(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 113689881 A
(43)申请公布日 2021.11.23

(21)申请号 202010420263.3

(22)申请日 2020.05.18

(71)申请人 北京中关村科金技术有限公司
地址 100000 北京市海淀区后屯南路26号4层5-03-2

(72)发明人 白安琪 蒋宁 赵立军 陈燕丽

(74)专利代理机构 北京万思博知识产权代理有限公司 11694
代理人 刘冀

(51)Int.Cl.
G10L 25/30(2013.01)
G10L 25/51(2013.01)
G10L 25/03(2013.01)
G10L 15/06(2013.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)
G06F 40/30(2020.01)

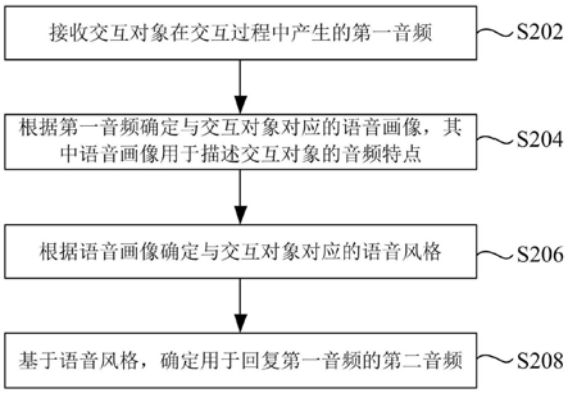
权利要求书2页 说明书9页 附图4页

(54)发明名称

针对语音画像进行音频交互的方法、装置以及存储介质

(57)摘要

本申请公开了一种针对语音风格进行音频交互的方法、装置以及存储介质。其中，该方法包括：接收交互对象在交互过程中产生的第一音频；根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；根据语音画像确定与交互对象对应的语音风格；以及基于语音风格，确定用于回复第一音频的第二音频。



CN 113689881 A

1.一种针对语音风格进行音频交互的方法,其特征在于,包括:

接收交互对象在交互过程中产生的第一音频;

根据所述第一音频确定与所述交互对象对应的语音画像,其中所述语音画像用于描述所述交互对象的音频特点;

根据所述语音画像确定与所述交互对象对应的语音风格;以及

基于所述语音风格,确定用于回复所述第一音频的第二音频。

2.根据权利要求1所述的方法,其特征在于,根据所述第一音频确定与所述交互对象对应的语音画像,包括:

从所述第一音频中获取与声音要素相关的第一声音信息;

从所述第一音频中获取与所述交互对象相关的第二声音信息,其中所述第二声音信息用于描述与所述交互对象对应的声音特征属性;以及

根据所述第一声音信息和所述第二声音信息,确定与所述交互对象对应的语音画像。

3.根据权利要求2所述的方法,其特征在于,从所述第一音频中获取与所述交互对象相关的第二声音信息,包括:

确定与所述第一音频对应的音频特征;

利用预先训练的用于预测用户声音特征的决策树模型对所述音频特征进行识别,确定与所述声音特征属性对应的属性值;以及

根据与所述声音特征属性对应的属性值,确定所述第二声音信息。

4.根据权利要求1所述的方法,其特征在于,根据所述语音画像确定与所述交互对象对应的语音风格,包括:

根据所述语音画像,确定与所述音频特点对应的特征向量;以及

利用预先训练的用于预测用户语音风格的模型对所述特征向量进行计算,确定与所述交互对象对应的语音风格。

5.根据权利要求1所述的方法,其特征在于,还包括:确定用于回复所述第一音频的文本信息,并且基于所述语音风格,确定用于回复所述第一音频的第二音频,包括:

基于所述语音风格以及所述文本信息确定所述第二音频。

6.根据权利要求1所述的方法,其特征在于,确定用于回复所述第一音频的第二音频,包括:

从预先设置的语音库中选择用于回复所述第一音频的第二音频。

7.根据权利要求2和3任意一项所述的方法,其特征在于,所述声音特征属性包括以下至少一项:

所述交互对象的性别、所述交互对象的年龄、所述交互对象的方言、所述交互对象的音质以及所述交互对象的情绪。

8.一种存储介质,其特征在于,所述存储介质包括存储的程序,其中,在所述程序运行时由处理器执行权利要求1至7中任意一项所述的方法。

9.一种针对语音风格进行音频交互的装置,其特征在于,包括:

音频接收模块,用于接收交互对象在交互过程中产生的第一音频;

画像确定模块,用于根据所述第一音频确定与所述交互对象对应的语音画像,其中所述语音画像用于描述所述交互对象的音频特点;

风格确定模块，用于根据所述语音画像确定与所述交互对象对应的语音风格；以及

音频确定模块，用于基于所述语音风格，确定用于回复所述第一音频的第二音频。

10.一种针对语音风格进行音频交互的装置，其特征在于，包括：

处理器；以及

存储器，与所述处理器连接，用于为所述处理器提供处理以下处理步骤的指令：

接收交互对象在交互过程中产生的第一音频；

根据所述第一音频确定与所述交互对象对应的语音画像，其中所述语音画像用于描述所述交互对象的音频特点；

根据所述语音画像确定与所述交互对象对应的语音风格；以及

基于所述语音风格，确定用于回复所述第一音频的第二音频。

针对语音画像进行音频交互的方法、装置以及存储介质

技术领域

[0001] 本申请涉及互联网人工智能技术领域，特别是涉及一种针对语音画像进行音频交互的方法、装置以及存储介质。

背景技术

[0002] TTS是机器将语言从文字载体转换到声音载体的过程，是人机对话、智能播报等系统中的关键模块。相关研究多立足于其两端：文本和语音，文本端要求语言单位分合恰当、定性准确、可理解度高，语音端则要求仿真能力强，这些的着眼点在于语音的社会属性与生理属性。因此语言单位切分、定性、句法分析、韵律边界预测等理论方案与实践技术受到了研究者们的广泛关注。随着语音合成技术的发展，单一风格的语音已经不能满足人们的需求，于是具有性别、年龄、地域方言等的差异化语音合成的软、硬件产品应运而生；更优者，提供了自主选择语音风格、形成语音混搭的定制化方案，这些均是对语音物理属性的关注。然而，现有的音频交互系统，缺乏对交互对象的音频进行分析，然后制定与交互对象的音频特色对应的回复音频。此外，现有技术在交互过程中可以根据用户的需求选择不同的语音风格，无法自主的根据对方的音频风格自动的确定与之对应的风格进行交互，因此影响交互对象的体验效果。

[0003] 针对上述的现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题，目前尚未提出有效的解决方案。

发明内容

[0004] 本公开的实施例提供了一种针对语音画像进行音频交互的方法、装置以及存储介质，以至少解决现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

[0005] 根据本公开实施例的一个方面，提供了一种针对语音风格进行音频交互的方法，包括：接收交互对象在交互过程中产生的第一音频；根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；根据语音画像确定与交互对象对应的语音风格；以及基于语音风格，确定用于回复第一音频的第二音频。

[0006] 根据本公开实施例的另一个方面，还提供了一种存储介质，存储介质包括存储的程序，其中，在程序运行时由处理器执行以上任意一项所述的方法。

[0007] 根据本公开实施例的另一个方面，还提供了一种针对语音风格进行音频交互的装置，包括：音频接收模块，用于接收交互对象在交互过程中产生的第一音频；画像确定模块，用于根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；风格确定模块，用于根据语音画像确定与交互对象对应的语音风格；以及音频确定模块，用于基于语音风格，确定用于回复第一音频的第二音频。

[0008] 根据本公开实施例的另一个方面，还提供了一种针对语音风格进行音频交互的装置，包括：处理器；以及存储器，与处理器连接，用于为处理器提供处理以下处理步骤的指令：接收交互对象在交互过程中产生的第一音频；根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；根据语音画像确定与交互对象对应的语音风格；以及基于语音风格，确定用于回复第一音频的第二音频。

[0009] 在本公开实施例中，服务器可以从交互对象的音频中分析出交互对象的语音画像，进而针对交互对象的语音画像选择合适的语音风格，最终利用该语音风格的话术与交互对象继续进行交互。与现有技术相比，本方案可以自动地对语音进行分析、预测以及选择，不需要交互对象自行选择话术风格。从而，可以在交互对象无感的情况下，拉近与交互对象的距离。进而，解决了现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

附图说明

[0010] 此处所说明的附图用来提供对本公开的进一步理解，构成本申请的一部分，本公开的示意性实施例及其说明用于解释本公开，并不构成对本公开的不当限定。在附图中：

[0011] 图1是用于实现根据本公开实施例1所述的方法的计算设备的硬件结构框图；

[0012] 图2是根据本公开实施例1所述的针对语音风格进行音频交互的方法的流程示意图；

[0013] 图3是根据本公开实施例1所述的语音分析过程的流程示意图；

[0014] 图4是根据本公开实施例1所述的模型训练的流程图；

[0015] 图5是根据本公开实施例1所述的语音风格预测的流程图；

[0016] 图6是根据本公开实施例1所述的针对语音风格进行音频交互的方法的整体流程示意图；

[0017] 图7是根据本公开实施例2所述的针对语音风格进行音频交互的装置的示意图；以及

[0018] 图8是根据本公开实施例3所述的针对语音风格进行音频交互的装置的示意图。

具体实施方式

[0019] 为了使本技术领域的人员更好地理解本公开的技术方案，下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述。显然，所描述的实施例仅仅是本公开一部分的实施例，而不是全部的实施例。基于本公开中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本公开保护的范围。

[0020] 需要说明的是，本公开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，以便这里描述的本公开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。

[0021] 实施例1

[0022] 根据本实施例，提供了一种针对语音画像进行音频交互的方法的实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0023] 本实施例所提供的方法实施例可以在服务器或者类似的计算设备中执行。图1示出了一种用于实现针对语音画像进行音频交互的方法的计算设备的硬件结构框图。如图1所示，计算设备可以包括一个或多个处理器(处理器可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器、以及用于通信功能的传输装置。除此以外，还可以包括：显示器、输入/输出接口(I/O接口)、通用串行总线(USB)端口(可以作为I/O接口的端口中的一个端口被包括)、网络接口、电源和/或相机。本领域普通技术人员可以理解，图1所示的结构仅为示意，其并不对上述电子装置的结构造成限定。例如，计算设备还可包括比图1中所示更多或者更少的组件，或者具有与图1所示不同的配置。

[0024] 应当注意到的是上述一个或多个处理器和/或其他数据处理电路在本文中通常可以被称为“数据处理电路”。该数据处理电路可以全部或部分的体现为软件、硬件、固件或其他任意组合。此外，数据处理电路可为单个独立的处理模块，或全部或部分的结合到计算设备中的其他元件中的任意一个内。如本公开实施例中所涉及到的，该数据处理电路作为一种处理器控制(例如与接口连接的可变电阻终端路径的选择)。

[0025] 存储器可用于存储应用软件的软件程序以及模块，如本公开实施例中的针对语音画像进行音频交互的方法对应的程序指令/数据存储装置，处理器通过运行存储在存储器内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的应用程序的针对语音画像进行音频交互的方法。存储器可包括高速随机存储器，还可包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至计算设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0026] 传输装置用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算设备的通信供应商提供的无线网络。在一个实例中，传输装置包括一个网络适配器(Network Interface Controller,NIC)，其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中，传输装置可以为射频(Radio Frequency,RF)模块，其用于通过无线方式与互联网进行通讯。

[0027] 显示器可以例如触摸屏式的液晶显示器(LCD)，该液晶显示器可使得用户能够与计算设备的用户界面进行交互。

[0028] 此处需要说明的是，在一些可选实施例中，上述图1所示的计算设备可以包括硬件元件(包括电路)、软件元件(包括存储在计算机可读介质上的计算机代码)、或硬件元件和软件元件两者的结合。应当指出的是，图1仅为特定具体实例的一个实例，并且旨在示出可存在于上述计算设备中的部件的类型。

[0029] 在上述运行环境下，根据本实施例的第一个方面，提供了一种针对语音画像进行音频交互的方法，该方法例如可以应用到机器人客服系统，系统的服务器可以根据交互对象的语音特征，自动选择适应的用于回复的语音风格。图2示出了该方法的流程示意图，参考图2所示，该方法包括：

[0030] S202：接收交互对象在交互过程中产生的第一音频；

[0031] S204：根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；

[0032] S206：根据语音画像确定与交互对象对应的语音风格；以及

[0033] S208：基于语音风格，确定用于回复第一音频的第二音频

[0034] 正如背景技术中所述的，随着语音合成技术的发展，单一风格的语音已经不能满足人们的需求，于是具有性别、年龄、地域方言等的差异化语音合成的软、硬件产品应运而生；更优者，提供了自主选择语音风格、形成语音混搭的定制化方案，这些均是对语音物理属性的关注。然而，现有的音频交互系统，缺乏对交互对象的音频进行分析，然后制定与交互对象的音频特色对应的回复音频。此外，现有技术在交互过程中可以根据用户的需求选择不同的语音风格，无法自主的根据对方的音频风格自动的确定与之对应的风格进行交互，因此影响交互对象的体验效果。

[0035] 针对背景技术中存在的技术问题，参考图3所示，本实施例技术方案在步骤S202中，服务器首先接收交互对象在交互过程中产生的第一音频。在一个具体实例中，第一音频例如可以是交互对象电话投诉的音频。

[0036] 进一步地，在步骤S204中，服务器对第一音频进行分析，确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点。例如：交互对象的性别、交互对象的年龄、交互对象的方言、交互对象的语速、交互对象的风格、交互对象的情绪、音量等共同构成与交互对象对应的语音画像。此外，还可以包含其他的与人物相关的语音特征，此处不做具体限定。在一个具体实例中，交互对象的语音画像为四川方言、语速较慢的老年男声。

[0037] 进一步地，在步骤S206中，服务器根据语音画像确定与交互对象对应的语音风格。例如：针对上述的交互对象的语音画像(即：四川方言、语速较慢的老年男声)，确定的语音风格为慢速的口语四川方言。

[0038] 最终，在步骤S208中，服务器基于语音风格，确定用于回复第一音频的第二音频。例如：服务器确定利用慢速的口语四川方言的第二音频与交互对象进行交互，即系统利用语音机器人采用慢速的口语四川方言的语音风格回复交互对象的投诉电话。

[0039] 从而通过这种方式，服务器可以从交互对象的音频中分析出交互对象的语音画像，进而针对交互对象的语音画像选择合适的语音风格，最终利用该语音风格的话术与交互对象继续进行交互。与现有技术相比，本方案可以自动地对语音进行分析、预测以及选择，不需要交互对象自行选择话术风格。从而，可以在交互对象无感的情况下，拉近与交互对象的距离。进而，解决了现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

[0040] 可选地，根据第一音频确定与交互对象对应的语音画像，包括：从第一音频中获取与声音要素相关的第一声音信息；从第一音频中获取与交互对象相关的第二声音信息，其

中第二声音信息用于描述与交互对象对应的声音特征属性;以及根据第一声音信息和第二声音信息,确定与交互对象对应的语音画像。

[0041] 具体地,参考图3所示,在根据第一音频确定与交互对象对应的语音画像的操作中,服务器从第一音频中获取与声音要素相关的第一声音信息,其中第一声音信息为声音的要素,例如:音速值、音量值等相关要素。并且服务器还从第一音频中获取与交互对象相关的第二声音信息,其中第二声音信息用于描述与交互对象对应的声音特征属性,在一个具体实例中,声音特征属性包括以下至少一项:交互对象的性别、交互对象的年龄、交互对象的方言、交互对象的音质以及交互对象的情绪。最终,服务器根据第一声音信息和第二声音信息,确定与交互对象对应的语音画像。从而通过这种方式,服务器可以对交互对象的第一音频从多个维度进行分析,最终确定交互对象的语音画像,因此对交互对象的声音分析的更加全面。

[0042] 可选地,从第一音频中获取与交互对象相关的第二声音信息,包括:确定与第一音频对应的音频特征;利用预先训练的用于预测用户声音特征的决策树模型对音频特征进行识别,确定与声音特征属性对应的属性值;以及根据与声音特征属性对应的属性值,确定第二声音信息。

[0043] 具体地,参考图3所示,在从第一音频中获取与交互对象相关的第二声音信息的操作中,服务器首先确定与第一音频对应的音频特征,其中该音频特征例如可以是MFCC特征或者FBank特征,还可以是其他方式提取的特征,此处不做具体限定。进一步地,服务器利用预先训练的用于预测用户声音特征的决策树模型对音频特征进行识别,确定与声音特征属性对应的属性值,声音特征属性例如包括:方言值(在实际操作中,对不同的方言设置不同的数值)、性别值(男、女对应不同的数值)、年龄值、音质值以及情绪值(不同的情绪对应不同的数值)。在实际操作中,可以训练多个决策树模型,分别预测每个声音特征属性对应的属性值。决策树模型例如可以是基于卷积神经网络或者其他的机器学习算法的。从而,可以通过模型确定第二声音信息,操作更加便捷,结果更加精准。

[0044] 可选地,根据语音画像确定与交互对象对应的语音风格,包括:根据语音画像,确定与音频特点对应的特征向量;以及利用预先训练的用于预测用户语音风格的模型对特征向量进行计算,确定与交互对象对应的语音风格。

[0045] 具体地,在根据语音画像确定与交互对象对应的语音风格的操作中,服务器首先根据语音画像,确定与音频特点对应的特征向量,参考图3所示,将上述的方言值、性别值、年龄值、音质值以及情绪值组成与音频特点对应的特征向量。进一步地,服务器利用预先训练的用于预测用户语音风格的模型对特征向量进行计算,确定与交互对象对应的语音风格。图4示出了语音风格模型训练的流程图,参考图4所示,语音风格模型训练过程中将用户的语音风格向量以及对应的语音风格标签(即用户的语音风格)作为模型训练数据,模型例如可以采用Xgboost算法,利用训练数据进行多分类训练,最终获得语音风格模型。在实际应用中,参考图5所示,将目标对象的特征向量输入到模型中,输出与交互对象对应的语音风格。

[0046] 可选地,还包括:确定用于回复第一音频的文本信息,并且基于语音风格,确定用于回复第一音频的第二音频,包括:基于语音风格以及文本信息确定第二音频。

[0047] 具体地,参考图6所示,服务器还确定用于回复第一音频的文本信息(对应于图6中

的回复语的普通话文本)，然后服务器基于语音风格以及文本信息确定第二音频。即：服务器基于所确定的语音风格将该文本信息生成第二音频，从而可以更加精准的对交互对象进行回复。

[0048] 可选地，确定用于回复第一音频的第二音频，包括：从预先设置的语音库中选择用于回复第一音频的第二音频。

[0049] 具体地，参考图6所示，在确定用于回复第一音频的第二音频的操作中，服务器从预先设置的语音库中选择用于回复第一音频的第二音频。即，从多种风格语音库中选择与确定文本表达相同语义，并且为用户偏爱风格(与交互对象对应的语音风格)的语音片段(第二音频)进行回复。其中语音库中有多种语音风格的音频片段，可以从语音库中寻找与交互对象的语音风格一致，且与第一音频对应的语音片段，作为回复语输出。例如：从语音库中选取相应风格的语音(慢速的口语四川方言)进行交互对象情绪安抚。从而通过这种方式，可以从现有的语音库中选择对应的第二音频，不需要生成，因此可以节省计算资源提高效率。

[0050] 此外，参考图1所示，根据本实施例的第二个方面，提供了一种存储介质。所述存储介质包括存储的程序，其中，在所述程序运行时由处理器执行以上任意一项所述的方法。

[0051] 从而根据本实施例，服务器可以从交互对象的音频中分析出交互对象的语音画像，进而针对交互对象的语音画像选择合适的语音风格，最终利用该语音风格的话术与交互对象继续进行交互。与现有技术相比，本方案可以自动地对语音进行分析、预测以及选择，不需要交互对象自行选择话术风格。从而，可以在交互对象无感的情况下，拉近与交互对象的距离。进而，解决了现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

[0052] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0053] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0054] 实施例2

[0055] 图7示出了根据本实施例所述的针对语音风格进行音频交互的装置700，该装置700与根据实施例1的第一个方面所述的方法相对应。参考图7所示，该装置700包括：音频接收模块710，用于接收交互对象在交互过程中产生的第一音频；画像确定模块720，用于根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；风格确定模块730，用于根据语音画像确定与交互对象对应的语音风格；以及音频确定模块

740，用于基于语音风格，确定用于回复第一音频的第二音频。

[0056] 可选地，画像确定模块720，包括：第一确定子模块，用于从第一音频中获取与声音要素相关的第一声音信息；第二确定子模块，用于从第一音频中获取与交互对象相关的第二声音信息，其中第二声音信息用于描述与交互对象对应的声音特征属性；以及画像确定子模块，用于根据第一声音信息和第二声音信息，确定与交互对象对应的语音画像。

[0057] 可选地，第二确定子模块，包括：特征确定单元，用于确定与第一音频对应的音频特征；特征识别单元，用于利用预先训练的用于预测用户声音特征的决策树模型对音频特征进行识别，确定与声音特征属性对应的属性值；以及声音信息确定单元，用于根据与声音特征属性对应的属性值，确定第二声音信息。

[0058] 可选地，风格确定模块730，包括：向量生成子模块，用于根据语音画像，确定与音频特点对应的特征向量；以及风格确定子模块，用于利用预先训练的用于预测用户语音风格的模型对特征向量进行计算，确定与交互对象对应的语音风格。

[0059] 可选地，装置700还包括：文本确定模块，用于确定用于回复第一音频的文本信息，并且音频确定模块740，包括：第一音频确定子模块，用于基于语音风格以及文本信息确定第二音频。

[0060] 可选地，音频确定模块740，还包括：音频选择子模块，用于从预先设置的语音库中选择用于回复第一音频的第二音频。

[0061] 可选地，声音特征属性包括以下至少一项：交互对象的性别、交互对象的年龄、交互对象的方言、交互对象的音质以及交互对象的情绪。

[0062] 从而根据本实施例，装置700可以从交互对象的音频中分析出交互对象的语音画像，进而针对交互对象的语音画像选择合适的语音风格，最终利用该语音风格的话术与交互对象继续进行交互。与现有技术相比，本方案可以自动地对语音进行分析、预测以及选择，不需要交互对象自行选择话术风格。从而，可以在交互对象无感的情况下，拉近与交互对象的距离。进而，解决了现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

[0063] 实施例3

[0064] 图8示出了根据本实施例所述的针对语音风格进行音频交互的装置800，该装置800与根据实施例1的第一个方面所述的方法相对应。参考图8所示，该装置800包括：处理器810；以及存储器820，与处理器810连接，用于为处理器810提供处理以下处理步骤的指令：接收交互对象在交互过程中产生的第一音频；根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点；根据语音画像确定与交互对象对应的语音风格；以及基于语音风格，确定用于回复第一音频的第二音频。

[0065] 可选地，根据第一音频确定与交互对象对应的语音画像，包括：从第一音频中获取与声音要素相关的第一声音信息；从第一音频中获取与交互对象相关的第二声音信息，其中第二声音信息用于描述与交互对象对应的声音特征属性；以及根据第一声音信息和第二声音信息，确定与交互对象对应的语音画像。

[0066] 可选地，从第一音频中获取与交互对象相关的第二声音信息，包括：确定与第一音频对应的音频特征；利用预先训练的用于预测用户声音特征的决策树模型对音频特征进行

识别，确定与声音特征属性对应的属性值；以及根据与声音特征属性对应的属性值，确定第二声音信息。

[0067] 可选地，根据语音画像确定与交互对象对应的语音风格，包括：根据语音画像，确定与音频特点对应的特征向量；以及利用预先训练的用于预测用户语音风格的模型对特征向量进行计算，确定与交互对象对应的语音风格。

[0068] 可选地，存储器820还用于为处理器810提供处理以下处理步骤的指令：确定用于回复第一音频的文本信息，并且基于语音风格以及文本信息确定第二音频。

[0069] 可选地，确定用于回复第一音频的第二音频，包括：从预先设置的语音库中选择用于回复第一音频的第二音频。

[0070] 可选地，声音特征属性包括以下至少一项：交互对象的性别、交互对象的年龄、交互对象的方言、交互对象的音质以及交互对象的情绪。

[0071] 从而根据本实施例，装置800可以从交互对象的音频中分析出交互对象的语音画像，进而针对交互对象的语音画像选择合适的语音风格，最终利用该语音风格的话术与交互对象继续进行交互。与现有技术相比，本方案可以自动地对语音进行分析、预测以及选择，不需要交互对象自行选择话术风格。从而，可以在交互对象无感的情况下，拉近与交互对象的距离。进而，解决了现有技术中存在的人机交互过程中缺乏对交互对象的音频特点进行分析，无法主动地根据交互对象的音频特色选择合适的音频交互风格，进而影响交互对象的体验效果的技术问题。

[0072] 上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

[0073] 在本发明的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0074] 在本申请所提供的几个实施例中，应该理解到，所揭露的技术内容，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

[0075] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0076] 另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

[0077] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备等)执行本发明各个实施例所述方法的全部或

部分步骤。而前述的存储介质包括:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,RandomAccess Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0078] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

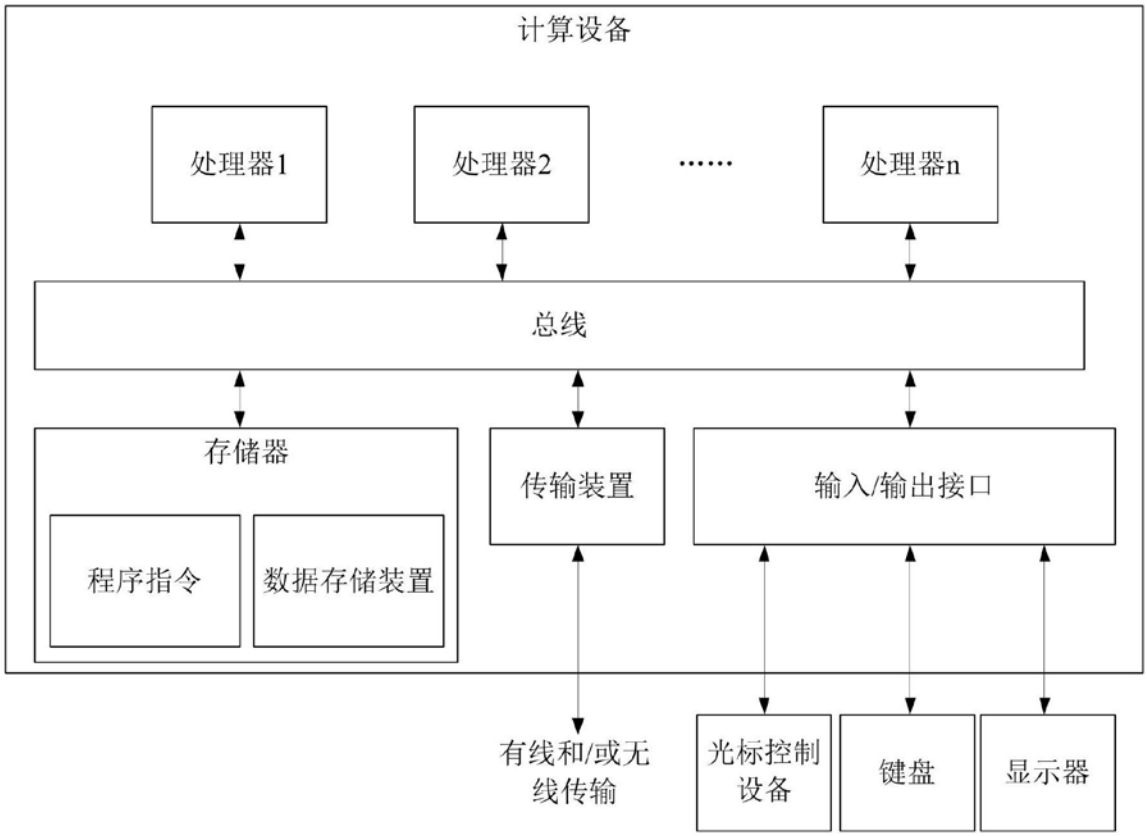

计算设备
处理器1
处理器2
……
处理器n
总线
存储器
程序指令
数据存储装置
传输装置
输入/输出接口
有线和/或无线传输
光标控制设备
键盘
显示器

图1

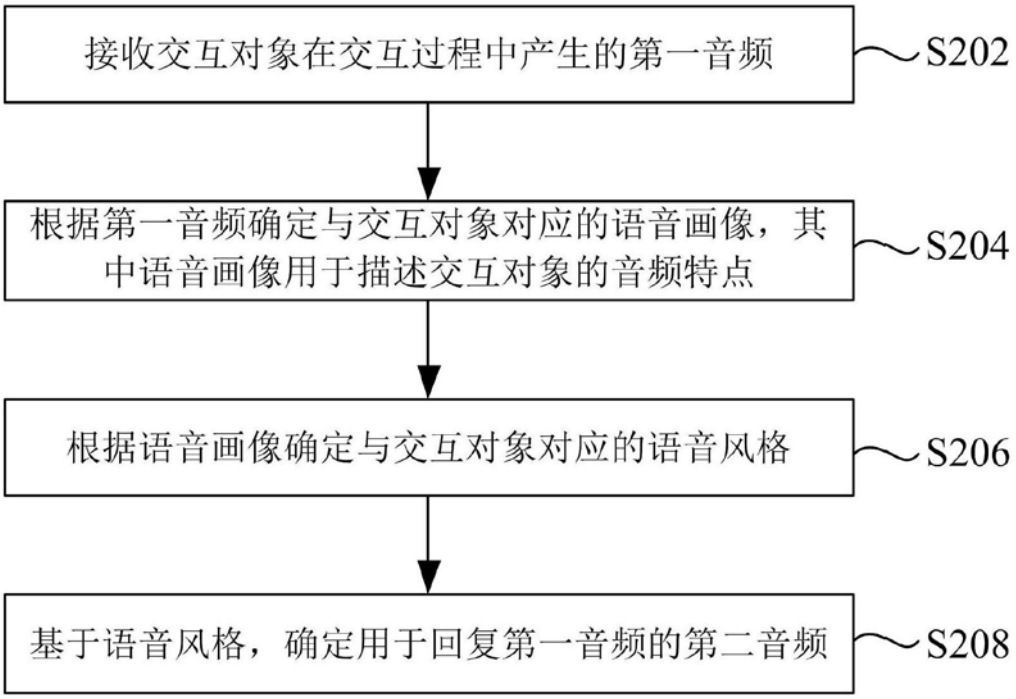

接收交互对象在交互过程中产生的第一音频
S202
根据第一音频确定与交互对象对应的语音画像，其中语音画像用于描述交互对象的音频特点
S204
根据语音画像确定与交互对象对应的语音风格
S206
基于语音风格，确定用于回复第一音频的第二音频
S208

图2

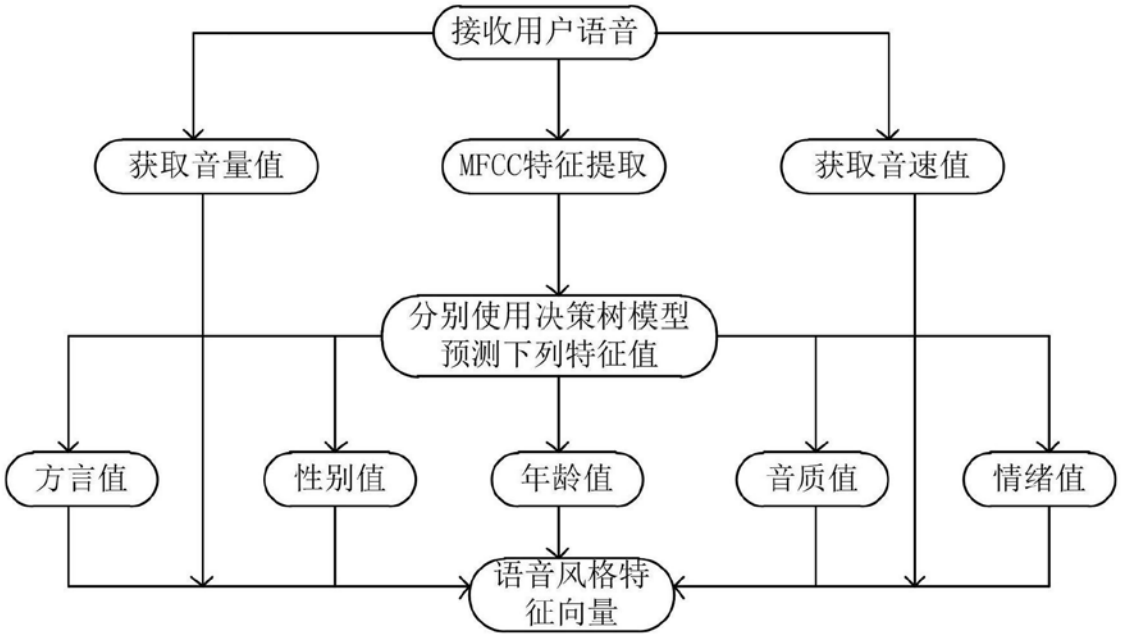

接收用户语音
获取音量值
MFCC特征提取
获取音速值
分别使用决策树模型预测下列特征值
方言值
性别值
年龄值
音质值
情绪值
语音风格特征向量

图3

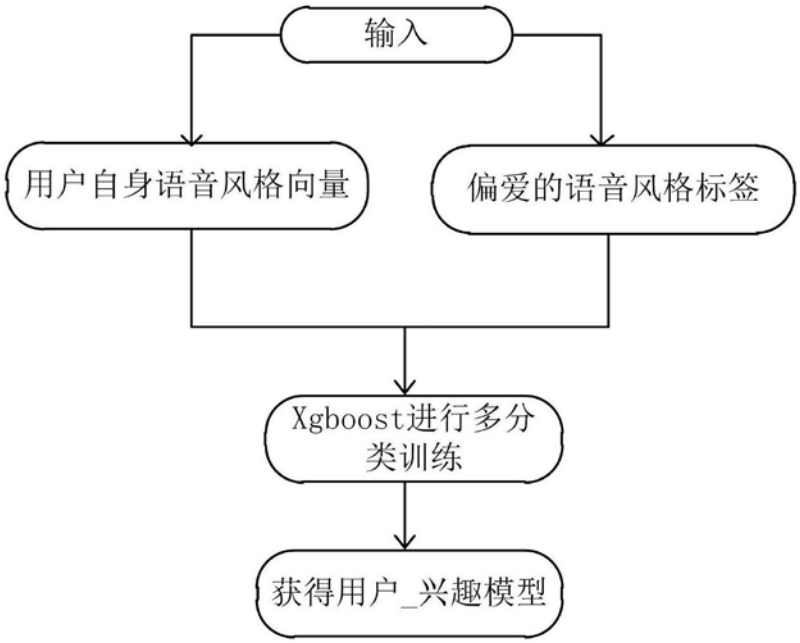

输入
用户自身语音风格向量
偏爱的语音风格标签
Xgboost进行多分类训练
获得用户_兴趣模型

图4

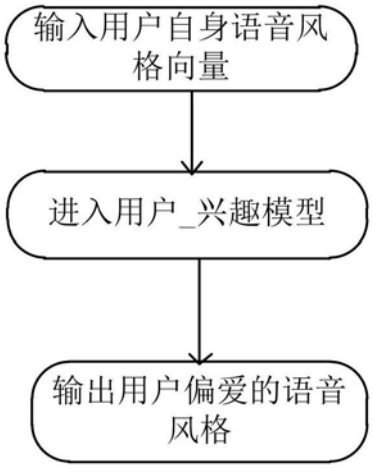

输入用户自身语音风格向量
进入用户_兴趣模型
输出用户偏爱的语音风格

图5

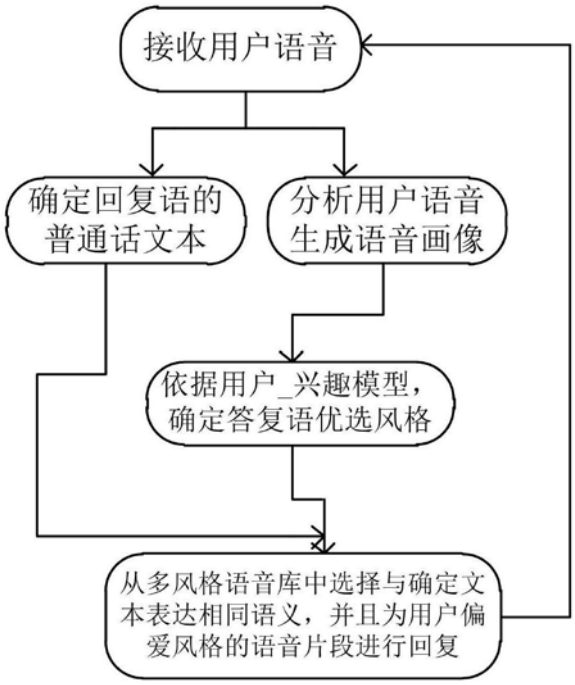

接收用户语音
确定回复语的普通话文本
分析用户语音生成语音画像
依据用户_兴趣模型，确定答复语优选风格
从多风格语音库中选择与确定文本表达相同语义，并且为用户偏爱风格的语音片段进行回复

图6

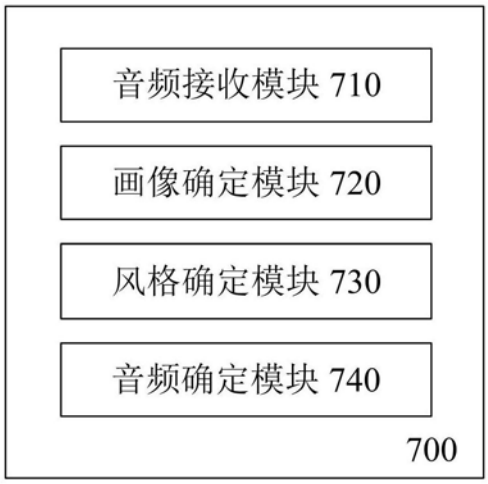

音频接收模块 710
画像确定模块 720
风格确定模块 730
音频确定模块 740
700

图7


处理器 810
存储器 820
800

图8