

语音通话的控制方法、装置、计算机可读介质及电子设备

本申请要求于 2021 年 1 月 21 日提交中国专利局、申请号为 202110083388.6, 发明名称为“语音通话的控制方法、装置、计算机可读介质及电子设备”的中国专利申请的优先权, 其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及计算机技术领域, 具体而言, 涉及一种语音通话的控制方法、装置、计算机可读介质及电子设备。

背景技术

在很多语音通话的场景中, 需要判别说话人的数量或者音色等, 以作之后的语音控制。相关技术中通过基于大量带标注的语音片段来训练说话人场景检测系统, 其中每个片段的标签为说话人的个数, 在进行测试时给定一个语音片断, 系统预测当前的说话人个数。这样的处理方式将为语音检测造成较长的延时, 尤其是在实时的通信场景中, 将大大降低语音识别的效率, 影响实时的语音控制效果。

技术内容

15 本申请的实施例提供了一种语音通话的控制方法、装置、计算机可读介质及电子设备, 进而至少在一定程度上可以保证语音人数的检测精度, 同时提高了语音人数的识别效率以及语音通话的控制效率。

本申请的其他特性和优点将通过下面的详细描述变得显然, 或部分地通过本申请的实践而习得。

20 根据本申请实施例的一个方面, 提供了一种语音通话的控制方法, 包括: 获取混合的通话语音, 所述混合通话语音包含至少一个分支语音; 对所述通话语音进行频域变换, 确定所述通话语音在频域中各个频点对应的能量信息; 基于神经网络对每个频点上的所述能量信息进行分离处理, 确定每个频点上对应各分支语音在该频点的全部能量中的能量占比; 基于每个频点上各分支语音的所述能量占比确定所述通话语音中所包含的分支
25 语音的数量; 基于所述分支语音的数量, 通过设定通话语音控制方式对所述语音通话进行控制。

根据本申请实施例的一个方面, 提供了一种语音通话的控制装置, 包括: 获取单元,

用于获取混合的通话语音，所述混合通话语音包含至少一个分支语音；变换单元，用于对所述通话语音进行频域变换，确定所述通话语音在频域中各个频点对应的能量信息；分离单元，用于基于神经网络对每个频点上的所述能量信息进行分离处理，确定每个频点上对应各分支语音在该频点的全部能量中的能量占比；数量单元，用于基于每个频点上各分支语音的所述能量占比确定所述通话语音中所包含的分支语音的数量；控制单元，用于基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制。

根据本申请实施例的一个方面，提供了一种计算机可读介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如本申请实施例中所述的语音通话的控制方法。

10 根据本申请实施例的一个方面，提供了一种电子设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现如本申请实施例中所述的语音通话的控制方法。

根据本申请实施例的一个方面，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。
15 计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行本申请各种实现方式中提供的语音通话的控制方法。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图简要说明

20 图 1 示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图；
图 2 示意性示出了根据本申请的一些实施例的会议系统的示意图；
图 3 示意性示出了根据本申请的一些实施例的语音通话的控制方法的流程图；
图 4 示意性示出了根据本申请的一些实施例的语音分离流程的示意图；
图 5 示意性示出了根据本申请的一些实施例的语音提取的示意图；
25 图 6 示意性示出了根据本申请的一些实施例的会议语音提取的示意图；
图 7 示意性示出了根据本申请的一些实施例的回声消除的应用场景图；
图 8 示意性示出了根据本申请的一些实施例的回声消除的示意图；
图 9 示意性示出了根据本申请的一些实施例的提取能量信息的流程图；
图 10 示意性示出了根据本申请的一些实施例的提取能量信息的示意图；
30 图 11 示意性示出了根据本申请的一些实施例的训练分离模型的流程图；

图 12 示意性示出了根据本申请的一些实施例的设置主讲人的界面图；

图 13 示意性示出了根据本申请的一些实施例的分配通信资源的界面图；

图 14 示意性示出了根据本申请的一些实施例的语音通话的控制装置的框图；

图 15 示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

5 具体实施方式

云计算(cloud computing)是一种计算模式,它将计算任务分布在大量计算机构成的资源池上,使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的,并且可以随时获取,按需使用,随时扩展,按使用付费。作为云计算的基础能力提供商,会建立云计算资源池(简称云平台,一般称为 IaaS(Infrastructure as a Service,基础设施即服务)平台,在资源池中部署多种类型的虚拟资源,供外部客户选择使用。云计算资源池中主要包括:计算设备(为虚拟化机器,包含操作系统)、存储设备、网络设备。按照逻辑功能划分,在 IaaS (Infrastructure as a Service,基础设施即服务)层上可以部署 PaaS(Platform as a Service,平台即服务)层, PaaS 层之上再部署 SaaS (Software as a Service,软件即服务)层,也可以直接将 SaaS 部署在 IaaS 上。PaaS 为软件运行的平台,如数据库、web 容器等。SaaS 为各式各样的业务软件,如 web 门户网站、短信群发器等。一般来说, SaaS 和 PaaS 相对于 IaaS 是上层。

云呼叫中心((Cloud Call Center))是基于云计算技术而搭建的呼叫中心系统,企业无需购买任何软、硬件系统,只需具备人员、场地等基本条件,就可以快速拥有属于自己的呼叫中心,软硬件平台、通信资源、日常维护与服务由服务器商提供。具有建设周期短、投入少、风险低、部署灵活、系统容量伸缩性强、运营维护成本低等众多特点;无论是电话营销中心、客户服务中心,企业只需按需租用服务,便可建立一套功能全面、稳定、可靠、座席可分布全国各地,全国呼叫接入的呼叫中心系统。

本申请实施例中可以通过云呼叫中心的方式呼叫中心系统,同时在该系统中嵌入语音通话的控制方法,来实现呼叫过程中的语音控制,进而使得云呼叫中心更加智能化,提高云呼叫中心的可靠性和安全性。

云会议是基于云计算技术的一种高效、便捷、低成本的会议形式。使用者只需要通过互联网界面,进行简单易用的操作,便可快速高效地与全球各地团队及客户同步分享语音、数据文件及视频,而会议中数据的传输、处理等复杂技术由云会议服务商帮助使

用者进行操作。目前国内云会议主要集中在以 SaaS (Software as a Service, 软件即服务) 模式为主体的服务内容, 包括电话、网络、视频等服务形式, 基于云计算的视频会议就叫云会议。在云会议时代, 数据的传输、处理、存储全部由视频会议厂家的计算机资源处理, 用户完全无需再购置昂贵的硬件和安装繁琐的软件, 只需打开浏览器, 登录相应界面, 就能进行高效的远程会议。云会议系统支持多服务器动态集群部署, 并提供多台高性能服务器, 大大提升了会议稳定性、安全性、可用性。近年来, 视频会议因能大幅提高沟通效率, 持续降低沟通成本, 带来内部管理水平升级, 而获得众多用户欢迎, 已广泛应用在政府、军队、交通、运输、金融、运营商、教育、企业等各个领域。毫无疑问, 视频会议运用云计算以后, 在方便性、快捷性、易用性上具有更强的吸引力, 必将激发视频会议应用新高潮的到来。

在云会议的应用场景中, 本申请实施例中可以将基于语音通话的控制方法应用与云会议中, 使得在云会议的过程中的语音通话更加清晰, 语音通信过程更加智能化, 进而提高会议的效率。

云社交 (Cloud Social) 是一种物联网、云计算和移动互联网交互应用的虚拟社交应用模式, 以建立著名的“资源分享关系图谱”为目的, 进而开展网络社交, 云社交的主要特征, 就是把大量的社会资源统一整合和评测, 构成一个资源有效池向用户按需提供服务。参与分享的用户越多, 能够创造的利用价值就越大。

人工智能 (Artificial Intelligence, AI) 是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能, 感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说, 人工智能是计算机科学的一个综合技术, 它企图了解智能的实质, 并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法, 使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科, 涉及领域广泛, 既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

语音技术 (Speech Technology) 的关键技术有自动语音识别技术 (ASR) 和语音合成技术 (TTS) 以及声纹识别技术。让计算机能听、能看、能说、能感觉, 是未来人机交互的发展方向, 其中语音成为未来最被看好的人机交互方式之一。机器学习 (Machine

Learning, ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习式教学习等技术。

随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。

相关技术中通过基于大量带标注的语音片段来训练说话人场景检测系统,其中每个片段的标签为说话人的个数,在进行测试时给定一个语音片断,系统预测当前的说话人个数。但是该方案在检测过程中需要较多的上下文信息才能判定当前的说话人个数,比如需要采集较长时间的语音片段作为数据基础,才能确定说话人个数,而这样的处理方式将为语音检测造成较长的延时,尤其是在实时的通信场景中,将大大降低语音识别的效率,影响实时的语音控制效果。

本申请实施例提供的方案涉及人工智能的语音技术和机器学习等技术,通过上述技术可以使得本申请实施例中的语音通话控制方法更加的精确,具体通过如下实施例进行说明:

图1示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图。

如图1所示,系统架构可以包括终端设备(如图1中所示智能手机101、平板电脑102和便携式计算机103中的一种或多种,当然也可以是台式计算机等等)、网络104和服务器105。网络104用以在终端设备和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线通信链路、无线通信链路等等。

应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器的。比如服务器105可以是多个服务器组成的服务器集群等。

需要说明的是,本实施例中的各个终端设备可以针对不同的通话集群,其中,通话集群中的参会者数量可以为一个、两个或者多个等等。示例性的,便携式计算机103中针对的通话集群中可以包含多个参会者,平板电脑102所针对的通话集群中可以包含其

他的参会者，用户还可以通过智能手机 101 参加到会议中来。

示例性的，在会议进行的过程中，多个用户或者一个用户可以使用终端设备进行会议通话。同时，服务器 105 可以通过网络 104 获取到终端设备之间的通话语音，对通话语音进行频域变换，确定通话语音在频域对应的能量信息；基于神经网络对能量信息进行分离处理，确定通话语音中所包含的各分支语音在通话语音中的能量占比；基于能量占比确定通话语音中所包含的分支语音的数量；基于分支语音的数量，通过设定通话语音控制方式对语音通话进行控制。

上述方案，在通话进行过程中实时获取通话语音，并对通话语音进行频域变换以确定通话语音在频域对应的能量信息，之后基于神经网络对能量信息进行分离处理，确定通话语音中所包含的各分支语音在通话语音中的能量占比，以基于能量占比确定通话语音中所包含的分支语音的数量，最后基于分支语音的数量，通过设定通话语音控制方式对语音通话进行控制，实现了针对语音通话过程中语音人数的即时检测，以及对语音通话的实时控制，既保证了语音人数的检测精度，同时提高了语音人数的识别效率以及语音通话的控制效率。

除此之外，如图 2 所示，本实施例中还可以只针对一个通话集群来处理，该通话集群中包括一个、两个或者多个参会者，通过上述语音通话的控制方法，可以检测到通话集群中实时的说话人数，并进行相应的通话控制，以保证通话过程中的语音质量，提高通话效率。

需要说明的是，本申请实施例所提供的语音通话的控制方法一般由服务器 105 执行，相应地，语音通话的控制装置一般设置于服务器 105 中。但是，在本申请的其它实施例中，终端设备也可以与服务器具有相似的功能，从而执行本申请实施例所提供的语音通话的控制方案。

需要说明的是，本实施例中的服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。

以下对本申请实施例的技术方案的实现细节进行详细阐述：

图 3 示出了根据本申请的一个实施例的语音通话的控制方法的流程图，该语音通话的控制方法可以由服务器来执行，该服务器可以是图 1 中所示的服务器。参照图 3 所示，

该语音通话的控制方法至少包括步骤 S110 至步骤 S150，详细介绍如下：

在步骤 S110 中，获取混合的通话语音，所述混合通话语音包含至少一个分支语音。

在本申请一些实施例中，可以在通话集群进行通话的过程中获取所述混合的通话语音，本实施例中的通话语音的长度不做限制，可以为实时获取到的一帧的通话语音，也可以是时长为一秒或者一分钟的通话语音等等。

示例性的，所述通话可以是实时通信会议场景，在所述实时通信会议过程中，可以实时采集通话语音，并基于采集到的通话语音进行相应的识别处理，以基于生成的识别结果进行之后的控制，实现针对通话语音的实时控制的效果。

在步骤 S120 中，对所述通话语音进行频域变换，确定所述通话语音在各个频点对应的能量信息。

在本申请一实施例中，本实施例中获取到的通话语音是时域的通话语音，是以时间为自变量，音量为因变量的语音信号。本实施例中在获取到通话语音之后，对通话语音进行频域变换，将时域的语音信号转化为频域的语音信号，以体现出通话语音在频域的能量信息。

如图 4 所示，基于语音分离的说话人场景分类框架，在“信号预处理”阶段，本申请实施例中通过获取到音频信号，并对音频信号进行预处理提取声学特征，生成音频信号对应的对数能量谱，作为通话语音在频域中各个频点对应的能量信息。

具体的，本实施例中的能量信息可以包括通话语音在各个频点对应的能量值、能量谱等信息。本实施例中可以通过能量信息来衡量各个频点的能量等属性，以基于能量信息来区别各个频点上对应的分支语音。

在步骤 S130 中，基于神经网络对每个频点上的能量信息进行分离处理，确定每个频点上对应的各分支语音在该频点的全部能量中的能量占比。

在本申请一实施例中，在获取到每个频点上的能量信息之后，基于神经网络的方式对能量信息进行分离处理。即将能量信息输入到已经训练好的分离模型中，得到各分支语音在该频点上的能量占比，也称为基于语音分离的频点系数，例如，在图 4 中，以两个分支语音为例，可以得到一个频点上的两个分支语音的频点系数：频点系数 P_A 和频点系数 P_B 。需要注意的是，图 4 中仅以两个分支语音为例，示出了一个频点上的两个频点系数，分别对应讲话人 A 和 B，实际应用中，一个频点上得到的频点系数与同时讲话的人数相关，不仅限于 2 个。此外，能量值不为 0 的各个频点上都会得到基于语音分离的频点系数。

具体的，本实施例中的频点系数用于表示一个频点上某个讲话人对应的能量在该频点的全部能量信息中的占比。本实施例中语音分离的思想是基于频域每个频点系数的方式，某个讲话人在某一个频点所占的能量占比与混合信号中预测的频点系数大小是成正比的。频点系数（P）的计算方式是：一个频点上某一说话人的语音能量值（E）比上该频点上混合说话人的语音能量值，假设两人（A 和 B），如下公式计算得到一个频点上讲话人 A 的频点系数：

$$P_A = \frac{E_A}{E_A + E_B}$$

在通过上式计算得到一个频点上的频点系数 P_A 和频点系数 P_B 之后，当 P_A 大于 P_B 时，该频点为说话人 A 主导，当 P_A 小于 P_B 时，该频点为说话人 B 主导。

上述方法，尤其是在多人同时说话的场景中，通过对能量信息进行分离，便可以确定每个频点上各个分支语音对应的能量占比，以基于能量占比来确定各个分支语音的分布情况，进而可以提高语音数量识别的准确性和实时性。

在步骤 S140 中，基于每个频点上各个分支语音的能量占比确定通话语音中所包含的分支语音的数量。

在本申请一实施例中，在确定了各个分支语音在通话语音中各个频点上的能量占比之后，本实施例中基于能量占比，通过求平均的方式来确定通话语音中所包含的分支语音的数量。

在本申请一实施例中，针对讲话人 A 的语音在每帧内各个频点对应的能量占比，将每帧内各个频点的能量占比进行求平均，得到一个一帧时间内稳定的能量均值，再基于设定的阈值来判定当前能量均值对应的是单人说话还是多人说话，最后输出当前帧的说话人数信息。比如，可以输出离散的 0（单人说话）或 1（多人说话）分别对应与各个帧当前的说话人数。

在本申请一实施例中，所采集到的通话语音由很多帧组成，并且一帧中存在多个频点，例如，频点个数可以为傅里叶变换的点数，设一帧中的频点个数为 f ， Fi 为某个讲话人在其中的第 i 个频点对应的能量占比，即频点系数，通过求平均的方式，得到该讲话人在该帧对应的能量占比均值为

$$\frac{1}{f} \sum_{i=0}^{f-1} Fi。$$

然后，可以将各个分支语音对应的平均值与阈值进行比较，从而确定同时讲话的人数。例如，以 2 人为例，因为 $P_A + P_B = 1$ ，所以只需要将 P_A 和 P_B 中的一个与阈值进行比

较即可，例如，以 P_A 为例，在实际的语音人数判定中，当该值是 0 或 1 时，表明当前说话的能量（ P_B 或者 P_A ）占满了全部的语音能量，因此只有某一个人在说话，即 B 或 A 在说话，此时为说话人个数为 1；如果该值是 0.5，则表明两个人都同时在说话，且此时说话能量大小相同，说话人个数为 2。通过将计算得到的能量占比均值和设定的阈值进行比较，便可以确定出当前的分支语音的数量。实际应用中，可以根据应用场景来设定阈值的具体取值。

本实施例中通过上述阈值检测方式来确定当前分支语音的数量，能实时在帧级很短的时间内对多说话人场景进行判断，实时处理语音流；并且可以使用多对多的标签，充分地利用语音信息，提升场景检测的准确率。

10 在步骤 S150 中，基于确定的分支语音的数量，通过设定通话语音控制方式对语音通话进行控制。

在本申请一实施例中，在确定了分支应用的数量之后，便可以结合当前分支语音的数量来确定当前的通话状况，进而通过设定的通话语音控制方式对语音通话进行控制，以实现对语音通话的实时控制，提高语音控制的精确性和实时性。

15 示例性的，如图 5 所示，在分支语音的数量为多个时，基于分支语音的数量，即滤除背景说话人，只提取主讲说话人的音频流，此时需要在前端检测当前说话人的个数，若当前说话人个数大于 1，则开启主讲说话人提取；若检测当前说话人个数为 1，则关闭语音提取，以避免对语音产生损伤。

本实施例中在提取语音的过程中，基于各频点上多个分支语音分别对应的能量占比，20 将能量占比中的最大值对应的分支语音识别为主讲人语音；从能量信息中确定主讲人语音对应的频率信息；基于频率信息从通话语音中提取出主讲人语音。

如图 6 所示，在多个用户都在说话的场景中，可以通过上述频率检测方式来确定其中的主讲人，如图 6 中的用户 4，并将主讲人的语音提取出来，或者过滤掉其余用户的语音，使得语音通话更加清晰。

25 通过上述方式可以在多人说话的场景下提取得到其中的一个主讲人的说话声音，使得通话过程中语音更加清楚，提高了通话质量和效果。

在图 7 所示的语音外放的情况下，通话人一方会从另一方回传过来的语音中听到自己的回声，而导致通话质量降低的问题。

为了避免这种情况发生，如图 8 所示，如果是通信对端与当前端轮流说话的单讲场景，即分支语音的数量为一个，则基于设定的单讲回声处理方式，识别分支语音的回传30

声音，并对回传声音进行单讲回声消除；

若是通信对端与当前端同时说话的双讲场景，即分支语音的数量为至少两个，则基于设定的双讲回声处理方式，分别识别分支语音对应的回传声音，并对回传声音进行双讲回声消除，在通信系统中尽量保证当前端的信号在回声消除过程中可以最大程度保留。

- 5 上述方案，在通话进行过程中，实时获取混合的通话语音，并对通话语音进行频域变换以确定通话语音在频域中各个频点对应的能量信息，之后基于神经网络对各个频点的能量信息进行分离处理，确定各个频点上对应的各分支语音的能量占比，以基于能量占比确定通话语音中所包含的分支语音的数量，最后基于分支语音的数量，通过设定通话语音控制方式对语音通话进行控制，实现了针对语音通话过程中语音人数的即时检测，
- 10 以及对语音通话的实时控制，既保证了语音人数的检测精度，同时提高了语音人数的识别效率以及语音通话的控制效率。

在本申请的一个实施例中，如图9所示，步骤S120中对通话语音进行频域变换，确定通话语音在频域对应的能量信息的过程，包括步骤S1210~步骤S1230，详细说明如下：

- 步骤S1210，对通话语音进行分帧处理，得到至少一帧的语音信息；
- 15 步骤S1220，对每帧语音信息进行频域变换，得到频域的语音能量谱；
- 步骤S1230，基于语音能量谱，确定通话语音在频域中每个频点对应的能量信息。

在本申请一实施例中，对每帧语音信息进行频域变换，得到频域的语音能量谱，包括：对每帧时域的语音信息进行傅里叶变换或者其他时域转换到频域的方式，得到语音信息对应于频域的语音能量谱。

- 20 本实施例中步骤S1230基于语音能量谱，确定通话语音在频域中各频点对应的能量信息，包括：对语音能量谱中各频点对应的幅度进行取模处理，得到语音能量谱对应的幅度谱；求取幅度谱的平方值，并对平方值进行对数运算，生成通话语音在频域中各个频点对应的能量信息。

- 如图10所示，在本申请一实施例中，对采集到的时域语音进行分帧、加窗处理；之后对每一帧分别进行N点傅里叶变换，求取傅里叶变换得到的N个频点的傅里叶变换系数，对其取模得到频域的幅度谱，再对获得的幅度谱求平方取对数，得到对数能量谱，即可得到语音在N个频点上的能量信息。
- 25

- 在本申请一实施例中神经网络包括长短时记忆神经网络；步骤S130中，基于神经网络对能量信息进行分离处理，确定通话语音中所包含的各个分支语音在通话语音中的能量占比，包括：将能量信息输入预设的语音分离模型中，进行基于长短期记忆神经网络
- 30

的卷积处理，确定各个频点上，通话语音中包含的各个分支语音在该频点对应的能量占比。

在本申请一实施例中，如图 11 所示，在对语音分离模型进行训练的过程中，方法还包括：

5 步骤 S1110，获取单一语音对应的第一语音样本，以及包含单一语音的混和语音对应的第二语音样本；

 步骤 S1120，从第一语音样本中提取出第一语音特征，从第二语音样本中提取出第二语音特征；

 步骤 S1130，将第二语音特征输入基于长短期记忆神经网络构建的语音分离模型中，确定从第二语音特征中分离出的预测语音，以及预测语音在第二语音样本中对应的预测能量占比；

10

 步骤 S1140，根据第一语音样本在第二语音样本中的实际能量占比，与预测能量占比之间的对比结果，更新语音分离模型的参数。

 在本申请一实施例中，首先构建训练所用的数据集获得两个音频库，单人说话语料库与多人说话语料库。其中，单人说话语料库中包括单一语音对应的第一语音样本，多人说话语料库就是分别随机抽取多个单人说话的片断进行叠加，然后分别对这两个数据库通过信号预处理提取语音中的对数能量谱特征，再经过分离模型得到每个有一段频点系数，再通过后处理得到分离后的语音。本实施例中将提取的对数能量谱特征作为输入，将该特征输入 2 层长短期记忆网络（Long Short-Term Memory, LSTM）以及一层输出层组成的分离模型中，得到频点系数。

15

20

 需要说明的是，本实施例中使用 LSTM 网络的原因是，它不仅考虑了当前时刻的输入，还赋予了网络对前面内容的一种记忆功能。同时，本实施例中的网络结构中的增加的输入门、输出门、遗忘门、细胞状态单元，使得 LSTM 在时序建模能力上得到显著的提升，能够记忆更多的信息，有效地抓住数据中的长时依赖。

25 在整个模型的训练中，训练集中包含多说话人和单说话人，使用多种单说话人语音相加得到多说话人语音，单说话人可以看成是与静音的混合。其中，从混合语音提取语音特征 a ，从对应的干净语音中提取语音特征 b ，以特征向量 a 作为训练输入，特征向量 b 作为训练目标，模型输出为频点系数 m 、 n 。通过后处理得到分离后的语音，以分离后的语音和干净语音的误差来指导 LSTM 模型进行训练，调整 LSTM 模型中的参数，从而

30 获得最终的分离模型，使得得到的分离模型更加精确和完整。

上述训练过程得到的语音分离模型，可以实时按帧级很短的时间内对多说话人场景进行判断，实时处理语音流；并且可以基于帧中各个频点对应的标签，通过多各频点对应于多个标签的方式，充分地利用语音信息，提升场景检测的准确率。

进一步的，本实施例中步骤 S1130 中将第二语音特征输入基于长短期记忆人工神经网络构建的语音分离模型中，确定从第二语音特征中分离出的预测语音的过程中，将得到的频点系数乘以混合语音的频谱，再经过逆傅里叶变换，结合混合信号的相位，可以得到分离后的语音信号。

步骤 S140 中，基于每个频点上各个分支语音的能量占比确定通话语音中所包含的分支语音的数量，包括：基于通话语音中所包含的分支语音在各频点对应的能量占比，求取各个分支语音的能量占比的平均值；基于各个分支语音的平均值和设定阈值，确定通话语音中所包含的分支语音的数量。

在本申请一实施例中，所采集到的通话语音由很多帧组成，并且一帧中存在多个频点个数，设一帧中的频点个数为 f ， Fi 为某个讲话人在其中的第 i 个频点对应的能量占比，即频点系数，通过求平均的方式，得到该讲话人在该帧对应的能量占比均值为

$$\frac{1}{f} \sum_{i=0}^{f-1} Fi。$$

15

在本申请一实施例中，在基于各个分支语音的平均值和设定阈值，确定通话语音中所包含的分支语音的数量，包括：若各个分支语音的平均值与设定阈值之间的差的绝对值小于差值阈值，则判定分支语音的数量为多个；若各个分支语音的平均值与设定阈值之间的差的绝对值大于或者等于差值阈值，则判定分支语音的数量为一个。

具体的，本实施例中以 2 人同时讲话为例，当平均值是越靠近 0 或者越靠近 1 时，单人说话的概率更大，越靠近 0.5 时，两人同时说话的概率更大，阈值的划定要根据具体的任务来划定。比如主说话人提取算法，为了避免单说话人时算法对语音的损伤，我们需要判断为单说话人的误检率低，此时阈值可以划定的靠近 0 或者 1。

如图 12 所示，在实际的会议应用场景中，可以在检测到当前的说话人数为多人时，将检测到的说话人或者与会人显示在界面中，并响应于用户触发的设定主讲人的指示，将其中的某一个人，或者多个人设定为主讲人，以对其余人的音频进行过滤，保证会议的通话质量。

如图 13 所示，对于会话控制服务器来说，当存在多个语音会话同时进行的过程中，

可以在为主讲人数较多的会议分配较多的通信资源，以保证通话的质量。

以下介绍本申请的装置实施例，可以用于执行本申请上述实施例中的语音通话的控制方法。可以理解的是，所述装置可以是运行于计算机设备中的一个计算机程序（包括程序代码），例如该装置为一个应用软件；该装置可以用于执行本申请实施例提供的方法中的相应步骤。对于本申请装置实施例中未披露的细节，请参照本申请上述的语音通话的控制方法的实施例。

图 14 示出了根据本申请的一个实施例的语音通话的控制装置的框图。

参照图 14 所示，根据本申请的一个实施例的语音通话的控制装置 1400，包括：获取单元 1410，用于获取混合的通话语音；变换单元 1420，用于对所述通话语音进行频域变换，确定所述通话语音在各个频点对应的能量信息；分离单元 1430，用于基于神经网络对每个频点上的所述能量信息进行分离处理，确定每个频点上所述通话语音中所包含的各个分支语音在所述通话语音中的能量占比；数量单元 1440，用于基于每个频点上各个分支语音的所述能量占比确定所述通话语音中所包含的分支语音的数量；控制单元 1450，用于基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制。

在本申请的一些实施例中，基于前述方案，所述变换单元 1420 包括：分帧单元，用于对所述通话语音进行分帧处理，得到至少一帧的语音信息；频域变换单元，用于对每帧所述语音信息进行频域变换，得到频域的语音能量谱；能量单元，用于基于所述语音能量谱，确定所述通话语音在频域中各频点对应的能量信息。

在本申请的一些实施例中，基于前述方案，所述频域变换单元用于：对时域的每帧所述语音信息进行傅里叶变换，得到所述语音信息对应于频域的语音能量谱。

在本申请的一些实施例中，基于前述方案，所述能量单元用于：对所述语音能量谱中各频点对应的幅度进行取模处理，得到所述语音能量谱对应的幅度谱；求取所述幅度谱的平方值，并对所述平方值进行对数运算，生成所述通话语音在频域中各频点对应的能量信息。

在本申请的一些实施例中，基于前述方案，所述神经网络包括长短时记忆神经网络；所述分离单元 1430 用于：将所述能量信息输入预设的语音分离模型中，进行基于长短期记忆神经网络的卷积处理，确定每个频点上对应的分支语音；基于每个频点上对应的各分支语音在该频点上的能量信息，确定每个频点上，所述各个分支语音在该频点上的能量占比。

在本申请的一些实施例中，基于前述方案，所述语音通话的控制装置 1400 还包括更新单元，用于：获取单一语音对应的第一语音样本，以及包含所述单一语音的混和语音对应的第二语音样本；从所述第一语音样本中提取出第一语音特征，从所述第二语音样本中提取出第二语音特征；将所述第二语音特征输入基于长短期记忆神经网络构建的语音分离模型中，确定从所述第二语音特征中分离出的预测语音，以及所述预测语音在所述第二语音样本中对应的预测能量占比；根据所述第一语音样本在所述第二语音样本中的实际能量占比，与所述预测能量占之间的对比结果，更新所述语音分离模型的参数。

在本申请的一些实施例中，基于前述方案，所述数量单元 1440 包括：平均单元，用于对于每个分支语音，基于该分支语音在各频点对应的能量占比，求取该分支语音的所述能量占比的平均值；语音数量单元，用于基于每个分支语音的所述平均值和设定阈值，确定所述通话语音中所包含的分支语音的数量。

在本申请的一些实施例中，基于前述方案，所述语音数量单元包括：第一数量单元，用于若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值小于差值阈值，则判定所述分支语音的数量为多个；第二数量单元，用于若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值大于或者等于所述差值阈值，则判定所述分支语音的数量为一个。

在本申请的一些实施例中，基于前述方案，所述控制单元 1450 包括提取单元，用于基于设定的语音提取方式，进行主讲人语音提取。

在本申请的一些实施例中，基于前述方案，所述提取单元，用于基于各频点上多个分支语音分别对应的能量占比，将所述能量占比中的最大值对应的分支语音识别为所述主讲人语音；从所述能量信息中确定所述主讲人语音对应的频率信息；基于所述频率信息从所述通话语音中提取出所述主讲人语音。

在本申请的一些实施例中，基于前述方案，所述分支语音的数量包括一个或者至少两个，所述控制单元 1450 用于：若所述分支语音的数量为一个，则基于设定的单讲回声处理方式，识别所述分支语音的回传声音，并对所述回传声音进行单讲回声消除；若所述分支语音的数量为至少两个，则基于设定的双讲回声处理方式，分别识别所述分支语音对应的回传声音，并对所述回传声音进行双讲回声消除。

图 15 示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

需要说明的是，图 15 示出的电子设备的计算机系统 1500 仅是一个示例，不应对本

申请实施例的功能和使用范围带来任何限制。

如图 15 所示,计算机系统 1500 包括中央处理单元(Central Processing Unit,CPU) 1501,其可以根据存储在只读存储器(Read-Only Memory,ROM) 1502 中的程序或者从
5 储存部分 1508 加载到随机访问存储器(Random Access Memory,RAM) 1503 中的程序而
执行各种适当的动作和处理,例如执行上述实施例中所述的方法。在 RAM 1503 中,还存
储有系统操作所需的各种程序和数据。CPU 1501、ROM 1502 以及 RAM 1503 通过总线 1504
彼此相连。输入/输出(Input /Output,I/O)接口 1505 也连接至总线 1504。

以下部件连接至 I/O 接口 1505:包括键盘、鼠标等的输入部分 1506;包括诸如阴极
射线管(Cathode Ray Tube,CRT)、液晶显示器(Liquid Crystal Display,LCD)等以
10 及扬声器等的输出部分 1507;包括硬盘等的储存部分 1508;以及包括诸如 LAN(Local
Area Network,局域网)卡、调制解调器等的网络接口卡的通信部分 1509。通信部分 1509
经由诸如因特网的网络执行通信处理。驱动器 1510 也根据需要连接至 I/O 接口 1505。
可拆卸介质 1511,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器
1510 上,以便于从其上读出的计算机程序根据需要被安装入储存部分 1508。

15 特别地,根据本申请的实施例,上文参考流程图描述的过程可以被实现为计算机软件
程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在计算机可读介
质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的计算机程序。在这
样的实施例中,该计算机程序可以通过通信部分 1509 从网络上被下载和安装,和/或从
可拆卸介质 1511 被安装。在该计算机程序被中央处理单元(CPU) 1501 执行时,执行本
20 申请的系统中限定的各种功能。

需要说明的是,本申请实施例所示的计算机可读介质可以是计算机可读信号介质或
者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质不限于电、磁、
光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读
存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计
25 算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储
器(Erasable Programmable Read Only Memory,EPR0M)、闪存、光纤、便携式紧凑磁
盘只读存储器(Compact Disc Read-Only Memory,CD-ROM)、光存储器件、磁存储器件、
或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储
程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。
30 而在本申请中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数

据信号,其中承载了计算机可读的计算机程序。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的计算机程序可以用任何适当的介质传输,包括但不限于:无线、有线等等,或者上述的任意合适的组合。

根据本申请的一个方面,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述各种实现方式中提供的方法。

作为另一方面,本申请还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该电子设备执行时,使得该电子设备实现上述实施例中所述的方法。

应当理解的是,本申请并不局限于上面已经描述并在附图中示出的精确结构,并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

权利要求书

1. 一种语音通话的控制方法，由服务器执行，包括：
获取混合的通话语音，所述混合通话语音包含至少一个分支语音；
对所述通话语音进行频域变换，确定所述通话语音在频域中各个频点对应的能量信息；
5 息；
基于神经网络对每个频点上的所述能量信息进行分离处理，确定每个频点上对应的各分支语音在该频点的全部能量中的能量占比；
基于每个频点上各分支语音的所述能量占比确定所述通话语音中所包含的分支语音的数量；
10 基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制。
2. 根据权利要求1所述的方法，其中，对所述通话语音进行频域变换，确定所述通话语音在频域中各个频点对应的能量信息，包括：
对所述通话语音进行分帧处理，得到至少一帧的语音信息；
对每帧所述语音信息进行频域变换，得到频域的语音能量谱；
15 基于所述语音能量谱，确定所述通话语音在频域中各个频点对应的能量信息。
3. 根据权利要求2所述的方法，其中，对每帧所述语音信息进行频域变换，得到频域的语音能量谱，包括：
对时域的每帧所述语音信息进行傅里叶变换，得到每帧所述语音信息对应于频域的语音能量谱。
- 20 4. 根据权利要求2所述的方法，其中，基于所述语音能量谱，确定所述通话语音在频域中各个频点对应的能量信息，包括：
对所述语音能量谱中各频点对应的幅度进行取模处理，得到所述语音能量谱对应的幅度谱；
求取所述幅度谱的平方值，并对所述平方值进行对数运算，生成所述通话语音在频
25 域中各个频点对应的能量信息。
5. 根据权利要求1所述的方法，其中，所述神经网络包括长短时记忆神经网络；
基于神经网络对每个频点上的所述能量信息进行分离处理，确定每个频点上对应的各个分支语音在该频点的全部能量中的能量占比，包括：
将所述能量信息输入预设的语音分离模型中，进行基于长短期记忆神经网络的卷积
30 处理，确定每个频点上对应的分支语音；
基于每个频点上对应的各分支语音在该频点上的能量信息，确定每个频点上，所述

各个分支语音在该频点上的能量占比。

6. 根据权利要求 5 所述的方法, 还包括:

获取单一语音对应的第一语音样本, 以及包含所述单一语音的混和语音对应的第二语音样本;

5 从所述第一语音样本中提取出第一语音特征, 从所述第二语音样本中提取出第二语音特征;

将所述第二语音特征输入基于长短期记忆人工神经网络构建的语音分离模型中, 确定从所述第二语音特征中分离出的预测语音, 以及所述预测语音在所述第二语音样本中对应的预测能量占比;

10 根据所述第一语音样本在所述第二语音样本中的实际能量占比, 与所述预测能量占之间的对比结果, 更新所述语音分离模型的参数。

7. 根据权利要求 5 所述的方法, 其中, 基于每个频点上各分支语音的所述能量占比确定所述通话语音中所包含的分支语音的数量, 包括:

15 对于每个分支语音, 基于该分支语音在各频点对应的能量占比, 求取该分支语音的所述能量占比的平均值;

基于每个分支语音的所述平均值和设定阈值, 确定所述通话语音中所包含的分支语音的数量。

8. 根据权利要求 7 所述的方法, 其中, 基于每个分支语音的所述平均值和设定阈值, 确定所述通话语音中所包含的分支语音的数量, 包括:

20 若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值小于差值阈值, 则判定所述分支语音的数量为多个;

若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值大于或者等于所述差值阈值, 则判定所述分支语音的数量为一个。

25 9. 根据权利要求 1-8 任一项所述的方法, 其中, 在确定所述分支语音的数量为多个时, 基于所述分支语音的数量, 通过设定通话语音控制方式对所述语音通话进行控制, 包括:

基于设定的语音提取方式, 进行主讲人语音提取。

10. 根据权利要求 9 所述的方法, 其中, 若确定所述分支语音的数量为多个, 基于设定的语音提取方式, 进行主讲人语音提取, 包括:

30 基于各频点上多个分支语音分别对应的能量占比, 将所述能量占比中的最大值对应的分支语音识别为所述主讲人语音;

从所述能量信息中确定所述主讲人语音对应的频率信息；

基于所述频率信息从所述通话语音中提取出所述主讲人语音。

11. 根据权利要求 1-8 任一项所述的方法，其中，所述分支语音的数量包括一个或者至少两个；

5 基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制，包括：

若所述分支语音的数量为一个，则基于设定的单讲回声处理方式，识别所述分支语音的回传声音，并对所述回传声音进行单讲回声消除；

10 若所述分支语音的数量为至少两个，则基于设定的双讲回声处理方式，分别识别所述分支语音对应的回传声音，并对所述回传声音进行双讲回声消除。

12. 一种语音通话的控制装置，包括：

获取单元，用于获取混合的通话语音，所述混合通话语音包含至少一个分支语音；

变换单元，用于对所述通话语音进行频域变换，确定所述通话语音在频域在各个频点对应的能量信息；

15 分离单元，用于基于神经网络对每个频点上的所述能量信息进行分离处理，确定每个频点上对应的各分支语音在该频点的全部能量中的能量占比；

数量单元，用于基于每个频点上各分支语音的所述能量占比确定所述通话语音中所包含的分支语音的数量；

20 控制单元，用于基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制。

13. 根据权利要求 12 所述的装置，其中，所述变换单元包括：

分帧单元，用于对所述通话语音进行分帧处理，得到至少一帧的语音信息；

频域变换单元，用于对每帧所述语音信息进行频域变换，得到频域的语音能量谱；

25 能量单元，用于基于所述语音能量谱，确定所述通话语音在频域中各个频点对应的能量信息。

14. 根据权利要求 13 所述的装置，其中，所述频域变换单元用于：

对时域的每帧所述语音信息进行傅里叶变换，得到每帧所述语音信息对应于频域的语音能量谱。

15. 根据权利要求 13 所述的装置，其中，所述能量单元用于：

30 对所述语音能量谱中各频点对应的幅度进行取模处理，得到所述语音能量谱对应的幅度谱；

求取所述幅度谱的平方值，并对所述平方值进行对数运算，生成所述通话语音在频域中各个频点对应的能量信息。

16. 根据权利要求 12 所述的装置，其中，所述神经网络包括长短期记忆神经网络；所述分离单元用于：

5 将所述能量信息输入预设的语音分离模型中，进行基于长短期记忆神经网络的卷积处理，确定每个频点上对应的分支语音；

基于每个频点上对应的各分支语音在该频点上的能量信息，确定每个频点上，所述各个分支语音在该频点上的能量占比。

17. 根据权利要求 16 所述的装置，还包括更新单元，用于：

10 获取单一语音对应的第一语音样本，以及包含所述单一语音的混和语音对应的第二语音样本；

从所述第一语音样本中提取出第一语音特征，从所述第二语音样本中提取出第二语音特征；

15 将所述第二语音特征输入基于长短期记忆人工神经网络构建的语音分离模型中，确定从所述第二语音特征中分离出的预测语音，以及所述预测语音在所述第二语音样本中对应的预测能量占比；

根据所述第一语音样本在所述第二语音样本中的实际能量占比，与所述预测能量占之间的对比结果，更新所述语音分离模型的参数。

18. 根据权利要求 16 所述的装置，其中，所述数量单元包括：

20 平均单元，用于对于每个分支语音，基于该分支语音在各频点对应的能量占比，求取该分支语音的所述能量占比的平均值；

语音数量单元，基于每个分支语音的所述平均值和设定阈值，确定所述通话语音中所包含的分支语音的数量。

19. 根据权利要求 18 所述的装置，其中，所述语音数量单元包括：

25 第一数量单元，用于若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值小于差值阈值，则判定所述分支语音的数量为多个；

第二数量单元，用于若每个分支语音的所述平均值与所述设定阈值之间的差的绝对值大于或者等于所述差值阈值，则判定所述分支语音的数量为一个。

20. 根据权利要求 12-19 任一项所述的装置，其中，所述控制单元包括：

30 提取单元，用于基于设定的语音提取方式，进行主讲人语音提取。

21. 根据权利要求 20 所述的装置，其中，所述提取单元用于：

基于各频点上多个分支语音分别对应的能量占比，将所述能量占比中的最大值对应的分支语音识别为所述主讲人语音；

从所述能量信息中确定所述主讲人语音对应的频率信息；

基于所述频率信息从所述通话语音中提取出所述主讲人语音。

- 5 22. 根据权利要求 12-19 任一项所述的装置，其中，所述分支语音的数量包括一个或者至少两个；

所述控制单元用于：

基于所述分支语音的数量，通过设定通话语音控制方式对所述语音通话进行控制，包括：

- 10 若所述分支语音的数量为一个，则基于设定的单讲回声处理方式，识别所述分支语音的回传声音，并对所述回传声音进行单讲回声消除；

若所述分支语音的数量为至少两个，则基于设定的双讲回声处理方式，分别识别所述分支语音对应的回传声音，并对所述回传声音进行双讲回声消除。

- 15 23. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求 1 至 11 中任一项所述的语音通话的控制方法。

24. 一种电子设备，包括：

一个或多个处理器；

- 20 存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现如权利要求 1 至 11 中任一项所述的语音通话的控制方法。