

基于多模态特征的语音对话处理方法、装置和电子设备

相关申请的交叉引用

5 本申请基于申请号为 202111337746.8、申请日为 2021 年 11 月 09 日的中国专利申请提出，并要求该中国专利申请的优先权，该中国专利申请的全部内容在此引入本申请作为参考。

技术领域

10 本申请涉及计算机技术领域，尤其涉及基于多模态特征的语音对话处理方法、装置和电子设备。

背景技术

在语音对话系统中，用户讲话时，语音对话系统需要判断在合适的时机接过话语权，即语音对话系统在听者和说话者的角色间来回转换，使得人机交互流畅自然。

15 目前，多数语音对话系统采用通过语音活动检测(Voice Activity Detection,VAD)识别用户静默时长的方式，当用户静默时长超过阈值（比如 0.8s~1s）时，系统接过话语权。但是，在这种固定静默时长的方式中，如果用户并未讲完且在思考中，然而静默时长超过阈值，这时系统响应就会过于迅速敏感；如果用户的交互迅速简明，这时系统仍然等待静默时长达到设定阈值才接过话语权，导致系统响应迟钝，可能造成用户重复回答。因此，如何确定语音对话系统何时接过话语权是目前亟需解决的问题。

20 发明内容

本申请提出一种基于多模态特征的语音对话处理方法、装置和电子设备。
本申请一方面实施例提出了一种基于多模态特征的语音对话处理方法，包括：在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，所述第一语音
25 信息包括静默段；根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息；根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息；获取所述第一语音信息的时间特征信息；根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入。

30 在本申请的一个实施例中，所述根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息，包括：对所述第一语音信息进行语音识别，以得到所述第一语音信息的文本信息；获取所述第一语音信息的历史上下文信息；将所述文本信息和所述历史上下文信息输入到语义表示模型中，以得到所述文本信息的语义特征信息。

35 在本申请的一个实施例中，所述根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息，包括：获取所述第一语音信息中在所述

静默段之前的第一预设时间长度的语音片段；按照第二预设时间长度，对所述语音片段进行分段，以得到多段语音片段；提取所述多段语音片段各自对应的声学特征信息，并分别对所述多段语音片段各自对应的声学特征信息进行拼接，以得到所述多段语音频段各自对应的拼接特征；将所述拼接特征输入到深度残差网络中，以得到所述第一语音信息的语音特征信息。

在本申请的一个实施例中，所述获取所述第一语音信息的时间特征信息，包括：获取所述第一语音信息的语音时长、语速和文本长度；将所述语音时长、所述语速和文本长度输入到预先训练好的多层感知机 MLP 模型，以得到所述第一语音信息的时间特征信息。

在本申请的一个实施例中，所述根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入，包括：将所述语义特征信息、所述语音特征信息和所述时间特征信息输入到多模态融合模型中；根据所述多模态融合模型的输出结果，确定所述用户是否结束语音输入。

在本申请的一个实施例中，还包括：在确定所述用户结束语音输入的情况下，确定所述第一语音信息所对应的第一回复语音信息，并输出所述第一回复语音信息。

在本申请的一个实施例中，还包括：在确定所述用户未结束语音输入的情况下，获取所述用户再次输入的第二语音信息；根据所述第一语音信息和所述第二语音信息，确定对应的第二回复语音信息，并输出所述第二回复语音信息。

本申请另一方面实施例提出了一种基于多模态特征的语音对话处理装置，包括：第一获取模块，用于在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，所述第一语音信息包括静默段；第一确定模块，用于根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息；第二确定模块，用于根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息；第二获取模块，用于获取所述第一语音信息的时间特征信息；第三确定模块，用于根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入。

在本申请的一个实施例中，所述第一确定模块，具体用于：对所述第一语音信息进行语音识别，以得到所述第一语音信息的文本信息；获取所述第一语音信息的历史上下文信息；将所述文本信息和所述历史上下文信息输入到语义表示模型中，以得到所述文本信息的语义特征信息。

在本申请的一个实施例中，所述第二确定模块，具体用于：获取所述第一语音信息中在所述静默段之前的第一预设时间长度的语音片段；按照第二预设时间长度，对所述语音片段进行分段，以得到多段语音片段；提取所述多段语音片段各自对应的声学特征信息，并分别对所述多段语音片段各自对应的声学特征信息进行拼接，以得到所述多段语音频段各自对应的拼接特征；将所述拼接特征输入到深度残差网络中，以得到所述第一语音信息的语音特征信息。

在本申请的一个实施例中，所述第二获取模块，具体用于：获取所述第一语音信息的语音时长、语速和文本长度；将所述语音时长、所述语速和文本长度输入到预先训练好的多层感知机 MLP 模型，以得到所述第一语音信息的时间特征信息。

在本申请的一个实施例中，所述第三确定模块，包括：多模态处理单元，用于将所述语义特征信息、所述语音特征信息和所述时间特征信息输入到多模态融合模型中；确定单元，用于根据所述多模态融合模型的输出结果，确定所述用户是否结束语音输入。

在本申请的一个实施例中，还包括：第一处理模块，用于在确定所述用户结束语音输入的情况下，确定所述第一语音信息所对应的第一回复语音信息，并输出所述第一回复语音信息。

在本申请的一个实施例中，还包括：第三获取模块，用于在确定所述用户未结束语音输入的情况下，获取所述用户再次输入的第二语音信息；第二处理模块，用于根据所述第一语音信息和所述第二语音信息，确定对应的第二回复语音信息，并输出所述第二回复语音信息。

本申请另一方面实施例提出了一种电子设备，包括：存储器，处理器；所述存储器中存储有计算机指令，当所述计算机指令被所述处理器执行时，实现本申请实施例的基于多模态特征的语音对话处理方法。

本申请另一方面实施例提出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本申请实施例公开的基于多模态特征的语音对话处理方法。

本申请另一方面实施例提出了一种计算机程序产品，当所述计算机程序产品中的指令处理器执行时实现本申请实施例中的基于多模态特征的语音对话处理方法。

上述可选方式所具有的其他效果将在下文中结合具体实施例加以说明。

附图说明

附图用于更好地理解本方案，不构成对本申请的限定。其中：

图1是根据本申请一个实施例的基于多模态特征的语音对话处理方法的流程示意图。

图2是根据本申请一个具体实施例的结合模型框架图对语音对话处理方法进行描述的示例图。

图3是根据本申请一个实施例的基于多模态特征的语音对话处理装置的结构示意图。

图4是根据本申请另一个实施例的基于多模态特征的语音对话处理装置的结构示意图。

图5是根据本申请一个实施例的电子设备的框图。

具体实施方式

下面详细描述本公开的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附

图描述的实施例是示例性的，旨在用于解释本申请，而不能理解为对本申请的限制。

下面参考附图描述本申请实施例的基于多模态特征的语音对话处理方法、装置和电子设备。

图1是根据本申请一个实施例的基于多模态特征的语音对话处理方法的流程示意图。

5 其中，需要说明的是，本申请实施例提供的基于多模态特征的语音对话处理方法的执行主体为基于多模态特征的语音对话处理装置，该基于多模态特征的语音对话处理装置可以由软件和/或硬件的方式实现。在本申请实施例中的基于多模态特征的语音对话处理装置语音对话系统中，该语音对话系统可以配置在电子设备中。电子设备可以包括终端设备或者服务器等。

10 如图1所示，该基于多模态特征的语音对话处理方法可以包括步骤101至步骤105。

步骤101，在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，第一语音信息包括静默段。

步骤102，根据第一语音信息的文本信息和第一语音信息的历史上下文信息，确定文本信息的语义特征信息。

15 在本申请的一个实施例中，可对第一语音信息进行语音识别，以得到第一语音信息的文本信息，并获取第一语音信息的历史上下文信息，以及将文本信息和历史上下文信息输入到语义表示模型中，以得到文本信息的语义特征信息。

20 在一些实施例中，为了可以捕获文本信息与历史上下文信息之间的长距离依赖关系，并基于长距离依赖关系准确地确定出文本信息的语义特征信息，上述语义表示模型可以为基于自注意力机制的转换器Transformer模型。

在一些实施例中，Transformer模型可以包括多层编码层。每层编码层中均包括基于Transformer的编码结构，对应编码结构对输入内容进行编码，并将输出结果输入至对应的下一层编码层进行处理。

25 在一些实施例中，获取第一语音信息的历史上下文信息的一种示例性的实施方式为：可获取在第一语音信息之前的多条历史语音对话信息，并根据多条历史语音对话信息，获取第一语音信息的历史上下文信息。

步骤103，根据第一语音信息中在静默段之前的语音片段，确定第一语音信息的语音特征信息。

30 在一些实施例中，可获取第一语音信息中在静默段之前的第一预设时间长度的语音片段；按照第二预设时间长度，对语音片段进行分段，以得到多段语音片段；提取多段语音片段各自对应的声学特征信息，并分别对多段语音片段各自对应的声学特征信息进行拼接，以得到多段语音片段各自对应的拼接特征；将拼接特征输入到深度残差网络中，以得到第一语音信息的语音特征信息。

35 在一些实施例中，第一预设时间长度是预先设置的，例如，上述第一预设时间长度可以为2秒，也就是说，可截取第一语音信息中在静默段之前的一段时长长度为2秒的语音片段。

在一些实施例中，第二预设时间长度是预先设置的，第一预设时间长度大于第二预设时长长度，例如，第一预设时间长度为 2 秒，上述第二预设时间长度可以为 50 毫秒（ms）。在一些实施例中，在获取 2 秒长度的语音片段后，可按照 50ms，对该语音片段进行分段，以得到多段语音片段，其中，每段语音频段场 50ms。

5 在一些实施例中，声学特征信息可以包括但不限于能量、音量、音高、过零率等。步骤 104，获取第一语音信息的时间特征信息。

在一些实施例中，可取第一语音信息的语音时长、语速和文本长度，将语音时长、语速和文本长度输入到预先训练好的多层感知机（Multi Layer Perceptron, MLP）模型，以得到第一语音信息的时间特征信息。

10 在一些实施例中，文本长度可以是基于第一语音信息对应的文本信息所确定出的。步骤 105，根据语义特征信息、语音特征信息和时间特征信息，确定用户是否结束语音输入。

在一些实施例中，为了可以准确确定出用户是否结束语音输入，可将语义特征信息、语音特征信息和时间特征信息输入到多模态融合模型中，根据多模态融合模型的输出结果，确定用户是否结束语音输入。

15 在一些实施例中，多模态融合模型在获取语义特征信息、语音特征信息和时间特征信息，可获取上述语义特征信息、语音特征信息和时间特征信息各自对应的权重，并基于权重，对语义特征信息、语音特征信息和时间特征信息进行加权处理，并将加权结果输入到多模态融合模型的激活函数中，以得到多模态融合模型的输出结果。

20 在一些实施例中，在多模态融合模型的输出结果指示用户结束语音输入的情况下，可确定用户结束语音输入，此时，可确定对话系统可以接过话语权。在另一些实施例中，在多模态融合模型的输出结果指示用户未结束语音输入的情况下，可确定用户未结束语音输入，此时，对话系统可继续倾听，并在确定用户输入结束后进行回复。

本申请实施例的基于多模态特征的语音对话处理方法，在与用户进行对话交互的过程中，结合用户当前输入的语音信息的文本信息和第一语音信息的历史上下文信息，确定文本信息的语义特征信息，根据第一语音信息中在静默段之前的语音片段，确定第一语音信息的语音特征信息，获取第一语音信息的时间特征信息；根据语义特征信息、语音特征信息和时间特征信息，确定用户是否结束语音输入。由此，在与用户进行对话交互的过程中，结合语义特征信息、语音特征信息和时间特征信息，准确确定出了系统是否可接过话语权。

30 基于上述实施例的基础上，为了使得对话系统准确对用户输入的语音信息进行回复，在一些实施例中，在确定用户结束语音输入的情况下，确定第一语音信息所对应的第一回复语音信息，并输出第一回复语音信息。

在另一些实施例中，在确定用户未结束语音输入的情况下，获取用户再次输入的第二语音信息；根据第一语音信息和第二语音信息，确定对应的第二回复语音信息，并输出第二回复语音信息。由此，结合用户当前输入的第一语音信息和再次输入的第二语音

信息进行准确回复。

为了使得本领域技术人员可以清楚了解本申请，下面结合图 2 对本申请实施例的方法进行进一步阐述。

通过图 2 可以看出，本申请实施例中在确定用户是否结束语音输入的过程中，使用了语音特征信息、语义特征信息和时间特征信息三个不同维度上的特征，来确定用户是否结束语音输入，即，本申请实施例中使用了语义特征信息、语音特征信息和时间特征信息三个不同维度上的特征，来确定对话系统是否可以接过话语权，即，确定对话系统是否输出对应回复。

下面分别对获取语义特征信息、语音特征信息和时间特征信息的过程进行描述。

1) 获取语义特征信息。

其中，语义特征信息来源于语音识别后的文本信息，它对于话语权决策的重要性是不言而喻的，尤其是考虑到“语义完整性”是话语权切换的基本要素，也就是说，在确定用户已经完整表述其意图之后，往往意味着系统可以接过话语权。而语义完整性一般也是结合上下文语境来判断的，例如下面的简单示例：

完成 Completed	未完成 Uncompleted
Sys:后天帮您送货可以吗? User: 可以	Sys: 后天帮您送货可以吗? User: 后天啊...

左边的例子中，用户进行了确定性的答复，语义明确，此时，对话系统可以接过话语权。右边的例子里用户出现了短暂的犹豫，但是根据用户当前输入的内容，可以确定用户没有讲完，此时，对话系统可以该选择继续倾听，等待用户把话讲完。

为了建模这种语义的完整性，在用户与对话系统进行对话交互的过程汇总，在获取用户当前输入的语音信息后，可对该语音信息进行语音识别，以得到当前文本信息，可对当前输入的语音信息的历史上上下文 context 信息以及当前文本信息进行编码，以得到该文本信息的语义特征信息。

在一些实施例中，可采用基于自注意力机制的 Transformer 模型对当前输入的语音信息的历史上上下文 context 信息以及当前文本信息所对应的文本信息进行编码。

其中，可以理解的是，Transformer 模型中的自注意力机制可以捕获历史上下文信息与文本信息之间的长距离依赖关系。最终语义特征表示为：

$$r^s = \text{Transformer}(e)$$

2)获取语音特征信息

可以理解的是，在对话过程中，一些语音特征例如音调的转变、音量的高低等都判断是否进行话语权切换的重要线索。因此，在与用户进行对话的过程中，在获取用户当前输入的语音信息后，可从语音信息中截取用户静默前的一段音频（2s），然后将其切分成固定长度的小段，即分帧（每帧 50ms）。接下来对每帧音频提取其对应的声学特征，如能量、音量、

音高、过零率等，并将其拼接成一维向量，得到每一帧音频的特征表示 f_i 。最后，可将序列帧的特征 $F=[f_1, f_2, \dots, f_n]$ 输入到一个多层的深度残差（Residual Network, ResNet）网络中，得到最终的语音特征表示：

$$r^a = \text{ResNet}(F)$$

5 3) 时间特征

需要理解的是，时间特征（例如语音片段的时长、语速、文本长度等）对于判断话语权的切换与否也是有一定作用的。比如在以系统为主导的外呼对话场景，多数情况下，在用户进行简短的回复之后系统便可接过话语权；而需要系统进行倾听的情况则大多是由于用户因为犹豫等因素而产生了比较长的答复，因此，为了可以准确确定出对话系统是否可以接过话语权，在与用户进行对话交互的过程中，可获取用户当前输入的语音信息的语音时长、语速和文本长度，并将语音时长、语速和文本长度分别进行分桶处理，并根据处理后的语音时长、语速和文本长度输入到 MLP 模型中，以得到语音信息的低维度的时间特征信息。

其中，通过多层感知网络提取其低维特征表示：

$$r^t = \text{MLP}(t)$$

4) 多模态特征融合

15 在一些实施例，在获取各个模态的特征表示，接下来通过将其输入到多模态融合模型，融合三种不同特征进行话语权的判断：

$$y = \sigma(W_s r^s + W_a r^a + W_t r^t + b)$$

其中， $\sigma(\cdot)$ 指的是 sigmoid 函数， y 即预测的二分类标签：1-表示用户结束讲话，系统接过话语权；0-则表示系统应该继续倾听用户回复， b 表示偏置值。

在一些实施例，上述多模态融合模型可以是基于前馈神经网络而建立的。

20 与上述几种实施例提供的基于多模态特征的语音对话处理方法相对应，本申请的一种实施例还提供一种基于多模态特征的语音对话处理装置，由于本申请实施例提供的基于多模态特征的语音对话处理装置与上述几种实施例提供的基于多模态特征的语音对话处理方法相对应，因此在基于多模态特征的语音对话处理方法的实施方式也适用于本申请实施例提供的基于多模态特征的语音对话处理装置。

25 图 3 是根据本申请一个实施例的基于多模态特征的语音对话处理装置的结构示意图。

如图 3 所示，该基于多模态特征的语音对话处理装置 300 包括第一获取模块 301、第一确定模块 302、第二确定模块 303、第二获取模块 304 和第三确定模块 305。

第一获取模块 301，用于在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，第一语音信息包括静默段。

30 第一确定模块 302，用于根据第一语音信息的文本信息和第一语音信息的历史上下文信息，确定文本信息的语义特征信息。

第二确定模块 303，用于根据第一语音信息中在静默段之前的语音片段，确定第一语音信息的语音特征信息。

第二获取模块 304，用于获取第一语音信息的时间特征信息。

第三确定模块 305，用于根据语义特征信息、语音特征信息和时间特征信息，确定用户是否结束语音输入。

在本申请的一个实施例中，第一确定模块 302，具体用于：对第一语音信息进行语音识别，以得到第一语音信息的文本信息；获取第一语音信息的历史上下文信息；将文本信息和历史上下文信息输入到语义表示模型中，以得到文本信息的语义特征信息。

在本申请的一个实施例中，第二确定模块 303，具体用于：获取第一语音信息中在静默段之前的第一预设时间长度的语音片段；按照第二预设时间长度，对语音片段进行分段，以得到多段语音片段；提取多段语音片段各自对应的声学特征信息，并分别对多段语音片段各自对应的声学特征信息进行拼接，以得到多段语音片段各自对应的拼接特征；将拼接特征输入到深度残差网络中，以得到第一语音信息的语音特征信息。

在本申请的一个实施例中，上述第二获取模块 304，具体用于：获取第一语音信息的语音时长、语速和文本长度；将语音时长、语速和文本长度输入到预先训练好的多层感知机 MLP 模型，以得到第一语音信息的时间特征信息。

在本申请的一个实施例中，在图 3 所示的装置实施例的基础上，如图 4 所示，上述第三确定模块 305，可以包括多模态处理单元 3051 和确定单元 3052。

多模态处理单元 3051，用于将语义特征信息、语音特征信息和时间特征信息输入到多模态融合模型中。

确定单元 3052，用于根据多模态融合模型的输出结果，确定用户是否结束语音输入。

在本申请的一个实施例中，如图 4 所示，该基于多模态特征的语音对话处理装置 300 还包括第一处理模块 306。

第一处理模块 306，用于在确定用户结束语音输入的情况下，确定第一语音信息所对应的第一回复语音信息，并输出第一回复语音信息。

在本申请的一个实施例中，如图 4 所示，该基于多模态特征的语音对话处理装置 300 还包括第三获取模块 307 和第二处理模块 308。

第三获取模块 307，用于在确定用户未结束语音输入的情况下，获取用户再次输入的第二语音信息。

第二处理模块 308，用于根据第一语音信息和第二语音信息，确定对应的第二回复语音信息，并输出第二回复语音信息。

本申请实施例的基于多模态特征的语音对话处理装置，在与用户进行对话交互的过程

中，结合用户当前输入的语音信息的文本信息和第一语音信息的历史上下文信息，确定文本信息的语义特征信息，根据第一语音信息中在静默段之前的语音片段，确定第一语音信息的语音特征信息，获取第一语音信息的时间特征信息；根据语义特征信息、语音特征信息和时间特征信息，确定用户是否结束语音输入。由此，在与用户进行对话交互的过程中，结合语义特征信息、语音特征信息和时间特征信息，准确确定出了系统是否可接过话语权。

根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

如图 5 所示，是根据本申请一个实施例的电子设备的框图。

如图 5 所示，该电子设备包括存储器 501、处理器 502 及存储在存储器 501 上并可在处理器 502 上运行的计算机指令。

10 处理器 502 执行指令时实现上述实施例中提供的基于多模态特征的语音对话处理方法。进一步地，电子设备还包括通信接口 503，用于存储器 501 和处理器 502 之间的通信。存储器 501，用于存放可在处理器 502 上运行的计算机指令。

存储器 501 可以包含高速 RAM 存储器，也可以还包括非易失性存储器（non-volatile memory），例如至少一个磁盘存储器。

15 处理器 502，用于执行程序时实现上述实施例中提供的基于多模态特征的语音对话处理方法。

如果存储器 501、处理器 502 和通信接口 503 独立实现，则通信接口 503、存储器 501 和处理器 502 可以通过总线相互连接并完成相互间的通信。总线可以是工业标准体系结构（Industry Standard Architecture，简称为 ISA）总线、外部设备互连（Peripheral Component，简称为 PCI）总线或扩展工业标准体系结构（Extended Industry Standard Architecture，简称为 EISA）总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示，图 5 中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

在一些实施例中，如果存储器 501、处理器 502 及通信接口 503，集成在一块芯片上实现，则存储器 501、处理器 502 及通信接口 503 可以通过内部接口完成相互间的通信。

25 处理器 502 可以是一个中央处理器（Central Processing Unit，简称为 CPU），或者是特定集成电路（Application Specific Integrated Circuit，简称为 ASIC），或者是被配置成实施本申请实施例的一个或多个集成电路。

本申请还提出一种计算机程序产品，当计算机程序产品中的指令处理器执行时实现本申请实施例的基于多模态特征的语音对话处理方法。

30 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或

多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

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
5 隐含地包括至少一个该特征。在本申请的描述中，“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。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且
10 本申请的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备（如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、
15 装置或设备取指令并执行指令的系统）使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），
20 可擦除可编程只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得程序，然后将其存储在计算机存储器中。

应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如，如果用硬件来实现和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

30 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通程序来指令相关的硬件完成，的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，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模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

5

上述提到的存储介质可以是只读存储器，磁盘或光盘等。尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

权利要求书

1、一种基于多模态特征的语音对话处理方法，包括：

在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，所述第一语音信息包括静默段；

5 根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息；

根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息；

获取所述第一语音信息的时间特征信息；

10 根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入。

2、如权利要求 1 所述的方法，其中，所述根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息，包括：

对所述第一语音信息进行语音识别，以得到所述第一语音信息的文本信息；

15 获取所述第一语音信息的历史上下文信息；

将所述文本信息和所述历史上下文信息输入到语义表示模型中，以得到所述文本信息的语义特征信息。

3、如权利要求 1 所述的方法，其中，所述根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息，包括：

20 获取所述第一语音信息中在所述静默段之前的第一预设时间长度的语音片段；

按照第二预设时间长度，对所述语音片段进行分段，以得到多段语音片段；

提取所述多段语音片段各自对应的声学特征信息，并分别对所述多段语音片段各自对应的声学特征信息进行拼接，以得到所述多段语音片段各自对应的拼接特征；

将所述拼接特征输入到深度残差网络中，以得到所述第一语音信息的语音特征信息。

25 4、如权利要求 1 所述的方法，其中，所述获取所述第一语音信息的时间特征信息，包括：

获取所述第一语音信息的语音时长、语速和文本长度；

将所述语音时长、所述语速和文本长度输入到预先训练好的多层感知机 MLP 模型，以得到所述第一语音信息的时间特征信息。

30 5、如权利要求 1 所述的方法，其中，所述根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入，包括：

将所述语义特征信息、所述语音特征信息和所述时间特征信息输入到多模态融合模型中；

根据所述多模态融合模型的输出结果，确定所述用户是否结束语音输入。

35 6、如权利要求 1 至 5 中任一项所述的方法，还包括：

在确定所述用户结束语音输入的情况下，确定所述第一语音信息所对应的第一回复

语音信息，并输出所述第一回复语音信息。

7、如权利要求1至5中任一项所述的方法，还包括：

在确定所述用户未结束语音输入的情况下，获取所述用户再次输入的第二语音信息；
根据所述第一语音信息和所述第二语音信息，确定对应的第二回复语音信息，并输

5 出所述第二回复语音信息。

8、一种基于多模态特征的语音交互处理装置，包括：

第一获取模块，用于在与用户进行对话交互的过程中，获取用户当前输入的第一语音信息，其中，所述第一语音信息包括静默段；

10 第一确定模块，用于根据所述第一语音信息的文本信息和所述第一语音信息的历史上下文信息，确定所述文本信息的语义特征信息；

第二确定模块，用于根据所述第一语音信息中在所述静默段之前的语音片段，确定所述第一语音信息的语音特征信息；

第二获取模块，用于获取所述第一语音信息的时间特征信息；

15 第三确定模块，用于根据所述语义特征信息、所述语音特征信息和所述时间特征信息，确定所述用户是否结束语音输入。

9、如权利要求8所述的装置，其中，所述第一确定模块，具体用于：

对所述第一语音信息进行语音识别，以得到所述第一语音信息的文本信息；

获取所述第一语音信息的历史上下文信息；

20 将所述文本信息和所述历史上下文信息输入到语义表示模型中，以得到所述文本信息的语义特征信息。

10、如权利要求8所述的装置，其中，所述第二确定模块，具体用于：

获取所述第一语音信息中在所述静默段之前的第一预设时间长度的语音片段；

按照第二预设时间长度，对所述语音片段进行分段，以得到多段语音片段；

25 提取所述多段语音片段各自对应的声学特征信息，并分别对所述多段语音片段各自对应的声学特征信息进行拼接，以得到所述多段语音片段各自对应的拼接特征；

将所述拼接特征输入到深度残差网络中，以得到所述第一语音信息的语音特征信息。

11、如权利要求8所述的装置，其中，所述第二获取模块，具体用于：

获取所述第一语音信息的语音时长、语速和文本长度；

30 将所述语音时长、所述语速和文本长度输入到预先训练好的多层感知机MLP模型，以得到所述第一语音信息的时间特征信息。

12、如权利要求8所述的装置，其中，所述第三确定模块，包括：

多模态处理单元，用于将所述语义特征信息、所述语音特征信息和所述时间特征信息输入到多模态融合模型中；

35 确定单元，用于根据所述多模态融合模型的输出结果，确定所述用户是否结束语音输入。

13、如权利要求8至12中任一项所述的装置，还包括：

第一处理模块，用于在确定所述用户结束语音输入的情况下，确定所述第一语音信息所对应的第一回复语音信息，并输出所述第一回复语音信息。

14、如权利要求 8 至 12 中任一项所述的装置，还包括：

5 第三获取模块，用于在确定所述用户未结束语音输入的情况下，获取所述用户再次输入的第二语音信息；

第二处理模块，用于根据所述第一语音信息和所述第二语音信息，确定对应的第二回复语音信息，并输出所述第二回复语音信息。

10 15、一种电子设备，包括：存储器，处理器；其中，所述存储器中存储有计算机指令，当所述计算机指令被所述处理器执行时，实现如权利要求 1 至 7 中任一项所述的基于多模态特征的语音对话处理方法。

16、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 至 7 中任一项所述的基于多模态特征的语音对话处理方法。

15 17、一种计算机程序产品，其中，当所述计算机程序产品中的指令处理器执行时，实现如权利要求 1 至 7 中任一项所述的基于多模态特征的语音对话处理方法。