(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112735385 B
(45)授权公告日 2024.05.31

(21)申请号 202011625225.8
(22)申请日 2020.12.30
(65)同一申请的已公布的文献号
申请公布号 CN 112735385 A
(43)申请公布日 2021.04.30
(73)专利权人 中国科学技术大学
地址 230022 安徽省合肥市金寨路96号
专利权人 科大讯飞股份有限公司
(72)发明人 王庆然 万根顺 高建清 刘聪 王智国 胡国平
(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 熊永强
(51)Int.Cl.
G10L 15/02(2006.01)
G10L 15/04(2013.01)
G10L 15/06(2013.01)
G10L 15/16(2006.01)
G10L 15/26(2006.01)

(56)对比文件
US 2006085188 A1,2006.04.20
CN 106251874 A,2016.12.21
CN 108509558 A,2018.09.07
CN 105913849 A,2016.08.31
CN 111816218 A,2020.10.23
CN 110689906 A,2020.01.14
WO 2018018906 A1,2018.02.01
US 2019341057 A1,2019.11.07
KR 20140076816 A,2014.06.23
CN 106611604 A,2017.05.03
CN 110136749 A,2019.08.16
CN 102074236 A,2011.05.25
母东杰;李悦;王建勋.基于尺度变换的数据转折点检测方法.控制工程.2018,(第01期),全文.

审查员 杜智慧

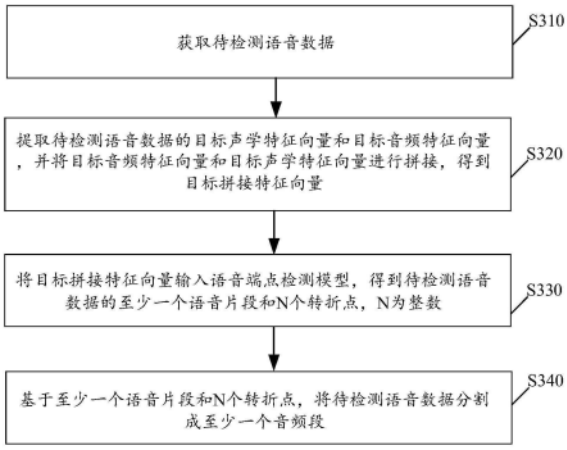
权利要求书3页 说明书14页 附图9页

(54)发明名称

语音端点检测方法、装置、计算机设备及存储介质

(57)摘要

本申请公开了一种语音端点检测方法、装置、计算机设备及存储介质,该方法包括:获取待检测语音数据;提取待检测语音数据的目标声学特征向量和目标音频特征向量,并将目标音频特征向量和目标声学特征向量进行拼接,得到目标拼接特征向量;将目标拼接特征向量输入语音端点检测模型,得到待检测语音数据的至少一个语音片段和N个转折点;基于该至少一个语音片段和N个转折点,将待检测语音数据分割成至少一个音频段。本申请通过融合语音数据的音频特征和声学特征得到语音片段和转折点,根据转折点和语音片段对语音数据进行分割,可以将包括多人说话的语音片段分割成多个单说话人的音频段,提高多人讨论场景下语音端点检测的准确性。



CN 112735385 B

1.一种语音端点检测方法,其特征在于,所述方法包括:

获取待检测语音数据;

提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数;

基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段,包括:若所述N 小于1,将所述至少一个语音片段确定为所述至少一个音频段;若所述N大于或等于1,对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段;

其中,所述对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段,包括:

当所述N等于1,若所述转折点满足第一条件时,删除所述转折点,并将所述至少一个语音片段确定为所述至少一个音频段,否则,根据所述转折点将所述第一语音片段进行分割,得到多个音频段,所述第一条件为所述转折点与第一语音片段边界的距离小于第一阈值,或所述转折点位于有效音频数据上,所述第一语音片段为所述转折点所在的语音片段;

在所述N大于1时,若任一转折点满足所述第一条件和第二条件时,删除所述任一转折点,否则保留所述任一转折点,并根据所述任一转折点对所述至少一个语音片段进行分割,得到所述多个音频段,所述第二条件为所述任一转折点与目标转折点的距离小于所述第一阈值,所述目标转折点为保留的转折点。

2.根据权利要求1所述的方法,其特征在于,所述语音端点检测模型通过以下方式预先训练得到的:

获取训练数据集,所述训练数据集包括多个说话人的多条训练数据;

提取所述多条训练数据中的每条训练数据的声学特征向量和音频特征向量,并将所述每条训练数据的所述音频特征向量和所述声学特征向量进行拼接,得到所述每条训练数据拼接特征向量;

将所述每条训练数据的拼接特征向量输入待训练语音端点检测模型进行训练,直至达到训练结束条件,得到所述语音端点检测模型。

3.根据权利要求2所述的方法,其特征在于,所述方法还包括:

获取多条携带标注信息的原始语音数据,所述标注信息包括每条原始语音数据中的至少一个说话人、每个所述说话人的起始点和结束点;

根据所述原始语音数据中每个所述说话人的起始端点和结束端点,将每条所述原始语音数据分割成至少一条子语音数据;

将所述多条原始语音数据的至少一条子语音数据拼接成多条样本数据;

标注所述每条样本数据中的每一帧音频的分类信息,得到所述多条训练数据。

4.根据权利要求3所述的方法,其特征在于,所述分类信息包括语音帧、非语音帧和转折帧;

所述将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,包括:

将所述目标拼接特征向量输入所述语音端点检测模型,得到第一后验概率、第二后验概率和第三后验概率,所述第一后验概率为每帧中包括语音帧的概率,所述第二后验概率为每帧中包括非语音帧的概率,所述第三后验概率为每帧中包括转折帧的概率;

根据所述第一后验概率和所述第二后验概率,确定所述待检测语音数据的所述至少一个语音片段;

根据所述第三后验概率和所述至少一个语音片段,确定所述待检测语音数据的所述N个转折点。

5.根据权利要求3所述的方法,其特征在于,所述分类信息包括转折帧和非转折帧;

所述将所述每条训练数据的拼接特征向量输入待训练语音端点检测模型,直至达到训练结束条件,得到所述语音端点检测模型,包括:

将所述每条训练数据的拼接特征向量输入所述待训练语音端点检测模型的共享层,得到所述每条训练数据中的每一帧音频的第四后验概率,所述共享层包括多个神经网络模型,所述第四后验概率为每帧音频中包括说话人的概率;

将所述每一帧音频的第四后验概率分别输入所述待训练语音端点检测模型的第一任务层和第二任务层,分别得到第五后验概率和第六后验概率,所述第五后验概率为每帧中包括非转折帧的概率,所述第六后验概率为每帧中包括转折帧的概率;

基于所述第五后验概率计算第一梯度,基于所述第六后验概率计算第二梯度,所述第一梯度为所述第一任务层的梯度,所述第二梯度为所述第二任务层的梯度;

根据所述第一梯度和所述第二梯度更新所述待训练语音端点检测模型的参数,直至达到训练结束条件,得到所述语音端点检测模型。

6.根据权利要求5所述的方法,其特征在于,所述将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,包括:

将所述目标拼接特征向量输入所述语音端点检测模型,得到第七后验概率和第八后验概率;其中,所述第七后验概率为每帧中包括非转折帧的后验概率,所述第八后验概率为每帧中包括转折帧的后验概率;

根据所述第七后验概率,确定所述待检测语音数据的所述至少一个语音片段;

根据所述至少一个语音片段和所述第八后验概率,确定所述待检测语音数据的所述N个转折点。

7.一种语音端点检测装置,其特征在于,所述装置包括:

获取单元,用于获取待检测语音数据;

提取单元,用于提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

检测单元,用于将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数;

分割单元,用于基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段,包括:若所述N 小于1,将所述至少一个语音片段确定为所述至少一个音频段;若所述N大于或等于1,对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段;

其中,所述对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段

进行分割,得到所述至少一个音频段,包括:

当所述N等于1,若所述转折点满足第一条件时,删除所述转折点,并将所述至少一个语音片段确定为所述至少一个音频段,否则,根据所述转折点将所述第一语音片段进行分割,得到多个音频段,所述第一条件为所述转折点与第一语音片段边界的距离小于第一阈值,或所述转折点位于有效音频数据上,所述第一语音片段为所述转折点所在的语音片段;

在所述N大于1时,若任一转折点满足所述第一条件和第二条件时,删除所述任一转折点,否则保留所述任一转折点,并根据所述任一转折点对所述至少一个语音片段进行分割,得到所述多个音频段,所述第二条件为所述任一转折点与目标转折点的距离小于所述第一阈值,所述目标转折点为保留的转折点。

8.一种计算机设备,其特征在于,所述计算机设备包括处理器、存储器、通信接口,以及一个或多个程序,所述一个或多个程序被存储在所述存储器中,并且被配置由所述处理器执行,所述程序包括用于执行如权利要求1-6任一项所述的方法中的步骤的指令。

9.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储用于电子数据交换的计算机程序,其中,所述计算机程序使得计算机执行如权利要求1-6任一项所述的方法。

语音端点检测方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请涉及语音信息处理技术领域,尤其涉及一种语音端点检测方法、装置、计算机设备及存储介质。

背景技术

[0002] 随着信息技术的发展,语音信息的应用越来越广泛。语音端点检测(Voice Activity Detection,VAD)技术是语音识别领域的重要技术,在一段长语音音频中获取真正想要提取的目标活跃语音片段,对提高语音识别的正确率至关重要。

[0003] VAD采用神经网络模型结合逻辑策略的方法,在提取输入语音音频的频域特征之后,通过实时的神经网络模型输出每一帧音频的后验信息,再通过逻辑策略的方法将每一帧音频的状态串起来解码,从而判断出每一个需要切割出来的语音片段。

[0004] VAD检测有效语音段的结束端点需要等待一段时间的纯静音时长,否则检测不出来有效的语音结束端点。但是,在嘈杂的多人讨论场景下,两人讨论问题的间隔可能不存在一段完全安静的片段,从而使得切割后的有效语音片段中可能包含多人的语音片段或者切割后的有效语音片段很长。而过长的语音片段或者有多个说话人的语音片段会使得语音识别不准确,同时可能会将包含两个说话人的语音片段中的第二个说话人的语音内容丢弃,只识别出第一个人的识别结果。因此,如何在多人讨论场景下提供语音端点检测的准确性是亟待解决的问题。

发明内容

[0005] 本申请实施例提供一种语音端点检测方法、装置、计算机设备及存储介质,能够提高多人讨论场景下的语音端点检测的准确性。

[0006] 第一方面,本申请实施例提供一种语音识别方法,该方法包括:

[0007] 获取待检测语音数据;

[0008] 提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

[0009] 将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数;

[0010] 基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段。

[0011] 第二方面,本申请实施例提供一种语音识别装置,该装置包括:

[0012] 获取单元,用于获取待检测语音数据;

[0013] 提取单元,用于提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

[0014] 检测单元,用于将所述目标拼接特征向量输入语音端点检测模型,得到所述待检

测语音数据的至少一个语音片段和N个转折点,所述N为整数;

[0015] 分割单元,用于基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段。

[0016] 第三方面,本申请实施例提供一种终端设备,包括处理器、存储器、通信接口以及一个或多个程序,其中,上述一个或多个程序被存储在上述存储器中,并且被配置由上述处理器执行,上述程序包括用于执行本申请实施例第一方面任一方法中的步骤的指令。

[0017] 第四方面,本申请实施例提供了一种计算机可读存储介质,其中,上述计算机可读存储介质存储用于电子数据交换的计算机程序,其中,上述计算机程序使得计算机执行如本申请实施例第一方面任一方法中所描述的部分或全部步骤。

[0018] 第五方面,本申请实施例提供了一种计算机程序产品,其中,上述计算机程序产品包括存储了计算机程序的非瞬时性计算机可读存储介质,上述计算机程序可操作来使计算机执行如本申请实施例第一方面任一方法中所描述的部分或全部步骤。该计算机程序产品可以为一个软件安装包。

[0019] 在本申请实施例提供的语音端点检测方法,获取待检测语音数据;提取待检测语音数据的目标声学特征向量和目标音频特征向量,并将目标音频特征向量和目标声学特征向量进行拼接,得到目标拼接特征向量;将目标拼接特征向量输入语音端点检测模型,得到待检测语音数据的至少一个语音片段和N个转折点;基于该至少一个语音片段和N个转折点,将待检测语音数据分割成至少一个音频段。本申请通过将融合语音数据的音频特征和声学特征输入到语音端点检测模型中,得到语音片段和转折点,根据转折点和语音片段对语音数据进行分割,可以将包括多人说话的语音片段分割成多个单说话人的音频段,实现在多人快速交换讨论的场景下,将传统VAD模块切不开的多人转换语音片段区分开来,以提高多人讨论场景下语音端点检测的准确性。

附图说明

[0020] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0021] 图1是本申请实施例提供的一种语音识别系统运行的网络架构图;

[0022] 图2为本申请实施例提供的一种语音识别系统的语音识别原理示意图;

[0023] 图3是本申请实施例提供的一种语音端点检测方法的流程示意图;

[0024] 图4a是本申请实施例提供的一种音频特征与声学特征拼接的示意图;

[0025] 图4b是本申请实施例提供的一种S330的具体流程示意图;

[0026] 图4c是本申请实施例提供的一种语音端点检测模型解码的示意图;

[0027] 图4d是本申请实施例提供的一种标注音频分类类型的示意图;

[0028] 图5是本申请实施例提供的另一种S330的具体流程示意图;

[0029] 图5a是本申请实施例提供的一种语音端点检测模型结构的示意图;

[0030] 图6是本申请实施例提供的一种语音端点检测模型训练的流程示意图;

[0031] 图6a是本申请实施例提供的另一种标注音频分类类型的示意图;

[0032] 图7a是本申请实施例提供的一种语音端点检测装置的功能单元组成框图；

[0033] 图7b是本申请实施例提供的另一种语音端点检测装置的功能单元组成框图；

[0034] 图8是本申请实施例提供的一种计算机设备的结构示意图。

具体实施方式

[0035] 下面结合附图,对本申请实施例进行详细说明。

[0036] 应理解,本申请实施例中涉及的“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B的情况,其中A,B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a,b,c,a-b,a-c,b-c,或a-b-c,其中a,b,c可以是单个,也可以是多个。

[0037] 以及,除非有相反的说明,本申请实施例提及“第一”、“第二”等序数词是用于对多个对象进行区分,不用于限定多个对象的顺序、时序、优先级或者重要程度。例如,第一信息和第二信息,只是为了区分不同的信息,而并不是表示这两种信息的内容、优先级、发送顺序或者重要程度等的不同。

[0038] 应理解,本申请提供的语音端点检测方法可以应用于终端设备中包含语音识别功能的系统或程序中,例如媒体内容平台,具体的,语音识别系统可以运行于如图1所示的网络架构中,如图1所示,是语音识别系统运行的网络架构图,如图可知,语音识别系统可以提供与多个信息源的语音识别,终端通过网络建立与服务器的连接,进而接收服务器发送的媒体内容,并对媒体内容中的语音进行还原并识别;可以理解的是,图1中示出了多种终端设备,在实际场景中可以有更多或更少种类的终端设备参与到语音识别的过程中,具体数量和种类因实际场景而定,此处不做限定,另外,图1中示出了一个服务器,但在实际场景中,也可以有多个服务器的参与,特别是在多内容应用交互的场景中,具体服务器数量因实际场景而定。

[0039] 本实施例中,服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请在此不做限制。

[0040] 应当注意的是,本实施例提供的语音端点检测方法也可以离线进行,即不需要服务器的参与,此时终端在本地与其他终端进行连接,进而进行终端之间的语音识别的过程。

[0041] 可以理解的是,上述语音识别系统可以运行于个人移动终端,例如:作为媒体内容平台这样的应用,也可以运行于服务器,还可以作为运行于第三方设备以提供语音识别,以得到信息源的语音识别处理结果;具体的语音识别系统可以是以一种程序的形式在上述设备中运行,也可以作为上述设备中的系统部件进行运行,还可以作为云端服务程序的一种,具体运作模式因实际场景而定,此处不做限定。

[0042] 请参阅图2,图2为本申请实施例提供的语音识别系统的语音识别原理示意图。语

音识别(Automatic Speech Recognition,ASR)所要解决的问题是让计算机能够"听懂"人类的语音,将语音转化成文本。如图1所示,该语音识别系统的识别过程包括前端处理过程和后端处理过程。其中,前端可以为能够与用户进行语音交互的各种智能终端,例如智能手机、智能音箱、智能电视、智能冰箱等,本实施例对智能终端的实现方式不做特别限制。后端可以为能够进行数据逻辑处理的服务器,本领域技术人员可以理解,该后端也可以为智能终端的处理器。预先训练好声学模型和语言模型可以存储到后端。其中,声学模型对应于语音到音节概率的计算,语音模型对应于音节到字概率的计算。

[0043] 在具体实现过程中,前端在接收到语音之后,对接收到的语音进行分帧处理,然后进行端点检测,在检测到语音段的起点后,对起点之后的语音帧进行降噪处理,然后进行特征提取直至检测到语音段的终点,根据提取到的声学特征、声学模型、语音模型进行解码,得到识别结果。在一种可能的实现方式中,解码是将声学模型、词典以及语言模型编译成一个网络。解码就是在这个动态网络空间中,基于最大后验概率,选择一条或多条最优路径作为识别结果(最优的输出字符序列)。

[0044] 由此可见,在此过程中端点检测尤其重要,决定了语音识别系统的输入。然而现有VAD检测有效语音段的结束端点需要等待一段时间的纯静音时长,否则检测不出来有效的语音结束端点。在嘈杂的多人讨论场景下,两人讨论问题的间隔可能不存在一段完全安静的片段,因此切割后的有效语音片段中可能包含多人的语音片段或者切割后的有效语音片段很长。而过长的语音片段或者有多个说话人的语音片段会使得语音识别不准确,同时可能会将包含两个说话人的语音片段中的第二个说话人的语音内容丢弃,只识别出第一个人的识别结果。

[0045] 基于此,本申请实施例提供一种语音端点检测方法,通过将融合语音数据的音频特征和声学特征输入到语音端点检测模型中,得到语音片段和转折点,根据转折点和语音片段对语音数据进行分割,可以将包括多人说话的语音片段分割成多个单说话人的音频段,实现在多人快速交换讨论的场景下,将传统VAD模块切不开的多人转换语音片段区分开来,以提高多人讨论场景下语音端点检测的准确性。

[0046] 下面分别从训练语音端点检测模型和使用语音端点检测模型两方面分别进行详细说明。

[0047] 请参阅图3,图3为本申请实施例提供的一种语音端点检测方法的流程示意图。如图3所示,该方法包括以下步骤。

[0048] S310、获取待检测语音数据。

[0049] 其中,所述待检测语音数据可包括至少一个说话人的多条语音数据,在具体实现过程中,可以采集至少一个用户的多条语音作为待检测语音数据,例如,电话会议过程中的语音数据,多人聊天过程中的语音数据。示例性地,所述待检测语音数据可以是从视频中的音频信息提取所得,例如短视频中的音频信息;具体形式因实际场景而定,此处不做限定。

[0050] S320、提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量。

[0051] 其中,在获取到待训练语音数据后,需要提取语音的目标音频特征来训练神经网络。对于常规的语音识别任务来说,只需要提取音频的频域特征,如梅尔倒谱系数(Mel-frequency Cepstral Coefficient,MFCC)和滤波器组特征(Filter Bank)等。但由于待检

测语音数据中包括不同说话人的声纹信息,为了区分出不同的说话人,因此还提取待检测语音数据的目标声学特征,如i-vector特征或者d-vector特征等。目标声学特征的提取可以采用滑动窗的方式来保证实时性,也可以直接取一整条语音片段来进行提取。进一步地,可以采用基于网络时延神经网络(Time Delay Neural Network,TDNN)结构的实时声纹提取模型提取每一待检测音频帧的声学特征。

[0052] 示例性地,提取各待检测音频帧的声学特征时,可以采用线性预测编码(Linear Predictive Coding,LPC)特征,MFCC特征,感知线性预测(Perceptual Linear Predictive,PLP)特征等,本申请实施例对声学特征的类型不进行限制,声学特征的提取即是将各待检测的音频帧转换为一个多维向量的过程。

[0053] 在提取了待检测语音数据的音频特征和声学特征后,需要构建对应的特征向量。可将从待检测语音数据中提取的目标声学特征向量和提取的目标音频特征拼接起来,构建新的多维目标拼接特征向量。

[0054] 进一步地,声学特征向量的维数一般较高,例如ivector特征维数为几百或上千,而音频特征向量维数较低,例如Filter Bank特征维数一般只有75维左右。因此,为了使目标音频特征和目标声学特征进行更好的拼接,可以在拼接前先对声学特征向量进行主成分分析(Principal Component Analysis,PCA),以将声学特征向量的有效维度降低在100维左右,从再将目标音频特征向量与降维后的目标声学特征向量进行首尾拼接,如图4a所示,从而加快训练速度。

[0055] S330、将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数。

[0056] 其中,将目标拼接特征向量输入预先训练好的语音端点检测模型中,该语音端点检测模型根据目标拼接特征向量中的目标音频特征向量将该待检测语音数据进行切割成至少一个语音片段,每个语音片段包括有效语音数据。然后根据目标拼接特征向量中的目标音频特征向量识别出每个语音片段中的每个说话人语音片段,得到每个语音片段中的说话人转折点。

[0057] 在一种可能的实现方式中,如图4b所示,上述S330,将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,包括以下步骤:

[0058] S41、将所述目标拼接特征向量输入所述语音端点检测模型,得到第一后验概率、第二后验概率和第三后验概率,所述第一后验概率为每帧中包括语音帧的概率,所述第二后验概率为每帧中包括非语音帧的概率,所述第三后验概率为每帧中包括转折帧的概率。

[0059] 在本申请实施例中,语音端点检测模型根据输入目标拼接特征向量进行逐帧三分类解码,即语音端点检测模型可对每一帧音频进行分类,分类的类型包括语音帧、非语音帧和转折帧。语音帧中为该帧中包括说话人的连续语音数据;非语音帧为该帧中不包括说话人语音数据,也可以称为噪音帧;在不同说话人的语音数据之间,一般会有一段比较短的静音帧,若两个不同说话人语音数据之间有一段静音帧,则将该静音帧和以及周围的部分语音帧为转折帧。例如,如图4c所示,第10帧-第60帧为一说话人的语音段,第61帧-第69帧为静音间隔,第70帧-第100帧为另一说话人的语音段,则将第51帧-第80帧确认为转折帧。示例性地,可以用0表示语音帧帧,用1表示非语音帧,用2表示转折帧。当然,本申请还可以使

用其他表示方法来表示音频的分类类型。

[0060] 其中,将目标拼接特征向量输入到语音端点检测模型,语音端点检测模型计算每一帧音频分类的后验概率,即分别计算每一帧音频的第一后验概率、第二后验概率和第三后验概率。所述第一后验概率、第二后验概率和第三后验概率的和为1,根据每一帧音频中第一后验概率、第二后验概率和第三后验概率的值可以确定该帧的类型,例如,若第一帧音频的第一后验概率为0.5、第二后验概率为0.23、第三后验概率为0.27,由于第一后验概率最大,即第一帧音频为语音帧的概率最大,因此将第一帧音频确定为语音帧。

[0061] 在本申请实施例中,该语音端点检测模型模型可以为深度神经网络模型,该深度神经网络模型例如可以是循环神经网络(Recurrent Neural Networks,RNN)。具体地,例如可以采用长短期记忆网络(longshort-termmemory,LSTM)模型,或者选通重复单元(Gated Recurrent Unit,GRU)模型,其中,GRU模型为LSTM模型的一种变体。针对LSTM模型,网络有多层RNN堆砌而成,最后的输出层是3个节点,softmax做为激活,采用交叉熵作为代价函数。每一帧都有分类类型的结果。

[0062] S42、根据所述第一后验概率和所述第二后验概率,确定所述待检测语音数据的所述至少一个语音片段。

[0063] 具体地,计算出待检测语音数据的每一帧音频的第一后验概率、第二后验概率和第三后验概率后,首先根据每一帧音频第一后验概率和第二后验概率的值判断每一帧中是否包括说话人语音数据,将第一后验概率大于第二后验概率的音频帧确定为语音帧,反之确认为非语音帧。然后根据语音帧和非语音帧,将待检测语音数据分割成至少一个语音片段,具体为:若连续M个音频帧均为非语音帧,且该M大于第一门限时,以该M个音频帧的两端为切割点进行切割,从而将待检测语音数据切割成包含说话人语音数据的至少一个语音片段,将一些无意义的非语音帧直接丢弃。

[0064] S43、根据所述第三后验概率和所述至少一个语音片段,确定所述待检测语音数据的所述N个转折点。

[0065] 其中,将待检测语音数据切割成至少一个语音片段后继续解析,根据语音片段中每一帧音频的第一后验概率和第三后验概率的值确定语音片段中转折帧,若语音片段中有连续P个音频帧均为转折帧,且该P大于第二门限时,则将该连续P个音频帧作为转折点,得到每个语音片段中的所有可能的转折点位置。

[0066] 需要说明的是,所述第一门限大于所述第二门限,所述第一门限和所述第二门限可以由语音识别系统进行设置,例如第二门限可以设置为10帧、15帧、20帧等;所述第一门限和所述第二门限也可以根据具体应用场景进行设置,本申请实施例对此不做限定。

[0067] 举例说明,一段语音数据如图4d中(a)所示,首先提取该语音数据的filterbank特征和声学特征,将filterbank特征和声学特征进行拼接,构建带声纹信息的拼接特征向量。然后将拼接特征向量输入到预先训练好的语音端点检测模型中,语音端点检测模型根据输入的拼接特征向量进行逐帧三分类解码。首先第一遍解码先用VAD解码策略,只解析语音数据中的语音帧和非语音帧,得到如图4d中(b)所示三个语音片段,其中第二个语音片段中有明显的说话人转折现象。然后在解析出的语音片段上继续解析,解析出语音数据中的转折帧,将满足预设阈值长度的取连续转折帧作为转折点,得到所有可能的转折点位置,如图4d中(c)中的A、B、C和D四个位置。

[0068] 在另一种可能的实现方式中,如图5所示,上述S330,将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,具体包括以下步骤:

[0069] S51、将所述目标拼接特征向量输入所述语音端点检测模型,得到第七后验概率和第八后验概率。

[0070] 在本申请实施例中,语音端点检测模型根据输入目标拼接特征向量进行逐帧解码,所述语音端点检测模型可以为Multi-task的实时深度学习网络。如图5a所示,所述语音端点检测模型包括共享层、第一任务层和第二任务层,所述共享层可以包括多层的卷积神经网络和循环神经网络,所述第一任务层用于执行VAD解码任务,所述第二任务层用于执行说话人转折点解码任务。

[0071] 具体地,语音端点检测模型可对每一帧音频进行分类,分类的类型包括转折帧和非转折帧。将目标拼接特征向量输入到语音端点检测模型的共享层后,所述共享层输出每一帧音频的包括说话人音频的后验概率和不包括说话人音频的后验概率,所述包括说话人音频的后验概率和不包括说话人音频的后验概率的和为1。

[0072] S52、根据所述第七后验概率,确定所述待检测语音数据的所述至少一个语音片段。

[0073] 其中,将得到每一帧的包括说话人音频的后验概率和不包括说话人音频的后验概率分别作为所述第一任务层和所述第二任务层的输入。所述第一任务层采用VAD解码策略,根据每一帧音频的说话人音频的后验概率,计算出每一帧音频数据中第七后验概率,所述第七后验概率为每帧中包括非转折帧的后验概率,从而根据第七后验概率的值解析出目标待检测语音数据中的语音帧和非语音帧。然后根据语音帧和非语音帧,将待检测语音数据分割成至少一个语音片段。所述待检测语音数据分割成至少一个语音片段的具体实现方式可参照上述描述,在此不再赘述。

[0074] S53、根据所述至少一个语音片段和所述第八后验概率,确定所述待检测语音数据的所述N个转折点。

[0075] 在本申请实施例中,第一任务层输出待检测语音数据的至少一个语音片段后,可将该至少一个语音片段作为第二任务层的输入。第二任务层根据该至少一个语音片段和每一帧的不包括说话人音频的后验概率,计算出每帧的第八后验概率,从而确定每个语音片段中转折点,得到每个语音片段中的所有可能的转折点位置。所述转折点的计算方式可参照上述描述,在此不再赘述。

[0076] 举例说明,一段语音数据如图4d中(a)所示,首先提取该语音数据的filterbank特征和声学特征,将filterbank特征和声学特征进行拼接,构建带声纹信息的拼接特征向量。然后将拼接特征向量输入到预先训练好的语音端点检测模型的共享层中,再将共享层的输出作为第一任务层的输入,经过第一任务层的VAD解码,得到如图4d中(b)所示三个语音片段,观察语音片段,整条音频确实被切割成三部分,但是由于传统VAD解码逻辑的缺陷,第二段语音段中有间隔较短的两个语音片段没有被切割开来,可能会导致语音识别框架的识别效果较差。最后将得到的三个语音片段和共享层的输出作为第二任务层的输入,经过说话人分离点检测任务解码后,得到如图4d中(c)中的A、B、C和D四个疑似的说话人转折点。

[0077] S340、基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至

少一个音频段。

[0078] 在本申请实施例中,在得到待检测语音数据的至少一个语音片段和N个转折点后,为了提高说话人分离的效果,还需要对所述N个转折点进行筛选,以去掉不合理的转折点,然后根据筛选后的转折点对所述至少一个语音片段进行分割,将待检测语音数据分割成至少一个音频段。

[0079] 可选的,所述基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段,包括:若所述N小于1,将所述至少一个语音片段确定为所述至少一个音频段;若所述N大于或等于1,对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段。

[0080] 其中,在经过语音端点检测模型后,若输出的转折点的数量N小于1,即待检测语音数据中未包括说话人转折点时,则直接将输出的语音片段确定为最后输出的音频段。当输出的转折点的数量N大于或等于1时,还需要根据过滤算法将这四个疑似的说话人转折点进行过滤,以过滤掉不合理的说话人转折点。

[0081] 可选的,所述对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段,包括:

[0082] 当所述N等于1,若所述转折点满足第一条件时,删除所述转折点,并将所述至少一个语音片段确定为所述至少一个音频段,否则,根据所述转折点将所述第一语音片段进行分割,得到多个音频段,所述第一条件为所述转折点与第一语音片段边界的距离小于第一阈值,或所述转折点位于有效音频数据上,所述第一语音片段为所述转折点所在的语音片段;在所述N大于1时,若任一转折点满足所述第一条件和第二条件时,删除所述第一转折点,否则保留所述转折点,并根据所述转折点对所述至少一个语音片段进行分割,得到所述多个音频段,所述第二条件为所述转折点与目标转折点的距离小于所述第一阈值,所述目标转折点为保留的转折点。

[0083] 具体地,N等于1时,若该转折点与任一语音片段边界的距离小于第一阈值或者该转折点位于语音帧上,则将该转折点删除,将输出的语音片段确定为最后输出的音频段;否则保留所述转折点,并根据所述转折点将语音片段进行分割,得到多个音频段。N大于1是,若转折点与任一语音片段边界的距离小于第一阈值,或者该转折点位于语音帧上,或者该转折点与保留下来的转折点的距离小于所述第一阈值时,将该转折点删除;否则保留该转折点,该转折点为N个转折点中的任一转折点;最后根据保留下来的转折点,将至少一个语音片段分割成多个音频段。例如,如图4d中(c)所示,转折点B不再语音片段中,转折点D位于语音帧上,因此将转折点B和转折点D删除,保留转折点A和转折点C,其中转折点A会将第一个语音片段切割成两个音频段,转折点C会将第二个语音片段切割成两个音频段,最终得到如图4d中(d)所示的五段音频段。

[0084] 可以看出,在本申请实施例提供的语音端点检测方法,通过将融合语音数据的音频特征和声学特征输入到语音端点检测模型中,得到语音片段和转折点,根据转折点和语音片段对语音数据进行分割,可以将包括多人说话的语音片段分割成多个单说话人的音频段,实现在多人快速交换讨论的场景下,将传统VAD模块切不开的多人转换语音片段区分开来,以提高多人讨论场景下语音端点检测的准确性。

[0085] 下面采用具体的实施例来说明本申请实施例通过语音端点检测过程。

[0086] 请参阅图6,图6为本申请实施例提供的一种语音端点检测方法的流程示意图。如图6所示,该方法包括如下步骤。

[0087] S61、获取训练数据集,所述训练数据集包括多个说话人的多条训练数据。

[0088] 其中,所述训练数据集可以是收集会议场景下的音频数据,一条音频数据可以包括多个说话人的语音数据,所述训练数据集也可以是收集多人聊天场景下的音频数据,还可以是电视直播场景、新闻节目场景下的音频数据,本申请实施例对此不做限定。

[0089] 可选的,所述方法还包括:获取多条携带标注信息的原始语音数据,所述标注信息包括每条原始语音数据中的至少一个说话人、每个所述说话人的起始点和结束点;根据所述原始语音数据中每个所述说话人的起始端点和结束端点,将每条所述原始语音数据分割成至少一条子语音数据;将所述多条原始语音数据的至少一条子语音数据拼接成多条样本数据;标注所述每条样本数据中的每一帧音频的分类信息,得到所述多条训练数据。

[0090] 其中,对于每一条的会议音频数据,用人工标注或者盲分声纹识别的方法将音频中不同说话人的身份以及每一个语音段的说话人身份信息标记出来,得到多条的携带说话人标注信息的原始语音数据。所述原始语音数据中可以只包括一个说话人的语音数据,也可以包括多个说话人的语音数据。

[0091] 对于只包括一个说话人的原始语音数据,可以直接将这些原始语音数据进行拼接,得到包括说话转折点的样本数据。所述原始语音数据的拼接方法可以按照自然语音顺承关系进行拼接,即根据多人对话时的对话顺序;也可以随机拼接。本申请实施例对比不做限定。

[0092] 对于包括多个说话人的原始语音数据,首先需要将该原始语音数据按照说话人说的每一段话将其分割成子句,具体为:根据人工标注的每个不同说话人的一句话的起始点和结束点,将原始语音数据切割成至少一条子语音数据,每条子语音数据为一个说话人的语音数据。然后选取不同说话人的具有前后转折关系的子语音数据来拼接,得到包括说话转折点的样本数据,比如上一条子语音数据是A说的,下一条子语音数据是B说的,则将这两条子语音数据拼接在一起作为一段典型的说话人转折的样本数据。同时也可以适当拼接一些上一条和下一条子语音数据均是同一个人说的子语音数据,作为训练反例,提高训练的鲁棒性。

[0093] 在本申请实施例中,拼接得到多条样本数据后,需要对样本数据进行标注,得到携带标注信息的训练数据,其标注信息用于指示待训练的每帧音频的分类类型的结果,即每帧音频被分类为转折帧还是非转折帧。示例性地,每帧音频的分类类型还可以包括常规语音帧、静音帧、转折开始帧和转折结束帧等,本申请实施例采用转折帧还是非转折帧的分类方式。如图6a所示,在两句拼接子语音数据之间,会有一段比较短的静音段(也可以人为拉长和缩短此静音段),该静音段用于指示不同说话人的转折帧。但是由于两条子语音数据之间的静音间隙周围的语音段内容也包含了转折信息,例如两个不同说话人的声纹特征等,因此在标注转折帧时,可将两条子语音数据中间的静音帧,以及周围的部分语音帧的内容也包含进去。其余部分则标记为非转折状态。示例性地,可以用1表示转折帧,用0表示非转折帧。当然,本申请还可以使用其他表示方法来表示音频的分类类型。

[0094] S62、提取所述多条训练数据中的每条训练数据的声学特征向量和音频特征向量,并将所述每条训练数据的所述音频特征向量和所述声学特征向量进行拼接,得到所述每条

训练数据拼接特征向量。

[0095] 其中,所述训练数据的特征提取、特征拼接的具体实现方式可参照待检测语音数据的特征提取、特征拼接的的具体实现方式,在此不再赘述。

[0096] S63、将所述每条训练数据的拼接特征向量输入待训练语音端点检测模型进行训练,直至达到训练结束条件,得到所述语音端点检测模型。

[0097] 其中,得到了拼接特征向量之后,就可以进入模型的训练步骤。语音端点检测模型的结构如图5a所示,在得到训练数据的拼接特征向量以及训练数据对应的标注信息后,对待训练语音端点检测模型进行训练,得到训练后的语音端点检测模型。

[0098] 可选的,所述将所述每条训练数据的拼接特征向量输入待训练语音端点检测模型,直至达到训练结束条件,得到所述语音端点检测模型,包括:

[0099] 将所述每条训练数据的拼接特征向量输入所述待训练语音端点检测模型的共享层,得到所述每条训练数据中的每一帧音频的第四后验概率,所述共享层包括多个神经网络模型,所述第四后验概率为每帧音频中包括说话人的概率;将所述每一帧音频的第四后验概率分别输入所述待训练语音端点检测模型的第一任务层和第二任务层,分别得到第五后验概率和第六后验概率,所述第五后验概率为每帧中包括非转折帧的概率,所述第六后验概率为每帧中包括转折帧的概率;基于所述第五后验概率计算第一梯度,基于所述第六后验概率计算第二梯度,所述第一梯度为所述第一任务层的梯度,所述第二梯度为所述第二任务层的梯度;根据所述第一梯度和所述第二梯度更新所述待训练语音端点检测模型的参数,直至达到训练结束条件,得到所述语音端点检测模型。

[0100] 具体地,第一任务层和第二任务层分别与语音端点检测模型中的共享层相连接,第一任务层和第二任务层输入共享层输出的后验概率后,各自输出对应的检测结果。第一任务层对共享层输出的后验概率进行VAD解码后,输出第五后验概率,根据第五概率的值可以确定对应帧是否为非转折帧;第二任务层对共享层输出的后验概率进行说话人分离点检测任务解码后,输出第六后验概率,根据第六概率的值可以确定对应帧是否为转折帧。然后根据每条每条训练数据的标注信息,分别计算第一任务层的第一梯度和第二任务层的第二梯度。

[0101] 其中,梯度loss值可以用于反向更新模型权重,所述梯度loss的计算公式为:$Loss_i=(y_i-f(x_i))^2$,其中y_i为每帧对应的标注信息,$f(x_i)$表示当前输入音频x_i计算得到的后验概率。因此根据梯度loss值的计算公式可分别计算出第一梯度$Loss_1$和第二梯度$Loss_2$。再将第一梯度$Loss_1$和第二梯度$Loss_2$分别乘以各自的梯度更新权重$Loss_weight_1$和$Loss_weight_2$形成最终的梯度$Loss_i$,即$Loss_weight_1*Loss_1+Loss_weight_2*Loss_2$。根据最终的梯度$Loss_i$对待训练的语音端点检测模型的参数进行更新,直到最终的梯度$Loss_i$收敛,即根据待训练的语音端点检测模型输出的第五后验概率和第六后验概率与标注信息的差别很小。从而得到训练后的语音端点检测模型。

[0102] 本实施例提供的语音端点检测方法,通过融合语音数据的音频特征和声学特征提取更丰富的语音抽象表征,采用多任务训练的方式训练具有时序性的深度神经网络,同时训练语音端点检测任务和说话人转折点检测任务,使得语音端点检测模型在多人会议讨论场景下将传统VAD模块切不开的多人转换语音片段区分开来,从而提高语音端点检测的准确性。

[0103] 上述主要从方法侧执行过程的角度对本申请实施例的方案进行了介绍。可以理解的是,终端设备为了实现上述功能,其包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到,结合本文中所提供的实施例描述的各示例的单元及算法步骤,本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0104] 本申请实施例可以根据上述方法示例对终端设备进行功能单元的划分,例如,可以对应各个功能划分各个功能单元,也可以将两个或两个以上的功能集成在一个处理单元中。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。需要说明的是,本申请实施例中对单元的划分是示意性的,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。

[0105] 请参阅图7a,图7a是本申请实施例提供的一种语音端点检测装置的功能单元组成框图,所述装置700包括:获取单元710、提取单元720、检测单元730和分割单元730,其中,

[0106] 获取单元710,用于获取待检测语音数据;

[0107] 提取单元720,用于提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

[0108] 检测单元730,用于将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数;

[0109] 分割单元740,用于基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段。

[0110] 可选的,所述装置700还包括训练单元750,其中,

[0111] 所述训练单元750,用于获取训练数据集,所述训练数据集包括多个说话人的多条训练数据;提取所述多条训练数据中的每条训练数据的声学特征向量和音频特征向量,并将所述每条训练数据的所述音频特征向量和所述声学特征向量进行拼接,得到所述每条训练数据拼接特征向量;将所述每条训练数据的拼接特征向量输入待训练语音端点检测模型进行训练,直至达到训练结束条件,得到所述语音端点检测模型。

[0112] 可选的,所述装置700还包括拼接单元760和标注单元770,其中,

[0113] 所述获取单元710,还包括:获取多条携带标注信息的原始语音数据,所述标注信息包括每条原始语音数据中的至少一个说话人、每个所述说话人的起始点和结束点;

[0114] 所述分割单元740,还用于根据所述原始语音数据中每个所述说话人的起始端点和结束端点,将每条所述原始语音数据分割成至少一条子语音数据;

[0115] 所述拼接单元760,用于将所述多条原始语音数据的至少一条子语音数据拼接成多条样本数据;

[0116] 所述标注单元770,用于标注所述每条样本数据中的每一帧音频的分类信息,得到所述多条训练数据。

[0117] 可选的,所述分类信息包括语音帧、非语音帧和转折帧;

[0118] 所述检测单元730具体用于:将所述目标拼接特征向量输入所述语音端点检测模

型,得到第一后验概率、第二后验概率和第三后验概率,所述第一后验概率为每帧中包括语音帧的概率,所述第二后验概率为每帧中包括非语音帧的概率,所述第三后验概率为每帧中包括转折帧的概率;根据所述第一后验概率和所述第二后验概率,确定所述待检测语音数据的所述至少一个语音片段;根据所述第三后验概率和所述至少一个语音片段,确定所述待检测语音数据的所述N个转折点。

[0119] 可选的,所述分类信息包括转折帧和非转折帧;

[0120] 如图7b所示,是本申请实施例提供的另一种语音端点检测装置700的功能单元组成框图,所述训练单元750具体用于:将所述每条训练数据的拼接特征向量输入所述待训练语音端点检测模型的共享层,得到所述每条训练数据中的每一帧音频的第四后验概率,所述共享层包括多个神经网络模型,所述第四后验概率为每帧音频中包括说话人的概率;将所述每一帧音频的第四后验概率分别输入所述待训练语音端点检测模型的第一任务层和第二任务层,分别得到第五后验概率和第六后验概率,所述第五后验概率为每帧中包括非转折帧的概率,所述第六后验概率为每帧中包括转折帧的概率;基于所述第五后验概率计算第一梯度,基于所述第六后验概率计算第二梯度,所述第一梯度为所述第一任务层的梯度,所述第二梯度为所述第二任务层的梯度;根据所述第一梯度和所述第二梯度更新所述待训练语音端点检测模型的参数,直至达到训练结束条件,得到所述语音端点检测模型。

[0121] 可选的,所述检测单元730具体用于:将所述目标拼接特征向量输入所述语音端点检测模型,得到第七后验概率和第八后验概率;根据所述第七后验概率,确定所述待检测语音数据的所述至少一个语音片段;根据所述至少一个语音片段和所述第八后验概率,确定所述待检测语音数据的所述N个转折点。

[0122] 可选的,所述分割单元740具体用于:若所述N小于1,将所述至少一个语音片段确定为所述至少一个音频段;若所述N大于或等于1,对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段。

[0123] 可选的,在对所述N个转折点进行过滤,根据过滤后的转折点对所述至少一个语音片段进行分割,得到所述至少一个音频段方面,所述分割单元740具体用于:

[0124] 当所述N等于1,若所述转折点满足第一条件时,删除所述转折点,并将所述至少一个语音片段确定为所述至少一个音频段,否则,根据所述转折点将所述第一语音片段进行分割,得到多个音频段,所述第一条件为所述转折点与第一语音片段边界的距离小于第一阈值,或所述转折点位于有效音频数据上,所述第一语音片段为所述转折点所在的语音片段;在所述N大于1时,若任一转折点满足所述第一条件和第二条件时,删除所述第一转折点,否则保留所述转折点,并根据所述转折点对所述至少一个语音片段进行分割,得到所述多个音频段,所述第二条件为所述转折点与目标转折点的距离小于所述第一阈值,所述目标转折点为保留的转折点。

[0125] 可以理解的是,本申请实施例的语音端点检测装置的各程序模块的功能可根据上述方法实施例中的方法具体实现,其具体实现过程可以参照上述方法实施例的相关描述,此处不再赘述。

[0126] 请参阅图8,图8是本申请实施例提供的一种计算机设备,该计算机设备包括:处理器、存储器、收发器,以及一个或多个程序。所述处理器、存储器和收发器通过通信总线相互连接。

[0127] 处理器可以是一个或多个中央处理器(central processing unit,CPU),在处理器是一个CPU的情况下,该CPU可以是单核CPU,也可以是多核CPU。

[0128] 所述一个或多个程序被存储在所述存储器中,并且被配置由所述处理器执行;所述程序包括用于执行以下步骤的指令:

[0129] 获取待检测语音数据;

[0130] 提取所述待检测语音数据的目标声学特征向量和目标音频特征向量,并将所述目标音频特征向量和所述目标声学特征向量进行拼接,得到目标拼接特征向量;

[0131] 将所述目标拼接特征向量输入语音端点检测模型,得到所述待检测语音数据的至少一个语音片段和N个转折点,所述N为整数;

[0132] 基于所述至少一个语音片段和所述N个转折点,将待检测语音数据分割成至少一个音频段。

[0133] 需要说明的是,本申请实施例的具体实现过程可参见上述方法实施例所述的具体实现过程,在此不再赘述。

[0134] 本申请实施例还提供一种计算机存储介质,其中,该计算机存储介质存储用于电子数据交换的计算机程序,该计算机程序使得计算机执行如上述方法实施例中记载的任一方法的部分或全部步骤。

[0135] 本申请实施例还提供一种计算机程序产品,上述计算机程序产品包括存储了计算机程序的非瞬时性计算机可读存储介质,上述计算机程序可操作来使计算机执行如上述方法实施例中记载的任一方法的部分或全部步骤。该计算机程序产品可以为一个软件安装包。

[0136] 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,本申请并不受所描述的动作顺序的限制,因为依据本申请,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申请所必须的。

[0137] 在上述实施例中,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。

[0138] 在本申请所提供的几个实施例中,应该理解到,所揭露的装置,可通过其它的方式实现。例如,以上所描述的装置实施例仅是示意性的,例如上述单元的划分,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,装置或单元的间接耦合或通信连接,可以是电性或其它的形式。

[0139] 上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申请实施例方案的目的。

[0140] 另外,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。

[0141] 上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器中。基于这样的理解,本申请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器中,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)执行本申请各个实施例方法的全部或部分步骤。而前述的存储器包括:U盘、ROM、RAM、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0142] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成,该程序可以存储于一计算机可读存储器中,存储器可以包括:闪存盘、ROM、RAM、磁盘或光盘等。

[0143] 以上对本申请实施例进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

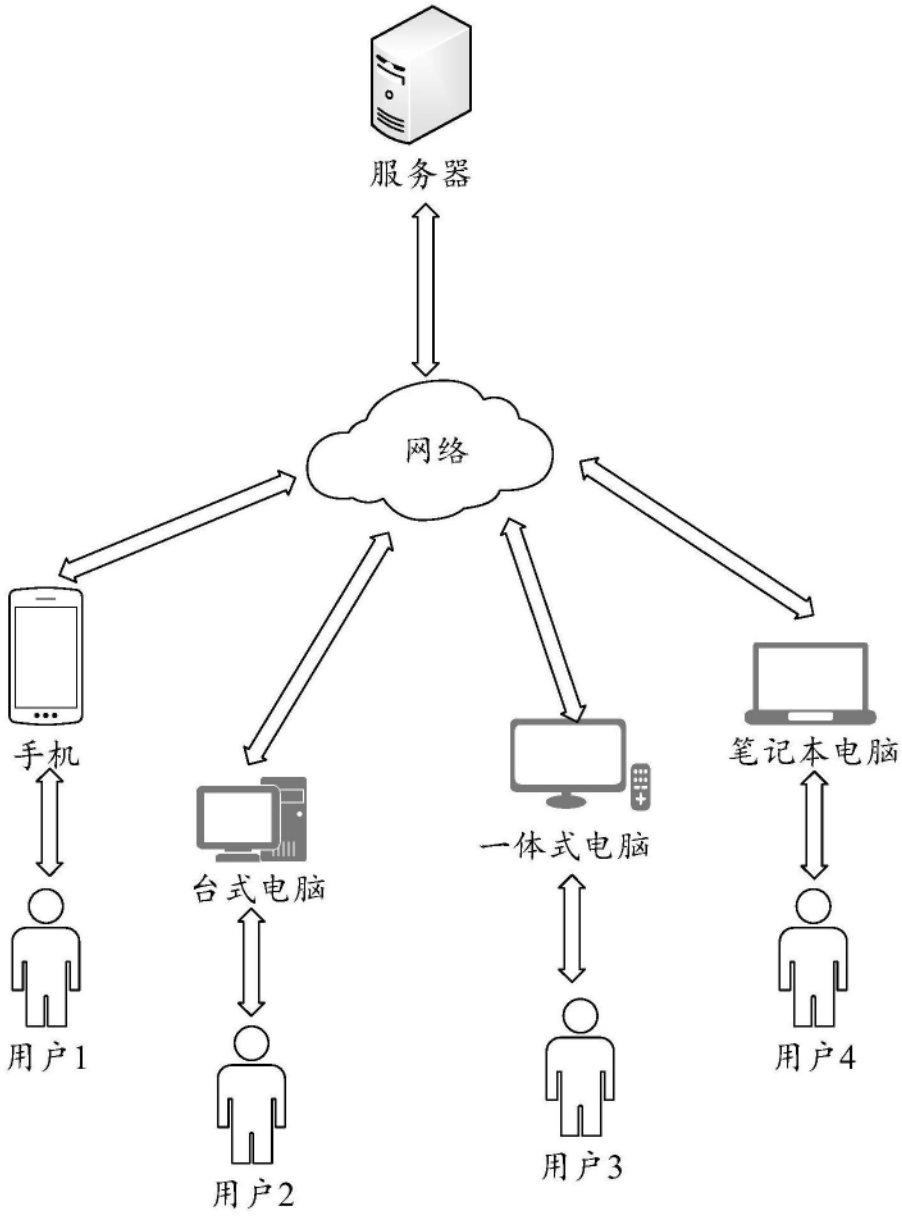

服务器
网络
手机
台式电脑
一体式电脑
笔记本电脑
用户1
用户2
用户3
用户4

图1

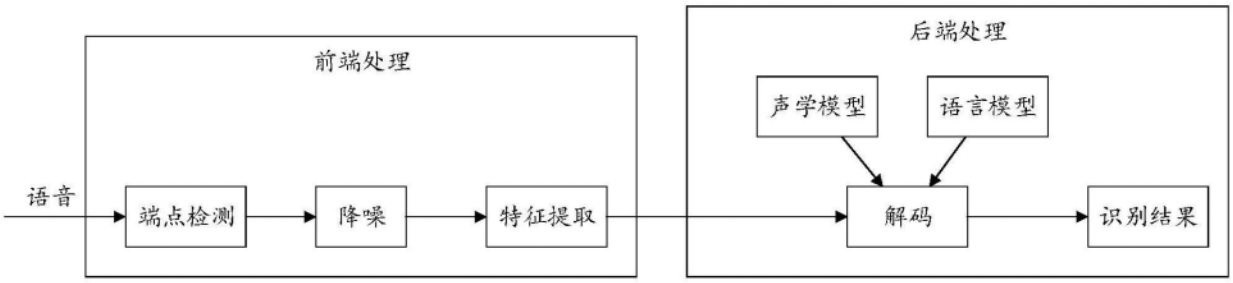

前端处理
后端处理
声学模型
语言模型
语音
端点检测
降噪
特征提取
解码
识别结果

图2

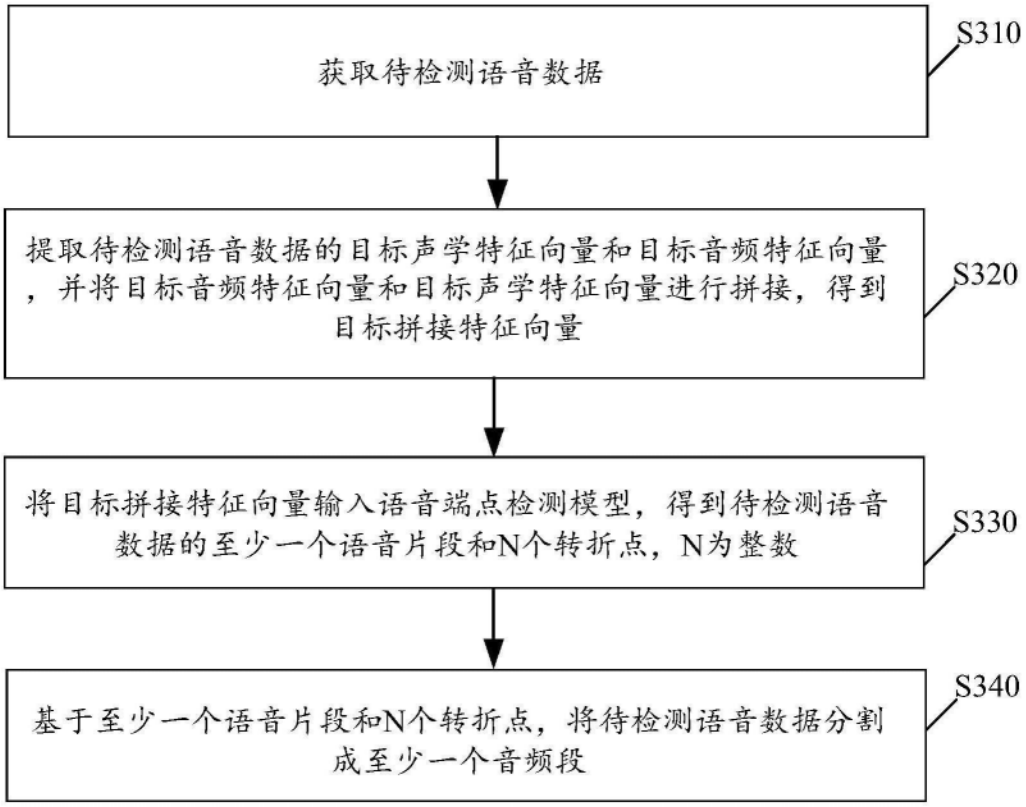

获取待检测语音数据
S310
提取待检测语音数据的目标声学特征向量和目标音频特征向量，并将目标音频特征向量和目标声学特征向量进行拼接，得到目标拼接特征向量
S320
将目标拼接特征向量输入语音端点检测模型，得到待检测语音数据的至少一个语音片段和N个转折点，N为整数
S330
基于至少一个语音片段和N个转折点，将待检测语音数据分割成至少一个音频段
S340

图3


75维FilterBank特征
100维主成分分析降维后的iVector特征

图4a

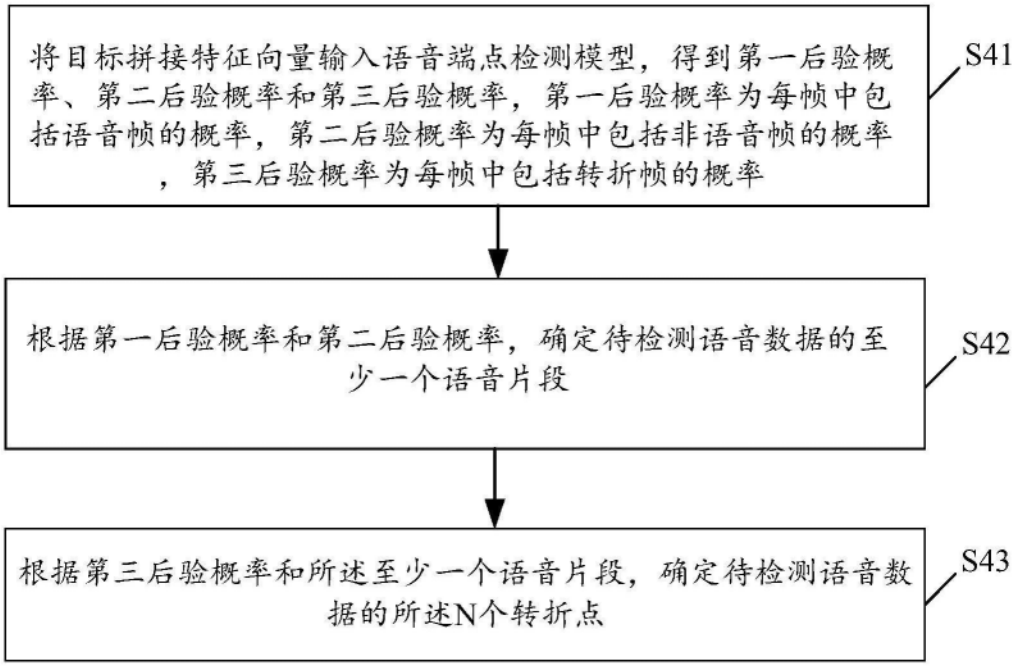

将目标拼接特征向量输入语音端点检测模型，得到第一后验概率、第二后验概率和第三后验概率，第一后验概率为每帧中包括语音帧的概率，第二后验概率为每帧中包括非语音帧的概率，第三后验概率为每帧中包括转折帧的概率
S41
根据第一后验概率和第二后验概率，确定待检测语音数据的至少一个语音片段
S42
根据第三后验概率和所述至少一个语音片段，确定待检测语音数据的所述N个转折点
S43

图4b

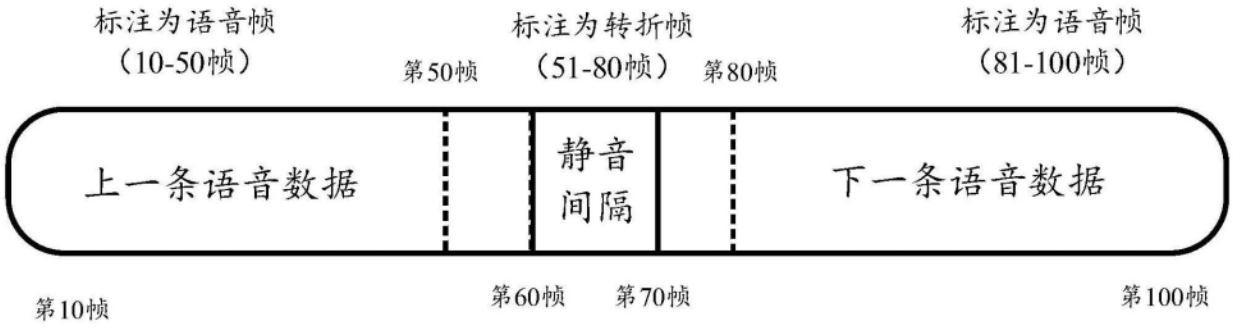

标注为语音帧
（10-50帧）
标注为转折帧
（51-80帧）
标注为语音帧
（81-100帧）
第50帧
第80帧
上一条语音数据
静音间隔
下一条语音数据
第10帧
第60帧
第70帧
第100帧

图4c

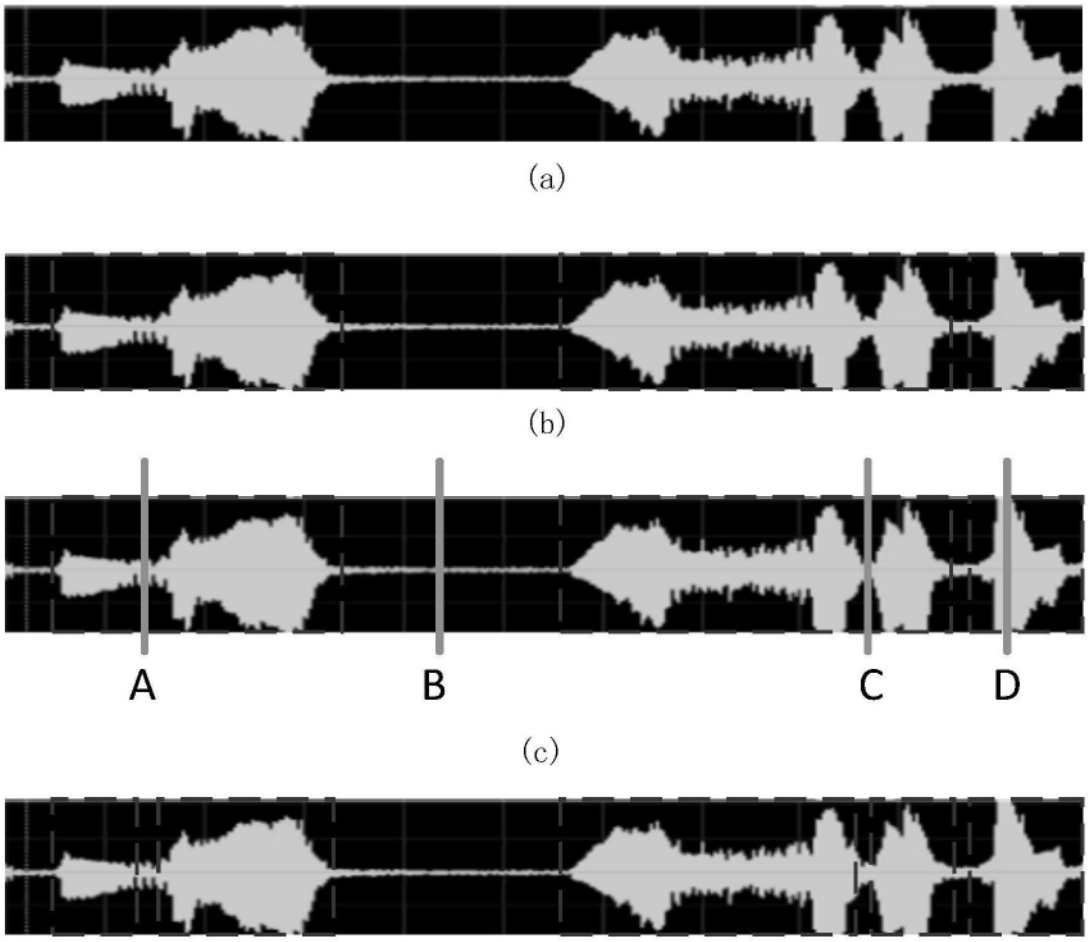


图4d

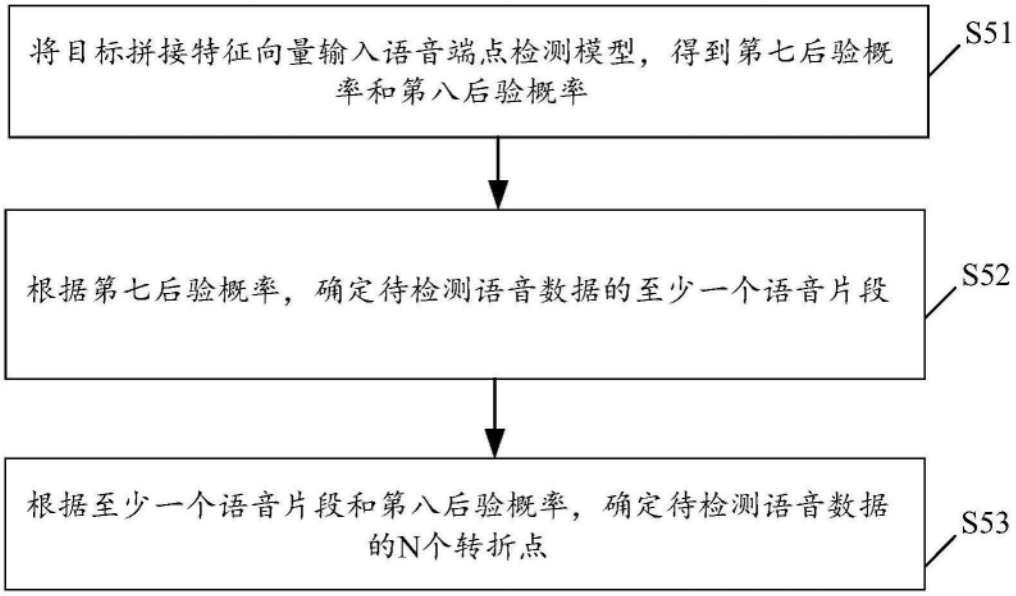

将目标拼接特征向量输入语音端点检测模型，得到第七后验概率和第八后验概率
S51
根据第七后验概率，确定待检测语音数据的至少一个语音片段
S52
根据至少一个语音片段和第八后验概率，确定待检测语音数据的N个转折点
S53

图5

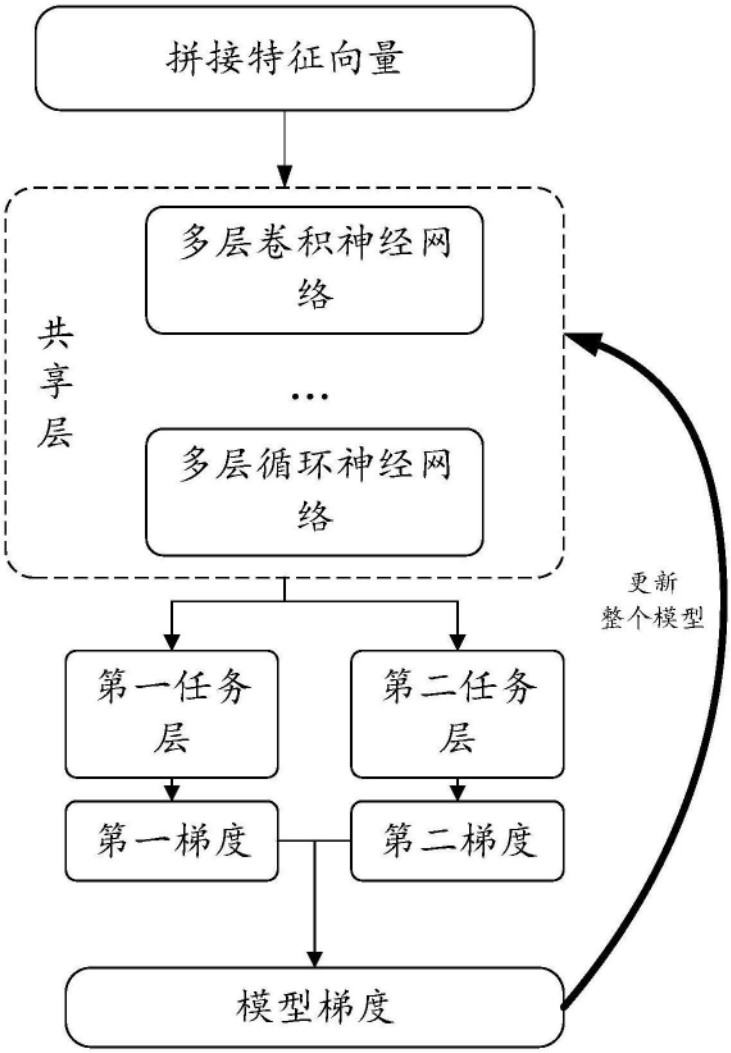

拼接特征向量
多层卷积神经网络
共享层
...
多层循环神经网络
更新整个模型
第一任务层
第二任务层
第一梯度
第二梯度
模型梯度

图5a

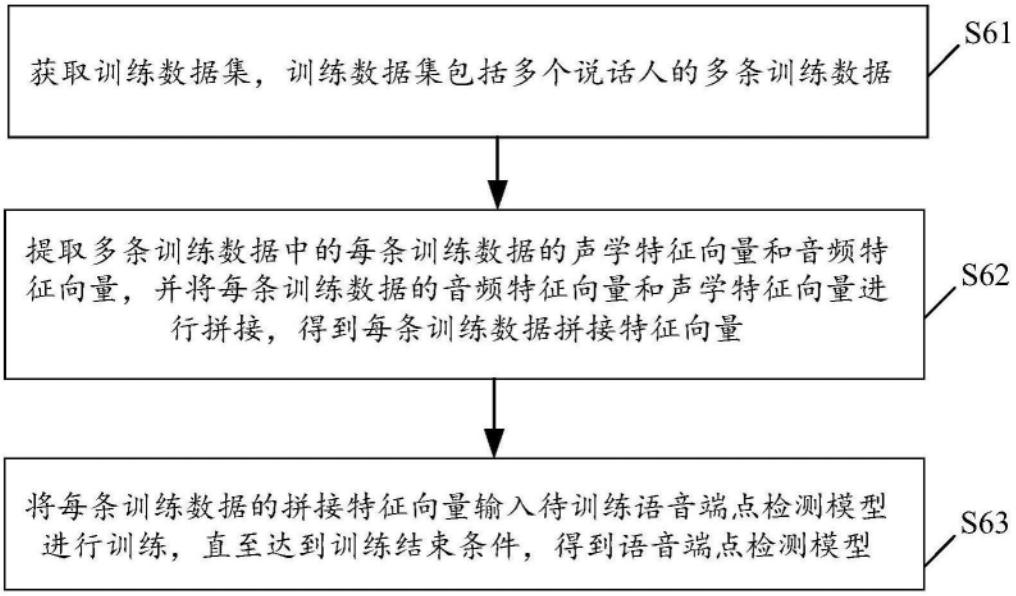

获取训练数据集，训练数据集包括多个说话人的多条训练数据
S61
提取多条训练数据中的每条训练数据的声学特征向量和音频特征向量，并将每条训练数据的音频特征向量和声学特征向量进行拼接，得到每条训练数据拼接特征向量
S62
将每条训练数据的拼接特征向量输入待训练语音端点检测模型进行训练，直至达到训练结束条件，得到语音端点检测模型
S63

图6

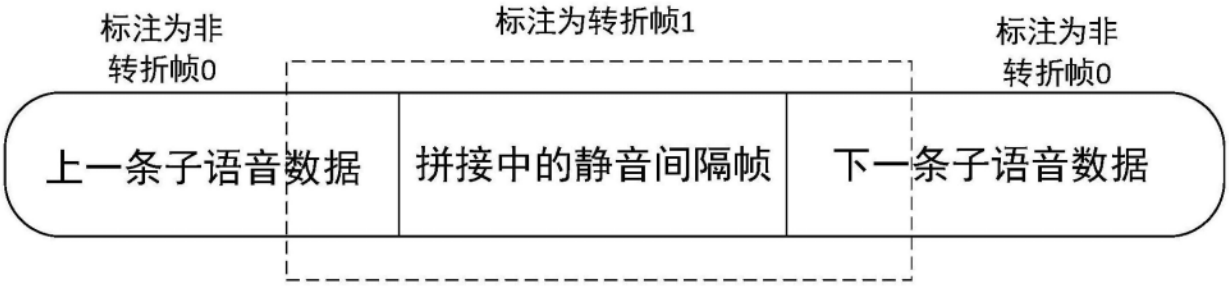

标注为非转折帧0
标注为转折帧1
标注为非转折帧0
上一条子语音数据
拼接中的静音间隔帧
下一条子语音数据

图6a

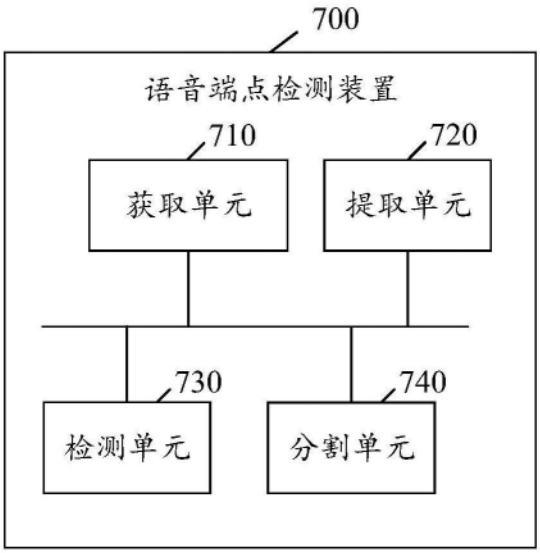

700
语音端点检测装置
710
获取单元
720
提取单元
730
检测单元
740
分割单元

图7a

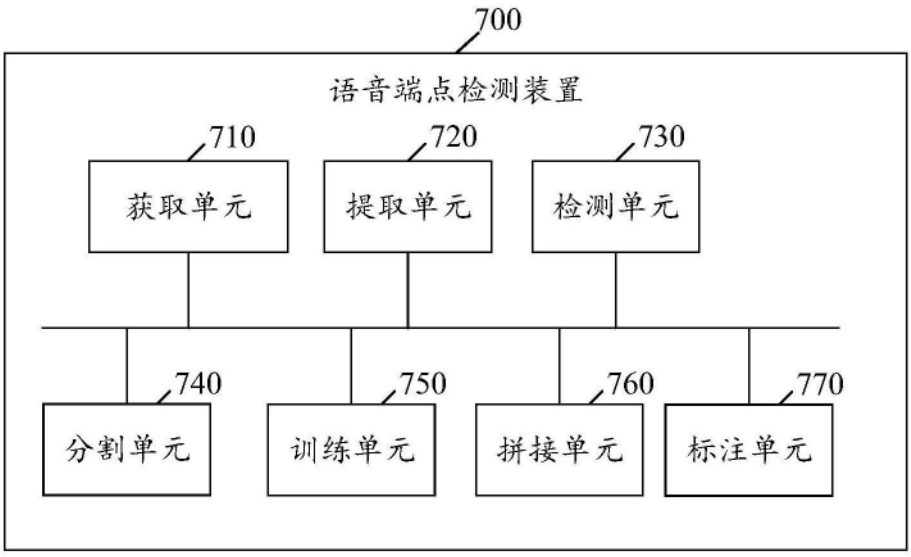

700
语音端点检测装置
710
获取单元
720
提取单元
730
检测单元
740
分割单元
750
训练单元
760
拼接单元
770
标注单元

图7b

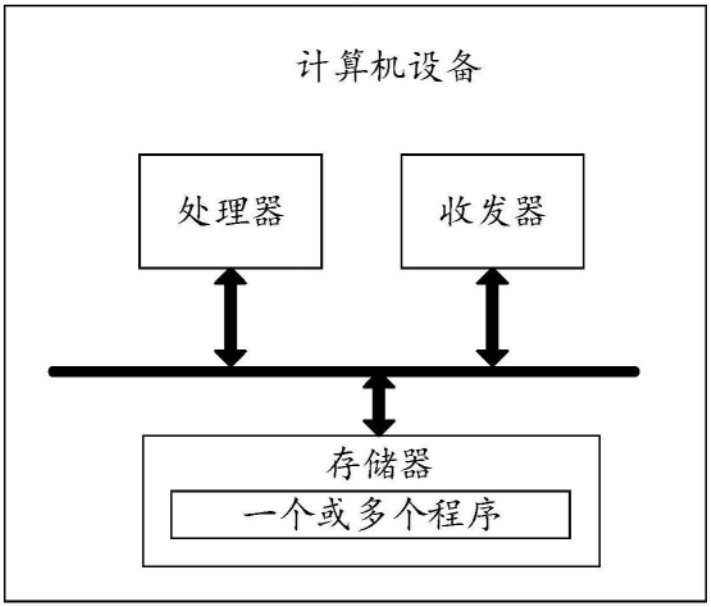


图8