



(12) 发明专利

(10) 授权公告号 CN 115440220 B

(45) 授权公告日 2025. 05. 16

(21) 申请号 202211073708.0

G10L 25/27 (2013.01)

(22) 申请日 2022.09.02

(56) 对比文件

(65) 同一申请的已公布的文献号

CN 109359177 A, 2019.02.19

申请公布号 CN 115440220 A

审查员 赵琳

(43) 申请公布日 2022.12.06

(73) 专利权人 京东科技信息技术有限公司

地址 100176 北京市大兴区经济技术开发区

区科创十一街18号院2号楼6层601

(72) 发明人 冯明超 王泽勋 杨久东 陈蒙

(74) 专利代理机构 北京品源专利代理有限公司

11332

专利代理师 马迪

(51) Int. Cl.

G10L 15/22 (2006.01)

G10L 15/18 (2013.01)

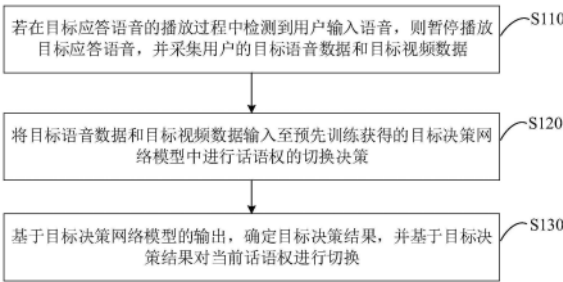
权利要求书2页 说明书11页 附图4页

(54) 发明名称

一种话语权切换方法、装置、设备和存储介质

(57) 摘要

本发明实施例公开了一种话语权切换方法、装置、设备和存储介质。该方法包括：若在目标应答语音的播放过程中检测到用户输入语音，则暂停播放目标应答语音，并采集用户的目标语音数据和目标视频数据；将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策；基于目标决策网络模型的输出，确定目标决策结果，并基于目标决策结果对当前话语权进行切换。通过本发明实施例的技术方案，可以提高话语权切换的准确性，从而提高人机对话的流畅性，提升用户体验。



1. 一种话语权切换方法,其特征在于,包括:

若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

确定所述目标应答语音对应的目标应答文本;

将所述目标应答文本、所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换;

其中,所述基于所述目标决策结果对当前话语权进行切换,包括:

若所述目标决策结果为话语权切换结果,则将当前话语权切换至用户,以停止播放所述目标应答语音并等待用户表述。

2. 根据权利要求1所述的方法,其特征在于,所述目标决策网络模型包括:文本特征提取子模型、语音特征提取子模型、视频特征提取子模型和决策子模型;

所述将所述目标应答文本、所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策,包括:

将所述目标应答文本输入至所述文本特征提取子模型中进行文本特征提取,获得提取出的目标文本特征信息;

将所述目标语音数据输入至所述语音特征提取子模型中进行语音特征提取,获得提取出的目标语音特征信息;

将所述目标视频数据输入至所述视频特征提取子模型中进行视频特征提取,获得提取出的目标视频特征信息;

将所述目标文本特征信息、所述目标语音特征信息和所述目标视频特征信息输入至所述决策子模型中进行话语权的切换决策,获得目标决策结果。

3. 根据权利要求2所述的方法,其特征在于,所述决策子模型包括:信息拼接层和全连接层;

所述将所述目标文本特征信息、所述目标语音特征信息和所述目标视频特征信息输入至所述决策子模型中进行话语权的切换决策,获得目标决策结果,包括:

将所述目标文本特征信息、所述目标语音特征信息和所述目标视频特征信息输入至所述信息拼接层中进行特征信息的拼接,获得拼接后的目标特征信息;

将所述目标特征信息输入至所述全连接层进行决策分类,获得目标决策结果。

4. 根据权利要求1所述的方法,其特征在于,所述目标决策网络模型的训练过程包括:

获取重叠交互样本数据和所述重叠交互样本数据对应的标签决策结果,其中,所述重叠交互样本数据包括:对话重叠交互时的样本应答文本以及用户的样本语音数据和样本视频数据;

将所述重叠交互样本数据输入至预设决策网络模型中进行话语权的切换决策,并基于所述预设决策网络模型的输出,获得输出决策结果;

基于所述输出决策结果和所述标签决策结果确定训练误差,并将所述训练误差反向传播至所述预设决策网络模型中,调整所述预设决策网络模型中的网络参数;

当满足预设收敛条件时,确定预设决策网络模型训练结束,获得目标决策网络模型。

5. 根据权利要求1-4任一项所述的方法,其特征在于,所述基于所述目标决策结果对当前话语权进行切换,包括:

若所述目标决策结果为话语权保留结果,则继续播放所述目标应答语音。

6. 根据权利要求5所述的方法,其特征在于,所述继续播放所述目标应答语音,包括:

若当前决策次数小于预设次数阈值,则基于预设采集时长,对当前采集时间段进行更新,并基于更新后的当前采集时间段返回执行所述采集用户的目标语音数据和目标视频数据的操作;

若当前决策次数等于预设次数阈值,则继续播放所述目标应答语音。

7. 一种话语权切换装置,其特征在于,包括:

数据采集模块,用于若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

切换决策模块,包括:目标应答文本确定单元,用于确定所述目标应答语音对应的目标应答文本;

切换决策单元,用于将所述目标应答文本、所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

话语权切换模块,用于基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换;

所述话语权切换模块,包括:话语权切换单元,用于若所述目标决策结果为话语权切换结果,则将当前话语权切换至用户,以停止播放所述目标应答语音并等待用户表述。

8. 一种电子设备,其特征在于,所述电子设备包括:

一个或多个处理器;

存储器,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,使得所述一个或多个处理器实现如权利要求1-6中任一所述的话语权切换方法。

9. 一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现如权利要求1-6中任一所述的话语权切换方法。

一种话语权切换方法、装置、设备和存储介质

技术领域

[0001] 本发明实施例涉及计算机技术,尤其涉及一种话语权切换方法、装置、设备和存储介质。

背景技术

[0002] 随着计算机技术的快速发展,可以利用智能设备进行人机对话,从而节省人力成本。在人机对话过程中往往会出现用户打断智能设备说话的情况,从而需要及时切换话语权,使得用户获得该话语权进行表述,进而提升用户体验。

[0003] 目前,通常是智能设备在播放应答语音过程中检测是否存在用户输入语音,若存在用户输入语音,则将话语权切换给用户。

[0004] 然而,在实现本发明过程中,发明人发现现有技术中至少存在如下问题:

[0005] 智能设备中的语音采集装置可能会出现误采集用户输入语音的情况,比如用户实际上没有说话但采集到了用户语音;或者,用户实际上说话了但真正意图上不是想要打断智能设备的语音播放操作,比如用户所说的不是关于打断的内容,或者用户仅是在和其他人交流等,从而都会导致话语权的误切换,进而降低了人机对话的流畅性,降低了用户体验。

发明内容

[0006] 本发明实施例提供了一种话语权切换方法、装置、设备和存储介质,以提高话语权切换的准确性,从而提高人机对话的流畅性,提升用户体验。

[0007] 第一方面,本发明实施例提供了一种话语权切换方法,包括:

[0008] 若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

[0009] 将所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

[0010] 基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换。

[0011] 第二方面,本发明实施例还提供了一种话语权切换装置,包括:

[0012] 数据采集模块,用于若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

[0013] 切换决策模块,用于将所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

[0014] 话语权切换模块,用于基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换。

[0015] 第三方面,本发明实施例还提供了一种电子设备,所述电子设备包括:

[0016] 一个或多个处理器;

[0017] 存储器,用于存储一个或多个程序;

[0018] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如本发明任意实施例所提供的话语权切换方法。

[0019] 第四方面,本发明实施例还提供了一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如本发明任意实施例所提供的话语权切换方法。

[0020] 上述发明中的一个实施例具有如下优点或有益效果:

[0021] 通过在目标应答语音的播放过程中检测到用户输入语音时,暂停播放目标应答语音,并采集用户的目标语音数据和目标视频数据,并将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策,基于目标决策网络模型的输出确定目标决策结果,从而通过利用目标决策网络模型,基于用户的目标语音数据和目标视频数据,可以更加准确地确定出用户真实意图上是否想要打断语音播放,从而基于目标决策结果可以更加准确地进行当前话语权的切换,提高了话语权切换的准确性,进而提高了人机对话的流畅性,提升了用户体验。

附图说明

[0022] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图做一简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0023] 图1是本发明一个实施例所提供的一种话语权切换方法的流程图;

[0024] 图2是本发明一个实施例所涉及的一种目标决策网络模型的结构示例图;

[0025] 图3是本发明一个实施例所涉及的一种人机对话的示例图;

[0026] 图4是本发明一个实施例提供的另一种话语权切换方法的流程图;

[0027] 图5是本发明一个实施例所涉及的另一种目标决策网络模型的结构示例图;

[0028] 图6是本发明一个实施例提供的一种话语权切换装置的结构示意图;

[0029] 图7是本发明一个实施例提供的一种电子设备的结构示意图。

具体实施方式

[0030] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是,此处所描述的具体实施例仅仅用于解释本发明,而非对本发明的限定。另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部结构。

[0031] 图1为本发明一个实施例所提供的一种话语权切换方法的流程图,本实施例可适用于在智能设备播放应答语音的过程中进行话语权切换的情况。该方法可以由话语权切换装置来执行,该装置可以由软件和/或硬件的方式来实现,集成于电子设备中。如图1所示,该方法具体包括以下步骤:

[0032] S110、若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放目标应答语音,并采集用户的目标语音数据和目标视频数据。

[0033] 其中,目标应答语音可以是指智能设备当前正在播报的应答语音。目标应答语音可以基于用户输入的询问语音进行确定。或者,在首次交互时,也可以将智能设备的主动问

候语音作为目标应答语音,以便在首次交互的主动问候语音的播放过程中检测是否存在用户打断的情况。目标语音数据可以是指用户输入的音频数据。目标视频数据可以是指采集的用户图像的数据,比如用户全身图像数据或者用户面部图像数据。

[0034] 具体地,在每次人机交互过程中,智能设备可以基于语音采集装置采集用户输入的当前询问语音,并在当前询问语音采集结束后,可以确定出与采集的当前询问语音相匹配的目标应答语音;或者,在首次交互时,也可以将智能设备的主动问候语音作为目标应答语音,并通过播放装置将目标应答语音进行播放。智能设备可以响应于目标应答语音的开始播放指令,开始实时检测目标应答语音的播放过程中是否存在用户输入语音,以便确定出在目标应答语音的播放过程中用户是否说话。若检测到用户输入语音,则可以立即暂停目标应答语音的播放操作,并利用语音采集装置采集用户的目标语音数据以及利用摄像头采集用户的目标视频数据。

[0035] 示例性地,本实施例可以基于预设采集方式,采集用户的目标语音数据和目标视频数据。例如,可以基于预设采集总时长确定出当前采集时间段,并采集当前采集时间段内用户的目标音频数据和目标视频数据。也可以将基于预设决策次数对预设采集总时长进行划分,获得每次采集的预设采集时长,并基于预设采集时长确定出当次采集的当前采集时间段,并采集当前采集时间段内用户的目标音频数据和目标视频数据。例如,预设采集总时长为10秒,预设次数阈值为5次,则每次采集的预设采集时长为2秒,从而可以以当前时刻为开始采集时刻,预设采集总时长为时间间隔确定出当前采集时间段,并采集当前采集时间段内,也就是10秒内的用户的目标音频数据和目标视频数据,以便后续可以进行一次统一的切换决策。或者,也可以以当前时刻为开始采集时刻,预设采集时长为时间间隔确定出当前采集时间段,并采集当前采集时间段内,也就是2秒内的用户的目标音频数据和目标视频数据,以便后续可以进行5次的切换决策。

[0036] 需要说明的是,在使用本公开各实施例公开的技术方案之前,均应当依据相关法律法规通过恰当的方式对本公开所涉及个人信息类型、使用范围、使用场景等告知用户并获得用户的授权。例如,本实施例中所采集的用户语音数据和用户视频数据均是提前告知用户并获得用户授权后采集的。

[0037] S120、将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策。

[0038] 其中,目标决策网络模型可以是预先基于样本数据训练获得的,用于决策出是否切换话语权的分类网络模型。

[0039] 具体地,可以将采集的目标语音数据和目标视频数据输入至目标决策网络模型中,使得目标决策网络模型可以结合目标语音数据和目标视频数据综合确定出用户当前说话的行为是否是真正想要打断智能设备语音播报操作的行为,从而提高话语权切换的准确性。

[0040] 示例性地,图2给出了一种目标决策网络模型的结构示例图。如图2所示,目标决策网络模型可以包括:语音特征提取子模型、视频特征提取子模型和决策子模型。其中,语音特征提取子模型可以是用于提取语音特征信息的任意一种网络架构。例如,语音特征提取子模型可以是但不限于Wav2Vec网络模型。视频特征提取子模型可以是用于提取视频特征信息的任意一种网络架构。例如,视频特征提取子模型可以是但不限于Video2Vec网络模

型。决策子模型可以是基于目标语音特征信息和目标视频特征信息进行决策分类的网络模型。

[0041] 示例性地, S120可以包括:将目标语音数据输入至语音特征提取子模型中进行语音特征提取,获得提取出的目标语音特征信息;将目标视频数据输入至视频特征提取子模型中进行视频特征提取,获得提取出的目标视频特征信息;将目标语音特征信息和目标视频特征信息输入至决策子模型中进行话语权的切换决策,获得目标决策结果。

[0042] 示例性地,如图2所示,决策子模型可以包括:信息拼接层和全连接层。相应地,将目标语音特征信息和目标视频特征信息输入至决策子模型中进行话语权的切换决策,获得目标决策结果,可以包括:将目标语音特征信息和目标视频特征信息输入至信息拼接层中进行特征信息的拼接,获得拼接后的目标特征信息;将目标特征信息输入至全连接层进行决策分类,获得目标决策结果。

[0043] 具体地,信息拼接层可以将输入的目标语音特征信息和目标视频特征信息进行向量维度的拼接处理,比如,目标语音特征信息对应的特征向量维度是 $(1, 1024)$,目标视频特征信息对应的特征向量维度是 $(1, 1024)$,则拼接后的目标特征信息对应的特征向量维度为 $1024+1024=2048$ 。全连接层可以将目标特征信息映射为一个二维向量 $[x1, x2]$,其中, $x1$ 表示需要切换话语权的概率,也就是用户真实想要打断的概率, $x2$ 表示无需切换话语权的概率,也就是用户真实不想打断的概率,两者之和等于1。全连接层可以将需要切换话语权的预测概率 $x1$ 进行输出,或者也可以判断 $x1$ 是否大于预设阈值,若是,则直接输出话语权切换结果,否则输出话语权保留结果。决策子模型通过利用信息拼接层和全连接层,可以更加便捷准确地确定出目标决策结果,进而提高了话语权切换效率和准确性。

[0044] S130、基于目标决策网络模型的输出,确定目标决策结果,并基于目标决策结果对当前话语权进行切换。

[0045] 其中,目标决策结果可以包括:话语权切换结果或者话语权保留结果。当前话语权可以是指当前具有说话权利的一端。本实施例是在智能设备播报语音过程中进行的话语权切换,从而当前话语权属于智能设备,而非用户。

[0046] 具体地,目标决策网络模型可以直接输出最终的目标决策结果,或者也可以输出需要切换话语权的预测概率。在输出预测概率时,需要检测输出的预测概率是否大于预设阈值(比如0.5),若是,则确定目标决策结果为话语权切换结果,否则确定目标决策结果为话语权保留结果。根据获得的目标决策结果可以准确地确定出是否需要当前话语权进行切换,进而提高了话语权切换的准确性。

[0047] 示例性地, S130中的“基于目标决策结果对当前话语权进行切换”,可以包括:若目标决策结果为话语权切换结果,则将当前话语权切换至用户;若目标决策结果为话语权保留结果,则继续播放目标应答语音。

[0048] 具体地,在目标决策结果为话语权切换结果时,表明用户是真正想要打断智能设备的播报操作,此时可以将当前话语权切换至用户,使得用户获得当前话语权,并且智能设备停止播放并等待用户表述,从而通过及时切换当前话语权可以避免导致用户的抵触情绪,提升了用户体验。在目标决策结果为话语权保留结果时,表明用户并不是真正想要打断智能设备的播报操作,比如是因为数据采集错误或者用户只是在和其他人交流,此时可以继续播放目标应答语音,提高人机对话的流畅性。

[0049] 示例性地,继续播放目标应答语音,可以包括:若当前决策次数小于预设次数阈值,则基于预设采集时长,对当前采集时间段进行更新,并基于更新后的当前采集时间段返回执行S110中的采集用户的目标语音数据和目标视频数据的操作;若当前决策次数等于预设次数阈值,则继续播放目标应答语音。

[0050] 其中,预设次数阈值可以是预先设置的允许切换决策的总次数。预设采集时长可以是指每次决策时所需要采集的用户数据的时间长度。预设采集时长可以基于预设采集总时长和预设次数阈值进行确定的。预设采集总时长可以是预先设置的,允许采集用户数据的总时间长度。需要说明的是,采集的用户语音数据和用户视频越多,则目标决策网络模型输出的决策结果越准确,但切换耗时越长,从而无法更加及时地切换话语权。

[0051] 具体地,本实施例可以基于较短的预设采集时长进行用户数据的多次采集,并基于当前采集到的所有目标语音数据和目标视频数据进行当次的切换决策,从而可以及时获得目标决策结果为话语权切换结果的情况,进而可以更加及时地切换话语权,无需等到预设采集总时长后再进行话语权切换,所以在保证切换准确性的前提下提高了切换效率,提升了用户体验。若当前决策次数等于预设次数阈值,则表明在预设采集总时长内每次决策出的目标决策结果均为话语权保留结果,此时可以继续播放目标应答语音,避免无时限地等待,从而提高了人机对话的流畅性,提升了用户体验。

[0052] 本实施例的技术方案,通过在目标应答语音的播放过程中检测到用户输入语音时,暂停播放目标应答语音,并采集用户的目标语音数据和目标视频数据,并将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策,基于目标决策网络模型的输出确定目标决策结果,从而通过利用目标决策网络模型,基于用户的目标语音数据和目标视频数据,可以更加准确地确定出用户真实意图上是否想要打断语音播放,从而基于目标决策结果可以更加准确地进行当前话语权的切换,提高了话语权切换的准确性,进而提高了人机对话的流畅性,提升了用户体验。

[0053] 在上述技术方案的基础上,针对于图2中的目标决策网络模型而言,目标决策网络模型的训练过程可以包括:获取重叠交互样本数据和重叠交互样本数据对应的标签决策结果;将重叠交互样本数据输入至预设决策网络模型中进行话语权的切换决策,并基于预设决策网络模型的输出,获得输出决策结果;基于输出决策结果和标签决策结果确定训练误差,并将训练误差反向传播至预设决策网络模型中,调整预设决策网络模型中的网络参数;当满足预设收敛条件时,确定预设决策网络模型训练结束,获得目标决策网络模型。

[0054] 其中,重叠交互样本数据可以包括:对话重叠交互时的用户的样本语音数据和样本视频数据。图3给出了一种人机对话的示例图。本实施例可以提取原有的对话日志,并可以基于对话日志中的用户和智能设备之间的对话时间戳,确定在智能设备播放应答语音的过程中存在用户说话时的重叠会话,比如,图3的虚线框中的对话为重叠对话。通过对重叠对话中的时间戳进行比较,确定出现交集的时间段,并将该时间段中的用户语音数据和用户视频数据分别作为样本语音数据和样本视频数据,比如,样本语音数据和样本视频数据分别为图3中的用户在14s~16s内的语音数据和视频数据。本实施例可以基于样本语音数据中的语义信息以及样本视频数据中的用户动作表情等信息进行人工标注,判断出用户是否是真正想要打断用户,从而获得准确的标签决策结果。

[0055] 具体地,可以基于训练函数,根据预设决策网络模型的输出决策结果和标签决策

结果确定训练误差,并将训练误差反向传播至预设决策网络模型中,调整预设决策网络模型中的网络参数,直到满足预设收敛条件,比如迭代次数达到预设次数或者训练误差收敛时,确定预设决策网络模型训练结束,此时可以将训练结束的预设决策网络模型作为目标决策网络模型。通过利用重叠交互样本数据和相应的标签决策结果进行模型训练,可以保证目标决策网络模型切换决策的准确性,进而保证话语权切换的准确性。

[0056] 图4为本发明实施例提供的另一种话语权切换方法的流程图,本实施例在上述各实施例的基础上,对步骤“将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策”进行了优化。其中与上述各实施例相同或相应的术语的解释在此不再赘述。

[0057] 参见图4,本实施例提供的另一种话语权切换方法具体包括以下步骤:

[0058] S410、若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放目标应答语音,并采集用户的目标语音数据和目标视频数据。

[0059] S420、确定目标应答语音对应的目标应答文本。

[0060] 具体地,本实施例可以基于语音转文本工具,对目标应答语音进行转换,获得转换后的目标应答文本。本实施例也可以基于预先存储的各个应答文本以及应答文本与应答语音之间的映射关系,确定出目标应答语音所对应的目标应答文本。

[0061] S430、将目标应答文本、目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策。

[0062] 具体地,可以将目标应答文本、采集的目标语音数据和目标视频数据输入至目标决策网络模型中,使得目标决策网络模型可以同时结合目标应答文本、目标语音数据和目标视频数据综合,更加准确地确定出用户当前说话的行为是否是真正想要打断智能设备语音播报操作的行为,从而进一步提高话语权切换的准确性。

[0063] 示例性地,图5给出了一种目标决策网络模型的结构示例图。如图5所示,该目标决策网络模型可以包括:文本特征提取子模型、语音特征提取子模型、视频特征提取子模型和决策子模型。其中,文本特征提取子模型可以是用于提取文本特征信息的任意一种网络架构。例如,文本特征提取子模型可以是但不限于BERT(Bidirectional Encoder Representation from Transformers)预训练的语言表征模型。语音特征提取子模型可以是用于提取语音特征信息的任意一种网络架构。例如,语音特征提取子模型可以是但不限于Wav2Vec网络模型。视频特征提取子模型可以是用于提取视频特征信息的任意一种网络架构。例如,视频特征提取子模型可以是但不限于Video2Vec网络模型。决策子模型可以是基于目标文本特征信息、目标语音特征信息和目标视频特征信息进行决策分类的网络模型。

[0064] 示例性地,S430可以包括:将目标应答文本输入至文本特征提取子模型中进行文本特征提取,获得提取出的目标文本特征信息;将目标语音数据输入至语音特征提取子模型中进行语音特征提取,获得提取出的目标语音特征信息;将目标视频数据输入至视频特征提取子模型中进行视频特征提取,获得提取出的目标视频特征信息;将目标文本特征信息、目标语音特征信息和目标视频特征信息输入至决策子模型中进行话语权的切换决策,获得目标决策结果。

[0065] 示例性地,如图5所示,决策子模型可以包括:信息拼接层和全连接层。相应地,将

目标文本特征信息、目标语音特征信息和目标视频特征信息输入至决策子模型中进行话语权的切换决策,获得目标决策结果,可以包括:将目标文本特征信息、目标语音特征信息和目标视频特征信息输入至信息拼接层中进行特征信息的拼接,获得拼接后的目标特征信息;将目标特征信息输入至全连接层进行决策分类,获得目标决策结果。

[0066] 具体地,信息拼接层可以将输入的目标文本特征信息、目标语音特征信息和目标视频特征信息进行向量维度的拼接处理,比如,目标文本特征信息对应的特征向量维度是(1,768)、目标语音特征信息对应的特征向量维度是(1,1024),目标视频特征信息对应的特征向量维度是(1,1024),则拼接后的目标特征信息对应的特征向量维度为 $768+1024+1024=2816$ 。全连接层可以将目标特征信息映射为一个二维向量 $[x1, x2]$,其中, $x1$ 表示需要切换话语权的概率,也就是用户真实想要打断的概率, $x2$ 表示无需切换话语权的概率,也就是用户真实不想打断的概率,两者之和等于1。全连接层可以将需要切换话语权的预测概率 $x1$ 进行输出,或者也可以判断 $x1$ 是否大于预设阈值,若是,则直接输出话语权切换结果,否则输出话语权保留结果。决策子模型通过利用信息拼接层和全连接层,可以更加便捷准确地确定出目标决策结果,进而提高了话语权切换效率和准确性。

[0067] S440、基于目标决策网络模型的输出,确定目标决策结果,并基于目标决策结果对当前话语权进行切换。

[0068] 本实施例的技术方案,通过将目标应答文本、目标语音数据和目标视频数据输入至目标决策网络模型中,使得目标决策网络模型根据目标应答文本、目标语音数据和目标视频数据,可以更加准确地进行话语权的切换决策,从而可以进一步提高话语权切换的准确性。

[0069] 在上述技术方案的基础上,针对于图5中的目标决策网络模型而言,目标决策网络模型的训练过程可以包括:获取重叠交互样本数据和重叠交互样本数据对应的标签决策结果;将重叠交互样本数据输入至预设决策网络模型中进行话语权的切换决策,并基于预设决策网络模型的输出,获得输出决策结果;基于输出决策结果和标签决策结果确定训练误差,并将训练误差反向传播至预设决策网络模型中,调整预设决策网络模型中的网络参数;当满足预设收敛条件时,确定预设决策网络模型训练结束,获得目标决策网络模型。

[0070] 其中,重叠交互样本数据可以包括:对话重叠交互时的样本应答文本以及用户的样本语音数据和样本视频数据。本实施例可以提取原有的对话日志,并可以基于对话日志中的用户和智能设备之间的对话时间戳,确定在智能设备播放应答语音的过程中存在用户说话时的重叠会话。将重叠会话中的应答语音对应的应答文本作为样本应答文本,比如,图3中智能设备在12s~16s内的应答语音对应的应答文本。通过对重叠对话中的时间戳进行比较,确定出现交集的时间段,并将该时间段中的用户语音数据和用户视频数据分别作为样本语音数据和样本视频数据,比如,样本语音数据和样本视频数据分别为图3中的用户在14s~16s内的语音数据和视频数据。本实施例可以基于样本应答文本和样本语音数据中的语义信息以及样本视频数据中的用户动作表情等信息进行人工标注,判断出用户是否是真正想要打断用户,从而获得准确的标签决策结果。

[0071] 具体地,可以基于训练函数,根据预设决策网络模型的输出决策结果和标签决策结果确定训练误差,并将训练误差反向传播至预设决策网络模型中,调整预设决策网络模型中的网络参数,直到满足预设收敛条件,比如迭代次数达到预设次数或者训练误差收敛

时,确定预设决策网络模型训练结束,此时可以将训练结束的预设决策网络模型作为目标决策网络模型,从而通过模型训练可以保证目标决策网络模型切换决策的准确性,进一步保证话语权切换的准确性。

[0072] 以下是本发明实施例提供的话语权切换装置的实施例,该装置与上述各实施例的话语权切换方法属于同一个发明构思,在话语权切换装置的实施例中未详尽描述的细节内容,可以参考上述话语权切换方法的实施例。

[0073] 图6为本发明实施例提供的一种话语权切换装置的结构示意图,本实施例可适用于在智能设备播放应答语音的过程中进行话语权切换的情况。如图6所示,该装置具体包括:数据采集模块610、切换决策模块620和话语权切换模块630。

[0074] 其中,数据采集模块610,用于若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;切换决策模块620,用于将所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;话语权切换模块630,用于基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换。

[0075] 本实施例的技术方案,通过在目标应答语音的播放过程中检测到用户输入语音时,暂停播放目标应答语音,并采集用户的目标语音数据和目标视频数据,并将目标语音数据和目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策,基于目标决策网络模型的输出确定目标决策结果,从而通过利用目标决策网络模型,基于用户的目标语音数据和目标视频数据,可以更加准确地确定出用户真实意图上是否想要打断语音播放,从而基于目标决策结果可以更加准确地进行当前话语权的切换,提高了话语权切换的准确性,进而提高了人机对话的流畅性,提升了用户体验。

[0076] 可选地,切换决策模块620,包括:

[0077] 目标应答文本确定单元,用于确定所述目标应答语音对应的目标应答文本;

[0078] 切换决策单元,用于将所述目标应答文本、所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策。

[0079] 可选地,所述目标决策网络模型包括:文本特征提取子模型、语音特征提取子模型、视频特征提取子模型和决策子模型;

[0080] 切换决策单元,具体用于:将所述目标应答文本输入至所述文本特征提取子模型中进行文本特征提取,获得提取出的目标文本特征信息;将所述目标语音数据输入至所述语音特征提取子模型中进行语音特征提取,获得提取出的目标语音特征信息;将所述目标视频数据输入至所述视频特征提取子模型中进行视频特征提取,获得提取出的目标视频特征信息;将所述目标文本特征信息、所述目标语音特征信息和所述目标视频特征信息输入至所述决策子模型中进行话语权的切换决策,获得目标决策结果。

[0081] 可选地,所述决策子模型包括:信息拼接层和全连接层;

[0082] 切换决策单元,还具体用于:将所述目标文本特征信息、所述目标语音特征信息和所述目标视频特征信息输入至所述信息拼接层中进行特征信息的拼接,获得拼接后的目标特征信息;将所述目标特征信息输入至所述全连接层进行决策分类,获得目标决策结果。

[0083] 可选地,该装置还包括:目标决策网络模型训练模块,用于:

[0084] 获取重叠交互样本数据和所述重叠交互样本数据对应的标签决策结果,其中,所

述重叠交互样本数据包括:对话重叠交互时的样本应答文本以及用户的样本语音数据和样本视频数据;将所述重叠交互样本数据输入至预设决策网络模型中进行话语权的切换决策,并基于所述预设决策网络模型的输出,获得输出决策结果;基于所述输出决策结果和所述标签决策结果确定训练误差,并将所述训练误差反向传播至所述预设决策网络模型中,调整所述预设决策网络模型中的网络参数;当满足预设收敛条件时,确定预设决策网络模型训练结束,获得目标决策网络模型。

[0085] 可选地,话语权切换模块630,包括:

[0086] 话语权切换单元,用于若所述目标决策结果为话语权切换结果,则将当前话语权切换至用户;

[0087] 话语权保留单元,用于若所述目标决策结果为话语权保留结果,则继续播放所述目标应答语音。

[0088] 可选地,话语权保留单元,具体用于:

[0089] 若当前决策次数小于预设次数阈值,则基于预设采集时长,对当前采集时间段进行更新,并基于更新后的当前采集时间段返回执行所述采集用户的目标语音数据和目标视频数据的操作;若当前决策次数等于预设次数阈值,则继续播放所述目标应答语音。

[0090] 本发明实施例所提供的话语权切换装置可执行本发明任意实施例所提供的话语权切换方法,具备执行话语权切换方法相应的功能模块和有益效果。

[0091] 值得注意的是,上述话语权切换装置的实施例中,所包括的各个单元和模块只是按照功能逻辑进行划分的,但并不局限于上述的划分,只要能够实现相应的功能即可;另外,各功能单元的具体名称也只是为了便于相互区分,并不用于限制本发明的保护范围。

[0092] 图7为本发明实施例提供的一种电子设备的结构示意图。图7示出了适于用来实现本发明实施方式的示例性电子设备12的框图。图7显示的电子设备12仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0093] 如图7所示,电子设备12以通用计算设备的形式表现。电子设备12的组件可以包括但不限于:一个或者多个处理器或者处理单元16,系统存储器28,连接不同系统组件(包括系统存储器28和处理单元16)的总线18。

[0094] 总线18表示几类总线结构中的一种或多种,包括存储器总线或者存储器控制器,外围总线,图形加速端口,处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说,这些体系结构包括但不限于工业标准体系结构(ISA)总线,微通道体系结构(MAC)总线,增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0095] 电子设备12典型地包括多种计算机系统可读介质。这些介质可以是任何能够被电子设备12访问的可用介质,包括易失性和非易失性介质,可移动的和不可移动的介质。

[0096] 系统存储器28可以包括易失性存储器形式的计算机系统可读介质,例如随机存取存储器(RAM) 30和/或高速缓存存储器32。电子设备12可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例,存储系统34可以用于读写不可移动的、非易失性磁介质(图7未显示,通常称为“硬盘驱动器”)。尽管图7中未示出,可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器,以及对可移动非易失性光盘(例如CD-ROM, DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下,每个驱动器可以通过一个或者多个数据介质接口与总线18相连。系统存储器28可以包括至少一个程序

产品,该程序产品具有一组(例如至少一个)程序模块,这些程序模块被配置以执行本发明各实施例的功能。

[0097] 具有一组(至少一个)程序模块42的程序/实用工具40,可以存储在例如系统存储器28中,这样的程序模块42包括但不限于操作系统、一个或者多个应用程序、其它程序模块以及程序数据,这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块42通常执行本发明所描述的实施例中的功能和/或方法。

[0098] 电子设备12也可以与一个或多个外部设备14(例如键盘、指向设备、显示器24等)通信,还可与一个或者多个使得用户能与该电子设备12交互的设备通信,和/或与使得该电子设备12能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口22进行。并且,电子设备12还可以通过网络适配器20与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器20通过总线18与电子设备12的其它模块通信。应当明白,尽管图中未示出,可以结合电子设备12使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0099] 处理单元16通过运行存储在系统存储器28中的程序,从而执行各种功能应用以及数据处理,例如实现本发实施例所提供的一种话语权切换方法步骤,该方法包括:

[0100] 若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

[0101] 将所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

[0102] 基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换。

[0103] 当然,本领域技术人员可以理解,处理器还可以实现本发明任意实施例所提供的话语权切换方法的技术方案。

[0104] 本实施例提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现如本发明任意实施例所提供的话语权切换方法步骤,该方法包括:

[0105] 若在目标应答语音的播放过程中检测到用户输入语音,则暂停播放所述目标应答语音,并采集用户的目标语音数据和目标视频数据;

[0106] 将所述目标语音数据和所述目标视频数据输入至预先训练获得的目标决策网络模型中进行话语权的切换决策;

[0107] 基于所述目标决策网络模型的输出,确定目标决策结果,并基于所述目标决策结果对当前话语权进行切换。

[0108] 本发明实施例的计算机存储介质,可以采用一个或多个计算机可读的介质的任意组合。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。计算机可读存储介质例如可以是但不限于: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(非穷举的列表)包括:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光

存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中,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。

[0109] 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程序代码。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。

[0110] 计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0111] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言,诸如Java、Smalltalk、C++,还包括常规的过程式程序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络,包括局域网(LAN)或广域网(WAN),连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0112] 本领域普通技术人员应该明白,上述的本发明的各模块或各步骤可以用通用的计算装置来实现,它们可以集中在单个计算装置上,或者分布在多个计算装置所组成的网络上,可选地,它们可以用计算机装置可执行的程序代码来实现,从而可以将它们存储在存储装置中由计算装置来执行,或者将它们分别制作成各个集成电路模块,或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样,本发明不限制于任何特定的硬件和软件的结合。

[0113] 注意,上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解,本发明不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此,虽然通过以上实施例对本发明进行了较为详细的说明,但是本发明不仅仅限于以上实施例,在不脱离本发明构思的情况下,还可以包括更多其他等效实施例,而本发明的范围由所附的权利要求范围决定。

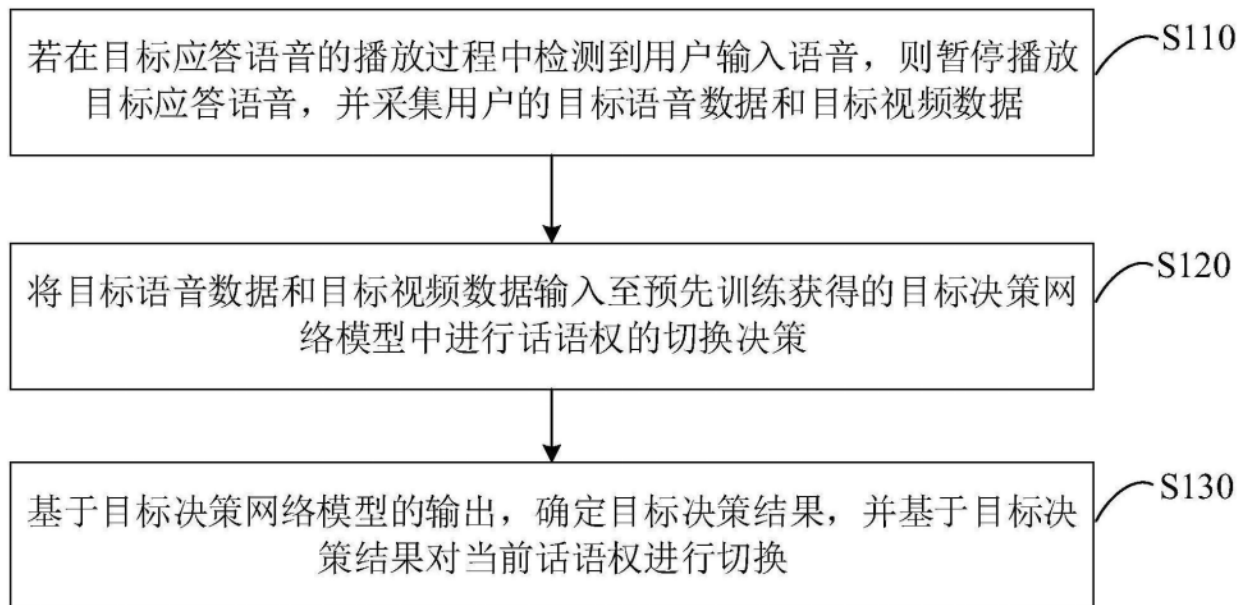


图1

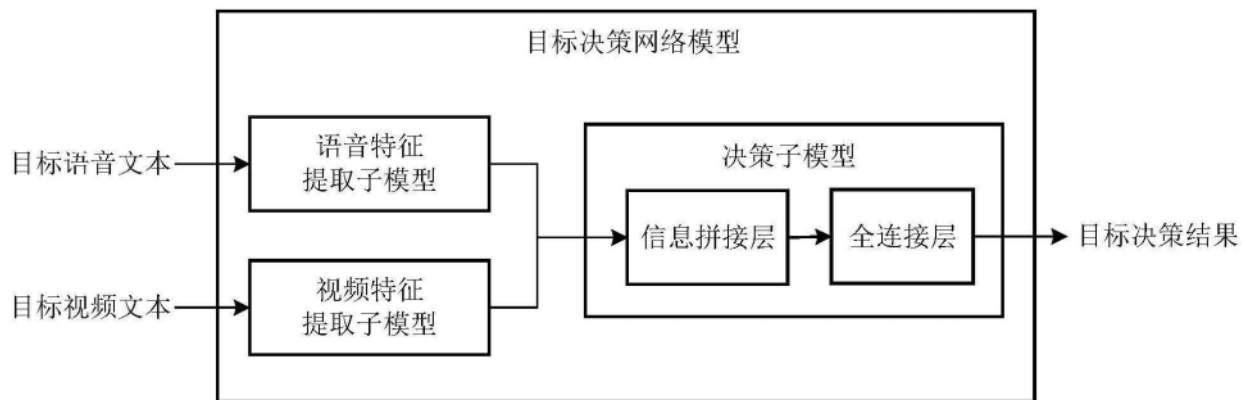


图2

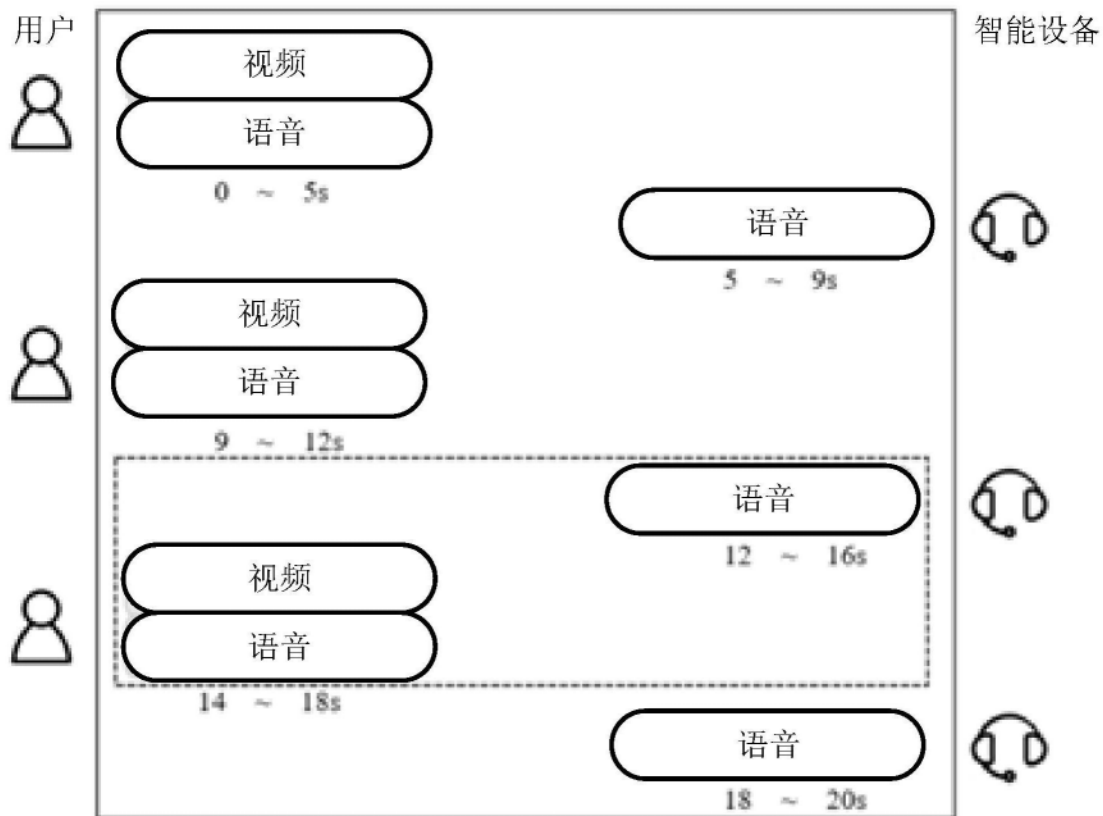


图3

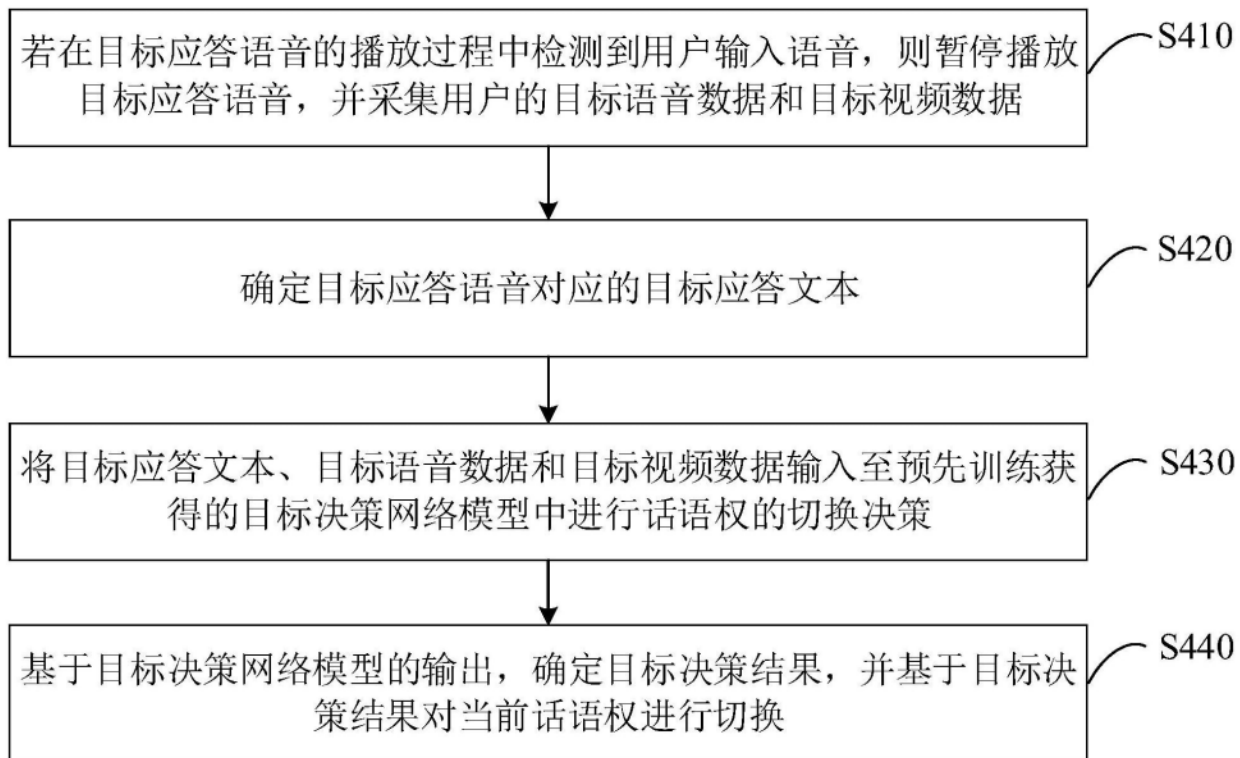


图4

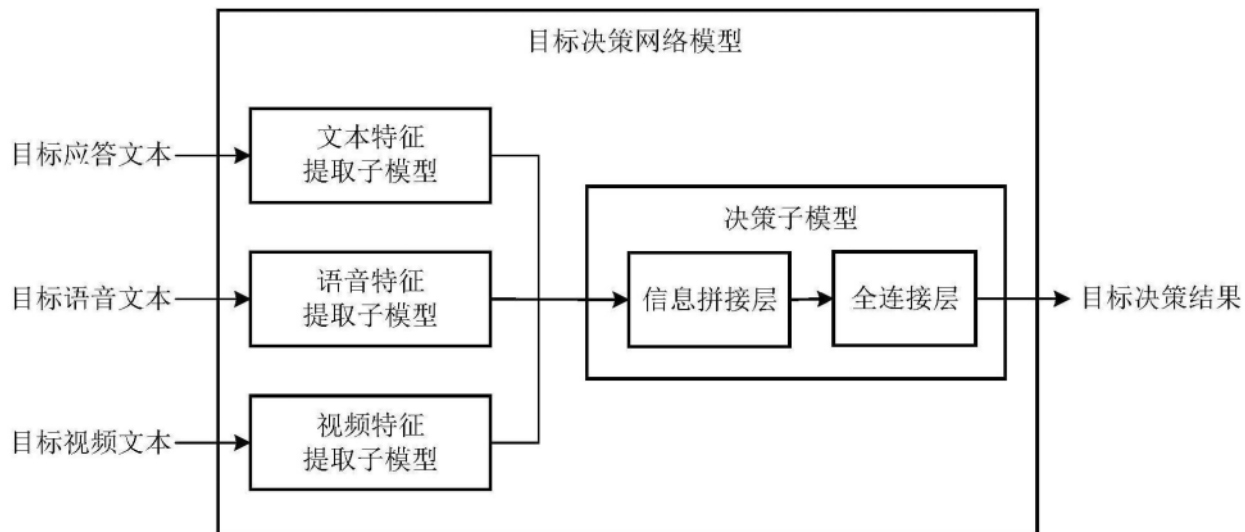


图5

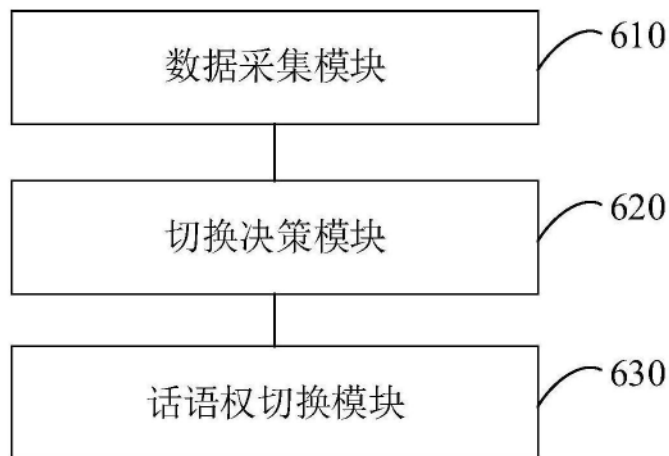


图6

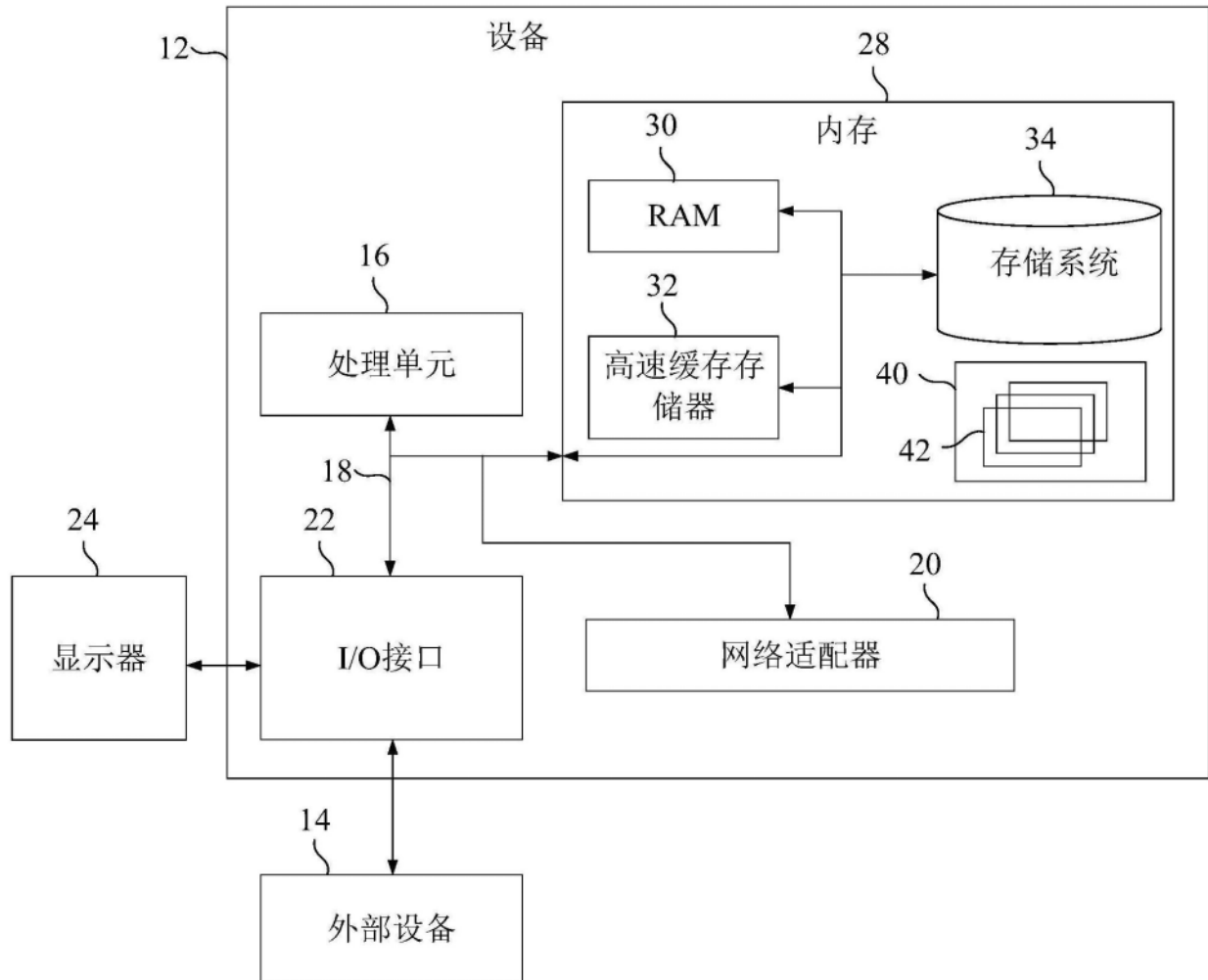


图7