(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 110085221 A
(43)申请公布日 2019.08.02

(21)申请号 201810079429.2

(22)申请日 2018.01.26

(71)申请人 上海智臻智能网络科技股份有限公司
地址 201803 上海市嘉定区金沙江西路1555弄398号7层

(72)发明人 王慧 余世经 朱频频

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G10L 15/22(2006.01)

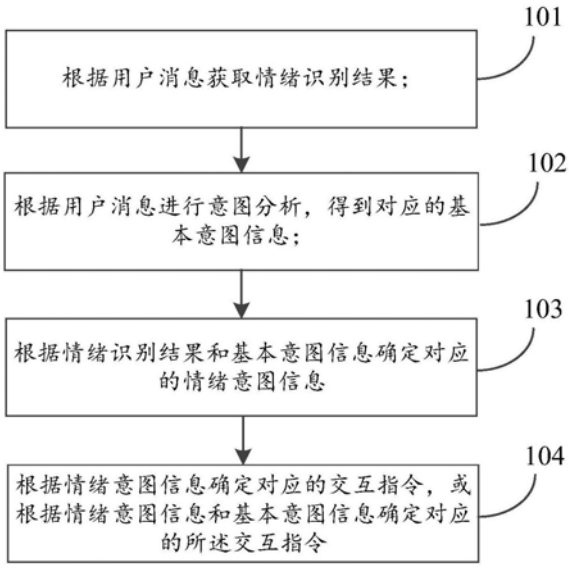
权利要求书2页 说明书16页 附图6页

(54)发明名称

语音情感交互方法、计算机设备和计算机可读存储介质

(57)摘要

本发明实施例提供了一种语音情感交互方法、计算机设备和计算机可读存储介质，解决了现有技术中的智能交互方式无法分析出用户消息的深层次意图以及无法提供更人性化的交互体验的问题。该语音情感交互方法包括：根据用户语音消息获取情绪识别结果，其中，情绪识别结果中至少包括音频情绪识别结果，或情绪识别结果中至少包括音频情绪识别结果和文本情绪识别结果；根据用户语音消息的文本内容进行意图分析，得到对应的基本意图信息；根据情绪识别结果和基本意图信息确定对应的情绪意图信息；以及根据情绪意图信息确定对应的交互指令，或根据情绪意图信息和基本意图信息确定对应的交互指令。



CN 110085221 A

1.一种语音情感交互方法,其特征在于,包括:

根据用户语音消息获取情绪识别结果,其中,所述情绪识别结果中至少包括音频情绪识别结果,或所述情绪识别结果中至少包括音频情绪识别结果和所述文本情绪识别结果;

根据所述用户语音消息的文本内容进行意图分析,得到对应的基本意图信息;

根据所述情绪识别结果和所述基本意图信息确定对应的情绪意图信息;

以及根据所述情绪意图信息确定对应的所述交互指令,或根据所述情绪意图信息和所述基本意图信息确定对应的所述交互指令。

2.根据权利要求1所述的语音情感交互方法,其特征在于,所述交互指令包括以下一种或多种情感呈现模态:文本输出情感呈现模态、乐曲播放情感呈现模态、语音情感呈现模态、图像情感呈现模态和机械动作情感呈现模态。

3.根据权利要求1所述的语音情感交互方法,其特征在于,所述情绪意图信息包括与所述情绪识别结果对应的情感需求信息;或,

所述情绪意图信息包括与所述情绪识别结果对应的所述情感需求信息以及所述情绪识别结果与所述基本意图信息的关联关系。

4.根据权利要求1所述的语音情感交互方法,其特征在于,所述根据用户语音消息获取情绪识别结果包括:

根据所述用户语音消息的音频数据获取音频情绪识别结果;以及,根据所述音频情绪识别结果确定所述情绪识别结果;

或,

根据所述用户语音消息的音频数据获取音频情绪识别结果,且根据所述用户语音消息的文本内容获取文本情绪识别结果;以及,根据所述音频情绪识别结果以及所述文本情绪识别结果确定所述情绪识别结果。

5.根据权利要求4所述的语音情感交互方法,其特征在于,所述音频情绪识别结果包括多个情绪分类中的一种或多种;或,所述音频情绪识别结果对应多维情感空间中的一个坐标点;

或,所述音频情绪识别结果以及所述文本情绪识别结果分别包括多个情绪分类中的一种或多种;或,所述音频情绪识别结果以及所述文本情绪识别结果分别对应多维情感空间中的一个坐标点;

其中,所述多维情感空间中的每个维度对应一个心理学定义的情感因素,每个所述情绪分类包括多个情绪强度级别。

6.根据权利要求4所述的语音情感交互方法,其特征在于,所述根据所述用户语音消息的音频数据获取音频情绪识别结果包括:

提取所述用户语音消息的音频特征向量,其中所述用户语音消息对应所述待识别音频流中的一段话;

将所述用户语音消息的音频特征向量与多个情绪特征模型进行匹配,其中所述多个情绪特征模型分别对应多个情绪分类中的一个;以及

将匹配结果为相匹配的所述情绪特征模型所对应的情绪分类作为所述用户语音消息的情绪分类。

7.根据权利要求1所述的语音情感交互方法,其特征在于,所述根据所述用户语音消息

的音频数据获取音频情绪识别结果进一步包括：

确定所述待识别音频流中的语音开始帧以及语音结束帧；以及

提取所述语音开始帧与所述语音结束帧之间的音频流部分作为所述用户语音消息。

8.根据权利要求7所述的语音情感交互方法，其特征在于，所述确定所述待识别音频流中的语音开始帧以及语音结束帧包括：

判断所述待识别音频流中的语音帧是发音帧还是非发音帧；

在上一段语音片段的所述语音结束帧之后或者当前未识别到第一段语音片段时，当有第一预设数量个语音帧被连续判断为发音帧时，将所述第一预设数量个语音帧中的第一个语音帧作为当前语音片段的所述语音开始帧；以及

在当前语音片段的所述语音开始帧之后，当有第二预设数量个语音帧被连续判断为非发音帧时，将所述第二预设数量个语音帧中的第一个语音帧作为当前语音片段的所述语音结束帧。

9.一种计算机设备，包括存储器、处理器以及存储在所述存储器上被所述处理器执行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至8中任一项所述方法的步骤。

10.一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，所述计算机程序被处理器执行时实现如权利要求1至8中任一项所述方法的步骤。

语音情感交互方法、计算机设备和计算机可读存储介质

技术领域

[0001] 本发明涉及智能交互技术领域，具体涉及一种语音情感交互方法、计算机设备和计算机可读存储介质。

背景技术

[0002] 随着人工智能技术的不断发展以及人们对于交互体验要求的不断提高，智能交互方式已逐渐开始替代一些传统的人机交互方式，并且已成为一个研究热点。然而，现有智能交互方式仅能大概分析出用户消息的语义内容，并无法识别用户当前的情绪状态，因而无法根据用户的情绪状态分析出用户消息所实际想要表达的深层次的情绪需求，也无法根据用户消息提供更人性化的交互体验。例如，对于一个正在赶时间的情绪状态为焦急的用户与一个刚开始做行程规划的情绪状态为平和的用户，在询问航班时间信息时所希望得到的回复方式肯定是有所不同的，而根据现有的基于语义的智能交互方式，不同的用户所得到的回复方式是相同的，例如只是把对应的航班时间信息程序给用户。

发明内容

[0003] 有鉴于此，本发明实施例提供了一种语音情感交互方法、计算机设备和计算机可读存储介质，解决了现有技术中的智能交互方式无法分析出用户消息的深层次意图以及无法提供更人性化的交互体验的问题。

[0004] 本发明一实施例提供的一种语音情感交互方法包括：

[0005] 根据用户消息获取情绪识别结果，所述用户消息中至少包括用户语音消息；

[0006] 根据所述用户语音消息的文本内容进行意图分析，得到对应的基本意图信息；以及

[0007] 根据所述情绪识别结果和所述基本意图信息确定对应的交互指令。

[0008] 本发明一实施例提供的一种智能交互装置包括：

[0009] 情绪识别模块，配置为根据用户消息获取情绪识别结果，所述用户消息中至少包括用户语音消息；

[0010] 基本意图识别模块，配置为根据所述用户语音消息的文本内容进行意图分析，得到对应的基本意图信息；以及

[0011] 交互指令确定模块，配置为根据所述情绪识别结果和所述基本意图信息确定对应的交互指令。

[0012] 本发明一实施例提供的一种计算机设备包括：存储器、处理器以及存储在所述存储器上被所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如前所述方法的步骤。

[0013] 本发明一实施例提供的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前所述方法的步骤。

[0014] 本发明实施例提供的一种语音情感交互方法、计算机设备和计算机可读存储介

质，在理解用户的基本意图信息的基础上，结合了基于用户语音消息获取的情绪识别结果，并进一步根据基本意图信息和情绪识别结果确定对应的情绪意图信息，并确定对应的带有情绪的交互指令，从而解决了现有技术中的智能交互方式无法分析出用户消息的深层次意图以及无法提供更人性化的交互体验的问题。

附图说明

[0015] 图1所示为本发明一实施例提供的一种语音情感交互方法的流程示意图。

[0016] 图2所示为本发明一实施例提供的语音情感交互方法中确定情绪识别结果的流程示意图。

[0017] 图3所示为本发明一实施例提供的语音情感交互方法中确定情绪识别结果的流程示意图。该实施例中的用户消息也至少包括用户语音消息，情绪识别。

[0018] 图4所示为本发明一实施例所提供的语音情感交互方法中根据用户语音消息的音频数据获取音频情绪识别结果的流程示意图。

[0019] 图5所示为本发明一实施例所提供的语音情感交互方法中建立情绪特征模型的流程示意图。

[0020] 图6所示为本发明一实施例所提供的语音情绪识别方法中提取用户语音消息的流程示意图。

[0021] 图7所示为本发明一实施例所提供的语音情感交互方法中确定语音开始帧以及语音结束帧的流程示意图。

[0022] 图8所示为本发明一实施例所提供的语音情感交互方法中检测发音帧或非发音帧的流程示意图。

[0023] 图9所示为本发明一实施例提供的语音情感交互方法中根据用户语音消息获取基本意图信息的流程示意图。

具体实施方式

[0024] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0025] 图1所示为本发明一实施例提供的一种语音情感交互方法的流程示意图。如图1所示，该语音情感交互方法包括如下步骤：

[0026] 步骤101：根据用户语音消息获取情绪识别结果，其中，所述情绪识别结果中至少包括音频情绪识别结果，或所述情绪识别结果中至少包括音频情绪识别结果和所述文本情绪识别结果。

[0027] 用户语音消息是指在与用户交互的过程中由用户输入的语音或获取到的与用户的交互意图和需求相关的语音信息。例如，在呼叫中心系统的客服交互场景中，用户消息的具体形式就可以包括用户发出的用户语音消息，此时的用户可能是客户端也可能是服务端；再例如在智能机器人交互场景中，用户消息就可包括用户通过该智能机器人的输入模块输入的信息(例如文字或语音等)，或该智能机器人的采集模块所采集到的用户的信息

(例如面部表情、动作姿势等)。本发明对用户语音消息的具体来源和具体形式不做限定。

[0028] 用户语音消息由于不同情绪状态的用户语音消息的音频数据会包括不同的音频特征,此时就可根据用户语音消息的音频数据获取音频情绪识别结果,并根据音频情绪识别结果确定情绪识别结果。

[0029] 根据该用户消息所获取到的情绪识别结果将在后续的过程中与基本意图信息进行结合,以推测用户的情绪意图。

[0030] 步骤102:根据用户消息进行意图分析,得到对应的基本意图信息。

[0031] 基本意图信息对应的是用户消息所直观反映出的意图,但并无法反映用户当前状态下的真实情绪需求,因此才需要结合情绪识别结果来综合确定用户消息所实际想要表达的深层次的意图和情绪需求。例如,对于一个正在赶时间的情绪状态为焦急的用户与一个刚开始做行程规划的情绪状态为平和的用户,当二者所发出的用户语音消息的内容同样为询问航班信息时,所得到的基本意图信息也是相同的,都为查询航班信息,但二者所需要的情绪需求显然是不同的。

[0032] 应当理解,根据用户消息的具体形式不同,基本意图信息的具体内容和获取方式也可有所不同。例如,当用户消息包括用户语音消息时,基本意图信息就可根据用户语音消息的文本内容进行意图分析得出,该基本意图信息对应的是用户语音消息的文本内容在语义层面所反映出的意图,并不会带有任何感情色彩。

[0033] 在本发明一实施例中,为了进一步提高所获取的基本意图信息的准确度,还可根据当前的用户语音消息,并结合过往的用户语音消息和/或后续的用户语音消息进行意图分析,得到对应的基本意图信息。例如,当前用户语音消息的意图中可能缺乏一些关键词和槽位(slot),但这些内容可通过过往的用户语音消息和/或后续的用户语音消息获取。例如,当前的用户语音消息的内容为“有什么特产?”时,其中的主语(slot)是缺失的,但通过结合过往的用户语音消息“常州天气如何?”即可提取“常州”作为主语,这样所最终获取的当前用户语音消息的基本意图信息就可为“常州有什么特产?”。

[0034] 步骤103:根据情绪识别结果和基本意图信息确定对应的情绪意图信息。

[0035] 步骤104:根据情绪意图信息确定对应的交互指令,或根据情绪意图信息和基本意图信息确定对应的所述交互指令。此时的情绪意图信息可以有具体的内容。

[0036] 情绪识别结果和基本意图信息与交互指令之间的对应关系可通过与学习过程建立。在本发明一实施例中,交互指令的内容和形式包括以下一种或多种情感呈现模态:文本输出情感呈现模态、乐曲播放情感呈现模态、语音情感呈现模态、图像情感呈现模态和机械动作情感呈现模态。然而应当理解,交互指令的具体情感呈现模态也可根据交互场景的需求而调整,本发明对交互指令的具体内容和形式并不做限定。

[0037] 具体而言,情绪意图信息的具体内容指的是带有感情色彩的意图信息,能在反映基本意图的同时反映用户消息的情绪需求,情绪意图信息与情绪识别结果和基本意图信息之间的对应关系可通过预学习过程预先建立。在本发明一实施例中,该情绪意图信息可包括与情绪识别结果对应的情感需求信息,或可包括与情绪识别结果对应的情感需求信息以及情绪识别结果与基本意图信息的关联关系。情绪识别结果与基本意图信息的关联关系可为预先设定(比如通过规则设定,或者逻辑判断)。例如,当情绪识别结果的内容为“焦急”,基本意图信息的内容为“挂失信用卡”时,确定出的情绪意图信息的内容就可包括情绪识别

结果与基本意图信息的关联关系:“挂失信用卡,用户很焦急,可能信用卡丢失或被盗”,同时所确定的情感需求信息就可为“安慰”。情绪识别结果与基本意图信息的关联关系也可以是基于特定训练过程得到的模型(比如训练好的端到端模型,可以通过输入情绪识别结果和基本意图信息直接输出情感意图)。这个训练模型可以是固定的深度网络模型(例如包括了预先设定好的规则),也可以通过在线学习不断更新(比如利用增强学习模型,在模型中设定目标函数和奖励函数,随着人机交互次数增加,该深度网络模型也可以不断更新演化)。

[0038] 然而应当理解,情绪意图信息也可仅作为映射关系的标识存在。情绪意图信息与交互指令之间的对应关系、以及情绪意图信息和基本意图信息与交互指令之间的对应关系也可通过预学习过程预先建立。

[0039] 应当理解,在一些应用场景下,是需要将对该情绪意图信息的回馈内容呈现出来的。例如在一些客服交互场景下,需要将根据客户的语音内容分析出的情绪意图信息呈现给客服人员,以起到提醒作用,此时就必然要确定对应的情绪意图信息,并将对该情绪意图信息的回馈内容呈现出来。然而在另外一些应用场景下,需要直接给出对应的交互指令,而并不需要呈现出对该情绪意图信息的回馈内容,此时也可根据情绪识别结果和基本意图信息直接确定对应的交互指令,而不用生成情绪意图信息。

[0040] 在本发明一实施例中,为了进一步提高所获取的情绪意图信息的准确度,也可以根据当前的用户语音消息的情绪识别结果和基本意图信息,并结合过往的用户语音消息和/或后续的用户语音消息的情绪识别结果和基本意图信息,确定对应的情绪意图信息。此时就需要实时记录当前的用户语音消息的情绪识别结果和基本意图信息,以便于在根据其他的用户语音消息确定情绪意图信息时作为参考。例如,当前的用户语音消息的内容为“没有银行卡怎么取钱?”,所获取情绪识别结果为“焦急”,但是根据当前的用户语音消息无法准确判断“焦急”情绪的原因。此时可以追溯过往的用户语音消息和/或后续的用户语音消息,结果发现过往的一个用户语音消息为“银行卡如何挂失?”,于是可以推测用户的情绪意图信息可为“银行卡丢失导致了情绪焦急,希望咨询如何挂失或者在无银行卡的情况下取钱”。这时候可以针对情绪意图信息生成交互指令,如播放如下安慰语音“无卡取款请按照如下步骤操作,请您不要着急,丢失银行卡还可以按照下述方法操作……”。

[0041] 在本发明一实施例中,为了进一步提高所获取的对应的交互指令的准确度,也可以根据当前的用户语音消息的情绪意图信息和基本意图信息,并结合过往的用户语音消息和/或后续的用户语音消息的情绪意图信息和基本意图信息,确定对应的交互指令。此时就需要实时记录当前的用户语音消息的情绪识别结果和基本意图信息,以便于在根据其他的用户语音消息确定交互指令时作为参考。

[0042] 由此可见,本发明实施例提供的语音情感交互方法,在理解用户的基本意图信息的基础上,结合了基于用户消息获取的情绪识别结果,并进一步推测用户的情绪意图,或直接根据基本意图信息和情绪识别结果给出带有情绪的交互指令,从而解决了现有技术中的智能交互方式无法分析出用户消息的深层次意图和情绪需求、以及无法提供更人性化的交互体验的问题。

[0043] 在本发明一实施例中,根据所述用户语音消息的音频数据获取音频情绪识别结果;以及,根据所述音频情绪识别结果确定所述情绪识别结果。在本发明另一实施例中,当

用户消息包括用户语音消息时，情绪识别结果可根据音频情绪识别结果以及文本情绪识别结果综合确定。具体而言，需要根据用户语音消息的音频数据获取音频情绪识别结果，并根据用户语音消息的文本内容获取文本情绪识别结果，然后根据音频情绪识别结果以及文本情绪识别结果综合确定情绪识别结果。然而如前所述，也可以仅根据音频情绪识别结果确定最终的情绪识别结果，本发明对此不作限定。

[0044] 进一步地，音频情绪识别结果包括多个情绪分类中的一种或多种；或，音频情绪识别结果对应多维情感空间中的一个坐标点；

[0045] 或，音频情绪识别结果以及所述文本情绪识别结果分别包括多个情绪分类中的一种或多种；或，音频情绪识别结果以及文本情绪识别结果分别对应多维情感空间中的一个坐标点；

[0046] 其中，多维情感空间中的每个维度对应一个心理学定义的情感因素，每个情绪分类包括多个情绪强度级别。

[0047] 应当理解，音频情绪识别结果和文本情绪识别结果可通过多种方式来表征。在本发明一实施例中，可采用离散的情绪分类的方式来表征情绪识别结果，此时音频情绪识别结果和文本情绪识别结果可分别包括多个情绪分类中的一种或多种。例如，在客服交互场景中，该多个情绪分类就可包括：满意分类、平静分类以及烦躁分类，以对应客服交互场景中用户可能出现的情绪状态；或者，该多个情绪分类可包括：满意分类、平静分类、烦躁分类以及生气分类，以对应客服交互场景中客服人员可能出现的情绪状态。然而应当理解，这些情绪分类的种类和数量可根据实际的应用场景需求而调整，本发明对情绪分类的种类和数量同样不做严格限定。在一进一步实施例中，每个情绪分类还可包括多个情绪强度级别。具体而言，情绪分类和情绪强度级别可以认为是两个维度参数，可以彼此独立(例如，每种情绪分类都有对应的N种情绪强度级别，例如轻度、中度和重度)，也可以有预设的对应关系(例如“烦躁”情绪分类包括三种情绪强度级别，轻度、中度和重度；而“满意”情绪分类只包括两种情绪强度级别，中度和重度)。由此可见，此时的情绪强度级别可以看做是情绪分类的一个属性参数，当通过情绪识别过程确定一种情绪分类时，也就确定了该情绪分类的情绪强度级别。

[0048] 在本发明另一实施例中，还可采用非离散的维度情绪模型的方式来表征情绪识别结果。此时音频情绪识别结果和文本情绪识别结果可分别对应多维情感空间中的一个坐标点，多维情感空间中的每个维度对应一个心理学定义的情感因素。例如，可采用PAD(Pleasure Arousal Dominanc)三维情绪模型。该模型认为情绪具有愉悦度，激活度和优势度三个维度，每种情绪都可通过这三个维度所分别对应的情感因素来表征。其中P代表愉悦度，表示个体情绪状态的正负特性；A代表激活度，表示个体的神经胜利激活水平；D代表优势度，表示个体对情景和他人的控制状态。

[0049] 应当理解，音频情绪识别结果和文本情绪识别结果也可采用其他的表征方式来表征，本发明对具体的表征方式并不做限定。

[0050] 图2所示为本发明一实施例提供的语音情感交互方法中确定情绪识别结果的流程示意图。该实施例中的用户消息至少包括用户语音消息，情绪识别结果需要根据音频情绪识别结果和文本情绪识别结果综合确定，且音频情绪识别结果和文本情绪识别结果分别包括多个情绪分类中的一种或多种，此时该确定情绪识别结果的方法可包括如下步骤：

[0051] 步骤201:如果音频情绪识别结果和文本情绪识别结果包括相同的情绪分类,则将相同的情绪分类作为情绪识别结果。

[0052] 步骤202:如果音频情绪识别结果和文本情绪识别结果没有包括相同的情绪分类,则将音频情绪识别结果和文本情绪识别结果共同作为情绪识别结果。

[0053] 应当理解,虽然在步骤202中限定了当音频情绪识别结果和文本情绪识别结果没有包括相同的情绪分类时,将音频情绪识别结果和文本情绪识别结果共同作为情绪识别结果,但在本发明的其他实施例中,也可采取更为保守的交互策略,例如直接生成报错信息或不输出情绪识别结果等,以免对交互过程造成误导,本发明对音频情绪识别结果和文本情绪识别结果没有包括相同的情绪分类时的处理方式并不做严格限定。

[0054] 图3所示为本发明一实施例提供的语音情感交互方法中确定情绪识别结果的流程示意图。该实施例中的用户消息也至少包括用户语音消息,情绪识别结果也需要根据

[0055] 音频情绪识别结果和文本情绪识别结果综合确定,且音频情绪识别结果和文本情绪识别结果分别包括多个情绪分类中的一种或多种,该确定情绪识别结果的方法可包括如下步骤:

[0056] 步骤301:计算音频情绪识别结果中情绪分类的置信度以及文本情绪识别结果中情绪分类的置信度。

[0057] 在统计学上,置信度也称为可靠度、置信水平、或置信系数。由于样本具有随机性,当利用抽样对总体参数作出估计时,所得出的结论总是不确定的。因此,可采用数理统计中的区间估计法来估计一个估计值与总体参数之间的误差在一定允许的范围以内的概率有多大,这个相应的概率即称作置信度。例如,假设预设的情绪分类与表征情绪分类的一个变量有关,即,根据该变量值的大小情绪分类可对应到不同的取值。当要获取语音情绪识别结果的置信度时,先通过多次的音频情绪识别/文本情绪识别过程得到该变量的多个测量值,然后将该多个测量值的均值作为一个估计值。再通过区间估计法来估计该估计值与该变量的真值之间的误差范围在一定范围内的概率,这个概率值越大说明这个估计值越准确,即当前的情绪分类的置信度越高。。

[0058] 步骤302:判断音频情绪识别结果中置信度最高的情绪分类与文本情绪识别结果中置信度最高的情绪分类是否相同。如果判断结果为是,则执行步骤303,否则执行步骤304。

[0059] 步骤303:将音频情绪识别结果中置信度最高的情绪分类或文本情绪识别结果中置信度最高的情绪分类作为情绪识别结果。

[0060] 例如,当音频情绪识别结果包括了满意分类(置信度为a1)和平静分类(置信度为a2),而文本情绪识别结果仅包括了满意分类(置信度为b1)时,且a1＞a2时,则将满意分类作为最终的情绪识别结果。

[0061] 步骤304:比较音频情绪识别结果中置信度最高的情绪分类的置信度与文本情绪识别结果中置信度最高的情绪分类的置信度。

[0062] 在本发明一实施例中,考虑到在实际的应用场景中,根据情绪识别的具体算法以及用户语音消息的类型和内容的限制,可选择音频情绪识别结果和文本情绪识别结果中的一个作为主要考虑的情绪识别结果输出,而将另一个作为辅助考虑的情绪识别结果输出,然后再利用置信度和情绪强度级别等因素来综合确定最终的情绪识别结果。应当理解,选

择音频情绪识别结果和文本情绪识别结果中的哪一个作为主要考虑的情绪识别结果输出可根据实际的场景而定。然而本发明对选择音频情绪识别结果和文本情绪识别结果中的哪一个作为主要考虑的情绪识别结果输出并不做限定。

[0063] 在本发明一实施例中，将音频情绪识别结果作为主要考虑的情绪识别结果输出，将文本情绪识别结果作为辅助考虑的情绪识别结果输出。此时，如果音频情绪识别结果中置信度最高的情绪分类的置信度大于文本情绪识别结果中置信度最高的情绪分类的置信度，执行步骤305；如果音频情绪识别结果中置信度最高的情绪分类的置信度小于文本情绪识别结果中置信度最高的情绪分类的置信度，执行步骤306；如果音频情绪识别结果中置信度最高的情绪分类的置信度等于文本情绪识别结果中置信度最高的情绪分类的置信度，执行步骤309。

[0064] 步骤305：将音频情绪识别结果中置信度最高的情绪分类作为情绪识别结果。

[0065] 由于选择了音频情绪识别结果作为主要考虑的情绪识别结果输出，因此本来就应优先考虑音频情绪识别结果中的情绪分类；再加上音频情绪识别结果中置信度最高的情绪分类的置信度大于文本情绪识别结果中置信度最高的情绪分类的置信度，因此就可选择主要考虑的音频情绪识别结果中可信度最高的情绪分类作为情绪识别结果。例如，当音频情绪识别结果包括了满意分类(置信度为a1)和平静分类(置信度为a2)，而文本情绪识别结果仅包括了平静分类(置信度为b1)时，a1＞a2且a1＞b1时，则将满意分类作为最终的情绪识别结果。

[0066] 步骤306：判断音频情绪识别结果中是否包括了文本情绪识别结果中置信度最高的情绪分类。如果判断结果为是，则执行步骤307；如果判断结果为否，则执行步骤309。

[0067] 例如，当音频情绪识别结果包括了满意分类(置信度为a1)和平静分类(置信度为a2)，而文本情绪识别结果仅包括了平静分类(置信度为b1)，a1＞a2且a1＜b1时，则需要判断一下音频情绪识别结果中是否包括了文本情绪识别结果中的置信度最高的平静分类。

[0068] 步骤307：进一步判断音频情绪识别结果中的文本情绪识别结果中置信度最高的情绪分类的情绪强度级别是否大于第一强度阈值。如果进一步判断的结果为是，则执行步骤308；否则执行步骤309。

[0069] 步骤308：将文本情绪识别结果中置信度最高的情绪分类作为情绪识别结果。

[0070] 执行到步骤308意味着文本情绪识别结果中的该置信度最高的情绪分类不仅可信度高，且情绪的倾向十分明显，因此可将文本情绪识别结果中置信度最高的情绪分类作为情绪识别结果。

[0071] 步骤309：将音频情绪识别结果中置信度最高的情绪分类作为情绪识别结果，或将音频情绪识别结果中置信度最高的情绪分类和文本情绪识别结果中置信度最高的情绪分类共同作为情绪识别结果。

[0072] 当音频情绪识别结果中置信度最高的情绪分类的置信度等于文本情绪识别结果中置信度最高的情绪分类的置信度，或音频情绪识别结果中并未包括文本情绪识别结果中置信度最高的情绪分类，或即使音频情绪识别结果中包括了文本情绪识别结果中置信度最高的情绪分类但该情绪分类的情绪强度级别不够高时，说明此时尚无法根据音频情绪识别结果和文本情绪识别结果输出一个统一的情绪分类作为最终的情绪识别结果。此时，在本发明一实施例中，考虑到选择了音频情绪识别结果作为主要考虑的情绪识别结果输出，因

此直接将音频情绪识别结果中置信度最高的情绪分类作为情绪识别结果即可。在本发明另一实施例中，也可将音频情绪识别结果和文本情绪识别结果共同作为情绪识别结果。并在后续的过程中结合过往的用户语音消息和/或后续的用户语音消息的情绪识别结果和基本意图信息，确定对应的情绪意图信息。

[0073] 在本发明一实施例中，音频情绪识别结果和文本情绪识别结果分别对应多维情感空间中的一个坐标点，此时就可将音频情绪识别结果和文本情绪识别结果在多维情感空间中的坐标点的坐标值进行加权平均处理，将加权平均处理后得到的坐标点作为情绪识别结果。例如，当采用PAD三维情绪模型时，音频情绪识别结果表征为(p1,a1,d1)，文本情绪识别结果表征为(p2,a2,d2)，那么最终的情绪识别结果就可表征为((p1+p2)/2,(a1+1.3*a2)/2,(d1+0.8*d2)/2)，其中的1.3和0.8为权重系数。采用非离散的维度情绪模型更便于以量化的方式计算出最终的情绪识别结果。然而应当理解，二者的组合方式并不限于上述的加权平均处理，本发明对当音频情绪识别结果和文本情绪识别结果分别对应多维情感空间中的一个坐标点时确定情绪识别结果的具体方式不做限定。

[0074] 图4所示为本发明一实施例所提供的语音情感交互方法中根据用户语音消息的音频数据获取音频情绪识别结果的流程示意图。如图4所示，该根据用户语音消息的音频数据获取音频情绪识别结果的流程包括：

[0075] 步骤401：提取待识别音频流中的用户语音消息的音频特征向量，其中用户语音消息对应待识别音频流中的一段话。

[0076] 音频特征向量包括至少一个音频特征在至少一个向量方向上的取值。这样其实是利用一个多维的向量空间来表征所有的音频特征，在该向量空间中，音频特征向量的方向和取值可看做是由很多个音频特征各自在不同的向量方向上的取值在向量空间内求和而成，其中每个音频特征在一个向量方向上的取值可看做音频特征向量的一个分量。包括了不同情绪的用户语音消息必然有着不同的音频特征，本发明正是利用不同情绪与不同音频特征之间的对应关系来识别用户语音消息的情绪的。具体而言，音频特征可包括以下几种中的一种或多种：能量特征、发音帧数特征、基音频率特征、共振峰特征、谐波噪声比特征以及梅尔倒谱系数特征。在本发明一实施例中，可在该向量空间内设置以下向量方向：比例值、均值、最大值、中值以及标准差。

[0077] 能量特征指的是用户语音消息的功率谱特征，可通过功率谱求和得到。计算公式可为：$E(k)=\sum_{j=0}^{N-1}P(k,j)$，其中E表示能量特征的取值，k代表帧的编号，j代表频率点的编号，N为帧长，P表示功率谱的取值。在本发明一实施例中，能量特征可包括短时能量一阶差分、和/或预设频率以下的能量大小。短时能量一阶差分的计算公式可为：

[0078] VE(k)＝(-2*E(k-2)-E(k-1)+E(k+1)+2*E(k+2))/3；

[0079] 预设频率以下的能量大小可通过比例值来衡量，例如500Hz以下频段能量占总能量的比例值的计算公式可为：

[0080] $$p1=\frac{\sum_{k=k1}^{k2}\sum_{j=1}^{j_{500}}P(k,j)}{\sum_{k=k1}^{k2}\sum_{j=1}^{N/2-1}P(k,j)}；$$

[0081] 其中j_{500}为500Hz对应的频点编号，k1为待识别的用户语音消息的语音开始帧的编号，k2为待识别的用户语音消息的语音结束帧的编号。

[0082] 发音帧数特征指的是用户语音消息内发音帧的数量大小，该发音帧的数量大小也可通过比例值来衡量。例如记该用户语音消息内发音帧和不发音帧的数量分别为n1和n2，则发音帧数和不发音帧数的比例为p2＝n1/n2，发音帧数和总帧数的比例为：p3＝n1/(n1+n2)。

[0083] 基音频率特征可采用基于线性预测(LPC)误差信号的自相关函数的算法来提取。基音频率特征可包括基音频率和/或基音频率一阶差分。基音频率的算法流程可如下：首先，计算发音帧x(k)的线性预测系数并计算线性预测估计信号$\bar{x}(k)$；其次，计算误差信号的自相关函数c1：$c1 = xcorr(x(k) - \bar{x}(k))$；然后，在对应基音频率为80-500Hz的偏移量范围内，寻找自相关函数的最大值，记录其对应的偏移量Δh。基音频率F0的计算公式为：F0＝Fs/Δh，其中Fs为采样频率。

[0084] 共振峰特征可采用基于线性预测的多项式求根的算法来提取，可包括第一共振峰、第二共振峰和第三共振峰，以及该三个共振峰的一阶差分。谐波噪声比(HNR)特征可采用基于独立分量分析(ICA)的算法来提取。梅尔倒谱(MFCC)系数特征可包括1-12阶梅尔倒谱系数，可采用通用的梅尔倒谱系数计算流程获取，在此不再赘述。

[0085] 应当理解，具体提取哪些音频特征向量可根据实际场景的需求而定，本发明对所提取音频特征向量所对应音频特征的种类、数量以及向量方向均不做限定。然而在本发明一实施例中，为了获得最优的情绪识别效果，可同时提取上述的六个音频特征：能量特征、发音帧数特征、基音频率特征、共振峰特征、谐波噪声比特征以及梅尔倒谱系数特征。例如，当同时提取上述的六个音频特征时，所提取的音频特征向量就可包括如下表1所示的173个分量，采用下表1的音频特征向量以及高斯模型(GMM)作为情绪特征模型来对casia汉语情绪语料库进行语音情绪识别的准确度可以达到74％至80％。

[0086] 表1

[0087]

分量编号	分量名称

[0088]

1-5	短时能量一阶差分的均值、最大值、最小值、中值、标准差
6	500Hz 以下频段能量占总能量的比例
7-8	发音帧数和不发音帧数的比例、发音帧数和总帧数的比例
9-13	基音频率的均值、最大值、最小值、中值、标准差
14-18	基音频率一阶差分的均值、最大值、最小值、中值、标准差
19-33	第一共振峰、第二共振峰以及第三共振峰各自的均值、最大值、最小值、中值、标准差
34-48	第一共振峰、第二共振峰以及第三共振峰各自一阶差分的均值、最大值、最小值、中值、标准差
49-53	谐波噪声比的均值、最大值、最小值、中值、标准差
54-113	1-12 阶梅尔倒谱系数的均值、最大值、最小值、中值、标准差
114-173	1-12 阶梅尔倒谱系数一阶差分的均值、最大值、最小值、中值、标准差

[0089] 在本发明一实施例中，待识别音频流可为客服交互音频流，用户语音消息对应待识别音频流中的一次用户输入语音段或一次客服输入语音段。由于客户交互过程往往是一问一答的形式，因此一次用户输入语音段就可对应一次交互过程中用户的一次提问或回答，而一次客服输入语音段就可对应一次交互过程中客服人员的一次提问或回答。由于一般认为用户或客服在一次提问或回答中能完整的表达情绪，因此通过将一次用户输入语音段或一次客服输入语音段作为情绪识别的单元，既能保证情绪识别的完整性，又能保证客服交互过程中情绪识别的实时性。

[0090] 步骤402：将用户语音消息的音频特征向量与多个情绪特征模型进行匹配，其中多个情绪特征模型分别对应多个情绪分类之一。

[0091] 这些情绪特征模型可通过对包括多个情绪分类对应的情绪分类标签的多个预设用户语音消息各自的音频特征向量进行预学习而建立，这样就相当于建立起了情绪特征模型与情绪分类之间的对应关系，每个情绪特征模型可对应一个情绪分类。如图5所示，该建立情绪特征模型的预学习过程可包括：首先将包括多个情绪分类对应的情绪分类标签的多个预设用户语音消息各自的音频特征向量进行聚类处理，得到预设情绪分类的聚类结果(S51)；然后，根据聚类结果，将每个聚类中的预设用户语音消息的音频特征向量训练为一个情绪特征模型(S52)。基于这些情绪特征模型，通过基于音频特征向量的匹配过程即可获得与当前用户语音消息对应的情绪特征模型，并进而获得对应的情绪分类。

[0092] 在本发明一实施例中，这些情绪特征模型可为混合高斯模型(GMM)(混合度可为5)。这样可先采用K-means算法对同一情绪分类的语音样本的情绪特征向量进行聚类，根据聚类结果计算出混合高斯模型的参数的初始值(迭代次数可为50)。然后再采用E-M算法训练出各类情绪分类对应的混合高斯模型(迭代次数为200)。当要利用这些混合高斯模型进行情绪分类的匹配过程时，可通过计算当前用户语音消息的音频特征向量分别与多个情绪特征模型之间的似然概率，然后通过衡量该似然概率来确定匹配的情绪特征模型，例如将似然概率大于预设阈值且最大的情绪特征模型作为匹配的情绪特征模型。

[0093] 应当理解，虽然在上面的描述中阐述了情绪特征模型可为混合高斯模型，但其实该情绪特征模型还可通过其他形式实现，例如支持向量机(SVM)模型、K最近邻分类算法(KNN)模型、马尔科夫模型(HMM)以及神经网络(ANN)模型等。

[0094] 在本发明一实施例中，该多个情绪分类可包括：满意分类、平静分类以及烦躁分类，以对应客服交互场景中用户可能出现的情绪状态。在另一实施例中，该多个情绪分类可包括：满意分类、平静分类、烦躁分类以及生气分类，以对应客服交互场景中客服人员可能出现的情绪状态。即，待识别音频流为客服交互场景中的用户客服交互音频流时，若当前用户语音消息对应一次客服输入语音段时，该多个情绪分类可包括：满意分类、平静分类以及烦躁分类；若当前用户语音消息对应一次用户输入语音段时，该多个情绪分类可包括：满意分类、平静分类、烦躁分类以及生气分类。通过上述的对用户以及客服的情绪分类，可以更简洁的适用于呼叫中心系统，减少计算量并满足呼叫中心系统的情绪识别需求。然而应当理解，这些情绪分类的种类和数量可根据实际的应用场景需求而调整。

[0095] 步骤403：将匹配结果为相匹配的情绪特征模型所对应的情绪分类作为用户语音消息的情绪分类。

[0096] 如前所述，由于情绪特征模型与情绪分类之间存在对应关系，因此当根据步骤402

的匹配过程确定了相匹配的情绪特征模型后，该匹配的情绪特征模型所对应的情绪分类便为所识别出的情绪分类。例如，当这些情绪特征模型为混合高斯模型时，该匹配过程就可通过衡量当前用户语音消息的音频特征向量分别与多个情绪特征模型之间的似然概率的方式实现，然后将似然概率大于预设阈值且最大的情绪特征模型所对应的情绪分类作为用户语音消息的情绪分类即可。

[0097] 由此可见，本发明实施例提供的一种语音情绪识别方法，通过提取待识别音频流中的用户语音消息的音频特征向量，并利用预先建立的情绪特征模型对所提取的音频特征向量进行匹配，从而实现了对用户语音消息的实时情绪识别。

[0098] 还应当理解，基于本发明实施例提供的语音情绪识别方法所识别出的情绪分类，还可进一步配合具体的场景需求实现更多灵活的二次应用。在本发明一实施例中，可实时显示当前识别出的用户语音消息的情绪分类，具体的实时显示方式可根据实际的场景需求而调整。例如，可以信号灯的不同颜色来表征不同的情绪分类，这样根据信号灯颜色的变化，可以实时的提醒客服人员和质检人员目前通话所处的情绪状态。在另一实施例中，还可统计预设时间段内的所识别出的用户语音消息的情绪分类，例如将通话录音的音频编号、用户语音消息的开始点和结束点的时间戳，以及情绪识别结果记录下来，最终形成一个情绪识别资料库，并统计出一段时间内各种情绪出现的次数和概率，做出曲线图或表格，用于企业评判一段时间内客服人员服务质量的参考依据。在另一实施例中，还可实时发送与所识别出的用户语音消息的情绪分类对应的情绪应答信息，这可适用于无人工值守的机器客服场景。例如，当实时识别出目前通话中用户已经处于“生气”状态时，则自动回复用户与“生气”状态对应的安抚话语，以平复用户心情，达到继续沟通的目的。至于情绪分类与情绪应答信息之间的对应关系可通过预学习过程预先建立。

[0099] 在本发明一实施例中，在提取待识别音频流中的用户语音消息的音频特征向量之前，需要先将用户语音消息从待识别音频流中提取出来，以便于后续以用户语音消息为单位进行情绪识别，该提取过程可以是实时进行的。

[0100] 图6所示为本发明一实施例所提供的语音情绪识别方法中提取用户语音消息的流程示意图。如图6所示，该用户语音消息的提取方法包括：

[0101] 步骤601：确定待识别音频流中的语音开始帧以及语音结束帧。

[0102] 语音开始帧为一个用户语音消息的开始帧，语音结束帧为一个用户语音消息的结束帧。当确定了语音开始帧和语音结束帧后，语音开始帧和语音结束帧之间的部分即为所要提取的用户语音消息。

[0103] 步骤602：提取语音开始帧与语音结束帧之间的音频流部分作为用户语音消息。

[0104] 在本发明一实施例中，如图7所示，可具体通过如下步骤确定待识别音频流中的语音开始帧以及语音结束帧：

[0105] 步骤801：判断待识别音频流中的语音帧是发音帧还是非发音帧。

[0106] 在本发明一实施例中，该发音帧或非发音帧的判断过程可基于对语音端点检测(VAD)判决参数以及功率谱均值的判断实现，如图8所示，具体如下：

[0107] 步骤8011：对待识别音频流进行分帧、加窗、预加重等预处理。窗函数可采用汉明窗，预加重系数可取0.97。记预处理后的第k帧信号为$x(k)=[x(k*N),x(k*N+1),...,x(k*N+N-1)]$，N为帧长，例如可取256。然而应当理解，是否需要进行预处理过程，以及需要经过哪

些预处理过程可根据实际的场景需求而定，本发明此不做限定。

[0108] 步骤8012：对预处理后的第k帧信号x(k)做离散傅里叶变换(DFT)并计算其功率谱，DFT长度取为和帧长一致：

[0109] $P(k,j)=|FFT(x(k))|^2, j=0,1,\ldots,N-1$；

[0110] 这里j代表频率点的编号。

[0111] 步骤8013：计算后验信噪比γ和先验信噪比ξ：

[0112] $$\gamma(k,j)=\min\left(\frac{P(k,j)}{\lambda(k,j)},0.0032\right);$$

[0113] $\xi(k,j)=\alpha\xi(k-1,j)+(1-\alpha)\max(\gamma(k,j)-1,0)$；

[0114] 这里的系数α＝0.98；λ为背景噪声功率谱，可以检测开始的最初5至10帧的功率谱算数平均值作为初始值；min()和max()分别为取最小函数和取最大函数；先验信噪比ξ(k,j)可初始化为0.98。

[0115] 步骤8014：计算似然比参数η：

[0116] $$\eta(k)=\exp\left(\frac{1}{N}\sum_{j=0}^{N-1}\left(\gamma(k,j)\frac{\xi(k,j)}{1+\xi(k,j)}-\log(1+\xi(k,j))\right)\right)。$$

[0117] 步骤8015：计算VAD判决参数Γ和功率谱均值ρ，

[0118] $$\Gamma(k)=\frac{0.2+0.9*\Gamma(k-1)}{0.8+0.1*\Gamma(k-1)}*\eta(k);\quad \rho(k)=\frac{1}{N^2}\sum_{j=0}^{N-1}P(k,j)$$。VAD判决参数可初始化为1。

[0119] 步骤8016：判断第k帧信号的VAD判决参数Γ(k)是否大于等于第一预设VAD阈值，并且ρ(k)是否大于等于预设功率均值阈值。在本发明一实施例中，该第一预设VAD阈值可为5，该预设功率均值阈值可为0.01。

[0120] 步骤8017：如果步骤8016中的两个判断的结果均为是，则将第k帧音频信号判定为发音帧。

[0121] 步骤8018：如果步骤8016中的两个判断中至少一个的结果为否，将第k帧音频信号判定为不发音帧，执行步骤8019。

[0122] 步骤8019：按下面公式更新噪声功率谱λ：

[0123] $\lambda(k+1,j)=\beta*\lambda(k,j)+(1-\beta)*P(k,j)$；

[0124] 这里的系数β为平滑系数，可取值为0.98。

[0125] 由此可见，通过不断循环如图5所示的方法步骤便可实时监测出待识别音频流中的发音帧和非发音帧。这些发音帧和非发音帧的识别结果是后续识别语音开始帧和语音结束帧的基础。

[0126] 步骤802：在确定上一段用户语音消息的所述语音结束帧之后或者当前用户语音消息为所述待识别音频流的第一段用户语音消息时，当有第一预设数量个语音帧被连续判断为发音帧时，将该第一预设数量个语音帧中的第一个语音帧作为当前用户语音消息的语音开始帧。

[0127] 在本发明一实施例中，可首先设置两个端点标志flag_start和flag_end，分别代表语音开始帧和语音结束帧的检测状态变量，ture和false分别代表出现和未出现。当flag_end＝ture时，则说明上一个用户语音消息的结束帧已经被确定，此时开始检测下一

个用户语音消息的开始帧。而当连续30帧信号的VAD判决参数满足大于等于第二预设阈值时，说明该30帧已经进入了一个用户语音消息，此时将该30帧中的第一个语音帧作为语音开始帧，flag_start＝ture；否则lag_start＝false。

[0128] 步骤803：在确定当前用户语音消息的所述语音开始帧之后，当有第二预设数量个语音帧被连续判断为非发音帧时，说明该第二预设数量个语音帧已经不属于该用户语音消息，此时将第二预设数量个语音帧中的第一个语音帧作为当前用户语音消息的语音结束帧。

[0129] 具体而言，仍沿用上面的例子，当flag_start＝ture时，则说明已经进入了一个用户语音消息且该用户语音消息的语音起始帧已经被确定，此时开始检查当前用户语音消息的结束帧。而当连续30帧信号的VAD判决参数满足小于第三预设阈值时，判定为当前用户语音消息结束，flag_end＝ture，对应30帧的第一帧为语音结束帧；否则flag_end＝false。

[0130] 在本发明一实施例中，为了进一步提高语音开始帧和语音结束帧的判断准确度，避免误判，可使得该第二预设阈值和第三预设阈值均大于前述发音帧和非发音帧识别过程中的第一预设阈值，例如该第二预设阈值可为40，该第三预设阈值可为20。

[0131] 由此可见，通过如图7所示的方法步骤，便可确定待识别音频流中的语音开始帧以及语音结束帧，并可提取语音开始帧和语音结束帧之间的用户语音消息进行情绪识别。

[0132] 应当理解，虽然在上述图7和图8的实施例描述中引入了一些计算系数、参数的初始值以及一些判断阈值，但这些计算系数、参数的初始值以及判断阈值可根据实际的应用场景而调整，本发明对这些计算系数、参数的初始值以及判断阈值的大小不做限定。

[0133] 图9所示为本发明一实施例提供的语音情感交互方法中根据用户语音消息获取基本意图信息的流程示意图。如图9所示，该获取基本意图信息的流程可包括如下步骤：

[0134] 步骤901：将用户语音消息的文本内容与语义知识库中多个预设的语义模板进行匹配以确定匹配的语义模板；其中语义模板与基本意图信息之间的对应关系预先建立在语义知识库中，同一意图信息对应一个或多个语义模板。

[0135] 应当理解，通过语义模板进行语义的匹配(如标准问、扩展问等语义模板)只是一种实现方式，用户输入的语音文本信息也可以直接通过网络提取字、词、句向量特征(可能加入attention机制)直接做匹配或分类。

[0136] 步骤902：获取与匹配的语义模板对应的基本意图信息。

[0137] 在本发明一实施例中，用户语音消息的文本内容可与语义知识库中的“标准问”对应，“标准问”是用来表示某个知识点的文字，主要目标是表达清晰，便于维护。这里的“问”不应被狭义地理解为“询问”，而应广义地来理解一“输入”，该“输入”具有对应的“输出”。用户在向智能交互机器输入时，最理想的情况是使用标准问，则机器的智能语义识别系统马上能够理解用户的意思。

[0138] 然而，用户往往并非使用的是标准问，而是标准问的一些变形的形式，即为扩展问。因此，对于智能语义识别而言，知识库里需要有标准问的扩展问，该扩展问与标准问表达形式有略微差异，但是表达相同的含义。因此，在本发明一进一步实施例中，语义模板为表示某一种语义内容的一个或多个语义表达式的集合，由开发人员根据预定的规则结合语义内容生成，即通过一个语义模板就可描述所对应语义内容的多种不同表达方式的语句，以应对用户语音消息的文本内容可能的多种变形。这样将用户消息的文本内容与预设的语

义模板进行匹配，避免了利用仅能描述一种表达方式的“标准问”来识别用户消息时的局限性。

[0139] 例如采用抽象语义对本体类属性做进一步抽象。一个类别的抽象语义通过一组抽象语义表达式的集合来描述一类抽象语义的不同表达，为表达更为抽象的语义，这些抽象语义表达式在组成元素上进行了扩充。

[0140] 应当理解，语义成分词的具体内容和词类，语义规则词的具体内容和词类以及语义符号的定义和搭配都可由开发人员根据该语音情感交互方法所应用的具体交互业务场景而预设，本发明对此并不做限定。

[0141] 在本发明一实施例中，根据用户语音消息的文本内容确定匹配的语义模板的过程可通过相似度计算过程实现。具体而言，计算用户语音消息的文本内容与多个预设的语义模板之间的多个文本相似度，然后将文本相似度最高的语义模板作为匹配的语义模板。相似度可采用如下计算方法中的一种或多种：编辑距离计算方法，n-gram计算方法，JaroWinkler计算方法以及Soundex计算方法。在一进一步实施例中，当识别出用户语音消息的文本内容中的语义成分词和语义规则词时，用户语音消息和语义模板中所包括语义成分词和语义规则词还可被转化成简化的文本字符串，以提高语义相似度计算的效率。

[0142] 在本发明一实施例中，如前所述，语义模板可由语义成分词和语义规则词构成，而这些语义成分词和语义规则词又与这些词语在语义模板中的词性以及词语之间的语法关系有关，因此该相似度计算过程可具体为：先识别出用户语音消息文本中的词语、词语的词性以及语法关系，然后根据词语的词性以及语法关系识别出其中的语义成分词和语义规则词，再将所识别出的语义成分词和语义规则词引入向量空间模型以计算用户语音消息的文本内容与多个预设的语义模板之间的多个相似度。在本发明一实施例中，可以如下分词方法中的一种或多种识别用户语音消息的文本内容中的词语、词语的词性以及词语之间的语法关系：隐马尔可夫模型方法、正向最大匹配方法、逆向最大匹配方法以及命名实体识别方法。

[0143] 在本发明一实施例中，如前所述，语义模板可为表示某一种语义内容的多个语义表达式的集合，此时通过一个语义模板就可描述所对应语义内容的多种不同表达方式的语句，以对应同一标准问的多个扩展问。因此在计算用户语音消息的文本内容与预设的语义模板之间的语义相似度时，需要计算用户语音消息的文本内容与多个预设的语义模板各自展开的至少一个扩展问之间的相似度，然后将相似度最高的扩展问所对应的语义模板作为匹配的语义模板。这些展开的扩展问可根据语义模板所包括的语义成分词和/或语义规则词和/或语义符号而获得。

[0144] 当然获取基本意图信息的方法并不限于此，用户输入的语音文本信息可以直接通过网络提取字、词、句向量特征(如可能加入attention机制)直接匹配或分类到基本意图信息来实现。

[0145] 由此可见，通过本发明实施例所提供的语音情感交互方法，可实现根据用户情绪状态不同而提供不同应答服务的智能交互方式，由此可大大提高智能交互的体验。例如，当本发明实施例所提供的语音情感交互方法应用在银行客服领域的实体机器人时，用户用语音对实体客服机器人说：“信用卡要挂失怎么办?”。实体客服机器人通过麦克风接收用户语音消息，并通过分析用户语音消息的音频数据得到音频情绪识别结果为“焦急”，并将音频

情绪识别结果作为最终的情绪识别结果;将用户语音消息转换为文本,得到客户的基本意图信息为“挂失信用卡”(这一步骤也可能需要涉及到结合过往或后续的用户语音消息和银行领域的语义知识库);然后,将情绪识别结果“焦急”与基本意图信息“挂失信用卡”联系在一起,得到情绪意图信息“挂失信用卡,用户很焦急,可能信用卡丢失或被盗”(这一步骤也可能需要涉及到结合过往或后续的用户语音消息和银行领域的语义知识库);确定对应的交互指令:屏幕输出信用卡挂失步骤,同时通过语音播报呈现情绪分类“安慰”,情绪强度级别为高,输出给用户符合该情绪指令的可能是音调轻快、中等语速的语音播报:“挂失信用卡的步骤请见屏幕显示,请您不要担心,如果是信用卡遗失或被盗,卡挂失后立刻冻结,不会对您的财产和信誉造成损失……”。

[0146] 在本发明一实施例中,一些应用场景(例如银行客服)也可能考虑交互内容的隐私性而避免语音播报操作,而改为以纯文本或动画的方式实现交互指令。这种交互指令的模态选择可根据应用场景而调整。

[0147] 应当理解,交互指令中对于情绪分类和情绪强度级别的呈现方式可通过调整语音播报的语速和语调等方式实现,本发明对此不做限定。

[0148] 再例如,当本发明实施例所提供的语音情感交互方法应用在智能终端设备的虚拟智能个人助理应用中时,用户对智能终端设备用语音说:“从家里到机场最快的路径是什么?”。虚拟智能个人助理应用通过智能终端设备的麦克风接收用户语音消息,并通过分析用户语音消息的音频数据得到音频情绪识别结果为“兴奋”;同时将用户语音消息转化为文本,并通过分析用户语音消息的文本内容得到文本情绪识别结果为“焦急”,经过逻辑判断将“兴奋”和“焦急”两种情绪分类同时作为了情绪识别结果。通过结合过往或后续的用户语音消息和本领域的语义知识库得到客户的基本意图信息为“获得用户从家到机场最快的路径导航”。由于虚拟智能个人助理应用将“焦急”与基本意图信息“获得用户从家到机场最快的路径导航”联系在一起得到的情绪意图信息为“获得用户从家到机场最快的路径导航,用户很焦急,可能担心误点飞机”;而将“兴奋”与基本意图信息联系在一起得到的情绪意图信息为“获得用户从家到机场最快的路径导航,用户很兴奋,可能马上要去旅行”;因此,这里会产生两种情绪意图信息,此时可结合过往或后续的用户语音消息,发现前面用户提到“我的航班是11点起飞,需要几点出发?”,于是判断用户的情绪识别结果为“焦急”,情绪意图信息为“获得用户从家到机场最快的路径导航,用户很焦急,可能担心误点飞机”。确定对应的交互指令:屏幕输出导航信息,同时通过语音播报呈现情绪分类“安慰”和“警示”,情绪强度级别分别为高,输出给用户符合该情绪指令的可能是音调平稳、中等语速的语音播报:“从您家庭住址到机场最快的路径规划完毕,请按屏幕显示进行导航,正常行驶预计可在1小时内到达机场,请您不要担心。另外提醒做好时间规划,注意行车安全,请勿超速行驶。”

[0149] 再例如,当本发明实施例所提供的语音情感交互方法应用在一种智能穿戴设备中时,用户在运动的时候对智能穿戴设备用语音说:“我现在的心跳情况如何?”。智能穿戴设备通过麦克风接收用户语音消息,并通过分析用户语音消息的音频数据得到音频情绪识别结果为PAD三维情绪模型向量(p1,a1,d1),通过分析用户语音消息的音频数据得到文本情绪识别结果为PAD三维情绪模型向量(p2,a2,d2),结合音频情绪识别结果和文本情绪识别结果得到最终的情绪识别结果(p3,a3,d3),表征了“担忧”和“紧张”的结合。与此同时,智能穿戴设备通过结合医疗健康领域的语义知识库得到客户的基本意图信息为“获得用户的心

跳数据”。接着，将情绪识别结果(p3,a3,d3)与基本意图“获得用户的心跳数据”联系在一起，得到情绪意图信息为“获得用户的心跳数据，用户表示担忧，可能当前有心跳过快等不适症状”。根据情绪意图信息和交互指令之间的对应关系确定交互指令：在输出心跳数据的同时呈现情绪(p6,a6,d6)，即“安慰”和“鼓励”的结合，情绪强度分别为高，同时启动实时监控心跳的程序持续10min，并以音调轻快、缓慢语速的语音播报：“您当前的心跳数据是每分钟150次，请您不要担心，该数据尚属于正常心跳范围。如有感到心跳过快等不适症状请放松心情做深呼吸进行调整。您以往的健康数据显示心脏工作良好，可以通过保持规律的锻炼增强心肺功能。”然后持续关注用户的情绪状态。如果5min后用户说“有些不舒服。”通过情绪识别过程得到情绪识别结果为三维情绪模型向量(p7,a7,d7)，表征了“痛苦”，则重新更新交互指令为：屏幕输出心跳数据，同时通过语音播报呈现情绪(p8,a8,d8)，即“警示”，情绪强度分别为高等，输出报警音，并以音调沉稳、缓慢语速的语音播报：“您当前的心跳数据是每分钟170次，已超过正常范围，请您停止运动，调整呼吸。如需求助请按屏幕。”

[0150] 本发明一实施例还提供一种计算机设备，包括存储器、处理器以及存储在存储器上被处理器执行的计算机程序，其特征在于，处理器执行计算机程序时实现如前任一实施例所述的语音情感交互方法。

[0151] 本发明一实施例还提供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，计算机程序被处理器执行时实现如前任一实施例所述的语音情感交互方法。该计算机存储介质可以为任何有形媒介，例如软盘、CD-ROM、DVD、硬盘驱动器、甚至网络介质等。

[0152] 应当理解，虽然以上描述了本发明实施方式的一种实现形式可以是计算机程序产品，但是本发明的实施方式的方法或装置可以被依软件、硬件、或者软件和硬件的结合来实现。硬件部分可以利用专用逻辑来实现；软件部分可以存储在存储器中，由适当的指令执行系统，例如微处理器或者专用设计硬件来执行。本领域的普通技术人员可以理解上述的方法和设备可以使用计算机可执行指令和/或包含在处理器控制代码中来实现，例如在诸如磁盘、CD或DVD-ROM的载体介质、诸如只读存储器(固件)的可编程的存储器或者诸如光学或电子信号载体的数据载体上提供了这样的代码。本发明的方法和装置可以由诸如超大规模集成电路或门阵列、诸如逻辑芯片、晶体管等的半导体、或者诸如现场可编程门阵列、可编程逻辑设备等的可编程硬件设备的硬件电路实现，也可以用由各种类型的处理器执行的软件实现，也可以由上述硬件电路和软件的结合例如固件来实现。

[0153] 应当理解，尽管在上文的详细描述中提及了装置的若干模块或单元，但是这种划分仅仅是示例性而非强制性的。实际上，根据本发明的示例性实施方式，上文描述的两个或更多模块/单元的特征和功能可以在一个模块/单元中实现，反之，上文描述的一个模块/单元的特征和功能可以进一步划分为由多个模块/单元来实现。此外，上文描述的某些模块/单元在某些应用场景下可被省略。

[0154] 应当理解，本发明实施例描述中所用到的限定词“第一”、“第二”和“第三”等仅用于更清楚的阐述技术方案，并不能用于限制本发明的保护范围。

[0155] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所作的任何修改、等同替换等，均应包含在本发明的保护范围之内。

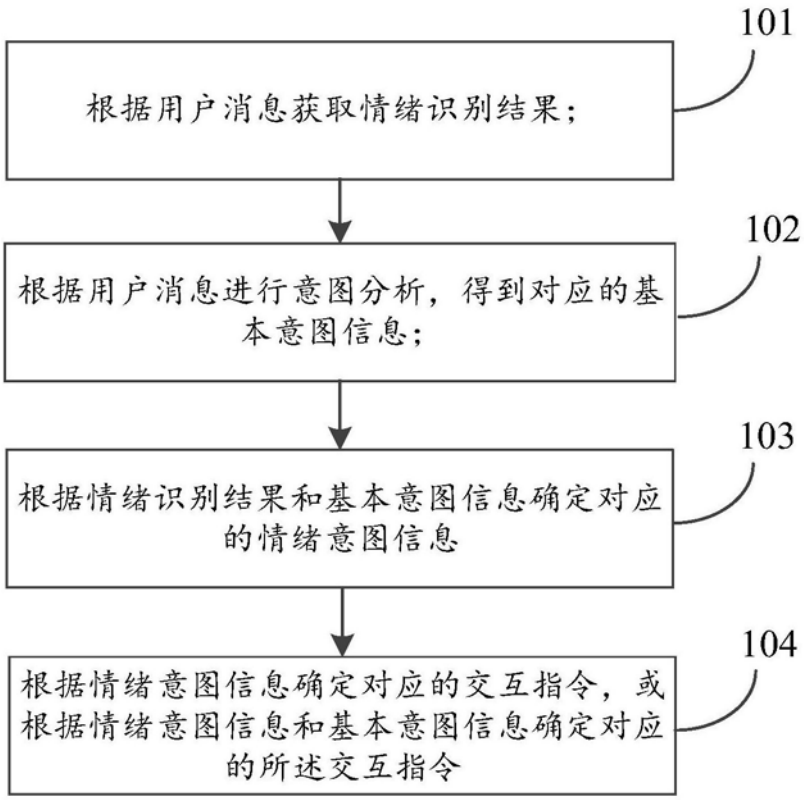

101
根据用户消息获取情绪识别结果；
102
根据用户消息进行意图分析，得到对应的基本意图信息；
103
根据情绪识别结果和基本意图信息确定对应的情绪意图信息
104
根据情绪意图信息确定对应的交互指令，或根据情绪意图信息和基本意图信息确定对应的所述交互指令

图1

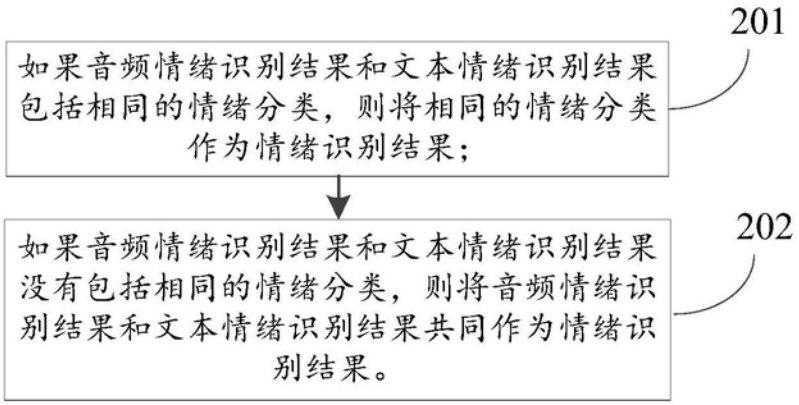

201
如果音频情绪识别结果和文本情绪识别结果包括相同的情绪分类，则将相同的情绪分类作为情绪识别结果；
202
如果音频情绪识别结果和文本情绪识别结果没有包括相同的情绪分类，则将音频情绪识别结果和文本情绪识别结果共同作为情绪识别结果。

图2

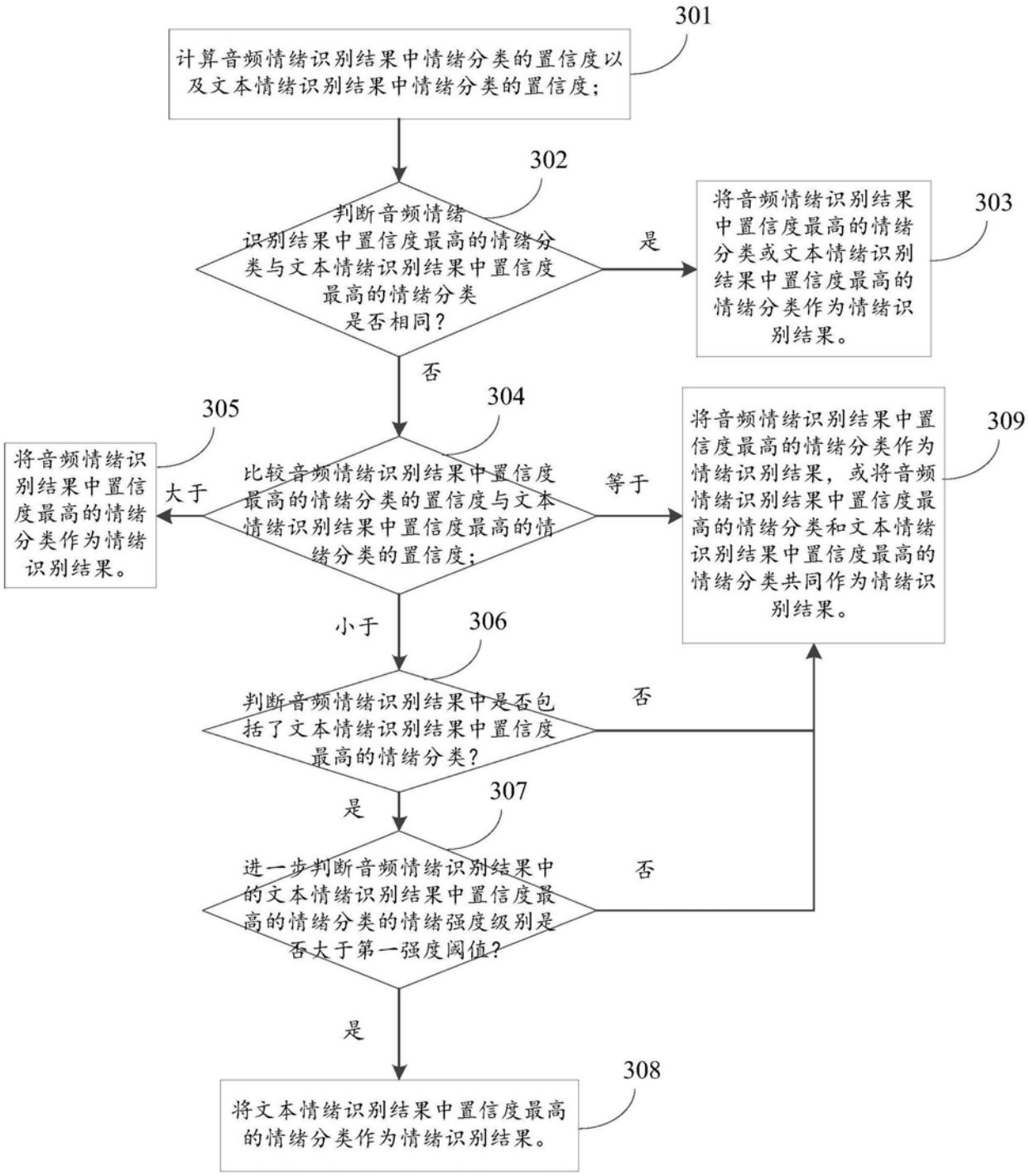

301
计算音频情绪识别结果中情绪分类的置信度以及文本情绪识别结果中情绪分类的置信度；
302
判断音频情绪识别结果中置信度最高的情绪分类与文本情绪识别结果中置信度最高的情绪分类是否相同？
是
303
将音频情绪识别结果中置信度最高的情绪分类或文本情绪识别结果中置信度最高的情绪分类作为情绪识别结果。
否
304
比较音频情绪识别结果中置信度最高的情绪分类的置信度与文本情绪识别结果中置信度最高的情绪分类的置信度；
大于
305
将音频情绪识别结果中置信度最高的情绪分类作为情绪识别结果。
等于
309
将音频情绪识别结果中置信度最高的情绪分类作为情绪识别结果，或将音频情绪识别结果中置信度最高的情绪分类和文本情绪识别结果中置信度最高的情绪分类共同作为情绪识别结果。
小于
306
判断音频情绪识别结果中是否包括了文本情绪识别结果中置信度最高的情绪分类？
否
是
307
进一步判断音频情绪识别结果中的文本情绪识别结果中置信度最高的情绪分类的情绪强度级别是否大于第一强度阈值？
否
是
308
将文本情绪识别结果中置信度最高的情绪分类作为情绪识别结果。

图3

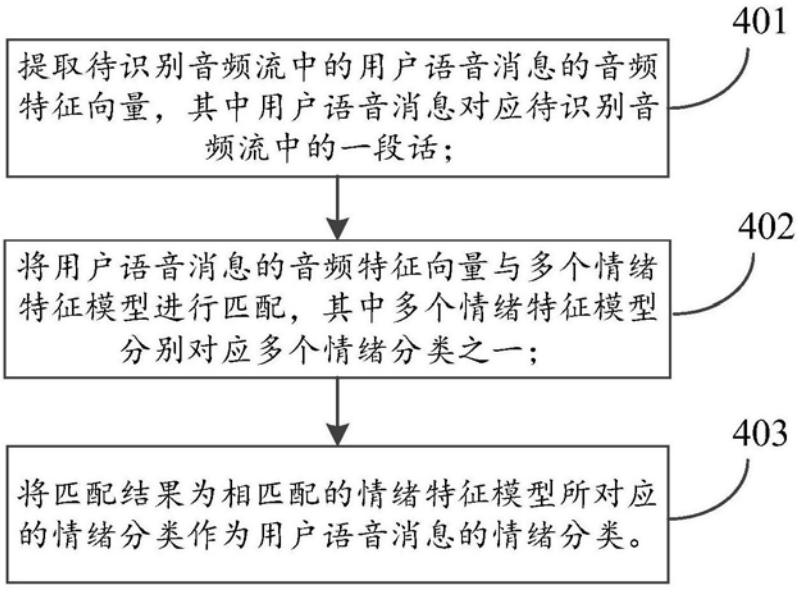

401
提取待识别音频流中的用户语音消息的音频特征向量，其中用户语音消息对应待识别音频流中的一段话；
402
将用户语音消息的音频特征向量与多个情绪特征模型进行匹配，其中多个情绪特征模型分别对应多个情绪分类之一；
403
将匹配结果为相匹配的情绪特征模型所对应的情绪分类作为用户语音消息的情绪分类。

图4

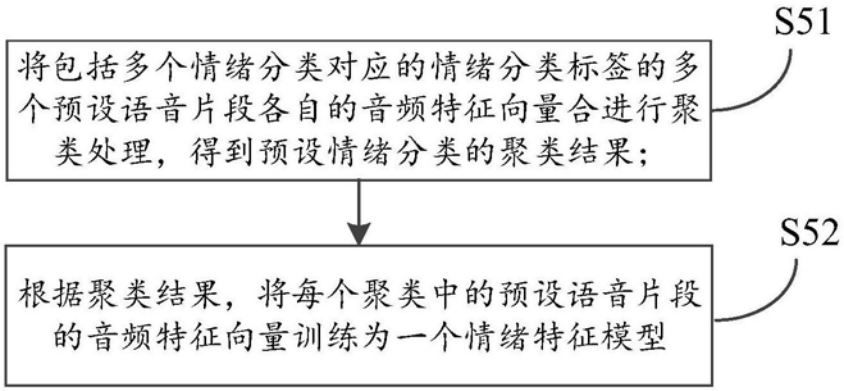

S51
将包括多个情绪分类对应的情绪分类标签的多个预设语音片段各自的音频特征向量合进行聚类处理，得到预设情绪分类的聚类结果；
S52
根据聚类结果，将每个聚类中的预设语音片段的音频特征向量训练为一个情绪特征模型

图5

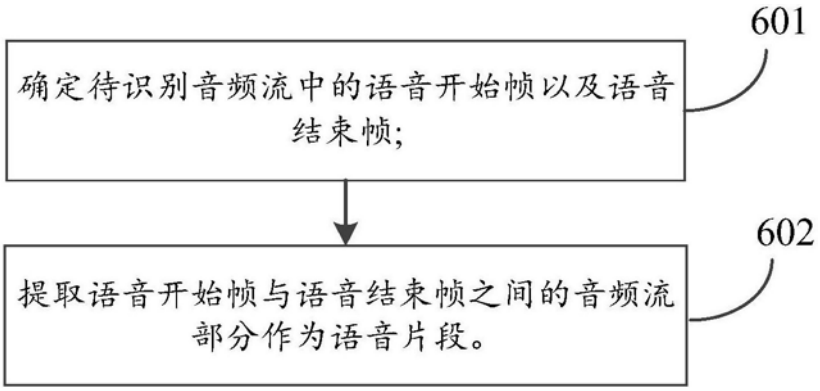

601
确定待识别音频流中的语音开始帧以及语音结束帧；
602
提取语音开始帧与语音结束帧之间的音频流部分作为语音片段。

图6

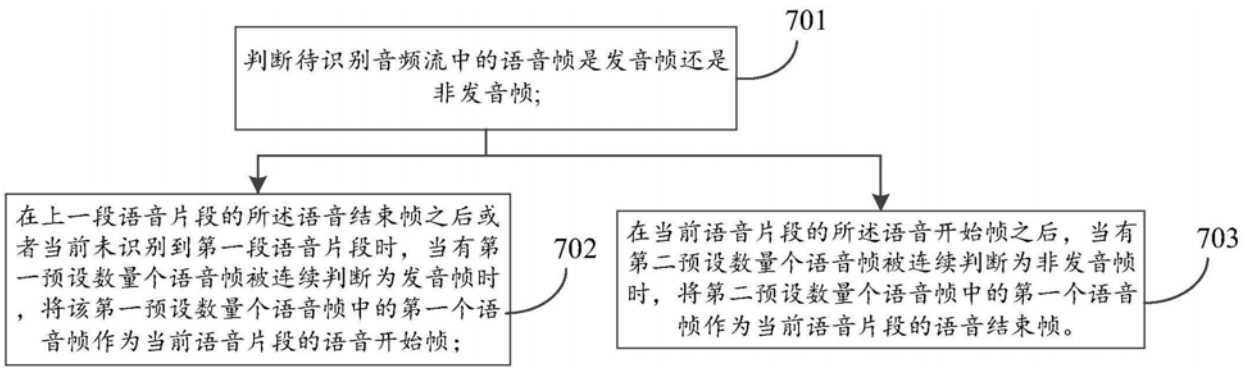

701
判断待识别音频流中的语音帧是发音帧还是非发音帧;
在上一段语音片段的所述语音结束帧之后或者当前未识别到第一段语音片段时，当有第一预设数量个语音帧被连续判断为发音帧时，将该第一预设数量个语音帧中的第一个语音帧作为当前语音片段的语音开始帧;
702
在当前语音片段的所述语音开始帧之后，当有第二预设数量个语音帧被连续判断为非发音帧时，将第二预设数量个语音帧中的第一个语音帧作为当前语音片段的语音结束帧。
703

图7

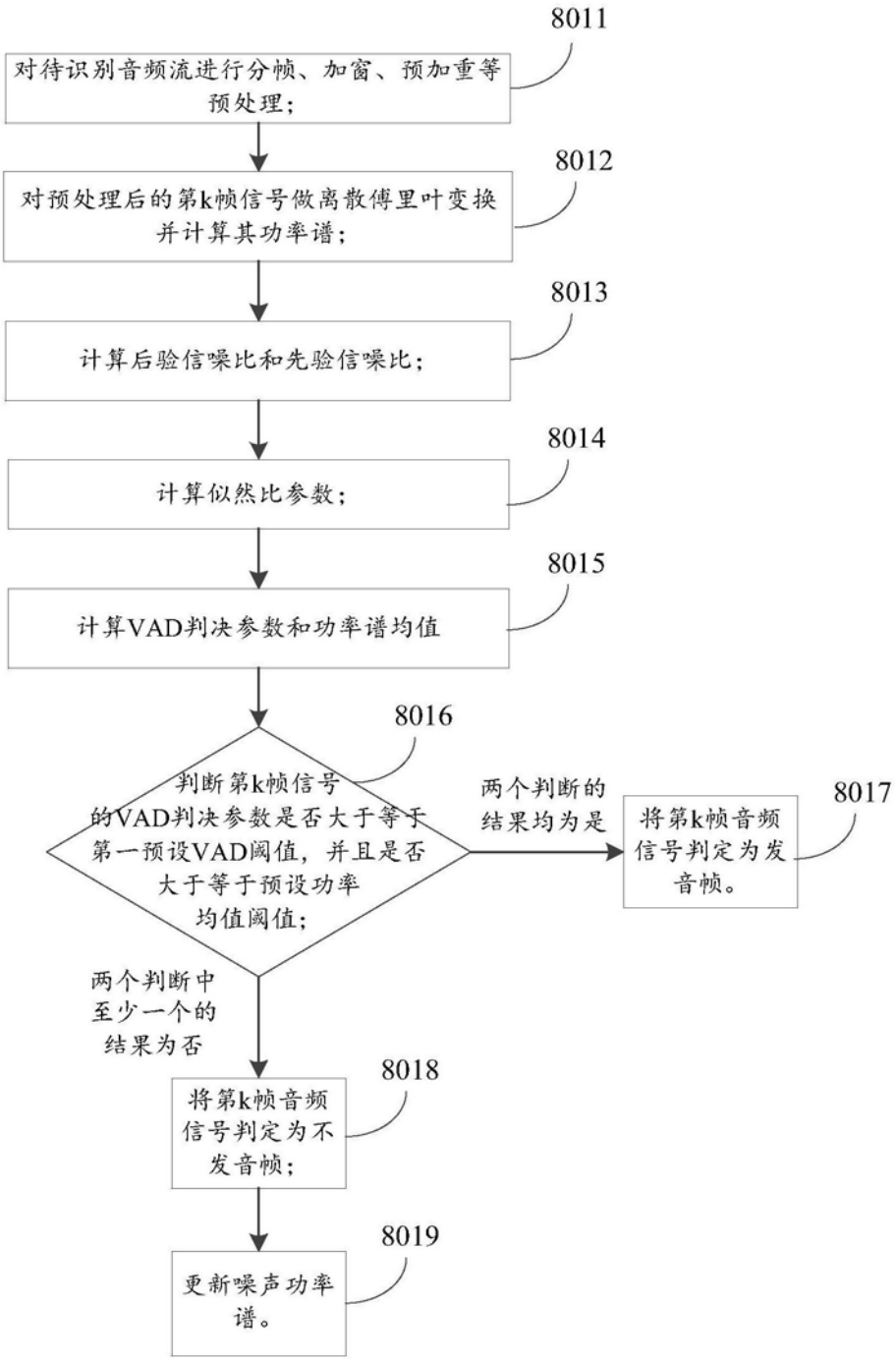

8011
对待识别音频流进行分帧、加窗、预加重等预处理；
8012
对预处理后的第k帧信号做离散傅里叶变换并计算其功率谱；
8013
计算后验信噪比和先验信噪比；
8014
计算似然比参数；
8015
计算VAD判决参数和功率谱均值
8016
判断第k帧信号的VAD判决参数是否大于等于第一预设VAD阈值，并且是否大于等于预设功率均值阈值；
两个判断的结果均为是
8017
将第k帧音频信号判定为发音帧。
两个判断中至少一个的结果为否
8018
将第k帧音频信号判定为不发音帧；
8019
更新噪声功率谱。

图8

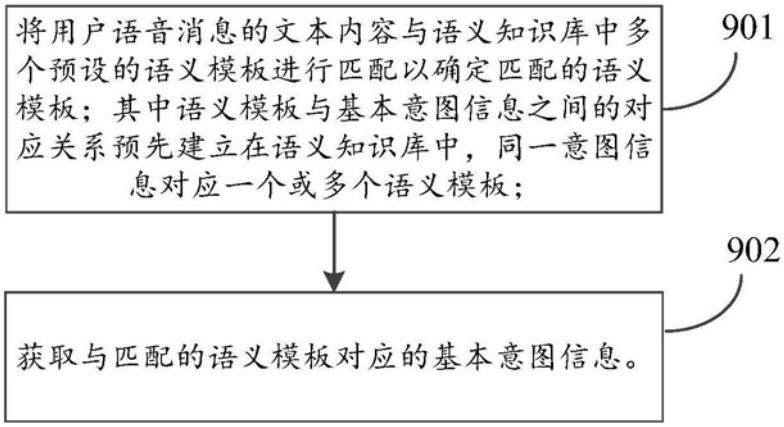

将用户语音消息的文本内容与语义知识库中多个预设的语义模板进行匹配以确定匹配的语义模板；其中语义模板与基本意图信息之间的对应关系预先建立在语义知识库中，同一意图信息对应一个或多个语义模板；
901
获取与匹配的语义模板对应的基本意图信息。
902

图9