(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114360535 B
(45)授权公告日 2023.01.31

(21)申请号 202111601277.6
(22)申请日 2021.12.24
(65)同一申请的已公布的文献号
申请公布号 CN 114360535 A
(43)申请公布日 2022.04.15
(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦二层
(72)发明人 吴文权 吴华
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 杜月
(51)Int.Cl.
*G10L 15/22*(2006.01)
*G10L 15/02*(2006.01)
*G10L 15/26*(2006.01)
*G10L 13/02*(2013.01)
*G10L 13/08*(2013.01)
*G10L 25/63*(2013.01)

(56)对比文件
CN 109215679 A,2019.01.15
CN 109215679 A,2019.01.15
CN 113434647 A,2021.09.24
CN 112786047 A,2021.05.11
CN 112286366 A,2021.01.29
CN 112434139 A,2021.03.02
CN 112528004 A,2021.03.19
US 2021280190 A1,2021.09.09

审查员 田树雪

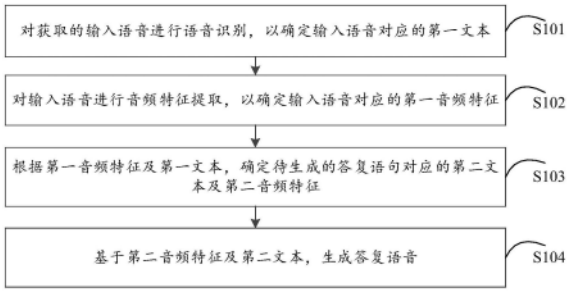
权利要求书3页 说明书11页 附图3页

(54)发明名称

# 语音对话的生成方法、装置、电子设备及存储介质

(57)摘要

本公开提供了一种语音对话的生成方法、装置、电子设备及存储介质，涉及计算机技术领域，尤其涉及语音技术、自然语言处理、计算机视觉等人工智能技术领域。具体实现方案为：对获取的输入语音进行语音识别，以确定输入语音对应的第一文本；对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征；根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征；基于第二音频特征及第二文本，生成答复语音。由此，根据输入语音对应的第一音频特征及第一文本，确定第二文本及第二音频特征，从而不仅提高了确定的第二文本的准确性，而且使生成的答复语音更加贴合输入语音对应的说话者的情绪。




CN 114360535 B

1.一种语音对话的生成方法,包括:

对获取的输入语音进行语音识别,以确定所述输入语音对应的第一文本;

对所述输入语音进行音频特征提取,以确定所述输入语音对应的第一音频特征;

根据所述第一音频特征及所述第一文本,确定待生成的答复语句对应的第二文本及第二音频特征,其中,将所述第一音频特征及第一文本输入预设的对话模型中,以获取所述待生成的答复语句对应的第二文本及第二音频特征;

基于所述第二音频特征及所述第二文本,生成答复语音;

所述预设的对话模型的生成方法包括:获取训练样本集,其中,所述训练样本集中包括输入文本及对应的音频特征,所述输入文本对应的标注答复文本及对应的音频特征标签,将所述输入文本及对应的音频特征输入初始对话模型中,以获取所述初始对话模型输出的预测答复文本及对应的预测音频特征,根据所述预测答复文本与标注答复文本之间的差异,及所述预测音频特征与音频特征标签之间的差异对所述初始对话模型进行修正,以生成所述预设的对话模型,其中,所述音频特征包括频率特征和幅值特征;

所述基于所述第二音频特征及所述第二文本,生成答复语音,包括:

获取所述输入语音对应的场景图像,所述场景图像包括所述输入语音的说话者所处的场景;

对所述场景图像进行视觉特征提取,以确定所述场景图像对应的视觉特征,其中,所述视觉特征为所述场景图像中包含的场景特征;

根据所述视觉特征,对所述第二文本和/或第二音频特征进行修正;

基于修正后的第二音频特征及所述第二文本,生成答复语音;

所述确定所述场景图像对应的视觉特征,包括:

对所述场景图像进行自动分割,划分出所述场景图像中包含的对象或颜色区域,对每个图像子块提取特征,并建立索引,从而得到所述场景图像中每个物体对应的空间关系特征,基于所述每个物体的种类及空间关系,确定所述场景图像对应的视觉特征。

2.如权利要求1所述的方法,其中,所述对所述输入语音进行音频特征提取,以确定所述输入语音对应的第一音频特征,包括:

根据所述输入语音中每帧语音对应的第一幅值,确定所述输入语音对应的第二幅值;

根据所述第二幅值所属的范围,确定所述输入语音对应的幅值特征。

3.如权利要求2所述的方法,其中,所述对所述输入语音进行音频特征提取,以确定所述输入语音对应的第一音频特征,包括:

对所述输入语音进行基音检测,以确定语音信号对应的频率值;

根据所述频率值所属的范围,确定所述输入语音对应的频率特征。

4.如权利要求1所述的方法,其中,所述获取所述输入语音对应的场景图像,包括:

响应于监测到采集的语音数据中包含用户语音的情况下,启动图像采集组件,以获取所述输入语音对应的场景图像;

或者,根据所述输入语音的获取时间,从采集的视频流中截取与所述输入语音对应的场景图像。

5.如权利要求1-3任一所述的方法,其中,在所述确定所述输入语音对应的第一音频特征之后,还包括:

根据所述第一音频特征及所述第一文本，确定待生成的答复语句对应的第二文本及所述第二文本中包含的表情符号；

在交互设备的显示屏幕上，显示所述第二文本及所述表情符号。

6.一种语音对话的生成装置，包括：

第一确定模块，用于对获取的输入语音进行语音识别，以确定所述输入语音对应的第一文本；

第二确定模块，用于对所述输入语音进行音频特征提取，以确定所述输入语音对应的第一音频特征；

第三确定模块，用于根据所述第一音频特征及所述第一文本，确定待生成的答复语句对应的第二文本及第二音频特征，其中，将所述第一音频特征及第一文本输入预设的对话模型中，以获取所述待生成的答复语句对应的第二文本及第二音频特征；

生成模块，用于基于所述第二音频特征及所述第二文本，生成答复语音；

所述预设的对话模型的生成方法包括：获取训练样本集，其中，所述训练样本集中包括输入文本及对应的音频特征，所述输入文本对应的标注答复文本及对应的音频特征标签，将所述输入文本及对应的音频特征输入初始对话模型中，以获取所述初始对话模型输出的预测答复文本及对应的预测音频特征，根据所述预测答复文本与标注答复文本之间的差异，及所述预测音频特征与音频特征标签之间的差异对所述初始对话模型进行修正，以生成所述预设的对话模型，其中，所述音频特征包括频率特征和幅值特征；

其中，所述生成模块，包括：

第一获取单元，用于获取所述输入语音对应的场景图像，所述场景图像包括所述输入语音的说话者所处的场景；

第一确定单元，用于对所述场景图像进行视觉特征提取，以确定所述场景图像对应的视觉特征，其中，所述视觉特征为所述场景图像中包含的场景特征；

修正单元，用于根据所述视觉特征，对所述第二文本和/或第二音频特征进行修正；

生成单元，用于基于修正后的第二音频特征及所述第二文本，生成答复语音；

所述确定所述场景图像对应的视觉特征，包括：

对所述场景图像进行自动分割，划分出所述场景图像中包含的对象或颜色区域，对每个图像子块提取特征，并建立索引，从而得到所述场景图像中每个物体对应的空间关系特征，基于所述每个物体的种类及空间关系，确定所述场景图像对应的视觉特征。

7.如权利要求6所述的装置，其中，所述第二确定模块，具体用于：

根据所述输入语音中每帧语音对应的第一幅值，确定所述输入语音对应的第二幅值；

根据所述第二幅值所属的范围，确定所述输入语音对应的幅值特征。

8.如权利要求7所述的装置，其中，所述第二确定模块，具体用于：

对所述输入语音进行基音检测，以确定语音信号对应的频率值；

根据所述频率值所属的范围，确定所述输入语音对应的频率特征。

9.如权利要求6所述的装置，其中，所述第一获取单元，具体用于：

响应于监测到采集的语音数据中包含用户语音的情况下，启动图像采集组件，以获取所述输入语音对应的场景图像；

或者，根据所述输入语音的获取时间，从采集的视频流中截取与所述输入语音对应的

场景图像。

10.如权利要求6-8任一所述的装置,其中,还包括:

第四确定模块,用于根据所述第一音频特征及所述第一文本,确定待生成的答复语句对应的第二文本及所述第二文本中包含的表情符号;

显示模块,用于在交互设备的显示屏幕上,显示所述第二文本及所述表情符号。

11.一种电子设备,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指令,所述指令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1-5中任一项所述的方法。

12.一种存储有计算机指令的非瞬时计算机可读存储介质,其中,所述计算机指令用于使所述计算机执行权利要求1-5中任一项所述的方法。

13.一种计算机程序产品,包括计算机指令,所述计算机指令被处理器执行时实现权利要求1-5中任一项所述方法的步骤。

# 语音对话的生成方法、装置、电子设备及存储介质

## 技术领域

[0001] 本公开涉及计算机技术领域，尤其涉及语音技术、自然语言处理、计算机视觉等人工智能技术领域，具体涉及一种语音对话的生成方法、装置、电子设备及存储介质。

## 背景技术

[0002] 随着人工智能技术地不断发展和完善，其已经在与人类日常生活相关的各个领域扮演着极其重要的作用。例如，人工智能已经在语音对话领域取得显著的进步。相关技术中，可以将语音信息转化为文本，并对文本进行语义分析以确定答复文本。由于相关技术中仅根据语音信息中包含的文本这一单一的特征，确定答复文本，从而可能导致最终确定的答复文本的准确性较低，因此，如何提高答复语句的准确性成为重点的研究方向。

## 发明内容

[0003] 本公开提供了一种语音对话的生成方法、装置、电子设备及存储介质。

[0004] 根据本公开的第一方面，提供了一种语音对话的生成方法，包括：

[0005] 对获取的输入语音进行语音识别，以确定所述输入语音对应的第一文本；

[0006] 对所述输入语音进行音频特征提取，以确定所述输入语音对应的第一音频特征；

[0007] 根据所述第一音频特征及所述第一文本，确定待生成的答复语句对应的第二文本及第二音频特征；

[0008] 基于所述第二音频特征及所述第二文本，生成答复语音。

[0009] 根据本公开的第二方面，提供了一种语音对话的生成装置，包括：

[0010] 第一确定模块，用于对获取的输入语音进行语音识别，以确定所述输入语音对应的第一文本；

[0011] 第二确定模块，用于对所述输入语音进行音频特征提取，以确定所述输入语音对应的第一音频特征；

[0012] 第三确定模块，用于根据所述第一音频特征及所述第一文本，确定待生成的答复语句对应的第二文本及第二音频特征；

[0013] 生成模块，用于基于所述第二音频特征及所述第二文本，生成答复语音。

[0014] 根据本公开的第三方面，提供了一种电子设备，包括：

[0015] 至少一个处理器；以及

[0016] 与所述至少一个处理器通信连接的存储器；其中，

[0017] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令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如第一方面所述的语音对话的生成方法。

[0018] 根据本公开第四方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行如第一方面所述的语音对话的生成方法。

[0019] 根据本公开的第五方面，提供了一种计算机程序产品，包括计算机指令，所述计算

机指令在被处理器执行时实现如第一方面所述的语音对话的生成方法的步骤。

[0020] 本公开提供的语音对话的生成方法、装置、电子设备及存储介质，存在如下有益效果：

[0021] 本公开实施例中，先对获取的输入语音进行语音识别，以确定输入语音对应的第一文本，之后对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征，再根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征，最后基于第二音频特征及第二文本，生成答复语音。由此，根据输入语音对应的第一音频特征及第一文本，确定答复语句对应的第二文本及第二音频特征，从而不仅提高了确定的第二文本的准确性，而且可以根据输入语句对应的情绪特征确定答复语句的情绪特征，从而使生成的答复语音更加贴合输入语音对应的说话者的情绪。

[0022] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

## 附图说明

[0023] 附图用于更好地理解本方案，不构成对本公开的限定。其中：

[0024] 图1是根据本公开一实施例提供的一种语音对话的生成方法的流程示意图；

[0025] 图2是根据本公开又一实施例提供的一种语音对话的生成方法的流程示意图；

[0026] 图3是根据本公开又一实施例提供的一种语音对话的生成方法的流程示意图；

[0027] 图4是根据本公开一实施例提供的一种语音对话的生成装置的结构示意图；

[0028] 图5是用来实现本公开实施例的语音对话的生成方法的电子设备的框图。

## 具体实施方式

[0029] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0030] 本公开实施例涉及计算机视觉、深度学习等人工智能技术领域。

[0031] 人工智能(Artificial Intelligence)，英文缩写为AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

[0032] 语音技术在计算机领域中的关键技术有自动语音识别技术(Automatic Speech Recognition,ASR)和语音合成技术(Text To Speech,TTS)。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，语音比其他的交互方式有更多的优势。

[0033] 自然语言处理是用计算机来处理、理解以及运用人类语言(如中文、英文等)，它是计算机科学与语言学的交叉学科，又常被称为计算语言学。由于自然语言是人类区别于其他动物的根本标志。没有语言，人类的思维也就无从谈起，所以自然语言处理体现了人工智能的最高任务与境界，也就是说，只有当计算机具备了处理自然语言的能力时，机器才算实现了真正的智能。

[0034] 计算机视觉，指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视

觉,并进一步做图形处理,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。

[0035] 本公开的技术方案中,所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0036] 图1是根据本公开一实施例提供的一种语音对话的生成方法的流程示意图;

[0037] 其中,需要说明的是,本实施例的语音对话的生成方法的执行主体为语音对话的生成装置,该装置可以由软件和/或硬件的方式实现,该装置可以配置在电子设备中,电子设备可以包括但不限于终端、服务器端等。

[0038] 如图1所示,该语音对话的生成方法包括:

[0039] S101:对获取的输入语音进行语音识别,以确定输入语音对应的第一文本。

[0040] 其中,获取的输入语音可以为需要根据语音中包含的内容生成相应的答复文本的语音。输入语音可以为一段连续的语音,例如一个句子、一段话等。

[0041] 可选的,可以通过语音采集设备,例如麦克风、声音传感器等获取输入语音,还可以通过从存储语音的存储空间中读取输入语音,本实施例对输入语音的获取方式不做限制。

[0042] 其中,第一文本是指输入语音中包含的文本,即将输入语音中包含的内容用文本的形式显示。

[0043] 本公开实施例中,语音识别用于把输入语音对应的语音信号转变为对应的第一文本。可选的,可以采用隐马尔可夫模型(Hidden Markov Model,HMM)对输入语音进行语音识别,以确定输入语音对应的第一文本;或者,也可以通过将获取的输入语音与语音数据库中语音进行比对,找到相同的语音,进而得到语音数据库中语音对应的话语文本作为输入语音对应的第一文本。本公开对此不做限定。

[0044] S102:对输入语音进行音频特征提取,以确定输入语音对应的第一音频特征。

[0045] 其中,第一音频特征可以为输入语音对应的语音信号的频率,幅值等信息。

[0046] 需要说明的是,语音信号的频率、幅值等特征可以反映出输入语音对应的说话者的情绪信息。比如,输入语音对应的语音信号的频率较高,表示说话者语速较快,情绪可能较为急躁;语音信号的频率正常时,表示说话者的情绪可能较为轻松。语音信号的幅值较高时,表示说话者的声音较大,情绪可能较为高涨时。语音信号对应的幅值较低时,表示说话者的声音较小,情绪可能较为低迷。

[0047] 可选的,可以采用快速傅里叶变换对输入语音进行音频特征提取,以确定输入语音对应的频率、幅值等。或者,也可以使用matlab工具中的max函数提取输入语音对应的幅值,使用pitch函数提取输入语音中的频率。本公开对此不做限定。

[0048] S103:根据第一音频特征及第一文本,确定待生成的答复语句对应的第二文本及第二音频特征。

[0049] 需要说明的是,本公开实施例中,在根据输入语音的第一音频特征,及第一文本确定答复语句时,不仅可以确定答复语句对应的第二文本,而且可以同时确定答复语句的第二音频特征,即可以确定播放答复语句时的情绪。

[0050] 其中,第二文本可以为根据第一音频特征及第一文本生成的,用于答复输入语音的文本。

[0051] 其中,第二音频特征可以为根据第一音频特征确定的,播放第二文本时的情绪特

征。比如，第一音频特征为频率较高、幅值也较高，表示输入语句对应的说话者的情绪较为暴躁，因此，答复语句对应的第二音频特征可以为频率适中、幅值适中，即采用较为舒缓的语调播放第二文本。

[0052] 可选的，可以将第一音频特征及第一文本输入预设的对话模型中，以获取待生成的答复语句对应的第二文本及第二音频特征。

[0053] 或者，也可以先提取第一文本中包含的关键词，之后根据第一文本中包含的关键词及第一音频特征，确定待生成的答复语句对应的第二文本及第二音频特征。

[0054] 本公开实施例中，根据输入语音对应的第一音频特征及第一文本，确定答复语句的第二文本，从而在输入语音对应的第一文本相同的情况下，若输入语音对应的第一音频特征不同，生成的答复语句对应的第二文本也不同，从而不仅提高了答复语句的准确性，而且使确定的答复语句更加贴合输入语音对应的说话者的情绪，提高了答复文本的多样性。

[0055] S104：基于第二音频特征及第二文本，生成答复语音。

[0056] 其中，答复语音为采用第二音频特征播放第二文本得到的语音。

[0057] 可选的，可以采用语音合成技术(Text to Speech，TTS)，将第二文本及第二音频特征相结合，生成答复语音。

[0058] 本公开实施例中，先对获取的输入语音进行语音识别，以确定输入语音对应的第一文本，之后对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征，再根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征，最后基于第二音频特征及第二文本，生成答复语音。由此，根据输入语音对应的第一音频特征及第一文本，确定答复语句对应的第二文本及第二音频特征，从而不仅提高了确定的第二文本的准确性，而且可以根据输入语句对应的情绪特征确定答复语句的情绪特征，从而使生成的答复语音更加贴合输入语音对应的说话者的情绪。

[0059] 图2是根据本公开又一实施例提供的一种语音对话的生成方法的流程示意图。如图2所示，该语音对话的生成方法包括：

[0060] S201：对获取的输入语音进行语音识别，以确定输入语音对应的第一文本。

[0061] 其中，步骤S201的具体实现形式，可参照本公开中其他各实施例中的详细描述，此处不再详细赘述。

[0062] S202：对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征。

[0063] 其中，第一音频特征可以包括幅值特征和频率特征。

[0064] 可选的，可以先根据输入语音中每帧语音对应的第一幅值，确定输入语音对应的第二幅值，之后根据第二幅值所属的范围，确定输入语音对应的幅值特征。

[0065] 其中，第一幅值可以为每帧语音对应的幅值中的最大值。

[0066] 其中，第二幅值可以为每帧语音对应的第一幅值中的最大值。即将输入语音对应的最大幅值作为输入语音对应的第二幅值。

[0067] 其中，幅值特征可以包括：高幅值、中幅值及低幅值等，本公开对此不做限定。需要说明的是，每个幅值特征对应不同的幅值范围，本公开实施例中，可以根据第二幅值所属的范围，确定输入语音对应的幅值特征。

[0068] 本公开实施例中，在获取输入语音中每帧语音对应的第一幅值之前，可以先对输入语音进行分帧处理，即将第一音频数据切分为固定长度的小段。

[0069] 可选的，可以采用任何可取的方式，确定输入语音中每帧语音对应的第一幅值，本公开对此不做限定。比如，可以将输入语音中的每帧语音进行傅里叶变换，以获取每帧语音对应的第一幅值。

[0070] 本公开实施例中，先根据输入语音中每帧语音对应的第一幅值，确定输入语音对应的第二幅值，之后根据第二幅值所属的范围，确定输入语音为高幅值、中幅值、或低幅值，从而可以用一个幅值特征表征输入音频对应的多个幅值，从而在不影响后续获取第二文本的准确性的情况下，降低了后续处理的计算量。

[0071] 可选的，可以对输入语音进行基音检测，以确定语音信号对应的频率值，之后根据频率值所属的范围，确定输入语音对应的频率特征。

[0072] 其中，频率特征可以包括：高频、中频及低频等，本公开对此不做限定。需要说明的是，每个频率特征对应不同的频率范围，本公开实施例中，可以根据频率值所属的范围，确定输入语音对应的频率特征。

[0073] 可选的，可以对输入语音进行基音检测，以获取输入语音对应的最大频率，并将最大的频率作为语音信号对应的频率值。或者，也可以将输入语音对应的平均频率作为语音信号对应的频率值。本公开对此不做限定。

[0074] 本公开实施例中，先对输入语音进行基音检测，以确定语音信号对应的频率值，之后根据频率值所属的范围，确定输入语音为高频率、中频率或低频率，从而可以用一个幅值特征代表输入音频对应的多个幅值，从而在不影响后续获取第二文本的准确性的情况下，降低了后续处理的计算量。

[0075] S203：根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二文本中包含的表情符号。

[0076] 其中，表情符号可以包括：开心、惊讶、害怕、担心、和蔼等，本公开对此不做限定。

[0077] 需要说明的是，第二文本中包含的表情符号可以为一个，也可以为多个，本公开对此不做限定。比如，第二文本中的第一句话对应一个表情符号，第二句话、及第三句话对应一个表情符号等。

[0078] 举例来说，若第一音频特征中包含的幅值特征为低频、频率特征也为低频，进而确定输入语音对应的说话者的情绪为悲伤，则生成的第二文本中包含的表情符号可以为担心。

[0079] 需要说明的是，上述示例只是简单的举例说明，不能作为本公开实施例中第一音频特征、第二文本中包含的表情符号的具体限定。

[0080] S204：在交互设备的显示屏幕上，显示第二文本及表情符号。

[0081] 其中，交互设备为可以与用户实现交互的电子设备。交互设备可以通过接收用户的交互请求，并对交互请求进行处理，以生成交互请求对应的结果，进而通过语音、文本等形式向用户展示结果。

[0082] 本公开实施例中，在确定了输入语音对应答复语句的第二文本及第二文本中包含的表情符号之后，可以在交互设备的显示屏幕上，显示第二文本及表情符号，从而可以使用户结合显示界面中包含的表情符号，阅读第二文本，从而不仅答复了用户的请求，从而实现了多角度的与用户的有效沟通。

[0083] 可选的，本公开实施例中，不仅可以在交互设备的显示屏幕上，显示第二文本及表

情符号，而且也可以使交互设备采用表情符号对应的语调播放第二文本。

[0084] 举例来说，若第二文本中包含的表情符号为担心，则交互设备不仅可以在显示屏幕上显示第二文本及担心的表情符号，而且可以用安慰的语调播放第二文本。

[0085] 本公开实施例中，先对获取的输入语音进行语音识别，以确定输入语音对应的第一文本，之后对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征，再根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二文本中包含的表情符号，最后在交互设备的显示屏幕上，显示第二文本及表情符号。由此，根据输入语音对应的第一音频特征及第一文本，确定答复语句对应的第二文本及第二文本中包含的表情符号，从而可以使确定的第二文本更加准确，而且还可以在交互设备的显示屏幕中显示第二语句及对应的表情符号，从而实现了多角度的与用户的有效沟通。

[0086] 图3是根据本公开又一实施例提供的一种语音对话的生成方法的流程示意图。如图3所示，该语音对话的生成方法包括：

[0087] S301：对获取的输入语音进行语音识别，以确定输入语音对应的第一文本。

[0088] S302：对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征。

[0089] 其中，步骤S301、步骤S302的具体实现形式，可参照本公开中其他各实施例中的详细描述，此处不再详细赘述。

[0090] S303：根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征。

[0091] 本公开实施例中，可以将第一音频特征及第一文本输入预设的对话模型中，以获取待生成的答复语句对应的第二文本及第二音频特征。

[0092] 可选的，获取预设的对话模型的具体步骤可以包括：获取训练样本集，其中，训练样本集中包含输入文本及对应的音频特征，输入文本对应的标注答复文本及对应的音频特征标签，之后将输入文本及对应的音频特征输入初始对话模型中，以获取初始对话模型输出的预测答复文本及对应的预测音频特征，之后再根据预测答复文本与标注答复文本之间的差异，及预测音频特征与音频特征标签之间的差异对初始对话模型进行修正，以生成预设的对话模型。

[0093] 可选的，训练样本集可以通过以下方式获得：首先从网络信息中自动挖掘大量的文本对话语料，并对文本对话语料进行人工配音，之后对配音的样本语音数据进行音频特征提取，以获取文本对话语料中包含的输入文本及对应的音频特征，标注答复文本及对应的音频特征标签。

[0094] 其中，音频特征可以包括频率特征和幅值特征。幅值特征可以包括：高幅值、中幅值及低幅值；频率特征可以包括：高频、中频及低频等。

[0095] 本公开实施例中，可以将配音得到的样本语音数据进行音频特征分析之后，将得到的频率及幅值按从大到小的顺序进行排序，进而将第一阈值范围内的频率标注为高频、将第二阈值范围内的频率标注为中频、第三阈值范围内的频率标注为低频；将第四阈值范围内的幅值标注为高幅值、将第五阈值范围内的幅值标注为中幅值、第六阈值范围内的幅值标注为低幅值。

[0096] 举例来说，若全部的样本语音数据对应的频率范围为[a,b]，则第一阈值范围可以为[b-10%*(b-a),b]，即将频率范围内最高的10%的频率标注为高频，第二阈值范围可以

为[a+10%*(b-a),b-10%*(b-a)],即将频率范围内10%-90%的频率标注为中频,第三阈值范围可以为[a,a+10%*(b-a)],即将频率范围内最低的10%的频率标注为低频。

[0097] 举例来说,若全部的样本语音数据对应的幅值范围为[c,d],第四阈值范围可以为[d-10%*(d-c),d],即将幅值范围内最高的10%的幅值标注为高幅值,第五阈值范围可以为[c+10%*(d-c),d-10%*(d-c)],即将幅值范围内10%-90%的幅值标注为中幅值,第六阈值范围可以为[c,c+10%*(d-c)],即将幅值范围内最低的10%的幅值标注为低幅值。

[0098] 需要说明的是,上述示例只是简单的举例说明,不能作为本公开实施例中第一阈值范围、第二阈值范围、第三阈值范围、第四阈值范围、第五阈值范围、第六阈值范围等的具体限定。

[0099] 将第四阈值范围内的幅值标注为高幅值、将第五阈值范围内的幅值标注为中幅值、第六阈值范围内的幅值标注为低幅值。

[0100] 可以理解的是,由于预设的对话模型不能学习到所有取值的频率或幅值,因此,本公开实施例中,可以将频率或幅值按范围划分为不同的等级,从而可以提高对话模型的泛化能力。

[0101] S304:获取输入语音对应的场景图像。

[0102] 其中,场景图像中可以包括输入语音的说话者所处的场景,比如,教室、餐厅、操场等等。可选的,场景图像中可以包含输入语音对应的说话者的人脸图像,也可以不包含不包含说话者的人脸图像,本公开对此不做限定。

[0103] 可选的,在监测到采集的语音数据中包含用户语音的情况下,启动图像采集组件,以获取输入语音对应的场景图像。

[0104] 其中,图像采集组件可以为交互设备中的具有拍照片功能的组件。比如,具有交互功能的手机设备、平板设备中包含的摄像头组件。

[0105] 或者,根据输入语音的获取时间,从采集的视频流中截取与输入语音对应的场景图像。

[0106] 可选的,用交互设备中包含的视频采集设备实时采集视频流,并将采集的视频流存储至存储器中,之后根据输入语音的获取时间,从采集的视频流中截取与输入语音对应的场景图像。

[0107] 本公开实施例中,在监测到采集的语音数据中包含用户语音的情况下,获取输入语音对应的场景图像;或者,根据输入语音的获取时间,从采集的视频流中截取与输入语音对应的场景图像,从而使获取的场景图像可以准确反映输入语音对应的说话者所处的场景。

[0108] S305:对场景图像进行视觉特征提取,以确定场景图像对应的视觉特征。

[0109] 其中,视觉特征可以为场景图像中包含的场景特征。比如,教师、餐厅、篮球场等。

[0110] 可选的,可以先对场景图像进行目标检测,以获取场景图像中包含的物体的种类及位置信息,进而根据各物体的种类及位置信息,确定场景图象描述的场景,即场景图像对应的视觉特征。

[0111] 或者,可以对场景图像进行自动分割,以划分出场景图像中包含的对象或颜色区域,之后对每个图像子块提取特征,并建立索引,从而得到场景图像中每个物体对应的空间关系特征,进而基于每个物体的种类及空间关系,确定场景图像描述的场景。

[0112] S306:根据视觉特征,对第二文本和/或第二音频特征进行修正。

[0113] 可以理解的是,在确定了视觉特征之后,可以根据视觉特征对第二文本,或第二音频特征进行修正,从而使修正后的第二文本,或第二音频特征更加准确,进而更加贴合输入语音对应的说话者的情绪。

[0114] S307:基于修正后的第二音频特征及第二文本,生成答复语音。

[0115] 其中,步骤S307的具体实现形式,可参照本公开其他各实施例中的详细描述,此处不再详细赘述。

[0116] 本公开实施例中,先对获取的输入语音进行语音识别,以确定输入语音对应的第一文本,之后对输入语音进行音频特征提取,以确定输入语音对应的第一音频特征,再根据第一音频特征及第一文本,确定待生成的答复语句对应的第二文本及第二音频特征,之后再根据场景图像对应的视觉特征对第二文本及第二音频特征进行修正,最后基于修正后的第二音频特征及第二文本,生成答复语音。由此,基于场景图像对应的视觉特征,对根据第一音频特征及第一文本生成的第二文本及第二音频特征进行修正,从而进一步提高了修正后的第二文本及第二音频特征的准确性,进而进一步提高了生成的答复语音的准确性,使得答复语音更加贴合输入语音对应的说话者的情绪。

[0117] 图4是根据本公开又一实施例提供的一种语音对话的生成装置的结构示意图,如图4所示,该语音对话的生成装置400,包括:第一确定模块410、第二确定模块420、第三确定模块430及生成模块440。

[0118] 第一确定模块410,用于对获取的输入语音进行语音识别,以确定输入语音对应的第一文本;

[0119] 第二确定模块420,用于对输入语音进行音频特征提取,以确定输入语音对应的第一音频特征;

[0120] 第三确定模块430,用于根据第一音频特征及第一文本,确定待生成的答复语句对应的第二文本及第二音频特征;

[0121] 生成模块440,用于基于第二音频特征及第二文本,生成答复语音。

[0122] 可选的,第二确定模块420,具体用于:

[0123] 根据输入语音中每帧语音对应的第一幅值,确定输入语音对应的第二幅值;

[0124] 根据第二幅值所属的范围,确定输入语音对应的幅值特征。

[0125] 可选的,第二确定模块420,具体用于:

[0126] 对输入语音进行基音检测,以确定语音信号对应的频率值;

[0127] 根据频率值所属的范围,确定输入语音对应的频率特征。

[0128] 可选的,生成模块440,包括:

[0129] 第一获取单元,用于获取输入语音对应的场景图像;

[0130] 第一确定单元,用于对场景图像进行视觉特征提取,以确定场景图像对应的视觉特征;

[0131] 修正单元,用于根据视觉特征,对第二文本和/或第二音频特征进行修正;

[0132] 生成单元,用于基于修正后的第二音频特征及第二文本,生成答复语音。

[0133] 可选的,第一获取单元,具体用于:

[0134] 响应于监测到采集的语音数据中包含用户语音的情况下,启动图像采集组件,以

获取输入语音对应的场景图像；

[0135] 或者，根据输入语音的获取时间，从采集的视频流中截取与输入语音对应的场景图像。

[0136] 可选的，还包括：

[0137] 第四确定模块，用于根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二文本中包含的表情符号；

[0138] 显示模块，用于在交互设备的显示屏幕上，显示第二文本及表情符号。

[0139] 需要说明的是，前述对语音对话的生成方法的解释说明也适用于本实施例的语音对话的生成装置，此处不再赘述。

[0140] 本公开实施例中，先对获取的输入语音进行语音识别，以确定输入语音对应的第一文本，之后对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征，再根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征，最后基于第二音频特征及第二文本，生成答复语音。由此，根据输入语音对应的第一音频特征及第一文本，确定答复语句对应的第二文本及第二音频特征，从而不仅提高了确定的第二文本的准确性，而且可以根据输入语句对应的情绪特征确定答复语句的情绪特征，从而使生成的答复语音更加贴合输入语音对应的说话者的情绪。

[0141] 根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0142] 图5示出了可以用来实施本公开的实施例的示例电子设备500的示意性框图。电子设备旨在表示各种形式的数字计算机，诸如，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，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备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、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0143] 如图5所示，设备500包括计算单元501，其可以根据存储在只读存储器(ROM)502中的计算机程序或者从存储单元508加载到随机访问存储器(RAM)503中的计算机程序，来执行各种适当的动作和处理。在RAM 503中，还可存储设备500操作所需的各种程序和数据。计算单元501、ROM 502以及RAM 503通过总线504彼此相连。输入/输出(I/O)接口505也连接至总线504。

[0144] 设备500中的多个部件连接至I/O接口505，包括：输入单元506，例如键盘、鼠标等；输出单元507，例如各种类型的显示器、扬声器等；存储单元508，例如磁盘、光盘等；以及通信单元509，例如网卡、调制解调器、无线通信收发机等。通信单元509允许设备500通过诸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。

[0145] 计算单元501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。计算单元501的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元501执行上文所描述的各个方法和处理，例如语音对话的生成。例如，在一些实施例中，语音对话的生成可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元508。在一些实施例中，计算机程序的部

分或者全部可以经由ROM 502和/或通信单元509而被载入和/或安装到设备500上。当计算机程序加载到RAM 503并由计算单元501执行时，可以执行上文描述的语音对话的生成的一个或多个步骤。备选地，在其他实施例中，计算单元501可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行语音对话的生成。

[0146] 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(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(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。

[0147] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码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或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码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。

[0148] 在本公开的上下文中，机器可读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程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、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。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(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。

[0149] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)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计算机。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。

[0150] 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(LAN)、广域网(WAN)、互联网及区块链网络。

[0151] 计算机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VPS服务("Virtual Private Server"，或简称"VPS")中，存在的管理难度大，业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器，或者是结合了区块链的服务器。

[0152] 本实施例中，先对获取的输入语音进行语音识别，以确定输入语音对应的第一文本，之后对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征，再根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征，最后基于第二音频特征及第二文本，生成答复语音。由此，根据输入语音对应的第一音频特征及第一文本，确定答复语句对应的第二文本及第二音频特征，从而不仅提高了确定的第二文本的准确性，而且可以根据输入语句对应的情绪特征确定答复语句的情绪特征，从而使生成的答复语音更加贴合输入语音对应的说话者的情绪。

[0153] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

[0154] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本公开的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。在本公开的描述中，所使用的词语“如果”及“若”可以被解释成为“在……时”或“当……时”或“响应于确定”或“在……情况下”。

[0155] 上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

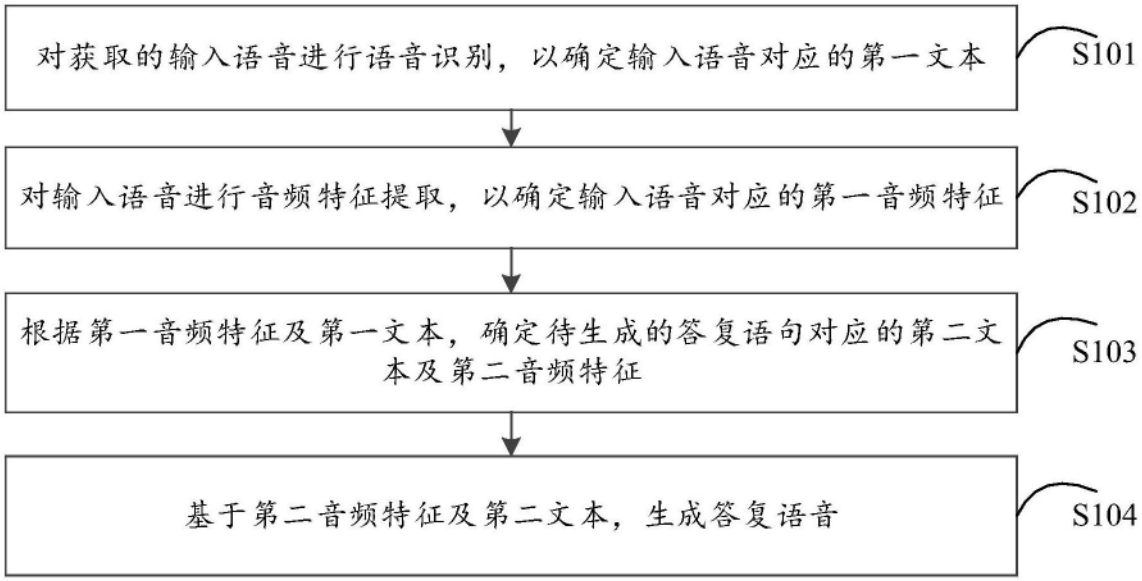

对获取的输入语音进行语音识别，以确定输入语音对应的第一文本
S101
对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征
S102
根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征
S103
基于第二音频特征及第二文本，生成答复语音
S104

图1

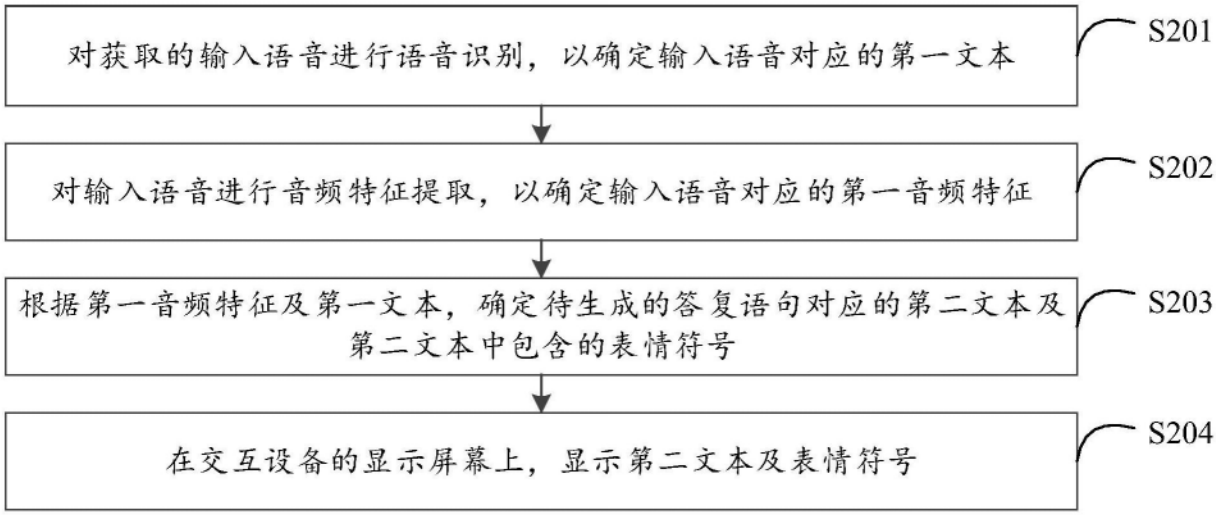

对获取的输入语音进行语音识别，以确定输入语音对应的第一文本
S201
对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征
S202
根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二文本中包含的表情符号
S203
在交互设备的显示屏幕上，显示第二文本及表情符号
S204

图2

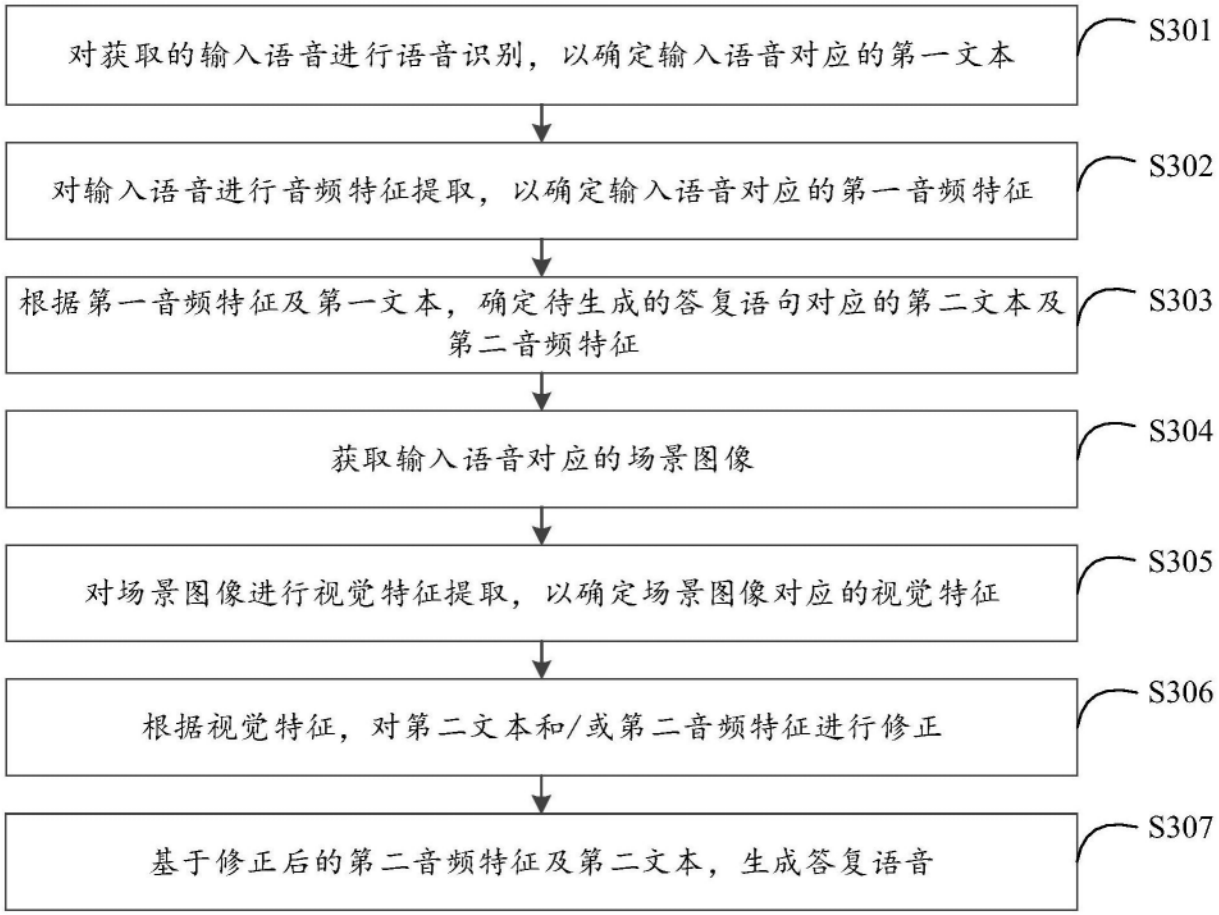

对获取的输入语音进行语音识别，以确定输入语音对应的第一文本
S301
对输入语音进行音频特征提取，以确定输入语音对应的第一音频特征
S302
根据第一音频特征及第一文本，确定待生成的答复语句对应的第二文本及第二音频特征
S303
获取输入语音对应的场景图像
S304
对场景图像进行视觉特征提取，以确定场景图像对应的视觉特征
S305
根据视觉特征，对第二文本和/或第二音频特征进行修正
S306
基于修正后的第二音频特征及第二文本，生成答复语音
S307

图3

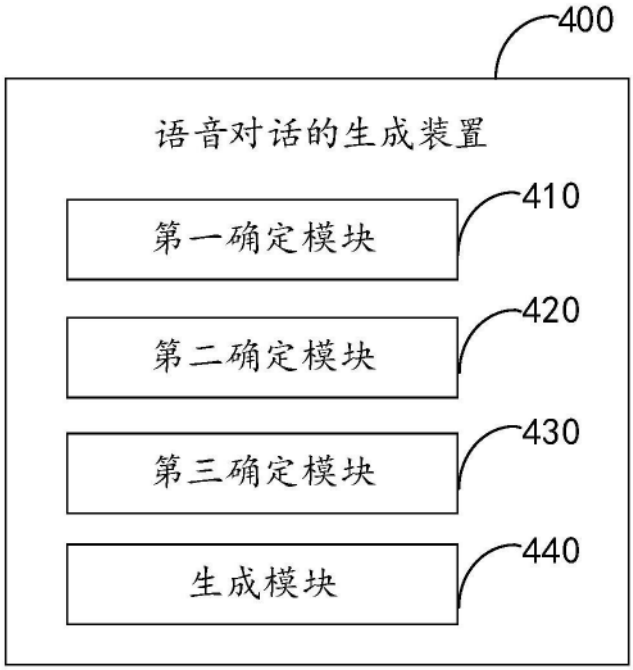

400
语音对话的生成装置
第一确定模块
410
第二确定模块
420
第三确定模块
430
生成模块
440

图4

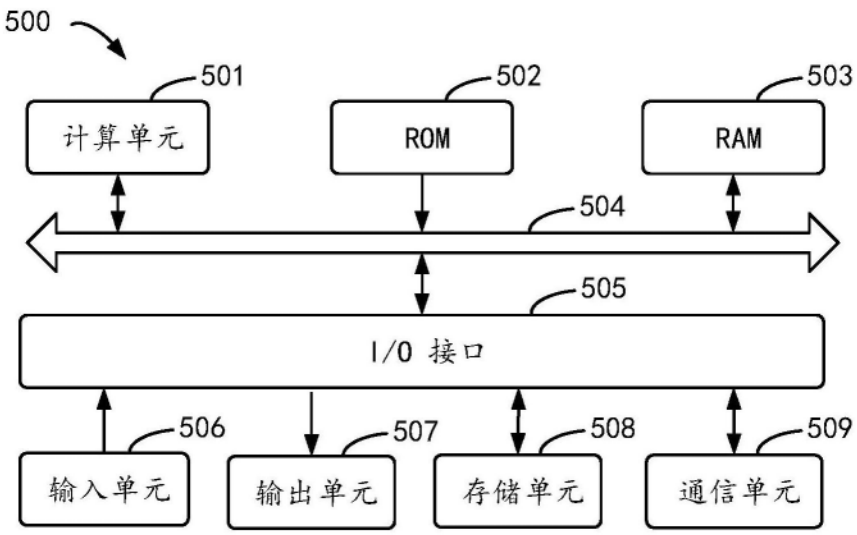

500
501
计算单元
502
ROM
503
RAM
504
505
I/O 接口
506
输入单元
507
输出单元
508
存储单元
509
通信单元

图5