说 明 书

直播语音播放方法、装置、设备及存储介质

本申请要求在2025年02月14日提交中国专利局、申请号为202510166685.5的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请涉及音频处理技术领域，例如涉及一种直播语音播放方法、装置、设备及存储介质。

背景技术

随着直播行业的火热发展和人工智能（Artificial Intelligence，AI）技术的不断成熟，使用数字人代替真人主播的新型直播也应运而生。在传统直播时，音频可以通过录音设备采集真人主播得到。但在数字人直播时，音频是通过从文本到语音（Text To Speech，TTS）技术生成的MP3。由于直播平台需要的音频必须是持续的语音流，如果直接将离散的MP3进行推流，会导致直播声音断断续续，使得用户在听觉方面体验感较差。

发明内容

本申请提供了一种直播语音播放方法、装置、设备及存储介质，实现了音频内容的实时打断与插播，满足了直播场景下灵活调整内容的需求，解决了数字人直播中音频不连续的问题，使得直播声音更加流畅自然。

根据本申请的一方面，提供了一种直播语音播放方法，该方法包括：

获取在直播过程中数字直播对象对应的待播放音频数据；

将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；

基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

根据本申请的另一方面，提供了一种直播语音播放装置，该装置包括：

音频数据获取模块，设置为获取在直播过程中数字直播对象对应的待播放音频数据；

碎片数据获得模块，设置为将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；

直播语音播放模块，设置为基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

根据本申请的另一方面，提供了一种电子设备，所述电子设备包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行本申请任一实施例所述的直播语音播放方法。

根据本申请的另一方面，提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使处理器执行时实现本申请任一实施例所述的直播语音播放方法。

附图说明

图 1 是根据本申请实施例一提供的一种直播语音播放方法的流程图；

图 2 是根据本申请实施例二提供的一种直播语音播放方法的流程图；

图 3 是根据本申请实施例三提供的一种直播语音播放装置的结构图；

图 4 是实现本申请实施例的直播语音播放方法的电子设备的结构示意图。

具体实施方式

本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”

以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

实施例一

图 1 为本申请实施例一提供的一种直播语音播放方法的流程图，本实施例可适用于为数字直播人进行直播语音配置的情况，该方法可以由直播语音播放装置来执行，该直播语音播放装置可以采用硬件和/或软件的形式实现，该直播语音播放装置可配置于电子设备中。如图 1 所示，该方法包括：

S101、获取在直播过程中数字直播对象对应的待播放音频数据。

其中，数字直播对象可以是指直播过程中利用数字技术创建的虚拟对象，如虚拟人物和虚拟玩偶等。数字直播对象能够模拟人类的表情、动作和声音，从而为用户提供逼真的互动体验。待播放音频数据可以是指用于生成直播语音的音频数据，如离散的 MP3 音频片段。

在直播场景下，数字人可以作为主播或嘉宾出现，与用户进行实时互动。为了实现这一功能，数字人背后的技术支持至关重要，其中就包括音频流生成服务。该生成服务能够将离散的 MP3 音频片段转换为持续的语音流，并支持实时内容插播，从而满足数字人直播中的音频需求。

本申请应用于服务端，服务端为调用方（如直播平台或数字人系统）提供直播音频生成服务。在获取直播过程中数字直播对象对应的待播放音频数据之前，服务端与调用方建立长连接，以便实时传输和处理待播放音频数据。示例性地，服务端可以基于 WebSocket 协议与调用方建立长连接，服务端向调用方发送音频获取请求，调用方将待播放的待播放音频数据发送给服务端，实现音频内容的实时传输和处理，以提高数据传输的效率和实时性。

S102、将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件。

其中，音频碎片数据可以是指待播放音频数据进行拆分后获得的音频碎片。

碎片索引文件可以是指音频碎片数据对应的文本索引文件，其可以用于将多个音频碎片数据按照顺序排列，以便播放器能够按需下载并连续播放，形成一个完整的音频流。

示例性地，所述音频碎片数据为传输流（Transport Stream，TS）碎片，所述碎片索引文件为M3U8索引文件。需要说明的是，TS碎片是基于超文本传输协议（Hypertext Transfer Protocol，HTTP）的直播流（HTTP Live Streaming，HLS）协议中音视频流被分割成的多个只有几秒长度的.ts 文件。这些文件包含了音视频流的部分内容，并且是按照顺序排列的，以便播放器能够按需下载并连续播放。M3U8是M3U（Moving Picture Experts Group Audio Layer 3 Uniform Resource Locator）文件的扩展版本，专门针对Unicode字符集进行了优化，增加了对UTF-8字符编码的支持，这使得它可以更好地处理多语言内容。

示例性地，服务端利用HLS技术将待播放音频数据拆分为更细粒度的音频碎片数据，并生成音频碎片数据对应的碎片索引文件。

S103、基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

示例性地，服务端按照碎片索引文件的顺序，通过FFmpeg工具将音频碎片数据推送到调用方的直播音频通道中，以使调用方根据多个所述音频碎片数据的排列顺序，实现数字直播对象的直播语音播放。

示例性地，在所述将多个所述音频碎片数据推送至直播音频通道中之后，还包括：根据所述碎片索引文件，将总索引文件进行更新处理。

在调用端设置有总索引文件，在将多个音频碎片数据推送至直播音频通道中之后，需要将碎片索引文件追加写入总索引文件中，可以方便地对各个索引文件进行统一管理和维护。例如，可以轻松地添加、删除或更新索引文件中的内容，而无需对每个音频碎片数据进行单独的操作，有助于降低管理成本和提高维护效率。同时，也使得音频流的更新和升级变得更加简单和方便。

在一些实施方式中，所述方法还包括：响应于所述音频碎片数据的未播放数量小于预设数量阈值，继续获取待播放音频数据，并实现所述数字直播对象

的直播语音播放。

其中，预设数量阈值可以根据实际情况或历史经验进行设定，示例性地，预设数量阈值可以为1。

也就是说，当音频碎片数据的未播放数量小于预设数量阈值时，或者直播语音播放还剩一句未播放时，服务端提前向调用方继续获取下一阶段的待播放音频数据，并重复上述处理步骤，以实现数字直播对象的直播语音的持续播放。

在一些实施方式中，所述方法，还包括：响应于未获取到待播放音频数据，将空白碎片数据推送至所述直播音频通道中。

也就是说，如果调用方未能及时提供新的待播放音频数据，服务端则将空白音频碎片推送到音频通道，确保音频推流不中断，提升了直播的稳定性和用户体验。

本申请实施例，通过获取在直播过程中数字直播对象对应的待播放音频数据；将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放，解决了数字人直播中音频不连续的问题，满足了直播场景下灵活调整音频内容的需求，使得数字人直播的直播声音更加流畅自然。

在上述各实施例的基础上，所述方法，还包括：接收音频插播指令，根据接收到的所述音频插播指令获取待插播音频数据；将所述待插播音频数据拆分处理后获得的插播碎片数据，插入至所述直播音频通道中，实现所述数字直播对象的插播语音播放。

其中，音频插播指令可以是指在预设直播音频播放过程中需要进行插播音频播放的指令。待插播音频数据可以是指待生成插播播放音频的音频数据。插播碎片数据可以是指待插播音频数据进行拆分处理的音频碎片。

示例性地，在接收到音频插播指令后，服务端向调用方请求获取与音频插播指令对应的待插播音频数据。服务端根据音频插播指令的要求，将待插播音频数据拆分为插播碎片数据，并生成插播索引文件。参照插播索引文件中插播

碎片数据的排列顺序，将插播碎片数据插入至调用方直播音频通道中，实现所述数字直播对象的插播语音播放。同时，还可以利用插播索引文件对调用方中的总索引文件进行更新。

在一些实施方式中，在将所述插播碎片数据插入至所述直播音频通道中之前，还包括：确定所述音频插播指令对应的插播类型，其中，所述插播类型包括打断式插播和临时式插播；响应于确定所述插播类型为打断式插播，将所述直播音频通道进行清空处理，并将所述插播碎片数据插入至清空后的直播音频通道中；响应于确定所述插播类型为临时式插播，基于所述音频插播指令对应的插播时间点，将所述插播碎片数据插入至所述直播音频通道中。

示例性地，如果音频插播指令要求打断式插播当前播放内容，调用方丢弃直播音频通道中未播放完毕的音频碎片数据，实现直播音频通道的清空，并将插播索引文件追加到总索引中。如果音频插播指令要求临时式插播当前播放内容，根据音频插播指令要求的插播时间点，将所述插播碎片数据插入至所述直播音频通道中，并将插播索引文件追加到总索引中，实现不打断插播。本申请实施例，通过支持打断当前播放内容进行插播，以及不打断当前内容实现插播，满足了直播场景下多样化的内容插播需求，提升了用户使用体验。

实施例二

图 2 为本申请实施例二提供的一种直播语音播放方法的流程图，本实施例为在上述各实施例的基础上的一可选实施方案。如图 2 所示，该方法包括：

在顺播阶段，服务端与调用方建立 WebSocket 长连接。

服务端向调用方发送音频获取请求，调用方将待播放音频数据发送给服务端。

服务端将待播放音频数据利用 HLS 技术拆分为音频碎片数据和碎片索引文件。

服务端将音频碎片数据存入待播放列表，并将碎片索引文件追加写入总索引。

服务端按照总索引文件的顺序将音频碎片数据用 FFmpeg 工具推送到音频

通道进行播放。

在当前音频通道仅剩一句未播放时，提前向调用方索取下一个待播放音频数据，并重复上述步骤。

在插播阶段，调用方向服务端发送音频插播指令。

服务端根据音频插播指令的要求，将待插播音频数据拆分为插播碎片数据及插播索引文件。

如果音频插播指令要求打断当前播放内容，调用方丢弃未播放完毕的音频碎片数据，并将插播碎片数据插入值音频通道，以及将插播索引文件追加到总索引中。

如果插播消息不需要打断当前内容，服务端将插播索引文件追加到总索引中。服务端按照更新后的总索引文件顺序推送插播碎片数据到音频通道。

在异常处理阶段，如果调用方未能及时提供新待播放音频数据，服务端将空白音频碎片推送到音频通道，防止推流中断。

实施例三

图 3 为本申请实施例三提供的一种直播语音播放装置的结构示意图。如图 3 所示，该装置包括：

音频数据获取模块 301，设置为获取在直播过程中数字直播对象对应的待播放音频数据；

碎片数据获得模块 302，设置为将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；

直播语音播放模块 303，设置为基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

本申请实施例，通过获取在直播过程中数字直播对象对应的待播放音频数据；将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；基于所述碎片索引文件，将多个所述音频碎片

数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放，解决了数字人直播中音频不连续的问题，满足了直播场景下灵活调整音频内容的需求，使得数字人直播的直播声音更加流畅自然。

在一些实施方式中，音频数据获取模块301，还设置为：

响应于所述音频碎片数据的未播放数量小于预设数量阈值，继续获取待播放音频数据，并实现所述数字直播对象的直播语音播放。

在一些实施方式中，直播语音播放模块303，还设置为：

响应于未获取到待播放音频数据，将空白碎片数据推送至所述直播音频通道中。

在一些实施方式中，所述装置还包括：直播语音插播模块。其中，

直播语音插播模块，设置为：

接收音频插播指令，根据接收到的所述音频插播指令获取待插播音频数据；

将所述待插播音频数据拆分处理后获得的插播碎片数据，插入至所述直播音频通道中，实现所述数字直播对象的插播语音播放。

在一些实施方式中，所述装置还包括：插播类型确定模块。其中，

插播类型确定模块，设置为：

在将所述插播碎片数据插入至所述直播音频通道中之前，确定所述音频插播指令对应的插播类型，其中，所述插播类型包括打断式插播和临时式插播；

响应于确定所述插播类型为打断式插播，将所述直播音频通道进行清空处理，并将所述插播碎片数据插入至清空后的直播音频通道中；

响应于确定所述插播类型为临时式插播，基于所述音频插播指令对应的插播时间点，将所述插播碎片数据插入至所述直播音频通道中。

在一些实施方式中，所述装置还包括索引文件更新模块。其中，

索引文件更新模块，设置为：

在所述将多个所述音频碎片数据推送至直播音频通道中之后，根据所述碎

片索引文件，将总索引文件进行更新处理。

在一些实施方式中，所述音频碎片数据为 TS 碎片，所述碎片索引文件为 M3U8 索引文件。

本申请实施例所提供的直播语音播放装置可执行本申请任意实施例所提供的直播语音播放方法，具备执行方法相应的功能模块和有益效果。

实施例四

图 4 示出了可以用来实施本申请的实施例的电子设备 10 的结构示意图。电子设备旨在表示各种形式的数字计算机，如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，如，个人数字处理、蜂窝电话、智能电话、可穿戴设备（如头盔、眼镜、手表等）和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能作为示例。

如图 4 所示，电子设备 10 包括至少一个处理器 11，以及与至少一个处理器 11 通信连接的存储器，如只读存储器（Read Only Memory，ROM）12、随机访问存储器（Random Access Memory，RAM）13 等，其中，存储器存储有可被至少一个处理器执行的计算机程序，处理器 11 可以根据存储在只读存储器（ROM）12 中的计算机程序或者从存储单元 18 加载到随机访问存储器（RAM）13 中的计算机程序，来执行各种适当的动作和处理。在 RAM 13 中，还可存储电子设备 10 操作所需的各种程序和数据。处理器 11、ROM 12 以及 RAM 13 通过总线 14 彼此相连。输入/输出（Input/Output，I/O）接口 15 也连接至总线 14。

电子设备 10 中的多个部件连接至 I/O 接口 15，包括：输入单元 16，例如键盘、鼠标等；输出单元 17，例如各种类型的显示器、扬声器等；存储单元 18，例如磁盘、光盘等；以及通信单元 19，例如网卡、调制解调器、无线通信收发机等。通信单元 19 允许电子设备 10 通过如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

处理器 11 可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器 11 包括中央处理单元（Central Processing Unit，CPU）、图形处理单元（Graphics

Processing Unit，GPU）、各种专用的AI计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器（Digital Signal Processor，DSP）、以及任何适当的处理器、控制器、微控制器等。处理器11执行上文所描述的各个方法和处理，例如直播语音播放方法。

在一些实施例中，直播语音播放方法可被实现为计算机程序，其被有形地包含于计算机可读存储介质，例如存储单元18。在一些实施例中，计算机程序的部分或者全部可以经由ROM 12和/或通信单元19而被载入和/或安装到电子设备10上。当计算机程序加载到RAM 13并由处理器11执行时，可以执行上文描述的直播语音播放方法的一个或多个步骤。在其他实施例中，处理器11可以通过其他任何适当的方式（例如，借助于固件）而被配置为执行直播语音播放方法。

本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列（Field Programmable Gate Array，FPGA）、专用集成电路（Application Specific Integrated Circuit，ASIC）、专用标准产品（Application-Specific Standard Product，ASSP）、系统级芯片（System on a Chip，SOC）、复杂可编程逻辑设备（Complex Programmable Logic Device，CPLD）、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

用于实施本申请的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器，使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

在本申请的上下文中，计算机可读存储介质可以是有形的介质，其可以包

含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。可选地，计算机可读存储介质可以是机器可读信号介质。机器可读存储介质可以包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、RAM、ROM、可擦除可编程只读存储器（Erasable Programmable Read Only Memory，EPROM 或快闪存储器）、光纤、便捷式紧凑盘只读存储器（Compact Disc Read Only Memory，CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

为了提供与用户的交互，可以在电子设备上实施此处描述的系统和技术，该电子设备具有：用于向用户显示信息的显示装置（例如，阴极射线管（Cathode Ray Tube，CRT）或者液晶显示器（Liquid Crystal Display，LCD）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或者、触觉输入）来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（Local Area Network，LAN）、广域网（Wide Area Network，WAN）、区块链网络和互联网。

计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器，

又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决了传统物理主机与虚拟专用服务器（Virtual Private Server，VPS）服务中，存在的管理难度大，业务扩展性弱的缺陷。

权 利 要 求 书

1. 一种直播语音播放方法，包括：

获取在直播过程中数字直播对象对应的待播放音频数据；

将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；

基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

2. 根据权利要求 1 所述的方法，还包括：

响应于所述音频碎片数据的未播放数量小于预设数量阈值，继续获取待播放音频数据，并实现所述数字直播对象的直播语音播放。

3. 根据权利要求 2 所述的方法，还包括：

响应于未获取到待播放音频数据，将空白碎片数据推送至所述直播音频通道中。

4. 根据权利要求 1 所述的方法，还包括：

接收音频插播指令，根据接收到的所述音频插播指令获取待插播音频数据；

将所述待插播音频数据拆分处理后获得的插播碎片数据，插入至所述直播音频通道中，实现所述数字直播对象的插播语音播放。

5. 根据权利要求 4 所述的方法，在将所述插播碎片数据插入至所述直播音频通道中之前，所述方法还包括：

确定所述音频插播指令对应的插播类型，其中，所述插播类型包括打断式插播和临时式插播；

响应于确定所述插播类型为打断式插播，将所述直播音频通道进行清空处理，并将所述插播碎片数据插入至清空后的直播音频通道中；

响应于确定所述插播类型为临时式插播，基于所述音频插播指令对应的插播时间点，将所述插播碎片数据插入至所述直播音频通道中。

6. 根据权利要求 1 所述的方法，在所述将多个所述音频碎片数据推送至直

播音频通道中之后，所述方法还包括：

根据所述碎片索引文件，将总索引文件进行更新处理。

7. 根据权利要求 1-6 任一项所述的方法，其中，所述音频碎片数据为传输流 TS 碎片，所述碎片索引文件为 M3U8 索引文件。

8. 一种直播语音播放装置，包括：

音频数据获取模块，设置为获取在直播过程中数字直播对象对应的待播放音频数据；

碎片数据获得模块，设置为将所述待播放音频数据进行拆分处理，获得多个音频碎片数据和所述音频碎片数据对应的碎片索引文件；

直播语音播放模块，设置为基于所述碎片索引文件，将多个所述音频碎片数据推送至直播音频通道中，并基于多个所述音频碎片数据的排列顺序，实现所述数字直播对象的直播语音播放。

9. 一种电子设备，所述电子设备包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求 1-7 中任一项所述的直播语音播放方法。

10. 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使处理器执行时实现权利要求 1-7 中任一项所述的直播语音播放方法。