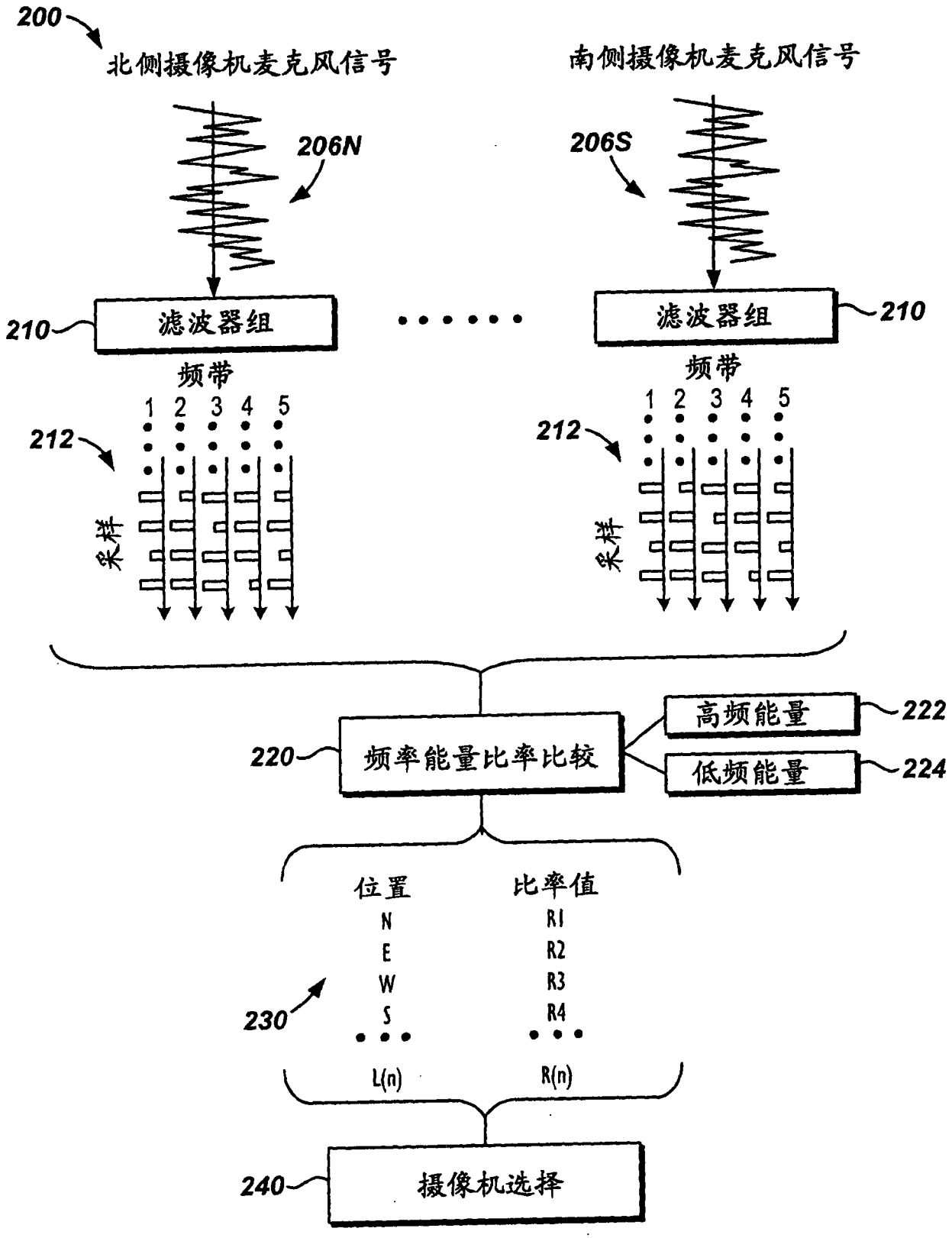

200
北侧摄像机麦克风信号
南侧摄像机麦克风信号
206N
206S
210
滤波器组
210
滤波器组
频带
1 2 3 4 5
212
采样
频带
1 2 3 4 5
212
采样
220
频率能量比率比较
高频能量
222
低频能量
224
位置
比率值
N
E
W
S
R1
R2
R3
R4
230
L(n)
R(n)
240
摄像机选择

FIG. 5A

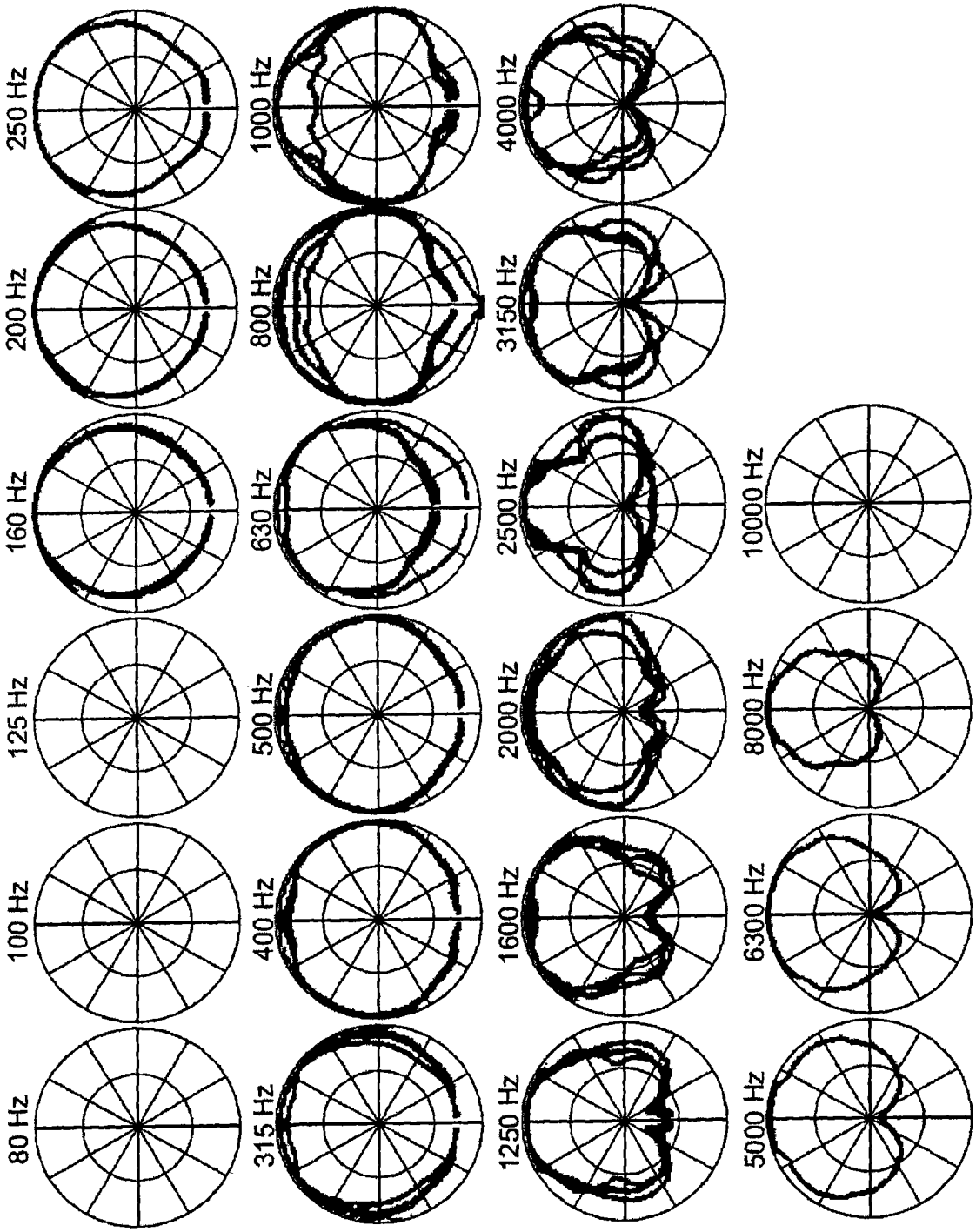

250 Hz
1000 Hz
4000 Hz
200 Hz
800 Hz
3150 Hz
160 Hz
630 Hz
2500 Hz
10000 Hz
125 Hz
500 Hz
2000 Hz
8000 Hz
100 Hz
400 Hz
1600 Hz
6300 Hz
80 Hz
315 Hz
1250 Hz
5000 Hz

图 5B

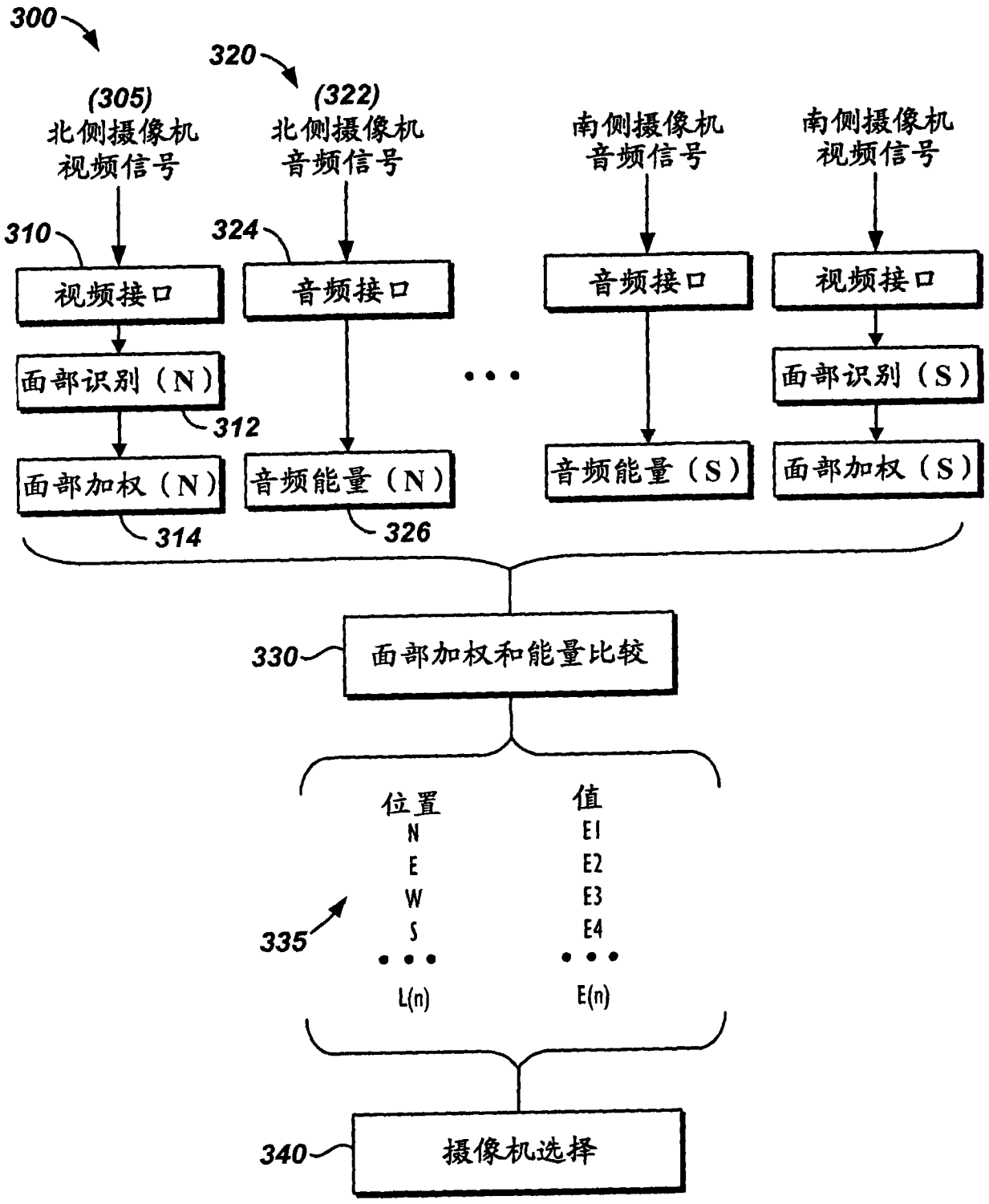

300
320
(305)
北侧摄像机
视频信号
(322)
北侧摄像机
音频信号
南侧摄像机
音频信号
南侧摄像机
视频信号
310
324
视频接口
音频接口
音频接口
视频接口
面部识别（N）
面部识别（S）
312
面部加权（N）
音频能量（N）
音频能量（S）
面部加权（S）
314
326
330
面部加权和能量比较
位置
值
N
E
W
S
E1
E2
E3
E4
335
L(n)
E(n)
340
摄像机选择

图 6

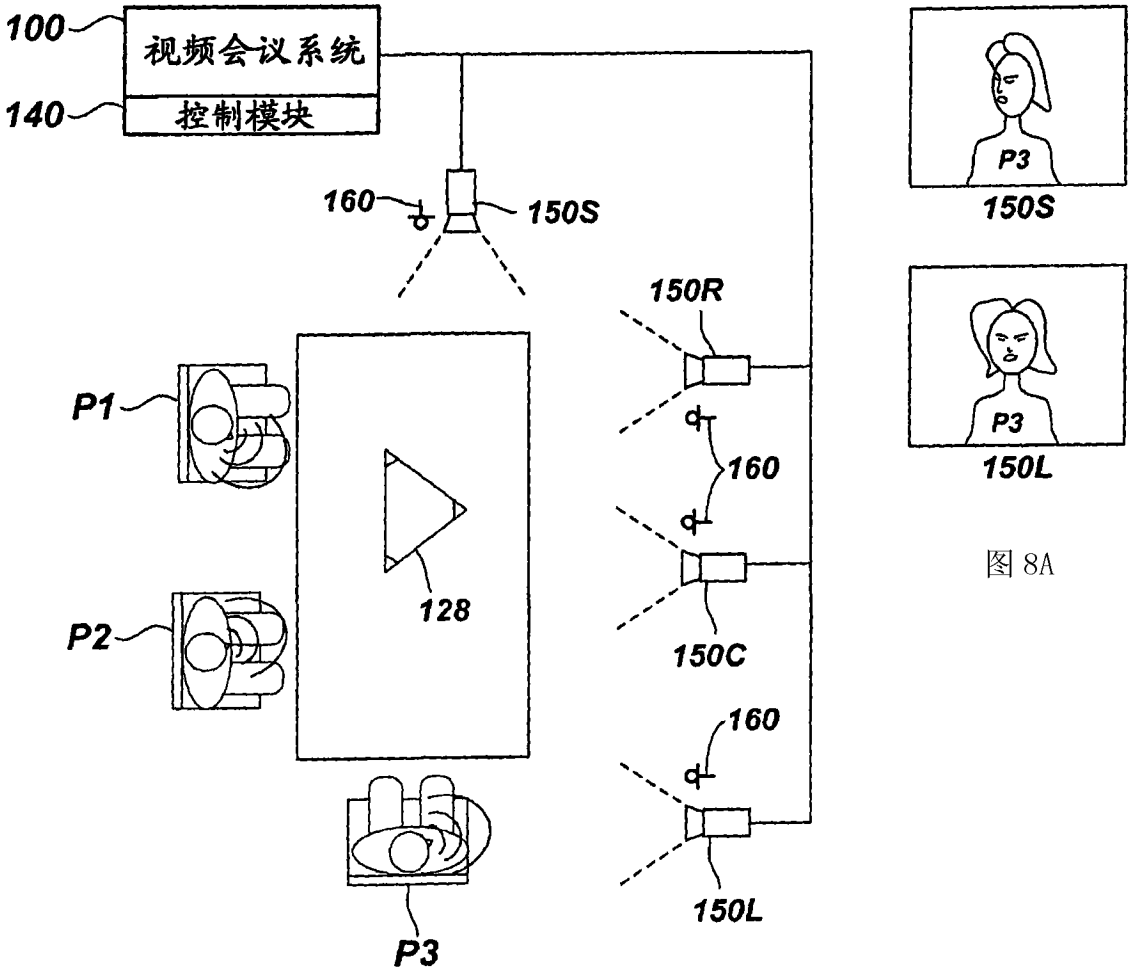

100
视频会议系统
140
控制模块
160
150S
150R
160
150C
160
150L
P1
P2
128
P3
P3
150S
P3
150L

图 8A

图 7

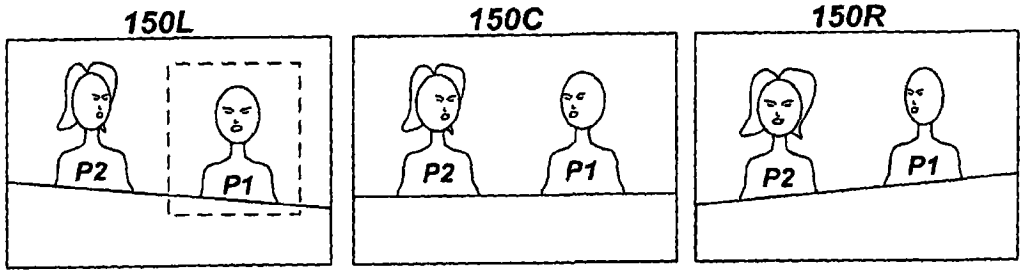

150L
150C
150R
P2
P1
P2
P1
P2
P1

图 8B

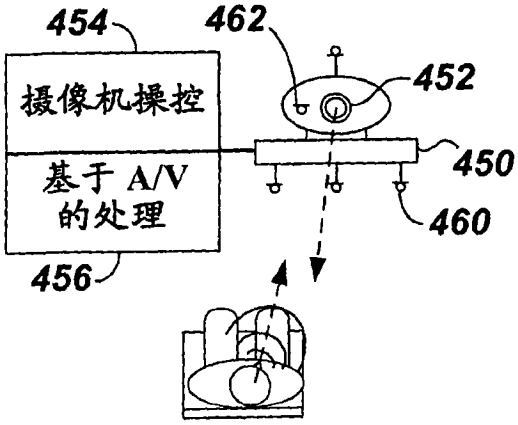


图 9A

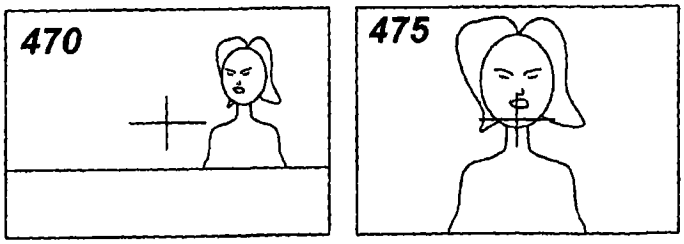


图 9B