



# (12) 发明专利申请

(10) 申请公布号 CN 114464204 A

(43) 申请公布日 2022. 05. 10

(21) 申请号 202210116757.1

(22) 申请日 2022.02.07

(71) 申请人 京东科技信息技术有限公司

地址 100176 北京市北京经济技术开发区  
科创十一街18号院2号楼6层601

(72) 发明人 杨久东 陈蒙 王培英 冯明超

(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 21/0208 (2013.01)

G10L 15/02 (2006.01)

G10L 15/16 (2006.01)

G10L 15/20 (2006.01)

G10L 15/22 (2006.01)

G10L 21/0216 (2013.01)

G10L 21/0224 (2013.01)

G10L 21/0232 (2013.01)

G10L 25/18 (2013.01)

G10L 25/24 (2013.01)

G10L 25/30 (2013.01)

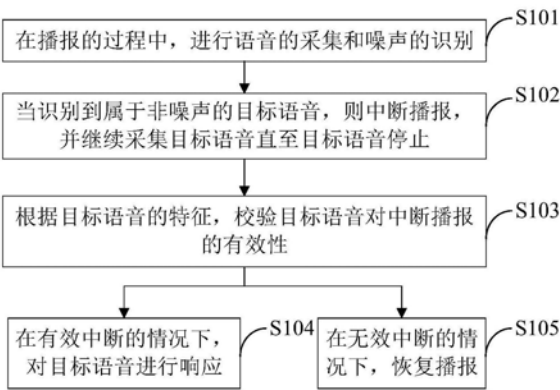
权利要求书3页 说明书11页 附图6页

## (54) 发明名称

语音处理方法、语音处理装置、电子设备和  
存储介质

## (57) 摘要

本发明提出一种语音处理方法、语音处理装置、电子设备和存储介质,其中,方法包括:在播报的过程中,进行语音的采集和噪声的识别;当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止;根据目标语音的特征,校验目标语音对中断播报的有效性;在有效中断的情况下,对目标语音进行响应;在无效中断的情况下,恢复播报。由此,本发明能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权。



1. 一种语音处理方法,其特征在于,包括以下步骤:  
在播报的过程中,进行语音的采集和噪声的识别;  
当识别到属于非噪声的目标语音,则中断播报,并继续采集所述目标语音直至所述目标语音停止;  
根据所述目标语音的特征,校验所述目标语音对中断播报的有效性;  
在有效中断的情况下,对所述目标语音进行响应;  
在无效中断的情况下,恢复播报。
2. 根据权利要求1所述的方法,其特征在于,所述根据所述目标语音的特征,校验所述目标语音对中断播报的有效性,包括:  
对所述目标语音进行特征提取,得到所述目标语音多维度的特征;其中,所述多维度的特征包括语义特征、时域特征和语音声学特征中的至少一个;  
对所述多维度的特征进行特征融合,以得到融合特征;  
将所述融合特征输入经过训练的第一分类模型,以根据所述第一分类模型的输出识别得到所述有效性。
3. 根据权利要求2所述的方法,其特征在于,所述对所述目标语音进行特征提取,得到所述目标语音多维度的特征,包括:  
采用多层感知器MLP模型对所述目标语音的音频进行时域特征提取,以得到所述时域特征;其中,所述时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个;  
采用长短期记忆神经网络LSTM模型对所述目标语音的语音识别文本进行语义特征提取,以得到所述语义特征;  
采用卷积神经网络CNN模型对所述目标语音的语谱图进行声学特征提取,以得到所述语音声学特征;其中,所述语音声学特征包括梅尔频率倒谱系数MFCC、能量和音调中的至少一个。
4. 根据权利要求2所述的方法,其特征在于,所述对所述多维度的特征进行特征融合,以得到融合特征,包括:  
将所述语义特征和所述语音声学特征拼接后输入第一全连接层,以融合得到中间特征;  
将所述中间特征和所述时域特征拼接后输入第二全连接层,以融合得到所述融合特征。
5. 根据权利要求1-4任一项所述的方法,其特征在于,所述在无效中断的情况下,恢复播报,包括:  
在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在所述停止点之前的最近一个停顿音素;  
从所述停顿音素开始恢复播报。
6. 根据权利要求1-4任一项所述的方法,其特征在于,所述在播报的过程中,进行语音的采集和噪声识别,包括:  
在播报的过程中,周期性进行语音的采集;  
对各周期采集的语音片段进行降噪处理;

根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率;

根据所述概率,对各周期采集的语音片段识别是否属于噪声;

将未识别为噪声的语音片段,作为所述非噪声的目标语音。

7. 一种语音处理装置,其特征在于,包括:

采集模块,用于在播报的过程中,进行语音的采集;

识别模块,用于在播报的过程中,进行噪声的识别;

中断模块,用于当识别到属于非噪声的目标语音,则中断播报,并继续通过采集模块采集所述目标语音直至所述目标语音停止;

检验模块,用于根据所述目标语音的特征,校验所述目标语音对中断播报的有效性;

响应模块,用于在有效中断的情况下,对所述目标语音进行响应;

恢复模块,用于在无效中断的情况下,恢复播报。

8. 根据权利要求7所述的装置,其特征在于,所述检验模块,包括:

提取单元,用于对所述目标语音进行特征提取,得到所述目标语音多维度的特征;其中,所述多维度的特征包括语义特征、时域特征和语音声学特征中的至少一个;

融合单元,用于对所述多维度的特征进行特征融合,以得到融合特征;

第一识别单元,用于将所述融合特征输入经过训练的第一分类模型,以根据所述第一分类模型的输出识别得到所述有效性。

9. 根据权利要求8所述的装置,其特征在于,所述提取单元,包括:

第一提取子单元,用于采用多层感知器MLP模型对所述目标语音的音频进行时域特征提取,以得到所述时域特征;其中,所述时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个;

第二提取子单元,用于采用长短期记忆人工神经网络LSTM模型对所述目标语音的语音识别文本进行语义特征提取,以得到所述语义特征;

第三提取子单元,用于采用卷积神经网络CNN模型对所述目标语音的语谱图进行声学特征提取,以得到所述语音声学特征;其中,所述语音声学特征包括梅尔频率倒谱系数MFCC、能量和音调中的至少一个。

10. 根据权利要求8所述的装置,其特征在于,所述融合单元,包括:

第一融合子单元,用于将所述语义特征和所述语音声学特征拼接后输入第一全连接层,以融合得到中间特征;

第二融合子单元,用于将所述中间特征和所述时域特征拼接后输入第二全连接层,以融合得到所述融合特征。

11. 根据权利要求7-10任一项所述的装置,其特征在于,所述恢复模块,包括:

查找单元,用于在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在所述停止点之前的最近一个停顿音素;

恢复单元,用于从所述停顿音素开始恢复播报。

12. 根据权利要求7-10任一项所述的装置,其特征在于,所述采集模块,包括:

采集单元,用于在播报的过程中,周期性进行语音的采集;

所述识别模块,包括:

降噪处理单元,用于对各周期采集的语音片段进行降噪处理;

预测单元,用于根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率;

第二识别单元,用于根据所述概率,对各周期采集的语音片段识别是否属于噪声;

确定单元,用于将未识别为噪声的语音片段,作为所述非噪声的目标语音。

13.一种电子设备,其特征在于,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1-6中任一项所述的语音处理方法。

14.一种非临时性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中任一所述的语音处理方法。

15.一种计算机程序产品,其特征在于,当所述计算机程序产品中的指令由处理器执行时,执行如权利要求1-6中任一所述的语音处理方法。

## 语音处理方法、语音处理装置、电子设备和存储介质

### 技术领域

[0001] 本发明涉及语音处理技术领域,尤其涉及一种语音处理方法、语音处理装置、电子设备和存储介质。

### 背景技术

[0002] 口语对话系统是一种用语音的方式进行人与机器之间的交流,常见于语音外呼和语音应答系统。

[0003] 在口语对话系统中,人与机器之间交流的顺畅性是非常重要的。在现实的场景中,很多时候机器在播报的时候,用户也开始了讲话,也即出现了用户打断机器播报,这个时候机器需要停止播报,把话语权交给用户。

[0004] 但是,在实际应用场景下,会出现各种各样的情况,比如用户在一个嘈杂的环境中,或者旁边有其他人在讲话,这时机器不应该停下来,应该继续播报。只有当出现目标用户真正的打断时,才应该停止播报交出话语权。所以,准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权,是口语对话系统中面对的非常重要的技术问题。

### 发明内容

[0005] 本发明旨在至少从一定程度上解决相关技术中的技术问题之一。

[0006] 本发明提出一种语音处理方法、语音处理装置和存储介质,其能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权。

[0007] 本发明第一方面实施例提出了一种语音处理方法,该方法包括以下步骤:

[0008] 在播报的过程中,进行语音的采集和噪声识别;

[0009] 当识别到属于非噪声的目标语音,则中断播报,并继续采集所述目标语音直至所述目标语音停止;

[0010] 根据所述目标语音的特征,校验所述目标语音对中断播报的有效性;

[0011] 在有效中断的情况下,对所述目标语音进行响应;

[0012] 在无效中断的情况下,恢复播报。

[0013] 在本发明第一方面实施例一种可能的实现方式中,所述根据所述目标语音的特征,校验所述目标语音对中断播报的有效性,包括:

[0014] 对所述目标语音进行特征提取,得到所述目标语音多维度的特征;其中,所述多维度的特征包括语义特征、时域特征和语音声学特征中的至少一个;

[0015] 对所述多维度的特征进行特征融合,以得到融合特征;

[0016] 将所述融合特征输入经过训练的第一分类模型,以根据所述第一分类模型的输出识别得到所述有效性。

[0017] 在本发明第一方面实施例一种可能的实现方式中,所述对所述目标语音进行特征提取,得到所述目标语音多维度的特征,包括:

[0018] 采用多层感知器MLP (Mutl-Layer Perception) 模型对所述目标语音的音频进行

时域特征提取,以得到所述时域特征;其中,所述时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个;

[0019] 采用长短期记忆人工神经网络LSTM(Long short-term Memory Network)模型对所述目标语音的语音识别文本进行语义特征提取,以得到所述语义特征;

[0020] 采用卷积神经网络CNN(Convolutional Neural Network)模型对所述目标语音的语谱图进行声学特征提取,以得到所述语音声学特征;其中,所述语音声学特征包括梅尔频率倒谱系数MFCC(Mel Frequency Cepstrum Coefficient)、能量和音调中的至少一个。

[0021] 在本发明第一方面实施例一种可能的实现方式中,所述对所述多维度的特征进行特征融合,以得到融合特征,包括:

[0022] 将所述语义特征和所述语音声学特征拼接后输入第一全连接层,以融合得到中间特征;

[0023] 将所述中间特征和所述时域特征拼接后输入第二全连接层,以融合得到所述融合特征。

[0024] 在本发明第一方面实施例一种可能的实现方式中,所述在无效中断的情况下,恢复播报,包括:

[0025] 在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在所述停止点之前的最近一个停顿音素;

[0026] 从所述停顿音素开始恢复播报。

[0027] 在本发明第一方面实施例一种可能的实现方式中,所述在播报的过程中,进行语音的采集和噪声识别,包括:

[0028] 在播报的过程中,周期性进行语音的采集;

[0029] 对各周期采集的语音片段进行降噪处理;

[0030] 根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率;

[0031] 根据所述概率,对各周期采集的语音片段识别是否属于噪声;

[0032] 将未识别为噪声的语音片段,作为所述非噪声的目标语音。

[0033] 本发明第二方面实施例提出了一种语音处理装置,该装置包括:

[0034] 采集模块,用于在播报的过程中,进行语音的采集;

[0035] 识别模块,用于在播报的过程中,进行噪声的识别;

[0036] 中断模块,用于当识别到属于非噪声的目标语音,则中断播报,并继续通过采集模块采集所述目标语音直至所述目标语音停止;

[0037] 检验模块,用于根据所述目标语音的特征,校验所述目标语音对中断播报的有效性;

[0038] 响应模块,用于在有效中断的情况下,对所述目标语音进行响应;

[0039] 恢复模块,用于在无效中断的情况下,恢复播报。

[0040] 在本发明第二方面实施例一种可能的实现方式中,所述检验模块,包括:

[0041] 提取单元,用于对所述目标语音进行特征提取,得到所述目标语音多维度的特征;其中,所述多维度的特征包括语义特征、时域特征和语音声学特征中的至少一个;

[0042] 融合单元,用于对所述多维度的特征进行特征融合,以得到融合特征;

[0043] 第一识别单元,用于将所述融合特征输入经过训练的第一分类模型,以根据所述第一分类模型的输出识别得到所述有效性。

[0044] 在本发明第二方面实施例一种可能的实现方式中,所述提取单元,包括:

[0045] 第一提取子单元,用于采用多层感知器MLP模型对所述目标语音的音频进行时域特征提取,以得到所述时域特征;其中,所述时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个;

[0046] 第二提取子单元,用于采用长短期记忆人工神经网络LSTM模型对所述目标语音的语音识别文本进行语义特征提取,以得到所述语义特征;

[0047] 第三提取子单元,用于采用卷积神经网络CNN模型对所述目标语音的语谱图进行声学特征提取,以得到所述语音声学特征;其中,所述语音声学特征包括梅尔频率倒谱系数MFCC、能量和音调中的至少一个。

[0048] 在本发明第二方面实施例一种可能的实现方式中,所述融合单元,包括:

[0049] 第一融合子单元,用于将所述语义特征和所述语音声学特征拼接后输入第一全连接层,以融合得到中间特征;

[0050] 第二融合子单元,将所述中间特征和所述时域特征拼接后输入第二全连接层,以融合得到所述融合特征。

[0051] 在本发明第二方面实施例一种可能的实现方式中,所述恢复模块,包括:

[0052] 查找单元,用于在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在所述停止点之前的最近一个停顿音素;

[0053] 恢复单元,用于从所述停顿音素开始恢复播报。

[0054] 在本发明第二方面实施例一种可能的实现方式中,所述采集模块,包括:

[0055] 采集单元,用于在播报的过程中,周期性进行语音的采集;

[0056] 所述识别模块,包括:

[0057] 降噪处理单元,用于对各周期采集的语音片段进行降噪处理;

[0058] 预测单元,用于根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率;

[0059] 第二识别单元,用于根据所述概率,对各周期采集的语音片段识别是否属于噪声;

[0060] 确定单元,用于将未识别为噪声的语音片段,作为所述非噪声的目标语音。

[0061] 本发明第三方面实施例提出了一种电子设备,其包括: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上述的语音处理方法。

[0062] 本发明第四方面实施例提出了一种非临时性计算机可读存储介质,其上存储有计算机程序,程序被处理器执行时实现上述的语音处理方法。

[0063] 本发明第五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,执行上述的语音处理方法。

[0064] 本发明实施例提供的技术方案至少带来已下有益效果:

[0065] 在播报的过程中,进行语音的采集和噪声的识别,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止,根据目标语音的特征,校验目标

语音对中断播报的有效性,其中,在有效中断的情况下,对目标语音进行响应;在无效中断的情况下,恢复播报。由此,本发明能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权。

[0066] 本发明附加的方面和优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本发明的实践了解到。

## 附图说明

[0067] 本发明上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解,其中:

[0068] 图1是本发明实施例所提供的一种语音处理方法的流程示意图;

[0069] 图2是本发明实施例所提供的一种语音处理方法的打断框架示意图;

[0070] 图3是本发明实施例所提供的另一种语音处理方法的流程示意图;

[0071] 图4是本发明实施例所提供的一种语音处理方法的打断模型结构图;

[0072] 图5是本发明实施例所提供的又一种语音处理方法的流程示意图;

[0073] 图6是本发明实施例所提供的一种语音处理装置的方框示意图;

[0074] 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的框图。

## 具体实施方式

[0075] 下面详细描述本发明的实施例,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,旨在用于解释本发明,而不能理解为对本发明的限制。

[0076] 下面参考附图描述本发明实施例的语音处理方法、语音处理装置、电子设备和存储介质。

[0077] 图1是本发明实施例所提供的一种语音处理方法的流程示意图。

[0078] 如图1所示,本发明实施例的语音处理方法,包括以下步骤:

[0079] S101,在播报的过程中,进行语音的采集和噪声的识别。

[0080] 在该实施例中,机器在播报的过程中,可通过设置在机器上的语音采集模块如麦克风阵列进行语音的采集。

[0081] 然后,构建噪声识别模型如基于朴素贝叶斯构建噪声识别模型,并结合图2说明,将采集到的语音片段输入到噪声识别模块,噪声识别模块给出当前语音片段的得分score。接着,判断当前语音片段的得分score与预设的语音片段的得分阈值T(如0.5)之间的关系,并根据两者之间的关系对当前语音片段进行判断。其中,如果 $score \geq T$ ,则确定当前语音片段不属于噪声,而是真正有效的语音,此时停止播报;如果 $score < T$ ,则将说明当前语音片段可能属于噪声,也可能不属于噪声,此时需要进一步对当前语音片段进行识别。例如,继续采用噪声识别模块,预测当前语音片段属于噪声的概率,并根据概率,识别当前语音片段是否属于噪声。例如,在当前语音片段属于噪声的概率不低于预设概率阈值的情况下,识别出当前语音片段属于噪声,这种情况下表示用户未打断机器;在当前语音片段属于噪声的概率低于预设概率阈值的情况下,识别出当前语音片段不属于噪声,此时,将未识别为噪声的语音片段,作为非噪声的目标语音。

[0082] S102,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止。

[0083] 在该实施例中,在识别到属于非噪声的目标语音的情况下,需要中断播报,再继续采集目标语音,直至用户说完,即可得到完整的目标语音。

[0084] 需要说明的是,在等待用户说完的过程中,即在还未得到完整的目标语音的情况下,机器一直处于播报中。

[0085] S103,根据目标语音的特征,校验目标语音对中断播报的有效性。

[0086] 在该实施例中,将完整的目标语音输入ASR模型,以得到具有目标语音特征的ASR识别文本,再将完整的目标语音和ASR识别文本送入到第一分类模型。由第一分类模型判断目标语音是有意义语音,还是无意义语音,进而确定目标用户的打断是真还是假,以此来校验目标语音对中断播报的有效性。其中,如果目标语音是有意义语音,则目标用户的打断为真,此时确定中断播报为有效中断;如果目标语音是无意义语音,则目标用户的打断为假,此时确定中断播报为无效中断。

[0087] S104,在有效中断的情况下,对目标语音进行响应。

[0088] S105,在无效中断的情况下,恢复播报。

[0089] 在该实施例中,如果是真打断,则进行后续正常流程,例如,可将目标用户的语音识别文本通过意图模型得到用户意图,再根据用户意图走对应流程,如根据采集到的目标语音,正常的理解目标用户的意图并响应。

[0090] 如果是假打断,则恢复播报。

[0091] 如果在播报过程中没有采集到任何声音,则机器不会发生打断,并在播报完成之后,采集用户的目标语音,正常的理解目标用户的意图并响应。

[0092] 举例说明下本发明可能的应用场景:在物流配送的情况下,机器播报的时候可能是在播报地址,在地址未播报完成的情况下,目标用户会出现抢答确定地址,即抢答,这种情况下就需要中断播报,并响应目标用户的确认指令;在地址未播报完成的情况下,目标用户可能并没有发出声音,这种就不需要打断,并在播报完成之后,采集用户的目标语音,正常的理解目标用户的意图并响应。

[0093] 本发明实施例的语音处理方法,在播报的过程中,进行语音的采集和噪声的识别,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止,根据目标语音的特征,校验目标语音对中断播报的有效性,其中,在有效中断的情况下,对目标语音进行响应;在无效中断的情况下,恢复播报。由此,本发明能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权。

[0094] 图3是本发明实施例所提供的另一种语音处理方法的流程示意图。

[0095] 如图3所述,本发明实施例的语音处理方法,包括以下步骤:

[0096] S301,在播报的过程中,进行语音的采集和噪声的识别。

[0097] S302,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止。

[0098] 需要说明的是,关于上述步骤S301和步骤S302的解释说明可参见上述实施例的相关部分,此处不再赘述。

[0099] S303,对目标语音进行特征提取,得到目标语音多维度的特征;其中,多维度的特

征包括语义特征、时域特征和语音声学特征中的至少一个。

[0100] 作为一种可选的实现方式,上述步骤S303的过程例如可以包括:采用多层感知器MLP模型对目标语音的音频进行时域特征提取,以得到时域特征;其中,时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个;采用长短期记忆人工神经网络LSTM模型对目标语音的语音识别文本进行语义特征提取,以得到语义特征;采用卷积神经网络CNN模型对目标语音的语谱图进行声学特征提取,以得到语音声学特征;其中,语音声学特征包括梅尔频率倒谱系数MFCC、能量和音调中的至少一个。

[0101] 在该实施例中,结合图4,在获取到目标语音之后,通过音频处理器获取目标语音的音频,并将目标语音的音频输入多层感知器MLP模型,以对时域特征提取,这样可以得到至少包括语音识别文本的长度、语音长度、语速和语音停顿间隔中一个的时域特征,其中,时域特征可以用向量 $(\alpha_1, \alpha_2, \dots, \alpha_i)$ 表示。

[0102] 在获取到目标语音,且获取到目标语音的语音识别文本(ASR识别文本)之后,将目标语音的语音识别文本输入长短期记忆人工神经网络LSTM模型,以对语义特征提取,这样可以得到语义特征,其中,语义特征可以用向量 $(\beta_1, \beta_2, \dots, \beta_m)$ 表示。例如,机器在播报“计划6月8日给您配送洗衣机,方便签收吗?”的语音,用户回答“6月8日方便签收”的情况下,目标语音的语音识别文本为“6月8日方便签收”。

[0103] 在获取到目标语音之后,获取目标语音的语谱图,并将目标语音的语谱图输入卷积神经网络CNN模型,以提取目标语音的语谱图的声学特征,可以得到至少包括梅尔频率倒谱系数MFCC、能量和音调中一个的语音声学特征,如获取到的语音声学特征用向量 $(r_1, r_2, \dots, r_n)$ 表示。

[0104] S304,对多维度的特征进行特征融合,以得到融合特征。

[0105] 作为一种可选的实现方式,上述步骤S304的过程例如可以包括:将语义特征和语音声学特征拼接后输入第一全连接层,以融合得到中间特征;将中间特征和时域特征拼接后输入第二全连接层,以融合得到融合特征。

[0106] 在该实施例中,继续结合图4,将表示语义特征的向量 $(\beta_1, \beta_2, \dots, \beta_m)$ 和表示语音声学特征的向量 $(r_1, r_2, \dots, r_n)$ 进行拼接,并在拼接后输入第一全连接层,以融合得到中间特征 $(\delta_1, \delta_2, \dots, \delta_j)$ ,接着将中间特征 $(\delta_1, \delta_2, \dots, \delta_j)$ 和表示时域特征的向量 $(\alpha_1, \alpha_2, \dots, \alpha_i)$ 进行拼接,并将拼接后的向量输入第二全连接层,以得到融合特征 $\hat{y}$ 。

[0107] S305,将融合特征输入经过训练的第一分类模型,以根据第一分类模型的输出识别得到有效性。

[0108] 在该实施例中,将融合特征 $\hat{y}$ 输入经过训练的第一分类模型,由第一分类模型对融合特征 $\hat{y}$ 进行识别,判断目标语音是有意义语音,还是无意义语音,进而确定目标用户的打断是真还是假,以此来校验目标语音对中断播报的有效性。其中,如果目标语音是有意义语音,则目标用户的打断为真,此时确定中断播报为有效中断;如果目标语音是无意义语音,则目标用户的打断为假,此时确定中断播报为无效中断。

[0109] S306,在有效中断的情况下,对目标语音进行响应。

[0110] 在该实施例中,如果是真打断,则进行后续正常流程,例如,可将目标用户的语音识别文本通过意图模型得到用户意图,再根据用户意图走对应流程,如根据采集到的目标语音,正常的理解目标用户的意图并响应。

[0111] S307,在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在停止点之前的最近一个停顿音素。

[0112] S308,从停顿音素开始恢复播报。

[0113] 在该实施例中,如果是假打断,则恢复播报。对于恢复播报,当机器在播报时,如果在某个点突然停止了播报,则在恢复播报的时候,需要找到一个恰当且自然的点重新开始播报。一般的方法是找到停止点前面最近的一个词的开始恢复,或者找到前面最近的一个标点符号,但是这些方法都显得不够自然。人的讲话都是抑扬顿挫且高低起伏的,基于此,本发明采用基于音素的方法,先找到中断播报的停止点前面最近的一个停顿音素,并从该处开始恢复播报,就显得更加自然了。

[0114] 如果在播报过程中没有采集到任何声音,则机器不会发生打断,并在播报完成之后,采集用户的目标语音,正常的理解目标用户的意图并响应。

[0115] 本发明实施例的语音处理方法,在播报的过程中,进行语音的采集和噪声的识别,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止,对目标语音进行特征提取,得到目标语音多维度的特征,对多维度的特征进行特征融合,以得到融合特征,将融合特征输入经过训练的第一分类模型,以根据第一分类模型的输出识别得到有效性,在有效中断的情况下,对目标语音进行响应,在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在停止点之前的最近一个停顿音素,从停顿音素开始恢复播报。由此,本发明能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权,另外,通过使用本发明恢复播报的方式,能够使得播报更加自然,进一步提高了用户体验感。

[0116] 图5是本发明实施例所提供的又一种语音处理方法的流程示意图。

[0117] 如图5所示,本发明实施例的语音处理方法,包括以下步骤:

[0118] S501,在播报的过程中,周期性进行语音的采集。

[0119] S502,对各周期采集的语音片段进行降噪处理。

[0120] S503,根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率。

[0121] S504,根据概率,对各周期采集的语音片段识别是否属于噪声。

[0122] S505,将未识别为噪声的语音片段,作为非噪声的目标语音。

[0123] 在该实施例中,机器在播报的过程中,可通过设置在机器上的语音采集模块如麦克风阵列,周期性进行语音的采集。

[0124] 然后,构建噪声识别模型如基于朴素贝叶斯构建噪声识别模型,并结合图2说明,将各周期采集到的语音片段输入到噪声识别模块,噪声识别模型先对各周期的语音片段进行降噪处理,再计算各周期语音片段的得分score。接着,判断各周期语音片段的得分score与预设的语音片段的得分阈值T(如,可以为0.5)之间的关系,并根据两者之间的关系对当前语音片段进行判断。其中,如果 $score \geq T$ ,则确定当前周期的语音片段不属于噪声,而是真正有效的语音,此时停止播报;如果 $score < T$ ,则将说明当前周期的语音片段可能属于噪声,也可能不属于噪声,此时进一步对当前周期的语音片段进行识别。例如,根据降噪前的语音片段和降噪后的语音片段,继续采用噪声识别模块,以预测当前周期的语音片段属于噪声的概率,并根据得到的概率,识别当前周期的语音片段是否属于噪声。其中,在当前周

期的语音片段属于噪声的概率不低于预设概率阈值的情况下,识别出当前周期的语音片段属于噪声,这种情况下表示用户未打断机器;在当期周期的语音片段属于噪声的概率低于预设概率阈值的情况下,识别出当期周期的语音片段不属于噪声,此时,将未识别为噪声的语音片段,作为非噪声的目标语音。

[0125] S506,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止。

[0126] S507,根据目标语音的特征,校验目标语音对中断播报的有效性。

[0127] S508,在有效中断的情况下,对目标语音进行响应。

[0128] S509,在无效中断的情况下,恢复播报。

[0129] 需要说明的是,关于上述步骤S506-步骤S509的解释说明可参见上述实施例的相关部分,此处不再赘述。

[0130] 本发明实施例的语音处理方法,在播报的过程中,周期性进行语音的采集,对各周期采集的语音片段进行降噪处理,根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率,根据概率,对各周期采集的语音片段识别是否属于噪声,将未识别为噪声的语音片段,作为非噪声的目标语音,当识别到属于非噪声的目标语音,则中断播报,并继续采集目标语音直至目标语音停止,根据目标语音的特征,校验目标语音对中断播报的有效性,其中,在有效中断的情况下,对目标语音进行响应;在无效中断的情况下,恢复播报。由此,本发明能够准确地判断是不是用户真正的打断,并在合适的时机停止播报交出话语权。

[0131] 为了实现上述实施例,本发明实施例提供了一种语音处理装置。

[0132] 图6是本发明实施例所提供的一种语音处理装置的方框示意图。

[0133] 如图6所示,本发明实施例的语音处理装置600,包括:采集模块601、识别模块602、中断模块603、检验模块604、响应模块605和恢复模块606。

[0134] 其中,采集模块601用于在播报的过程中,进行语音的采集。识别模块602用于在播报的过程中,进行噪声的识别。中断模块603用于当识别到属于非噪声的目标语音,则中断播报,并继续通过采集模块601采集目标语音直至目标语音停止。检验模块604用于根据目标语音的特征,校验目标语音对中断播报的有效性。响应模块605用于在有效中断的情况下,对目标语音进行响应。恢复模块606用于在无效中断的情况下,恢复播报。

[0135] 在本发明实施例一种可能的实现方式中,检验模块604,包括:提取单元、融合单元和第一识别单元。其中,提取单元用于对目标语音进行特征提取,得到目标语音多维度的特征;其中,多维度的特征包括语义特征、时域特征和语音声学特征中的至少一个。融合单元用于对多维度的特征进行特征融合,以得到融合特征。第一识别单元用于将融合特征输入经过训练的第一分类模型,以根据第一分类模型的输出识别得到有效性。

[0136] 在本发明实施例一种可能的实现方式中,提取单元,包括:第一提取子单元、第二提取子单元和第三提取子单元。其中,第一提取子单元用于采用多层感知器MLP模型对目标语音的音频进行时域特征提取,以得到时域特征;其中,时域特征包括语音识别文本的长度、语音长度、语速和语音停顿间隔中的至少一个。第二提取子单元用于采用长短期记忆人工神经网络LSTM模型对目标语音的语音识别文本进行语义特征提取,以得到语义特征。第三提取子单元用于采用卷积神经网络CNN模型对目标语音的语谱图进行声学特征提取,以

得到语音声学特征;其中,语音声学特征包括梅尔频率倒谱系数MFCC、能量和音调中的至少一个。

[0137] 在本发明实施例一种可能的实现方式中,融合单元,包括:第一融合子单元和第二融合子单元。其中,第一融合子单元用于将语义特征和语音声学特征拼接后输入第一全连接层,以融合得到中间特征。第二融合子单元用于将中间特征和时域特征拼接后输入第二全连接层,以融合得到融合特征。

[0138] 在本发明实施例一种可能的实现方式中,恢复模块606,包括:查找单元和恢复单元。其中,查找单元用于在无效中断的情况下,根据中断播报的停止点,在所播报的语音中查找在停止点之前的最近一个停顿音素。恢复单元用于从停顿音素开始恢复播报。

[0139] 在本发明实施例一种可能的实现方式中,采集模块601,包括:采集单元,采集单元用于在播报的过程中,周期性进行语音的采集。

[0140] 在本发明实施例一种可能的实现方式中,识别模块602,包括:降噪处理单元、预测单元、第二识别单元和确定单元。其中,降噪处理单元用于对各周期采集的语音片段进行降噪处理。预测单元用于根据降噪前的语音片段和降噪后的语音片段,采用噪声识别模型,预测各周期采集的语音片段属于噪声的概率。第二识别单元用于根据概率,对各周期采集的语音片段识别是否属于噪声。确定单元用于将未识别为噪声的语音片段,作为非噪声的目标语音。

[0141] 在此需要说明的是,本发明实施例提供的上述语音处理装置,能够实现上述图1、图3和图5方法实施例所实现的所有方法步骤,且能够达到相同的技术效果,在此不再对本实施例中的与方法实施例相同的部分及有益效果进行具体赘述。

[0142] 为了实现上述实施例,本发明还提出了一种电子设备。

[0143] 其中,该电子设备包括至少一个处理器;以及与至少一个处理器通信连接的存储器;其中,存储器存储有可被至少一个处理器执行的指令,指令被至少一个处理器执行,以使至少一个处理器能够执行上述图1、图3和图5实施例所述的。

[0144] 为了实现上述实施例,本发明还提出一种非临时性计算机可读存储介质。

[0145] 其中,该非临时性计算机可读存储介质上存储有计算机程序,该程序被处理器执行时实现上述图1、图3和图5实施例所述的语音处理方法。

[0146] 为了实现上述实施例,本发明还提出一种计算机程序产品。

[0147] 其中,当计算机程序产品中的指令由处理器执行时,执行上述图1、图3和图5实施例所述的语音处理方法。

[0148] 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的框图。其中,该电子设备可以为上述实施例中的机器。图7显示的电子设备7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。

[0149] 如图7所示,电子设备7以通用计算设备的形式表现。电子设备7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。

[0150] 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(Industry Standard

Architecture;以下简称:ISA)总线,微通道体系结构(Micro Channel Architecture;以下简称:MAC)总线,增强型ISA总线、视频电子标准协会(Video Electronics Standards Association;以下简称:VESA)局域总线以及外围组件互连(Peripheral Component Interconnection;以下简称:PCI)总线。

[0151] 电子设备7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7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。

[0152] 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(Random Access Memory;以下简称:RAM)30和/或高速缓存存储器32。电子设备7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:光盘只读存储器(Compact Disc Read Only Memory;以下简称:CD-ROM)、数字多功能只读光盘(Digital Video Disc Read Only Memory;以下简称: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。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。

[0153] 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。

[0154] 电子设备7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计算机系统/服务器12交互的设备通信,和/或与使得该计算机系统/服务器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7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(Local Area Network;以下简称:LAN),广域网(Wide Area Network;以下简称: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7的其它模块通信。应当明白,尽管图7中未示出,可以结合电子设备7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。

[0155] 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,例如实现前述实施例中提及的方法。

[0156] 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发明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。

[0157] 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。在本发明的描述中,“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。

[0158] 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,并且本发明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,这应被本发明的实施例所属技术领域的技术人员所理解。

[0159] 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、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),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所述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所述程序,然后将其存储在计算机存储器中。

[0160] 应当理解,本发明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它们的组合来实现: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。

[0161] 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,所述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。

[0162] 此外,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。所述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,也可以存储在一个计算机可读存储介质中。

[0163] 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发明的实施例,可以理解的是,上述实施例是示例性的,不能理解为对本发明的限制,本领域的普通技术人员在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

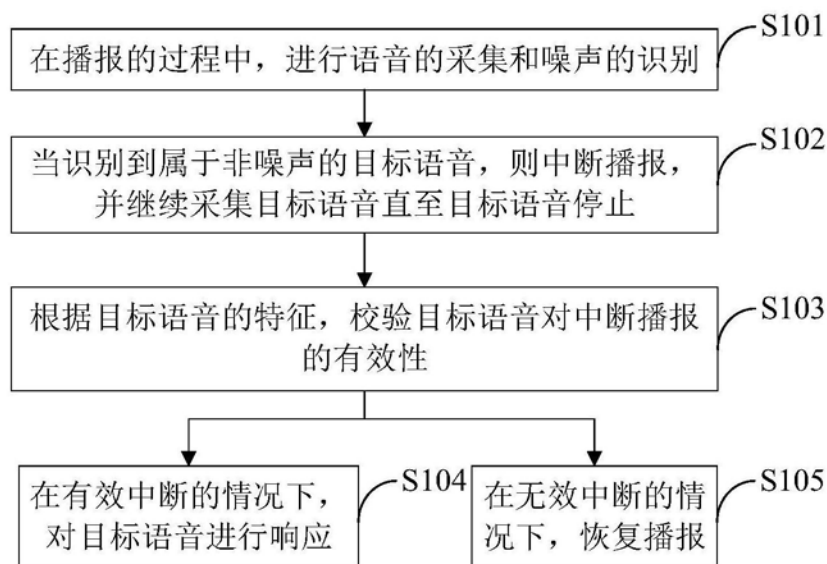


图1

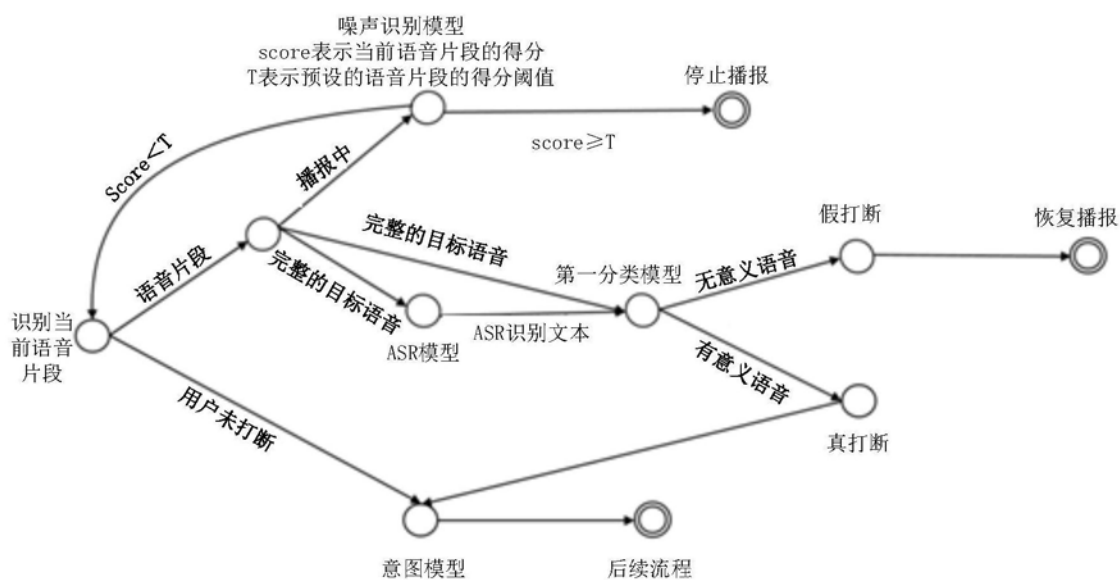


图2

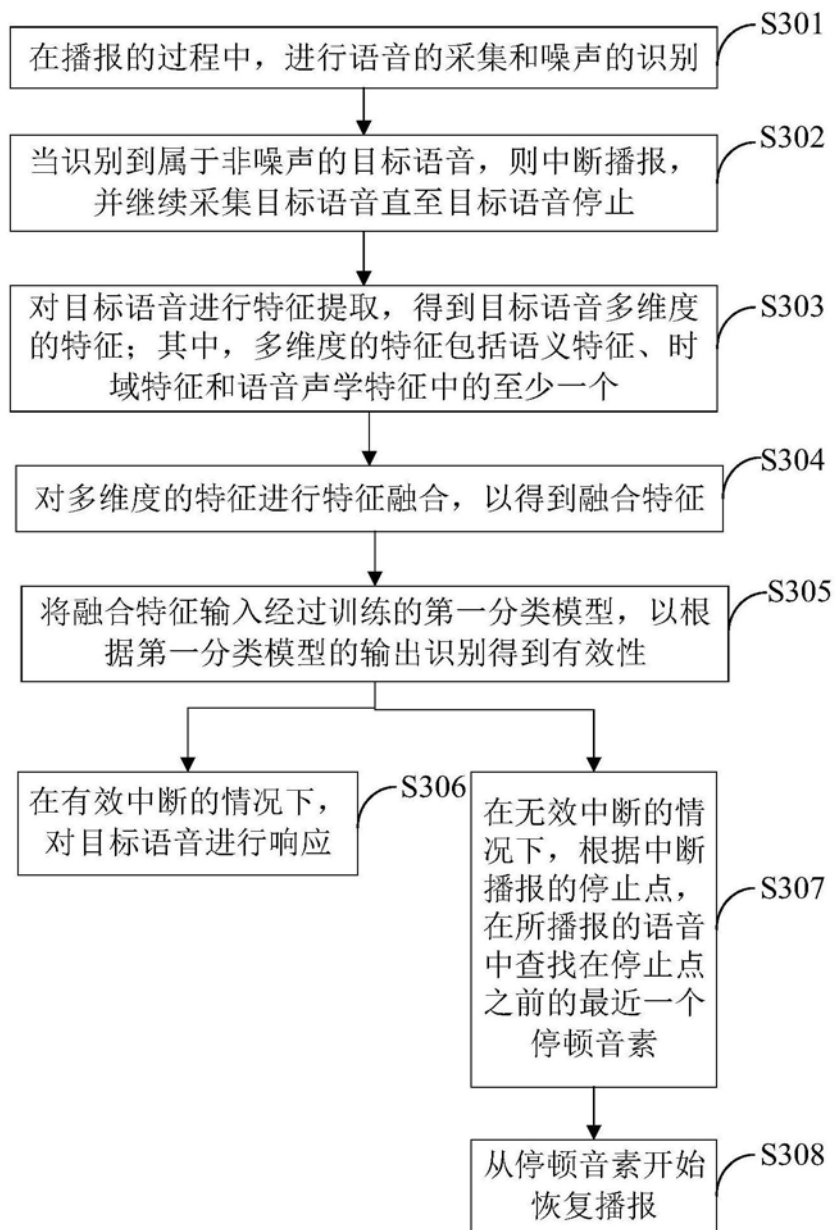


图3

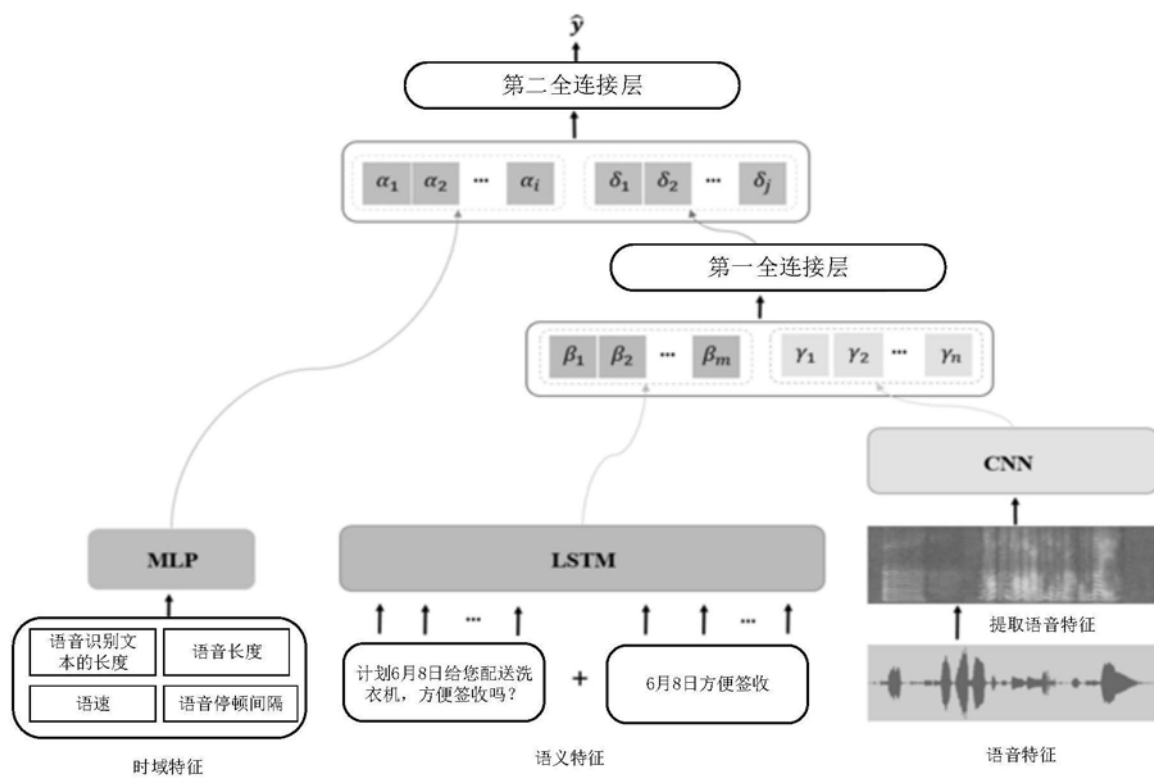


图4

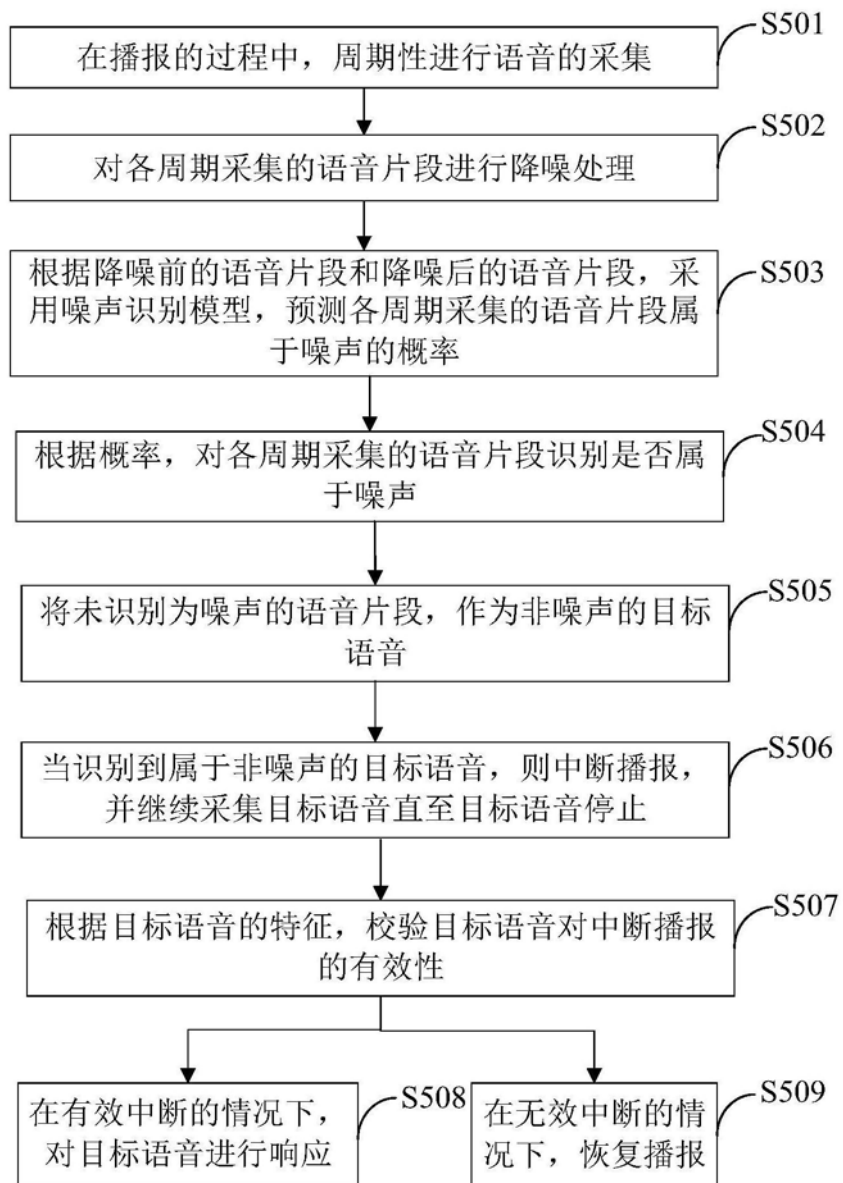


图5

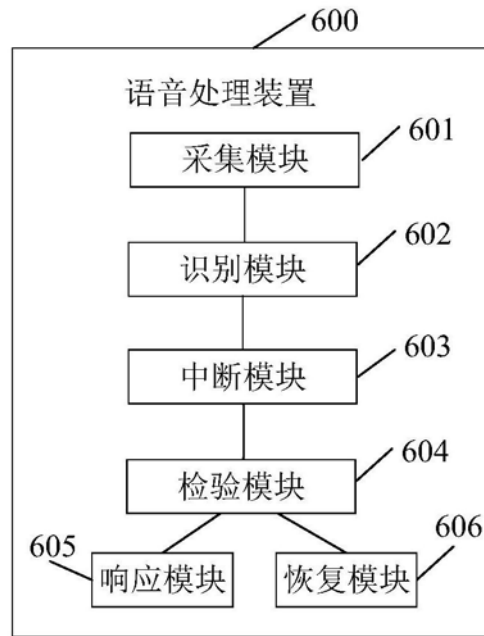


图6

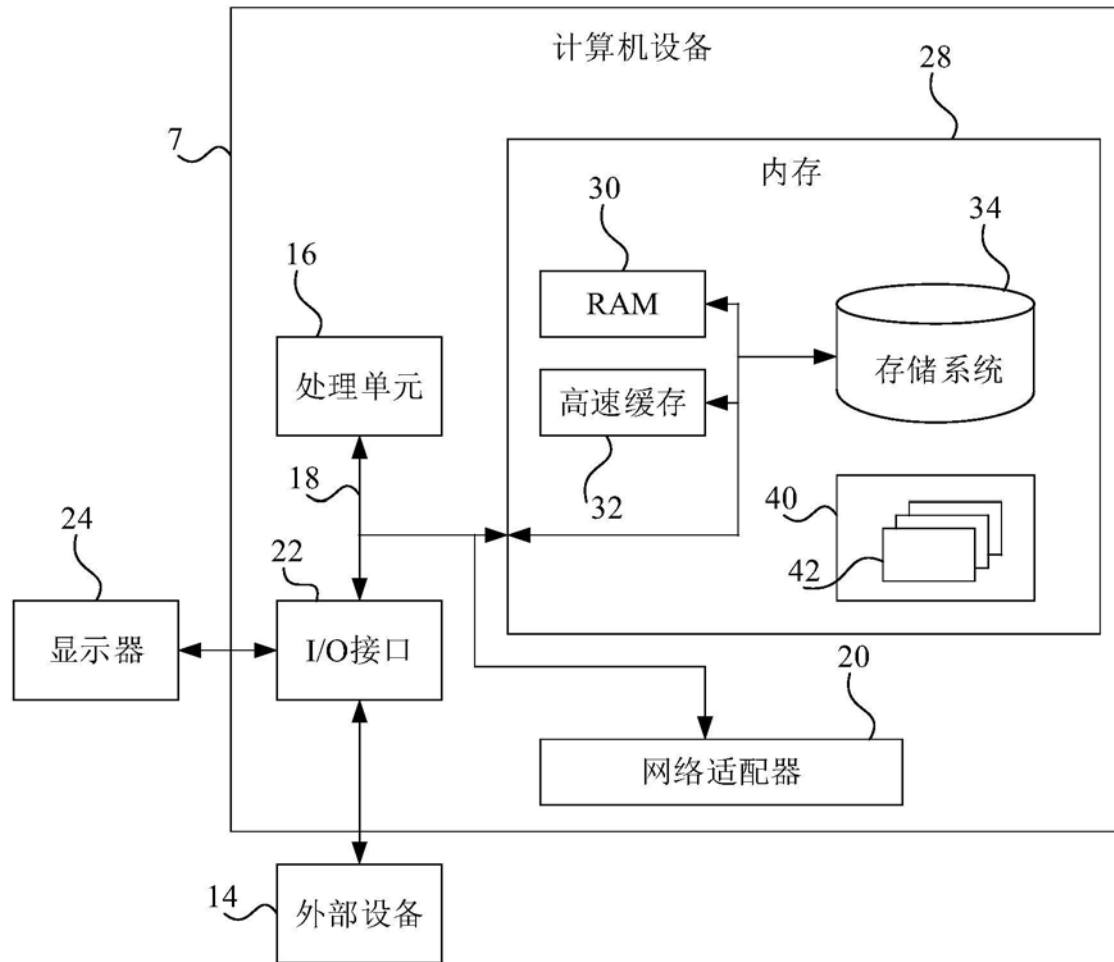


图7