(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 113409768 A
(43)申请公布日 2021.09.17

(21)申请号 202011119857.7

(22)申请日 2020.10.19

(71)申请人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层
申请人 北京语言大学

(72)发明人 付凯奇 林炳怀 张劲松 解焱陆 冯晓莉 王丽园

(74)专利代理机构 深圳市隆天联鼎知识产权代理有限公司 44232
代理人 叶虹

(51)Int.Cl.
G10L 15/02(2006.01)
G10L 25/78(2013.01)
G10L 25/51(2013.01)
G10L 25/60(2013.01)

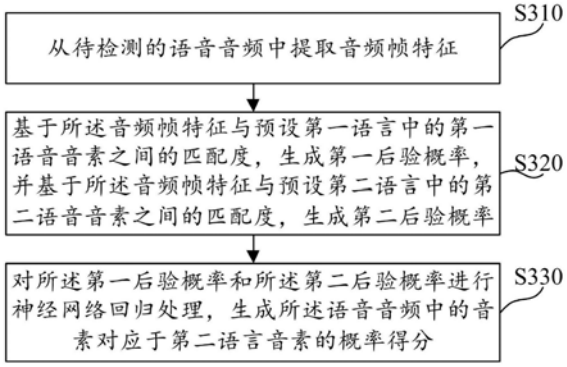
权利要求书3页 说明书17页 附图11页

(54)发明名称

发音检测方法、装置及计算机可读介质

(57)摘要

本申请的实施例基于人工智能中的语音技术和机器学习方法，提供了一种发音检测方法、装置及计算机可读介质。该发音检测方法包括：从待检测的语音音频中提取音频帧特征；基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率；对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分。本申请实施例的技术方案可以得到精确的发音检测结果，提高发音检测的精确性和发音者的练习效率。



CN 113409768 A

1.一种发音检测方法,其特征在于,包括:

从待检测的语音音频中提取音频帧特征;

基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度,生成第一后验概率,并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度,生成第二后验概率;

对所述第一后验概率和所述第二后验概率进行神经网络回归处理,生成所述语音音频中的音素对应于第二语言音素的概率得分。

2.根据权利要求1所述的方法,其特征在于,从待检测的语音音频中提取音频帧特征,包括:

对所述语音音频进行信号增强处理,生成增强语音;

基于设定帧长对所述增强语音进行分帧处理,生成语音序列;

基于设定窗口长度对所述语音序列进行加窗处理,生成加窗语音序列;

对所述加窗语音序列进行傅里叶变换,生成频域语音信号;

对所述频域语音信号进行滤波处理,生成所述音频帧特征。

3.根据权利要求2所述的方法,其特征在于,对所述语音音频进行信号增强处理,生成增强语音,包括:

获取所述语音音频中第一时刻对应的第一信号、所述第一时刻之前的第二时刻对应的第二信号;

基于设定的信号系数和所述第二信号,计算所述第二信号对应的加权信号;

基于所述第一信号强度与所述加权信号之间的差值,生成所述第一时刻对应的增强信号;

将所述语音音频中各时刻对应的增强信号进行组合,得到所述增强语音。

4.根据权利要求1所述的方法,其特征在于,基于所述音频帧特征与预设的第一语音音素之间的匹配度,生成第一后验概率,包括:

将所述音频帧特征输入基于第一语言样本训练得到的第一声学模型,输出所述音频帧特征与所述第一语音音素的匹配度对应的第一后验概率;

基于所述语音音频中各音素对应的波形,识别所述音素对应的始末时刻;

基于所述音素对应的始末时刻和所述音频帧特征对应的时间帧信息,确定所述音素中包含的音频帧特征;

对所述音素中包含的音频帧特征对应的第一后验概率进行均值计算,生成所述音素对应于所述第一语言音素的第一后验概率。

5.根据权利要求4所述的方法,其特征在于,将所述音频帧特征输入基于第一语言样本训练得到的第一声学模型,输出所述音频帧特征与所述第一语音音素的匹配度对应的第一后验概率之前,还包括:

获取基于第一语言生成的第一语音样本、以及所述第一语音样本对应的第一语音文本,并获取基于第二语言生成的第二语音样本、以及所述第二语音样本对应的第二语音文本;

基于时延神经网络构建用于识别音频中所包含音素的声学模型;

将所述第一语音样本输入所述声学模型中,并基于输出的第一音素与所述第一语音文

本得到的第一损失函数，对所述声学模型的参数进行调整，得到所述第一声学模型；

将所述第二语音样本输入所述声学模型中，并基于输出的第二音素与所述第二语音文本得到的第二损失函数，对所述声学模型的参数进行调整，得到第二声学模型。

6.根据权利要求1所述的方法，其特征在于，基于所述音频帧特征与预设的第二语音音素之间的匹配度，生成第二后验概率，包括：

将所述音频帧特征输入基于第二语言样本训练得到的第二声学模型，输出所述音频帧特征与所述第二语音音素的匹配度对应的第二后验概率；

基于所述语音音频进行识别，确定所述音素对应的始末时刻；

基于所述语音音频的波形，识别所述音素对应的始末时刻；

基于所述音素对应的始末时刻，对所述音素中各所述音频帧特征对应的第二后验概率进行均值计算，确定所述音素对应于所述第二语言音素的第二后验概率。

7.根据权利要求1所述的方法，其特征在于，对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分，包括：

对所述第一后验概率和所述第二后验概率进行拼接，得到概率特征；

对所述概率特征进行神经网络回归处理，生成所述语音音频中的音素对应于所述第二语言音素的概率得分。

8.根据权利要求1所述的方法，其特征在于，基于所述概率得分确定各音素对应的发音准确等级，包括：

基于所述音素对应于第二语言音素的概率得分，确定所述音素与所述第二语言音素之间的置信度；

基于所述置信度与设定的置信度阈值，确定所述语音音频中各音素对应的发音准确等级。

9.根据权利要求8所述的方法，其特征在于，基于所述音素对应于第二语言音素的概率得分，确定所述音素与所述第二语言音素之间的置信度，包括：

从所述音素对应于第二语言音素的概率得分中，确定最大概率得分；

计算指定音素对应于所述第二语言音素的概率得分与所述最大概率得分之间的比值；

基于所述比值确定所述指定音素与所述第二语言音素之间的置信度。

10.根据权利要求1所述的方法，其特征在于，对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分之后，还包括：

基于所述概率得分确定各音素对应的发音准确等级，并基于所述发音准确等级对应的显示方式显示所述音素对应的文本。

11.根据权利要求10所述的方法，其特征在于，基于所述发音准确等级对应的显示方式显示所述音素对应的文本，包括：

获取所述语音音频对应的文本；

基于所述语音音频中的音素，对所述文本进行切词，生成各音素对应的文本；

基于各音素对应的发音准确等级，通过所述发音准确等级对应的显示方式显示所述音素对应的文本。

12.根据权利要求10所述的方法，其特征在于，基于所述概率得分确定各音素对应的发音准确等级，并基于所述发音准确等级对应的显示方式显示所述音素对应的文本之后，所述方法还包括：

从各所述音素对应的发音准确等级中，查询发音准确等级最低的目标音素；

获取所述目标音素对应的发音示教信息，其中，所述发音示教信息包括以下信息中的至少一个：音标文本、正确读法以及示意视频；

展示所述发音示教信息。

13.根据权利要求12所述的方法，其特征在于，从各所述音素对应的发音准确等级中，查询发音准确等级最低的目标音素之后，所述方法还包括：

从所述第二语言对应的词句库中获取包含所述目标音素的目标词句；

展示所述目标词句；

获取用户基于所述目标词句发送的练习音频；

对所述练习音频进行检测，得到所述目标音频对应的发音准确等级。

14.一种发音检测装置，其特征在于，包括：

提取单元，用于从待检测的语音音频中提取音频帧特征；

概率单元，用于基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率；

得分单元，用于对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分；

显示单元，用于基于所述概率得分确定各音素对应的发音准确等级，并基于所述发音准确等级对应的显示方式显示所述音素对应的文本。

15.一种计算机可读介质，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13中任一项所述的发音检测方法。

发音检测方法、装置及计算机可读介质

技术领域

[0001] 本申请涉及计算机技术领域，具体而言，涉及一种发音检测方法、装置及计算机可读介质。

背景技术

[0002] 在应用于教育的很多语言学习软件中，都是通过获取用户发出的语音，来进行识别，以判断用户的发音水准，并在发错音或者发不准的时候执行对应的教学。但是在很多情况下，相关技术中的识别方式仅仅是针对语音的音素来识别当前语音的发音情况，并未考虑到用户的用语习惯和水准等信息，而造成发音检测结果不够客观、不精确的问题，进而可能影响学习者的学习效率和积极性。

发明内容

[0003] 本申请的实施例提供了一种发音检测方法、装置及计算机可读介质，进而至少在一定程度上可以得到精确的发音检测结果，提高发音检测的精确性和发音者的练习效率。

[0004] 本申请的其他特性和优点将通过下面的详细描述变得显然，或部分地通过本申请的实践而习得。

[0005] 根据本申请实施例的一个方面，提供了一种发音检测方法，包括：从待检测的语音音频中提取音频帧特征；基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率；对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分。

[0006] 根据本申请实施例的一个方面，提供了一种发音检测装置，包括：提取单元，用于从待检测的语音音频中提取音频帧特征；概率单元，用于基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率；得分单元，用于对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分。

[0007] 在本申请的一些实施例中，基于前述方案，所述提取单元包括：增强单元，用于对所述语音音频进行信号增强处理，生成增强语音；分帧单元，用于基于设定帧长对所述增强语音进行分帧处理，生成语音序列；加窗单元，用于基于设定窗口长度对所述语音序列进行加窗处理，生成加窗语音序列；变换单元，用于对所述加窗语音序列进行傅里叶变换，生成频域语音信号；滤波单元，用于对所述频域语音信号进行滤波处理，生成所述音频帧特征。

[0008] 在本申请的一些实施例中，基于前述方案，所述增强单元用于：获取所述语音音频中第一时刻对应的第一信号、所述第一时刻之前的第二时刻对应的第二信号；基于设定的信号系数和所述第二信号，计算所述第二信号对应的加权信号；基于所述第一信号强度与所述加权信号之间的差值，生成所述第一时刻对应的增强信号；将所述语音音频中各时刻

对应的增强信号进行组合，得到所述增强语音。

[0009] 在本申请的一些实施例中，基于前述方案，所述概率单元包括：第一模型单元，用于将所述音频帧特征输入基于第一语言样本训练得到的第一声学模型，输出所述音频帧特征与所述第一语音音素的匹配度对应的第一后验概率；第一时刻单元，用于基于所述语音音频中各音素对应的波形，识别所述音素对应的始末时刻；第一特征单元，用于基于所述音素对应的始末时刻和所述音频帧特征对应的时间帧信息，确定所述音素中包含的音频帧特征；第一概率单元，用于对所述音素中包含的音频帧特征对应的第一后验概率进行均值计算，生成所述音素对应于所述第一语言音素的第一后验概率。

[0010] 在本申请的一些实施例中，基于前述方案，所述发音检测装置还用于：获取基于第一语言生成的第一语音样本、以及所述第一语音样本对应的第一语音文本，并获取基于第二语言生成的第二语音样本、以及所述第二语音样本对应的第二语音文本；基于时延神经网络构建用于识别音频中所包含音素的声学模型；将所述第一语音样本输入所述声学模型中，并基于输出的第一音素与所述第一语音文本得到的第一损失函数，对所述声学模型的参数进行调整，得到所述第一声学模型；将所述第二语音样本输入所述声学模型中，并基于输出的第二音素与所述第二语音文本得到的第二损失函数，对所述声学模型的参数进行调整，得到第二声学模型。

[0011] 在本申请的一些实施例中，基于前述方案，所述概率单元包括：第二模型单元，用于将所述音频帧特征输入基于第二语言样本训练得到的第二声学模型，输出所述音频帧特征与所述第二语音音素的匹配度对应的第二后验概率；第二时刻单元，用于基于所述语音音频的波形，识别所述音素对应的始末时刻；第二特征单元，用于基于所述音素对应的始末时刻和所述音频帧特征对应的时间帧信息，确定所述音素中包含的音频帧特征；第二概率单元，用于基于所述音素对应的始末时刻，对所述音素中各所述音频帧特征对应的第二后验概率进行均值计算，确定所述音素对应于所述第二语言音素的第二后验概率。

[0012] 在本申请的一些实施例中，基于前述方案，所述得分单元用于：对所述第一后验概率和所述第二后验概率进行拼接，得到概率特征；对所述概率特征进行神经网络回归处理，生成所述语音音频中的音素对应于所述第二语言音素的概率得分。

[0013] 在本申请的一些实施例中，基于前述方案，所述显示单元包括：置信度单元，用于基于所述音素对应于第二语言音素的概率得分，确定所述音素与所述第二语言音素之间的置信度；等级确定单元，用于基于所述置信度与设定的置信度阈值，确定所述语音音频中各音素对应的发音准确等级。

[0014] 在本申请的一些实施例中，基于前述方案，所述置信度单元用于：从所述音素对应于第二语言音素的概率得分中，确定最大概率得分；计算指定音素对应于所述第二语言音素的概率得分与所述最大概率得分之间的比值；基于所述比值确定所述指定音素与所述第二语言音素之间的置信度。

[0015] 在本申请的一些实施例中，基于前述方案，所述发音检测装置还包括显示单元，用于基于所述概率得分确定各音素对应的发音准确等级，并基于所述发音准确等级对应的显示方式显示所述音素对应的文本。

[0016] 在本申请的一些实施例中，基于前述方案，所述显示单元用于：获取所述语音音频对应的文本；基于所述语音音频中的音素，对所述文本进行切词，生成各音素对应的文本；

基于各音素对应的发音准确等级，通过所述发音准确等级对应的显示方式显示所述音素对应的文本。

[0017] 在本申请的一些实施例中，基于前述方案，所述发音检测装置还用于：从各所述音素对应的发音准确等级中，查询发音准确等级最低的目标音素；获取所述目标音素对应的发音示教信息，其中，所述发音示教信息包括以下信息中的至少一个：音标文本、正确读法以及示意视频；展示所述发音示教信息。

[0018] 在本申请的一些实施例中，基于前述方案，所述发音检测装置还用于：从所述第二语言对应的词句库中获取包含所述目标音素的目标词句；展示所述目标词句；获取用户基于所述目标词句发送的练习音频；对所述练习音频进行检测，得到所述目标音频对应的发音准确等级。

[0019] 根据本申请实施例的一个方面，提供了一种计算机可读介质，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上述实施例中所述的发音检测方法。

[0020] 根据本申请实施例的一个方面，提供了一种电子设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现如上述实施例中所述的发音检测方法。

[0021] 根据本申请实施例的一个方面，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述各种可选实现方式中提供的发音检测方法。

[0022] 在本申请的一些实施例所提供的技术方案中，在获取到待检测的语音音频之后，从语音音频中提取音频帧特征，基于音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并同时基于音频帧特征与预设的第二语言中第二语音音素之间的匹配度，生成第二后验概率，之后对第一后验概率和第二后验概率进行神经网络回归处理，生成语言音频中的音素对应于第二语言音素的概率得分，以基于概率得分确定各音素对应的发音准确等级，通过基于发音准确等级确定对应的显示方式，以在终端中显示音素对应的文本。通过上述方式可以避免第一语言的发音习惯对第二语言的发音检测结果的影响，有效区别第一语言和第二语言中的相似音素，进而得到精确的发音检测结果，提高发音检测的精确性和发音者的练习效率。

[0023] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

[0024] 此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。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。在附图中：

[0025] 图1示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图；

[0026] 图2示意性示出了根据本申请的一个实施例的基于云平台的系统架构的示意图；

[0027] 图3示意性示出了根据本申请的一个实施例的发音检测方法的流程图；

[0028] 图4示意性示出了根据本申请的一个实施例的获取语音音频的示意图；

[0029] 图5示意性示出了根据本申请的一个实施例的获取语音音频的示意图；

[0030] 图6示意性示出了根据本申请的一个实施例的获取语音音频的示意图；

[0031] 图7示意性示出了根据本申请的一个实施例的发音检测的示意图；

[0032] 图8示意性示出了根据本申请的一个实施例的显示发音准确等级的示意图；

[0033] 图9示意性示出了根据本申请的一个实施例的从待检测的语音音频中提取音频帧特征；

[0034] 图10示意性示出了根据本申请的一个实施例的构建声学模型的流程图；

[0035] 图11示意性示出了根据本申请的一个实施例的声学模型的示意图；

[0036] 图12示意性示出了根据本申请的一个实施例的生成第一后验概率的流程图；

[0037] 图13示意性示出了根据本申请的一个实施例的显示语音音频对应的文本的示意图；

[0038] 图14示意性示出了根据本申请的一个实施例的显示语音音频示教的示意图；

[0039] 图15示意性示出了根据本申请的一个实施例的语音练习的示意图；

[0040] 图16示意性示出了根据本申请的一个实施例的发音检测方法的框图；

[0041] 图17示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

具体实施方式

[0042] 现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例；相反，提供这些实施方式使得本申请将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。

[0043] 此外，所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施例中。在下面的描述中，提供许多具体细节从而给出对本申请的实施例的充分理解。然而，本领域技术人员将意识到，可以实践本申请的技术方案而没有特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知方法、装置、实现或者操作以避免模糊本申请的各方面。

[0044] 附图中所示的方框图仅仅是功能实体，不一定必须与物理上独立的实体相对应。即，可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

[0045] 附图中所示的流程图仅是示例性说明，不是必须包括所有的内容和操作/步骤，也不是必须按所描述的顺序执行。例如，有的操作/步骤还可以分解，而有的操作/步骤可以合并或部分合并，因此实际执行的顺序有可能根据实际情况改变。

[0046] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理

技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0047] 语音技术(Speech Technology)的关键技术有自动语音识别技术(ASR)和语音合成技术(TTS)以及声纹识别技术。让计算机能听、能看、能说、能感觉,是未来人机交互的发展方向,其中语音成为未来最被看好的人机交互方式之一。自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。机器学习(Machine Learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0048] 本申请实施例提供的方案涉及人工智能的语音技术、自然语言处理以及机器学习等技术,通过基于预先训练得到的第一语言对应的第一语音模型、第二语言对应的第二语音模型,通过自然语言处理识别得到用户发出的语音音频中相比于第一语言音素和第二语言音素之间的匹配度,进而基于这两个匹配度通过机器学习的方式确定语音音频基于第二语言音素的匹配度,以确定该语音音频的发音准确度,进而显示在用户终端上,以提高了语音音频检测和示教的准确性。

[0049] 图1示出了可以应用本申请实施例的技术方案的示例性系统架构的示意图。

[0050] 如图1所示,系统架构可以包括终端设备(如图1中所示智能手机101、平板电脑102和便携式计算机103中的一种或多种,当然也可以是台式计算机等等)、网络104和服务器105。网络104用以在终端设备和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线通信链路、无线通信链路等等。

[0051] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。比如服务器105可以是多个服务器组成的服务器集群等。

[0052] 用户可以使用终端设备通过网络104与服务器105交互,以接收或发送消息等。服务器105可以是提供各种服务的服务器。例如用户利用终端设备103(也可以是终端设备101或102)向服务器105上传了从待检测的语音音频中提取音频帧特征;基于音频帧特征与预设第一语言中的第一语音音素之间的匹配度,生成第一后验概率,并基于音频帧特征与预设第二语言中的第二语音音素之间的匹配度,生成第二后验概率;对第一后验概率和第二后验概率进行神经网络回归处理,生成语音音频中的音素对应于第二语言音素的概率得分;基于概率得分确定各音素对应的发音准确等级,并将发音准确等级发送至终端设备中,通过基于发音准确等级确定对应的显示方式,以在终端中显示音素对应的文本。

[0053] 上述方案中,在获取到待检测的语音音频之后,从语音音频中提取音频帧特征,基于音频帧特征与预设第一语言中的第一语音音素之间的匹配度,生成第一后验概率,并同

时基于音频帧特征与预设的第二语言中第二语音音素之间的匹配度，生成第二后验概率，之后对第一后验概率和第二后验概率进行神经网络回归处理，生成语言音频中的音素对应于第二语言音素的概率得分，以基于概率得分确定各音素对应的发音准确等级，并将发音准确等级发送至终端设备中，通过基于发音准确等级确定对应的显示方式，以在终端中显示音素对应的文本。通过上述方式可以避免第一语言的发音习惯对第二语言的发音检测结果的影响，有效区别第一语言和第二语言中的相似音素，进而得到精确的发音检测结果，提高发音检测的精确性和发音者的练习效率。

[0054] 需要说明的是，本申请实施例所提供的发音检测方法一般由服务器105执行，相应地，发音检测装置一般设置于服务器105中。但是，在本申请的其它实施例中，终端设备也可以与服务器具有相似的功能，从而执行本申请实施例所提供的发音检测的方案。

[0055] 图2为本申请实施例提供的一种基于云平台的系统架构的示意图。

[0056] 云计算(cloud computing)是一种计算模式，它将计算任务分布在大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和信息服务。提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的，并且可以随时获取，按需使用，随时扩展，按使用付费。作为云计算的基础能力提供商，会建立云计算资源池，简称云平台，一般称为基础设施即服务(Infrastructure as a Service，IaaS)平台，在资源池中部署多种类型的虚拟资源，供外部客户选择使用。云计算资源池中主要包括：计算设备(为虚拟化机器，包含操作系统)、存储设备、网络设备。按照逻辑功能划分，在IaaS层上可以部署平台即服务(Platform as a Service，PaaS)层，PaaS层之上再部署软件即服务(Software as a Service，SaaS)层，也可以直接将SaaS部署在IaaS上。PaaS为软件运行的平台，如数据库、web容器等。SaaS为各式各样的业务软件，如web门户网站、短信群发器等。一般来说，SaaS和PaaS相对于IaaS是上层。

[0057] 云计算(cloud computing)指IT基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需资源；广义云计算指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需服务。这种服务可以是IT和软件、互联网相关，也可是其他服务。云计算是网格计算(Grid Computing)、分布式计算(DistributedComputing)、并行计算(Parallel Computing)、效用计算(Utility Computing)、网络存储(Network Storage Technologies)、虚拟化(Virtualization)、负载均衡(Load Balance)等传统计算机和网络技术发展融合的产物。

[0058] 随着互联网、实时数据流、连接设备多样化的发展，以及搜索服务、社会网络、移动商务和开放协作等需求的推动，云计算迅速发展起来。不同于以往的并行分布式计算，云计算的产生从理念上将推动整个互联网模式、企业管理模式发生革命性的变革。

[0059] 如图2所示，在本实施例中的系统架构中，云204中存储有第一语言音素库和第二语言音素库，其存储的方式可以是通过第一语言对应的第一语言模型来存储第一语言音素库，通过第二语言对应的第二语言模型来存储第二语言音素库。

[0060] 系统架构中还包括智能手机201、平板电脑202以及便携式计算机203等终端设备，除此之外，还可以是其他终端设备。在终端设备获取到到待检测的语音音频之后，从语音音频中提取音频帧特征，基于音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并同时基于音频帧特征与预设的第二语言中第二语音音素之间的匹配

度，生成第二后验概率，之后对第一后验概率和第二后验概率进行神经网络回归处理，生成语言音频中的音素对应于第二语言音素的概率得分，最后基于概率得分确定各音素对应的发音准确等级，以基于发音准确等级确定对应的显示方式，并基于该显示方式显示音素对应的文本。通过上述方式可以避免第一语言的发音习惯对第二语言的发音检测结果的影响，有效区别第一语言和第二语言中的相似音素，进而得到精确的发音检测结果，提高发音检测的精确性和发音者的练习效率。

[0061] 本实施例中，云对应的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。

[0062] 以下对本申请实施例的技术方案的实现细节进行详细阐述：

[0063] 图3示出了根据本申请的一个实施例的发音检测方法的流程图，该发音检测方法可以由服务器来执行，该服务器可以是图1中所示的服务器，也可以通过中端设备直接执行。参照图3所示，该发音检测方法至少包括步骤S310至步骤S330，详细介绍如下：

[0064] 在步骤S310中，从待检测的语音音频中提取音频帧特征。

[0065] 图4～图6为本申请实施例提供的一种获取语音音频的示意图。

[0066] 如图4所示，在本申请的一个实施例中，先获取待检测的语音音频，其获取的方式是在应用程序的界面中先显示待朗读的文本，例如，“hello”；再由用户触发“点击开始跟读”的按钮。

[0067] 如图5所示，在检测到用户触发“点击开始跟读”的按钮时，开始获取语音音频，并显示“请靠近麦克风大声朗读”，以提示用户朗读的方式。同时显示“点击结束”按钮，以在用户朗读完毕之后，点击该按钮，以获取到结束获取音频的指示。

[0068] 如图6所示，为了对用户朗读的时长有所限制，且提高朗读检测的效率，本实施例中还可以在朗读过程中进行倒计时处理，在界面中显示倒计时的时长，例如，在采集语音音频的过程中显示“倒计时2秒”。通过上述方式，可以提醒用户及时朗读，提高语音音频获取的效率和检测的效率。

[0069] 图7为本申请实施例提供的一种发音检测的示意图。

[0070] 如图7所示，在本申请的一个实施例中，分为客户端710和服务器端720两个部分。客户端710部分介绍来用户在软件上进行发音练习操作，软件记录下用户的练习音频后，将其传至服务器端720，服务器进行发音偏误的检测后，将偏误进行回传给用户，并提示用户修改意见。服务器端描述了接收到用户发音练习的音频后，对用户发音进行音素级别的发音偏误检测的全过程，也同时说明了在服务器端检测出来发音偏误信息后，将其回传给客户端，以便用户进行下次练习。

[0071] 具体的，在通过客户端710获取到语音音频之后，在服务器端720从语音音频中进行特征提取730，提取出音频的帧级别特征740。具体的，本实施例中的音频帧特征用于表示在语音音频中的每一帧对应的音频特征，即一组能代表用户发音帧层级的特征序列。示例性的，本实施例中提取音频帧的方式可以通过滤波的方式得到，也可以通过语音识别的方式得到。

[0072] 需要说明的是，本实施例中通过提取帧级别的音频特征，作为音频的帧级别特征，以对帧级别特征进行强制对齐来确定每个音频特征对应的播放时刻，进而确定一个音素所包含的所有音频特征对应的播放时间段，以基于各个音频特征对应的后验概率，确定该播放时间段对应的音素整体的后验概率。通过上述音频的帧级别特征提取，可以将音频检测和识别的精度精确到帧级别，更加增强了发音的检测精度。

[0073] 除了上述从音频中提取帧级别特征的方法之外，还可以从中提取各个音素对应的时间段的音频特征，以直接识别该音素对应的后验概率，通过这种方式可以增加音频数据处理的效率，进而提高音频检测和识别的效率。

[0074] 在步骤S320中，基于音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率。

[0075] 如图7所示，在本申请的一个实施例中，在获取到帧级别特征740，即音频帧特征之后，通过强制对齐750的方式确定语音音频中的各个音素对应的播放时间。在终端设备或者服务器获取到语音音频之后，并不确定语音音频为第一语言还是第二语言，这种情况下，需要通过两种语言的识别模型来检测当前的语音音频与两种模型中各自对应的语音音素之间的匹配度或者相似度，即当前语音音频对应于设定语音音素的后验概率。如图7所示，本实施例中基于声学模型来识别音段级别的后验概率，即通过第一语言(First language，L1)声学模型760来确定音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率771，以及通过第二语言(Second Language，L2)声学模型762来确定音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率772。

[0076] 需要说明的是，本实施例中的第一语言可以为用户的母语，第二语言可以为用户正在练习的外语。例如，在一个以中文为母语的小孩练习英语的场景下，本实施例中的第一语言对应为汉语，第二语言对应为英语。

[0077] 除此之外，本实施例中第一语言可以为用户的常用语，第二语言可以为用户的联系语言，即，用户平常用第一语言来交流，在练习过程中，是通过第二语言来进行练习的。

[0078] 在步骤S330中，对第一后验概率和第二后验概率进行神经网络回归处理，生成语音音频中的音素对应于第二语言音素的概率得分。

[0079] 在本申请的一个实施例中，在获取到第一后验概率和第二后验概率之后，基于图7中的深度神经网络(Deep Neural Networks，DNN)780的方式，通过第一后验概率和第二概率进行回归处理，之后基于发音良好度(Goodness ofPronunciation，GOP)790的方式确定语音音频中的音素对应于第二语言音素的概率得分。本实施例中的概率得分用于表示用户所发出的语音音频与第二语言之间的相似度，且该相似度是基于第一语言与第二语言中相同或者相似的音素，通过对第二后验概率中第一后验概率对应的第一语言的发音习惯进行滤除，进而得到只针对于第二语言音素的概率得分。

[0080] 在实际的发音过程中，以第一语言为母语的发音人，可能会在讲出第二语言的语音时，携带第一语言的发音习惯，即通过第一语言的发音方式来讲出第二语言中对应的音素，这种情况下将导致发音检测的偏误。因此，本实施例中通过计算后验概率的方式来捕获发音者在第二语言的语音中与第一语言相同或相似的音位，以基于第二语言对应的第二后验概率，滤除第二语言中第一后验概率对应的第一语言的发音内容，并基于滤除之后得到

的语音音素来计算语音音频中的音素对应于第二语言音素的概率得分。通过上述方式，避免了第一语言的发音习惯对第二语言的发音判断的影响，提高了发音检测的精确度和客观性。

[0081] 具体的，本实施例中对第一后验概率和第二后验概率进行神经网络回归处理，生成语音音频中的音素对应于第二语言音素的概率得分的具体方式可以是通过Sigmoid函数对第一后验概率和第二后验概率进行逻辑回归，得到概率得分。

[0082] 在本申请一实施例中，在步骤S330之后，还可以包括步骤S340，基于概率得分确定各音素对应的发音准确等级，并基于发音准确等级对应的显示方式显示音素对应的文本。

[0083] 如图7所示，在本申请的一个实施例中，预设有各个发音准确等级对应的概率得分阈值，确定是否发音偏误711。具体的，本实施例中的发音准确等级可以包括：精确、良好、合格、错误以及漏发等状态对应的等级，且每个发音准确等级有其对应的显示方式，例如颜色、深浅或者文字大小等等。本实施例中在生成概率得分之后，基于发音准确等级对应的显示方式，在终端界面中显示音素对应的文本。

[0084] 图8为本申请实施例提供的一种显示发音准确等级的示意图。

[0085] 如图8所示，在生成发音准确等级之后，可以通过星标的数量来显示整个语音音频对应的总等级的高低。并且，本实施例中的方法可以识别到语音音频中的各音素的对应的发音准确等级，以通过不同的显示方式来显示。例如，在对“Good afternoon”对应的语音音频进行识别之后，得到其中“Good”发音精确，则通过加粗的方式来显示，“after”发音存在偏差，则通过灰色的方式来显示，“noon”发音合格，则通过一般的显示方式来显示。通过上述显示方式可以明确表示出来用户对各个音素的发音状态，提高用户的练习效率。

[0086] 在本申请的一个实施例中，如图9所示，步骤S310中从待检测的语音音频中提取音频帧特征的过程，包括如下步骤：

[0087] 步骤S910，对语音音频进行信号增强处理，生成增强语音。

[0088] 在本申请的一个实施例中，通过对学习者的语音进行预加重等预处理，其原理主要是要对语音信号的高频进行一定程度的增强，去除口腔辐射的影响。具体的，步骤S910中对语音音频进行信号增强处理，生成增强语音的过程具体包括：

[0089] 步骤S9101，获取语音音频中第一时刻对应的第一信号、第一时刻之前的第二时刻对应的第二信号；

[0090] 步骤S9102，基于设定的信号系数和第二信号，计算第二信号对应的加权信号；

[0091] 步骤S9103，基于第一信号强度与加权信号之间的差值，生成第一时刻对应的增强信号；

[0092] 步骤S9104，将语音音频中各时刻对应的增强信号进行组合，得到增强语音。

[0093] 具体的，在连续信号中通过n来表示语音的播放时刻，本实施例中第一时刻为n，第一时刻之前的第二时刻对应的第二信号n-1。第一时刻对应的第一信号为x(n)、第一时刻之前的第二时刻对应的第二信号x(n-1)；基于设定的信号系数α和第二信号x(n-1)，计算第二信号对应的加权信号αx(n-1)；基于第一信号强度与加权信号之间的差值，生成第一时刻对应的增强信号y(n)＝x(n)-αx(n-1)；最后将语音音频中各时刻对应的增强信号进行组合，得到增强语音。

[0094] 步骤S920，基于设定帧长对增强语音进行分帧处理，生成语音序列。

[0095] 在本申请的一个实施例中，然后会对信号进行分帧等操作。示例性的，以25ms为帧长、10ms为帧移，将一段若干秒的发音分解成一组25ms长的语音段序列。

[0096] 步骤S930，基于设定窗口长度对语音序列进行加窗处理，生成加窗语音序列。

[0097] 在本申请的一个实施例中，对上述步骤中得到的语音段序列里的每一个小段语音进行加窗处理，可以通过加汉明窗的方式进行。

[0098] 步骤S940，对加窗语音序列进行傅里叶变换，生成频域语音信号。

[0099] 在本申请的一个实施例中，对每一小段语音进行傅里叶变换，这样就可以将语音信号从时域变换到了频域。

[0100] 步骤S950，对频域语音信号进行滤波处理，生成音频帧特征。

[0101] 在本申请的一个实施例中，将这一组在频域上的语音帧序列分别按帧进行梅尔滤波提取成后续模型可用的特征，本质是一个信息压缩和抽象的过程。这个阶段可抽取的特征多种多样，如频谱特征(梅尔频率倒谱系数MFCC、过滤器组FBANK、分组级协议PLP等)、频率特征(基频、共振峰等)、时域特征(时长特征)、能量特征等等。本案实验所用特征为40维的FBANK特征。经过了这个模块之后，学习者一段发音便变成了一组能代表其发音的特征序列，也就是图中所说的帧层级的特征。

[0102] 在本申请的一个实施例中，如图10所示，本实施例中发音检测方法还包括：

[0103] 步骤S1010，获取基于第一语言生成的第一语音样本、以及第一语音样本对应的第一语音文本，并获取基于第二语言生成的第二语音样本、以及第二语音样本对应的第二语音文本。

[0104] 步骤S1020，基于时延神经网络构建用于识别音频中所包含音素的声学模型；

[0105] 步骤S1030，将第一语音样本输入声学模型中，并基于输出的第一音素与第一语音文本得到的第一损失函数，对声学模型的参数进行调整，得到第一声学模型；

[0106] 步骤S1040，将第二语音样本输入声学模型中，并基于输出的第二音素与第二语音文本得到的第二损失函数，对声学模型的参数进行调整，得到第二声学模型。

[0107] 图11为本申请实施例提供的一种声学模型的示意图。

[0108] 在本申请的一个实施例中，学习者在学习第二语言(Second Language，L2)发音时，对于L2中和学习者母语，即第一语言(First language，L1)相似音素，会使用L1的音素进行替代，这是构成发音偏误的重要原因之一。

[0109] 为了避免这种相似音素造成的混淆检测的问题，本实施例中基于以第一语言为本地语言的用户朗读的第二语言数据1110以及以第一语言为本地语言的用户朗读的第一语言数据1120，即本地L1语音语料库和本地L2语音语料库，在输入层引入这两种语料库作为训练数据，以保证训练的完整性和精确性。在输出层分别设置汉语及英语的语音识别任务，通过时延神经网络1150中的迁移学习机制构成的共享隐藏层，获得具有汉语和英语发音泛化能力的声学模型，即第一声学模型1140和第二声学模型1130。

[0110] 本实施例中利用不同的数据和任务可能存在内在的关联性，利用深度神经网络的隐含层级参数设法获取这种关联性，就可以把从一个任务中获得的知识运用到另外一个任务的解决中。通过利用多任务多语言迁移学习方法，把与目标任务—学习者的英语发音偏误检测具有较强关联性的数据尽量涵盖进来，构建出具有语言泛化能力的声学模型。

[0111] 在本申请的一个实施例中，如图12所示，步骤S320中基于音频帧特征与预设的第

一语音音素之间的匹配度,生成第一后验概率的过程,包括如下步骤:

[0112] 步骤S3210,将音频帧特征输入基于第一语言样本训练得到的第一声学模型,输出音频帧特征与第一语音音素的匹配度对应的第一后验概率;

[0113] 步骤S3220,基于语音音频中各音素对应的波形,识别音素对应的始末时刻;

[0114] 步骤S3230,基于音素对应的始末时刻和音频帧特征对应的时间帧信息,确定音素中包含的音频帧特征;

[0115] 步骤S3240,对音素中包含的音频帧特征对应的第一后验概率进行均值计算,生成音素对应于第一语言音素的第一后验概率。

[0116] 在本申请的一个实施例中,将音频帧特征输入基于第一语言样本训练得到的第一声学模型,输出音频帧特征与第一语音音素的匹配度对应的第一后验概率,这个概率代表了学习者的每一帧发音和声学模型中第一语言样本的音素分布的匹配度。基于语音识别框架和强制对齐技术,将给定的语音和文本进行音素级别的对齐,这样就可以知道在语音段中每个音素的起始时间和结束时间。基于音素对应的始末时刻和音频帧特征对应的时间帧信息,确定音素中包含的音频帧特征;最后对音素中包含的音频帧特征对应的第一后验概率进行均值计算,求取概率的平均值,作为音素对应于第一语言音素的第一后验概率。本实施例中引入L1的后验概率特征是为了更好的区分L1和L2中相同或相似的音位。结合两种特征最后通过DNN回归来得到最后在L2音素集上的概率得分。

[0117] 具体的,在计算第一后验概率时,当帧层级的语音特征输入到声学模型后,就可以得到每一帧的后验概率,这个概率代表了学习者的每一帧发音和声学模型中音素分布的匹配度。因为声学模型常用母语者数据进行训练,所以也就可以当成是以母语者的视角来看学习者发成了什么样子,本实施例所采用的声学模型为一个基于隐马尔可夫模型-时延神经网络HMM-TDNN的语音识别框架的一部分,其原理如下:

[0118] $$\hat{w} = \arg\max P(w|x) = \arg\max \frac{P(x|w)P(w)}{P(x)} = \arg\max P(x|w)P(w)$$

[0119] 其中,p(x|w)表示声学模型部分,w表示发音文本,表示为学习者当前的发音,即第二语言对应的语音音频,概率p(x|w)则表征了若学习者是想发出当前文本所代表的音素发音的好坏程度。

[0120] 在本申请的一个实施例中,步骤S320中基于音频帧特征与预设的第二语音音素之间的匹配度,生成第二后验概率的过程,包括如下步骤:将音频帧特征输入基于第二语言样本训练得到的第二声学模型,输出音频帧特征与第二语音音素的匹配度对应的第二后验概率;基于语音音频的波形,识别音素对应的始末时刻;基于音素对应的始末时刻和音频帧特征对应的时间帧信息,确定音素中包含的音频帧特征;基于音素对应的始末时刻,对音素中各音频帧特征对应的第二后验概率进行均值计算,确定音素对应于第二语言音素的第二后验概率。

[0121] 在本申请的一个实施例中,将音频帧特征输入基于第二语言样本训练得到的第二声学模型,输出音频帧特征与第二语音音素的匹配度对应的第二后验概率,这个概率代表了学习者的每一帧发音和声学模型中第二语言样本的音素分布的匹配度。基于语音识别框架和强制对齐技术,将给定的语音和文本进行音素级别的对齐,这样就可以知道在语音段中每个音素的起始时间和结束时间。基于音素对应的始末时刻和音频帧特征对应的时间帧

信息，确定音素中包含的音频帧特征；最后对音素中包含的音频帧特征对应的第二后验概率进行均值计算，求取概率的平均值，作为音素对应于第二语言音素的第二后验概率。

[0122] 在本申请的一个实施例中，步骤S330中对第一后验概率和第二后验概率进行神经网络回归处理，生成语音音频中的音素对应于第二语言音素的概率得分的过程，包括如下步骤：对第一后验概率和第二后验概率进行拼接，得到概率特征；对概率特征进行神经网络回归处理，生成语音音频中的音素对应于第二语言音素的概率得分。

[0123] 在本申请的一个实施例中，根据第二个模块中得到用户发音中每个音素的时间段和第三个模块中通过L1和L2声学模型得到每帧上的音素后验概率，进一步求出每个音素段上在两个声学模型中的音素后验概率。L1后验特征代表了用户对于该音素的发音在L1音素集上每个音素的得分情况，L2后验特征代表了用户对于该音素的发音在L2音素集上每个音素的得分情况。这样对于L1和L2中相似的音位。我们可以在利用L2后验特征的情况下结合L1后验特征进行辅助加强检测。

[0124] 在本申请的一个实施例中，步骤S340中基于概率得分确定各音素对应的发音准确等级的过程，包括如下步骤：基于音素对应于第二语言音素的概率得分，确定音素与第二语言音素之间的置信度；基于置信度与设定的置信度阈值，确定语音音频中各音素对应的发音准确等级。

[0125] 具体的，在本申请的一个实施例中，一个概率样本的置信区间是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度。置信区间给出的是被测量参数的测量值的可信程度。本实施例中基于置信度与设定的置信度阈值，确定语音音频中各音素对应的发音准确等级。

[0126] 在本申请的一个实施例中，基于音素对应于第二语言音素的概率得分，确定音素与第二语言音素之间的置信度，包括：从音素对应于第二语言音素的概率得分中，确定最大概率得分；计算指定音素对应于第二语言音素的概率得分与最大概率得分之间的比值；基于比值确定指定音素与第二语言音素之间的置信度。

[0127] 具体的，本模块是根据DNN神经网络输出的音素层级的后验概率和对应的音素级别的对齐信息，通过发音良好度(Goodness ofPronunciation，GOP)算法，经过对比用户本该发的音素和用户实际发的音素的概率大小，即可对每一个发音是否发生了偏误进行判断。从音素对应于第二语言音素的概率得分P(p)中，确定最大概率得分P(q)；计算指定音素对应于第二语言音素的概率得分与最大概率得分之间的比值为：

[0128] $$GOP = \frac{P(p)}{\max_{q \in s} P(q)}$$

[0129] 其中，p代表了当前发音音素；P(p)代表了DNN输出的当前音素的概率；S代表了整个音素集；q代表了DNN输出的最大概率所对应的音素；P(q)代表了DNN输出的最大概率。GOP打分之后，通过阈值判断用户发音中音素的偏误情况，然后将当前发音哪个音素是偏误，误发成了什么样子返回给客户端的用户。

[0130] 通过上述流程，我们可得知，在学习者的发音中，哪个音素发生了偏误。在这其中，如何得到最后的音素得分就非常的重要，音素得分的高低直接影响系统的偏误判断。

[0131] 在本申请的一个实施例中，步骤S340中基于发音准确等级对应的显示方式显示音素对应的文本的过程，包括：获取语音音频对应的文本；基于语音音频中的音素，对文本进

行切词,生成各音素对应的文本;基于各音素对应的发音准确等级,通过发音准确等级对应的显示方式显示音素对应的文本。

[0132] 图13为本申请实施例显示语音音频对应的文本的示意图。

[0133] 如图13所示,本实施例中先获取语音音频对应的文本:Good afternoon;基于语音音频中的音素,对文本进行切词,得到各音素分别对应的文本“Good”、“after”以及“noon”,基于各音素对应的发音准确等级,确定其对应的显示方式,即加粗、灰度以及正常显示,并基于这些显示方式来显示各音素对应的文本。

[0134] 在本申请的一个实施例中,基于概率得分确定各音素对应的发音准确等级,并基于发音准确等级对应的显示方式显示音素对应的文本之后,方法还包括:从各音素对应的发音准确等级中,查询发音准确等级最低的目标音素;获取目标音素对应的发音示教信息,其中,发音示教信息包括以下信息中的至少一个:音标文本、正确读法以及示意视频;展示发音示教信息。

[0135] 图14为本申请实施例显示语音音频示教的示意图。

[0136] 如图14所示,在确定了用户对于某一个词句的发音等级,并在界面1410中显示出来之后,可基于具体的发音情况进行针对性的示教,也可以针对词句中全部的发音进行示教。例如图14中针对“Good afternoon”中的各个音素在界面1420中进行示教。本实施例中的示教信息包括音标文本、正确读法以及示意视频中的至少一个。通过对词句进行示教可以提高用户的联系效率,提高用户的学习和练习效果。

[0137] 在本申请的一个实施例中,从各音素对应的发音准确等级中,查询发音准确等级最低的目标音素之后,方法还包括:从第二语言对应的词句库中获取包含目标音素的目标词句;展示目标词句;获取用户基于目标词句发送的练习音频;对练习音频进行检测,得到目标音频对应的发音准确等级。

[0138] 图15为本申请实施例提供的一种语音练习的示意图。

[0139] 如图15所示,在从各音素对应的发音准确等级界面1510中,查询发音准确等级最低的目标音素之后,从第二语言对应的词句库中获取包含目标音素的目标词句;展示目标词句1520;获取用户基于目标词句发送的练习音频“fool”并显示在界面1530中;对练习音频进行检测,得到目标音频对应的发音准确等级。通过上述强化练习的方式,进一步的提高用户的练习效果和发音的准确性。

[0140] 本实施例在检测中国K12儿童英语发音音素等错误率指标上,相较于传统只使用本地L2语音语料库的系统整体性能相对改善8.82%。在辅音的改善非常明显,但元音表现并不突出,这是因为L1和L2中大部分相似或相同音位都存在于L1的辅音中,在辅音Z、JH、F等音素上性能改善相对超过20%以上,表明本方案可以有效区别在中国学习者的英语发音偏误检测系统中L1和L2相同或相似的音位,提高发音偏误检测模型的鲁棒性。和产品结合后,因为学习者往往对于相似的发音很容易发错,英语君能更精确地检测出来学习者发音中和母语相似的发音,让基于发音质量的打分更有据可循。从而让孩子们可以把有限的注意力集中在最重要的偏误改正上。这样他们就可以更为高效地更有信心地改善口语能力。

[0141] 以下介绍本申请的装置实施例,可以用于执行本申请上述实施例中的发音检测方法。可以理解的是,所述装置可以是运行于计算机设备中的一个计算机程序(包括程序代码),例如该装置为一个应用软件;该装置可以用于执行本申请实施例提供的方法中的相应

步骤。对于本申请装置实施例中未披露的细节，请参照本申请上述的发音检测方法的实施例。

[0142] 图16示出了根据本申请的一个实施例的发音检测装置的框图。

[0143] 参照图16所示，根据本申请的一个实施例的发音检测装置1600，包括：提取单元1610，用于从待检测的语音音频中提取音频帧特征；概率单元1620，用于基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率；得分单元1630，用于对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分。

[0144] 在本申请的一些实施例中，基于前述方案，所述提取单元1610包括：增强单元，用于对所述语音音频进行信号增强处理，生成增强语音；分帧单元，用于基于设定帧长对所述增强语音进行分帧处理，生成语音序列；加窗单元，用于基于设定窗口长度对所述语音序列进行加窗处理，生成加窗语音序列；变换单元，用于对所述加窗语音序列进行傅里叶变换，生成频域语音信号；滤波单元，用于对所述频域语音信号进行滤波处理，生成所述音频帧特征。

[0145] 在本申请的一些实施例中，基于前述方案，所述增强单元用于：获取所述语音音频中第一时刻对应的第一信号、所述第一时刻之前的第二时刻对应的第二信号；基于设定的信号系数和所述第二信号，计算所述第二信号对应的加权信号；基于所述第一信号强度与所述加权信号之间的差值，生成所述第一时刻对应的增强信号；将所述语音音频中各时刻对应的增强信号进行组合，得到所述增强语音。

[0146] 在本申请的一些实施例中，基于前述方案，所述概率单元1620包括：第一模型单元，用于将所述音频帧特征输入基于第一语言样本训练得到的第一声学模型，输出所述音频帧特征与所述第一语音音素的匹配度对应的第一后验概率；第一时刻单元，用于基于所述语音音频中各音素对应的波形，识别所述音素对应的始末时刻；第一特征单元，用于基于所述音素对应的始末时刻和所述音频帧特征对应的时间帧信息，确定所述音素中包含的音频帧特征；第一概率单元，用于对所述音素中包含的音频帧特征对应的第一后验概率进行均值计算，生成所述音素对应于所述第一语言音素的第一后验概率。

[0147] 在本申请的一些实施例中，基于前述方案，所述发音检测装置1600还用于：获取基于第一语言生成的第一语音样本、以及所述第一语音样本对应的第一语音文本，并获取基于第二语言生成的第二语音样本、以及所述第二语音样本对应的第二语音文本；基于时延神经网络构建用于识别音频中所包含音素的声学模型；将所述第一语音样本输入所述声学模型中，并基于输出的第一音素与所述第一语音文本得到的第一损失函数，对所述声学模型的参数进行调整，得到所述第一声学模型；将所述第二语音样本输入所述声学模型中，并基于输出的第二音素与所述第二语音文本得到的第二损失函数，对所述声学模型的参数进行调整，得到第二声学模型。

[0148] 在本申请的一些实施例中，基于前述方案，所述概率单元1620包括：第二模型单元，用于将所述音频帧特征输入基于第二语言样本训练得到的第二声学模型，输出所述音频帧特征与所述第二语音音素的匹配度对应的第二后验概率；第二时刻单元，用于基于所述语音音频的波形，识别所述音素对应的始末时刻；第二特征单元，用于基于所述音素对应

的始末时刻和所述音频帧特征对应的时间帧信息，确定所述音素中包含的音频帧特征；第二概率单元，用于基于所述音素对应的始末时刻，对所述音素中各所述音频帧特征对应的第二后验概率进行均值计算，确定所述音素对应于所述第二语言音素的第二后验概率。

[0149] 在本申请的一些实施例中，基于前述方案，所述得分单元1630用于：对所述第一后验概率和所述第二后验概率进行拼接，得到概率特征；对所述概率特征进行神经网络回归处理，生成所述语音音频中的音素对应于所述第二语言音素的概率得分。

[0150] 在本申请的一些实施例中，基于前述方案，所述显示单元包括：置信度单元，用于基于所述音素对应于第二语言音素的概率得分，确定所述音素与所述第二语言音素之间的置信度；等级确定单元，用于基于所述置信度与设定的置信度阈值，确定所述语音音频中各音素对应的发音准确等级。

[0151] 在本申请的一些实施例中，基于前述方案，所述置信度单元用于：从所述音素对应于第二语言音素的概率得分中，确定最大概率得分；计算指定音素对应于所述第二语言音素的概率得分与所述最大概率得分之间的比值；基于所述比值确定所述指定音素与所述第二语言音素之间的置信度。

[0152] 在本申请的一些实施例中，基于前述方案，所述发音检测装置还包括显示单元，用于基于所述概率得分确定各音素对应的发音准确等级，并基于所述发音准确等级对应的显示方式显示所述音素对应的文本。

[0153] 在本申请的一些实施例中，基于前述方案，所述显示单元用于：获取所述语音音频对应的文本；基于所述语音音频中的音素，对所述文本进行切词，生成各音素对应的文本；基于各音素对应的发音准确等级，通过所述发音准确等级对应的显示方式显示所述音素对应的文本。

[0154] 在本申请的一些实施例中，基于前述方案，所述发音检测装置1600还用于：从各所述音素对应的发音准确等级中，查询发音准确等级最低的目标音素；获取所述目标音素对应的发音示教信息，其中，所述发音示教信息包括以下信息中的至少一个：音标文本、正确读法以及示意视频；展示所述发音示教信息。

[0155] 在本申请的一些实施例中，基于前述方案，所述发音检测装置1600还用于：从所述第二语言对应的词句库中获取包含所述目标音素的目标词句；展示所述目标词句；获取用户基于所述目标词句发送的练习音频；对所述练习音频进行检测，得到所述目标音频对应的发音准确等级。

[0156] 图17示出了适于用来实现本申请实施例的电子设备的计算机系统的结构示意图。

[0157] 需要说明的是，图17示出的电子设备的计算机系统1700仅是一个示例，不应对本申请实施例的功能和使用范围带来任何限制。

[0158] 如图17所示，计算机系统1700包括中央处理单元(Central Processing Unit，CPU)1701，其可以根据存储在只读存储器(Read-Only Memory，ROM)1702中的程序或者从储存部分1708加载到随机访问存储器(Random Access Memory，RAM)1703中的程序而执行各种适当的动作和处理，例如执行上述实施例中所述的方法。在RAM 1703中，还存储有系统操作所需的各种程序和数据。CPU 1701、ROM 1702以及RAM 1703通过总线1704彼此相连。输入/输出(Input/Output，I/O)接口1705也连接至总线1704。

[0159] 以下部件连接至I/O接口1705：包括键盘、鼠标等的输入部分1706；包括诸如阴极

射线管(Cathode Ray Tube,CRT)、液晶显示器(Liquid Crystal Display,LCD)等以及扬声器等的输出部分1707;包括硬盘等的储存部分1708;以及包括诸如LAN(Local Area Network,局域网)卡、调制解调器等的网络接口卡的通信部分1709。通信部分1709经由诸如因特网的网络执行通信处理。驱动器1710也根据需要连接至I/O接口1705。可拆卸介质1711,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器1710上,以便于从其上读出的计算机程序根据需要被安装入储存部分1708。

[0160] 特别地,根据本申请的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。例如,本申请的实施例包括一种计算机程序产品,其包括承载在计算机可读介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的计算机程序。在这样的实施例中,该计算机程序可以通过通信部分1709从网络上被下载和安装,和/或从可拆卸介质1711被安装。在该计算机程序被中央处理单元(CPU)1701执行时,执行本申请的系统中限定的各种功能。

[0161] 需要说明的是,本申请实施例所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于:具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(Erasable Programmable Read Only Memory,EPROM)、闪存、光纤、便携式紧凑磁盘只读存储器(Compact Disc Read-Only Memory,CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的计算机程序。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的计算机程序可以用任何适当的介质传输,包括但不限于:无线、有线等等,或者上述的任意合适的组合。

[0162] 附图中的流程图和框图,图示了按照本申请各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。其中,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0163] 描述于本申请实施例中所涉及到的单元可以通过软件的方式实现,也可以通过硬件的方式来实现,所描述的单元也可以设置在处理器中。其中,这些单元的名称在某种情况

下并不构成对该单元本身的限定。

[0164] 根据本申请的一个方面，提供了一种计算机程序产品或计算机程序，该计算机程序产品或计算机程序包括计算机指令，该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令，处理器执行该计算机指令，使得该计算机设备执行上述各种可选实现方式中提供的方法。

[0165] 作为另一方面，本申请还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该电子设备执行时，使得该电子设备实现上述实施例中所述的方法。

[0166] 应当注意，尽管在上文详细描述中提及了用于动作执行的设备的若干模块或者单元，但是这种划分并非强制性的。实际上，根据本申请的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0167] 通过以上的实施方式的描述，本领域的技术人员易于理解，这里描述的示例实施方式可以通过软件实现，也可以通过软件结合必要的硬件的方式来实现。因此，根据本申请实施方式的技术方案可以以软件产品的形式体现出来，该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM，U盘，移动硬盘等)中或网络上，包括若干指令以使得一台计算设备(可以是个人计算机、服务器、触控终端、或者网络设备等)执行根据本申请实施方式的方法。

[0168] 本领域技术人员在考虑说明书及实践这里公开的实施方式后，将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。

[0169] 应当理解的是，本申请并不局限于上面已经描述并在附图中示出的精确结构，并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求来限制。

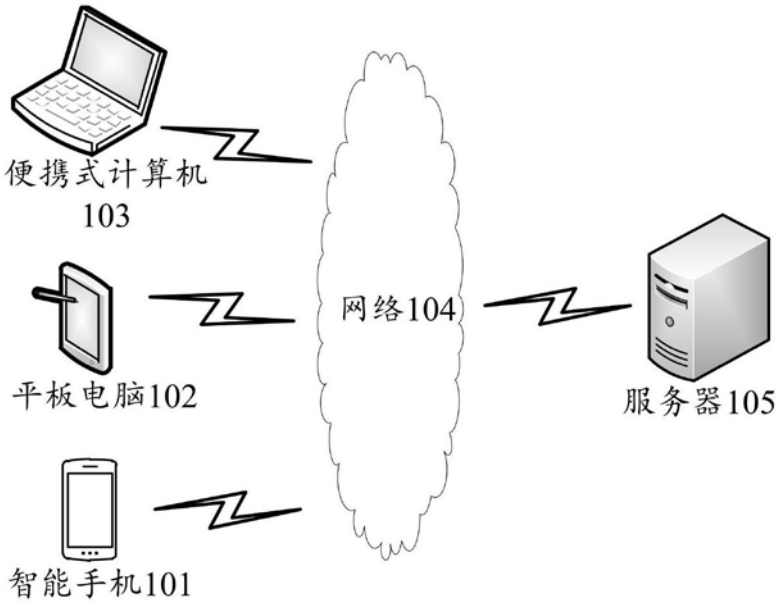

便携式计算机
103
平板电脑102
智能手机101
网络104
服务器105

图1

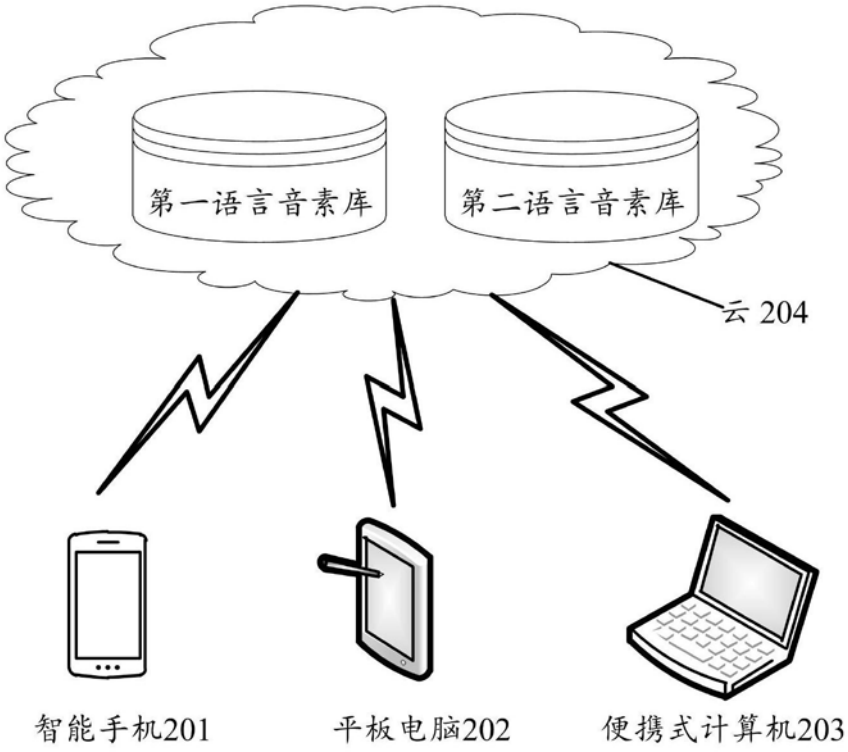

第一语言音素库
第二语言音素库
云 204
智能手机201
平板电脑202
便携式计算机203

图2

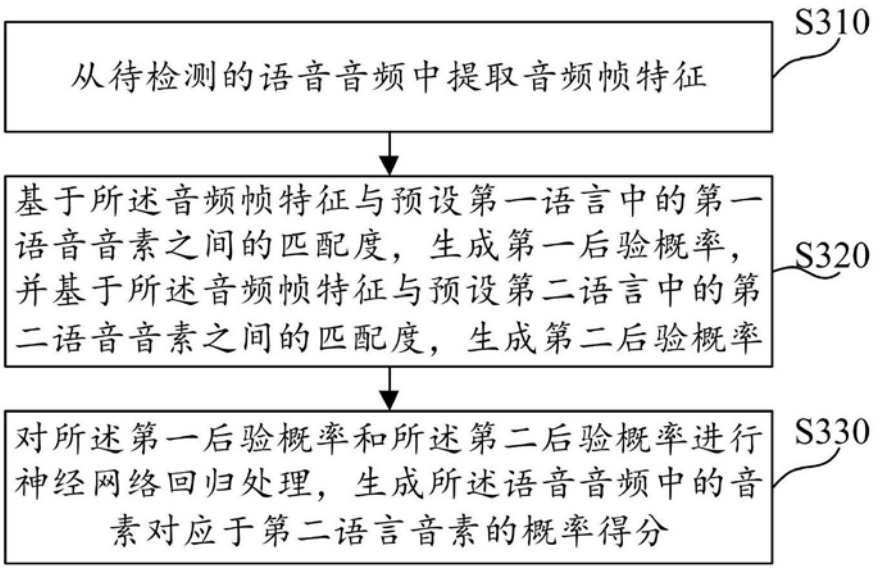

S310
从待检测的语音音频中提取音频帧特征
基于所述音频帧特征与预设第一语言中的第一语音音素之间的匹配度，生成第一后验概率，并基于所述音频帧特征与预设第二语言中的第二语音音素之间的匹配度，生成第二后验概率
S320
对所述第一后验概率和所述第二后验概率进行神经网络回归处理，生成所述语音音频中的音素对应于第二语言音素的概率得分
S330

图3


单词跟读1/2
hello
你好
点击开始跟读

图4

图5

图6

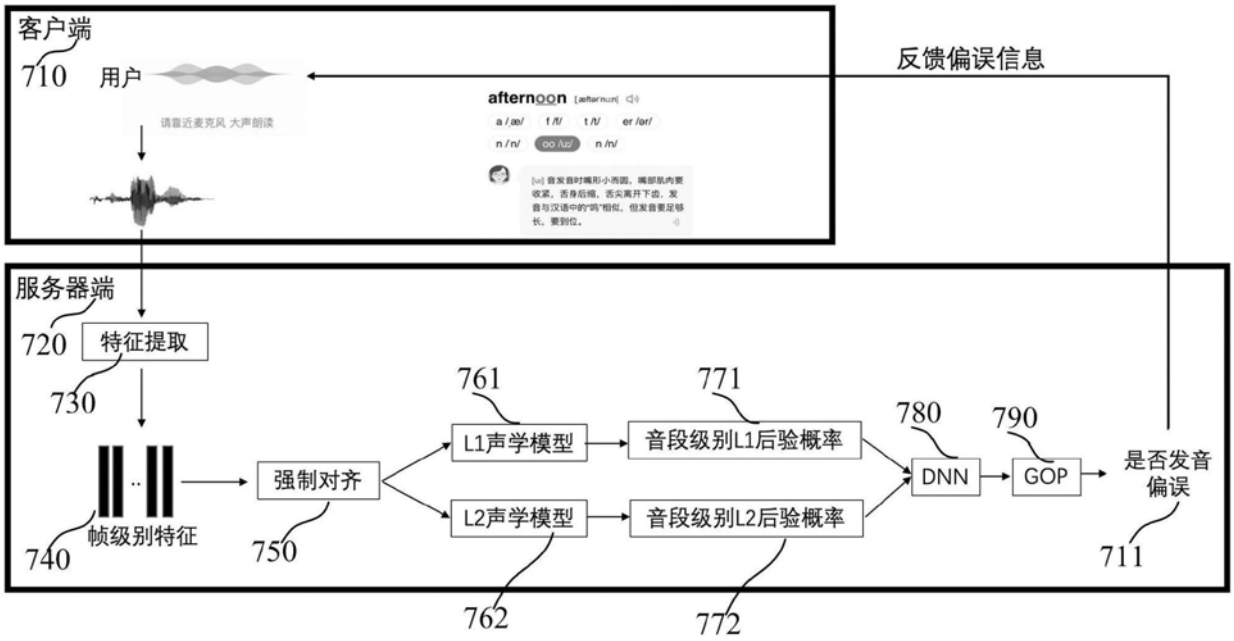

客户端
710
用户
请靠近麦克风 大声朗读
afternoon
反馈偏误信息
服务器端
720
特征提取
730
740
帧级别特征
强制对齐
750
761
L1声学模型
762
L2声学模型
771
音段级别L1后验概率
772
音段级别L2后验概率
780
DNN
790
GOP
是否发音偏误
711

图7


单词跟读1/2
Good afternoon.
下午好。
注意afternoon中[ər]的发音
重做
下一题

图8

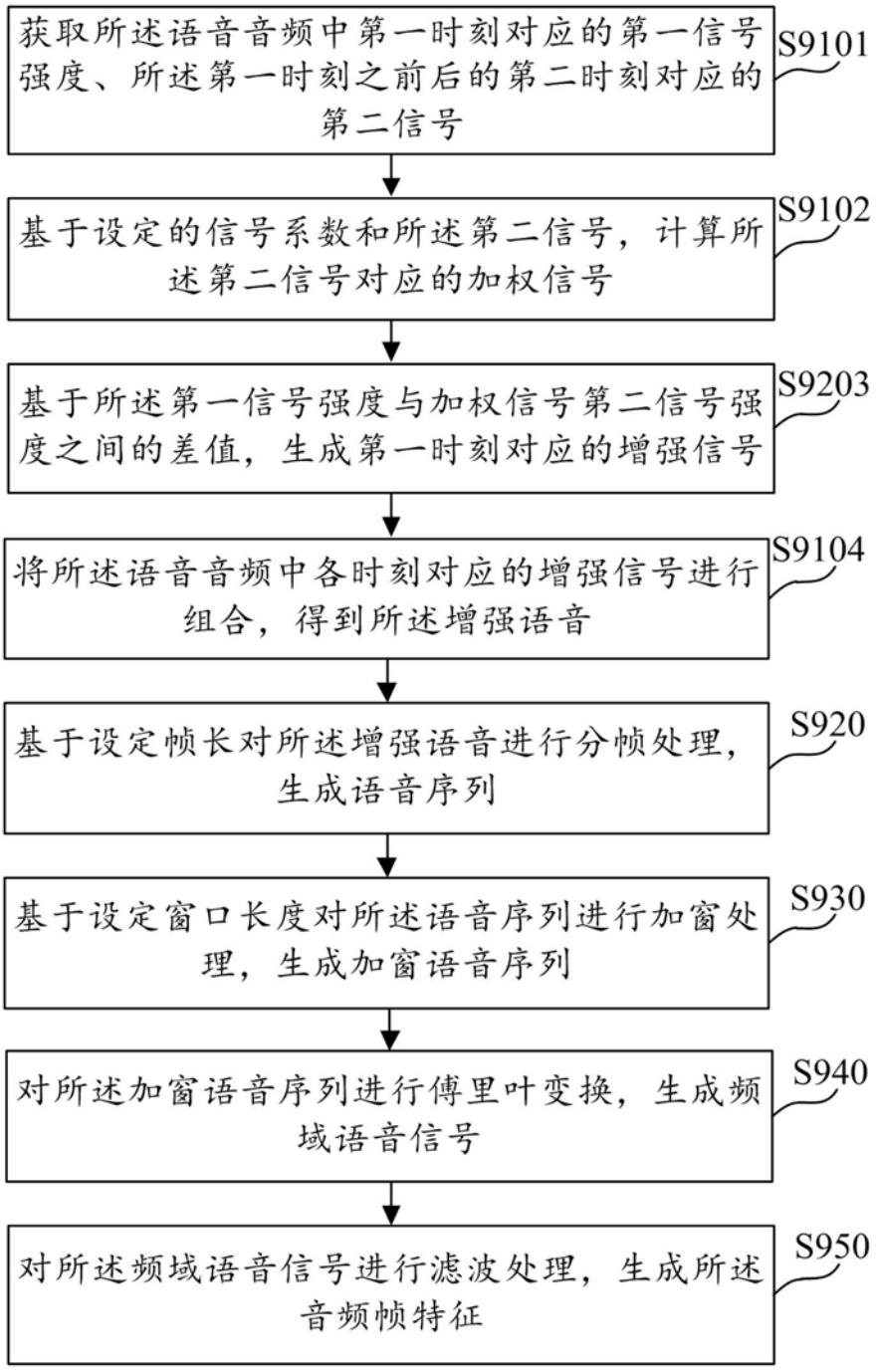

获取所述语音音频中第一时刻对应的第一信号强度、所述第一时刻之前后的第二时刻对应的第二信号
S9101
基于设定的信号系数和所述第二信号，计算所述第二信号对应的加权信号
S9102
基于所述第一信号强度与加权信号第二信号强度之间的差值，生成第一时刻对应的增强信号
S9203
将所述语音音频中各时刻对应的增强信号进行组合，得到所述增强语音
S9104
基于设定帧长对所述增强语音进行分帧处理，生成语音序列
S920
基于设定窗口长度对所述语音序列进行加窗处理，生成加窗语音序列
S930
对所述加窗语音序列进行傅里叶变换，生成频域语音信号
S940
对所述频域语音信号进行滤波处理，生成所述音频帧特征
S950

图9

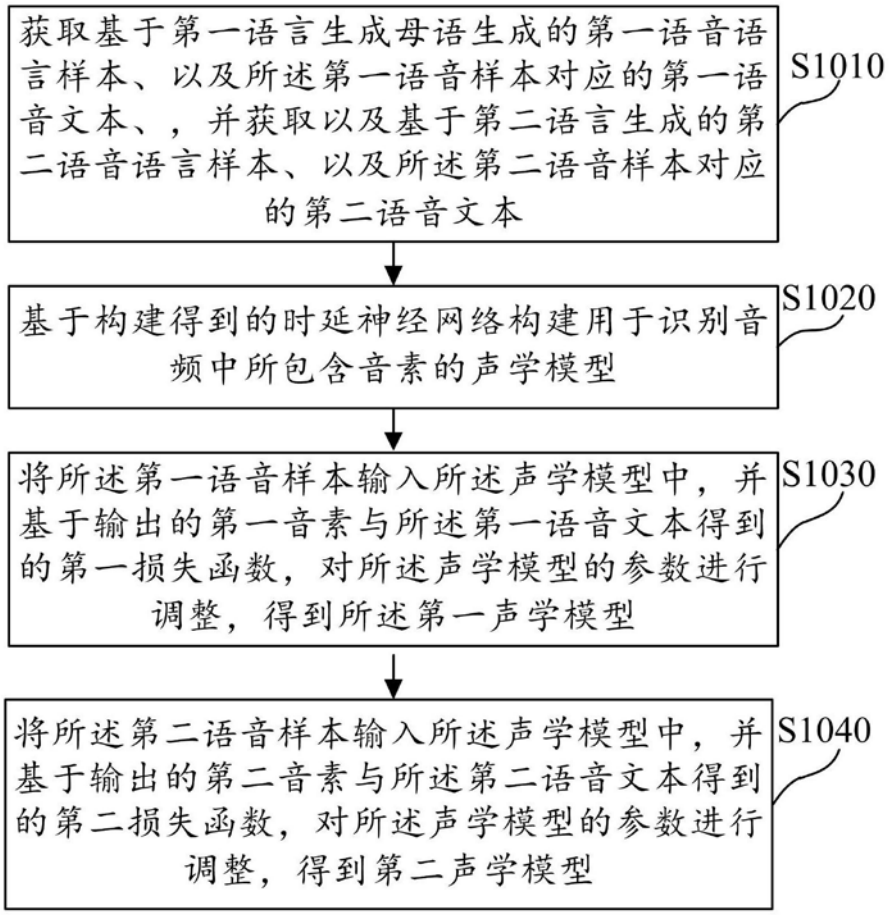

获取基于第一语言生成母语生成的第一语音语言样本、以及所述第一语音样本对应的第一语音文本、，并获取以及基于第二语言生成的第二语音语言样本、以及所述第二语音样本对应的第二语音文本
S1010
基于构建得到的时延神经网络构建用于识别音频中所包含音素的声学模型
S1020
将所述第一语音样本输入所述声学模型中，并基于输出的第一音素与所述第一语音文本得到的第一损失函数，对所述声学模型的参数进行调整，得到所述第一声学模型
S1030
将所述第二语音样本输入所述声学模型中，并基于输出的第二音素与所述第二语音文本得到的第二损失函数，对所述声学模型的参数进行调整，得到第二声学模型
S1040

图10

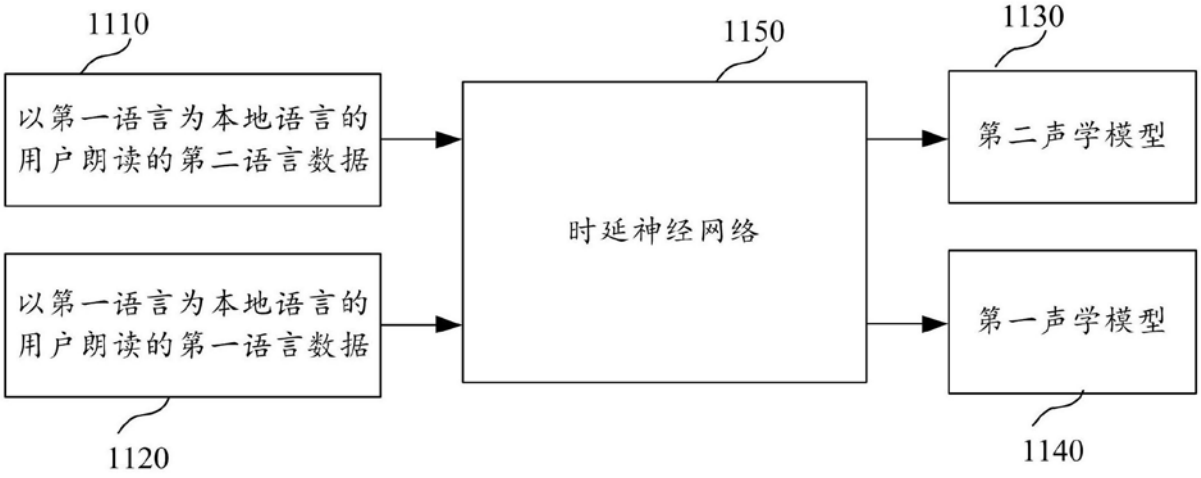

1110
以第一语言为本地语言的用户朗读的第二语言数据
1120
以第一语言为本地语言的用户朗读的第一语言数据
1150
时延神经网络
1130
第二声学模型
第一声学模型
1140

图11

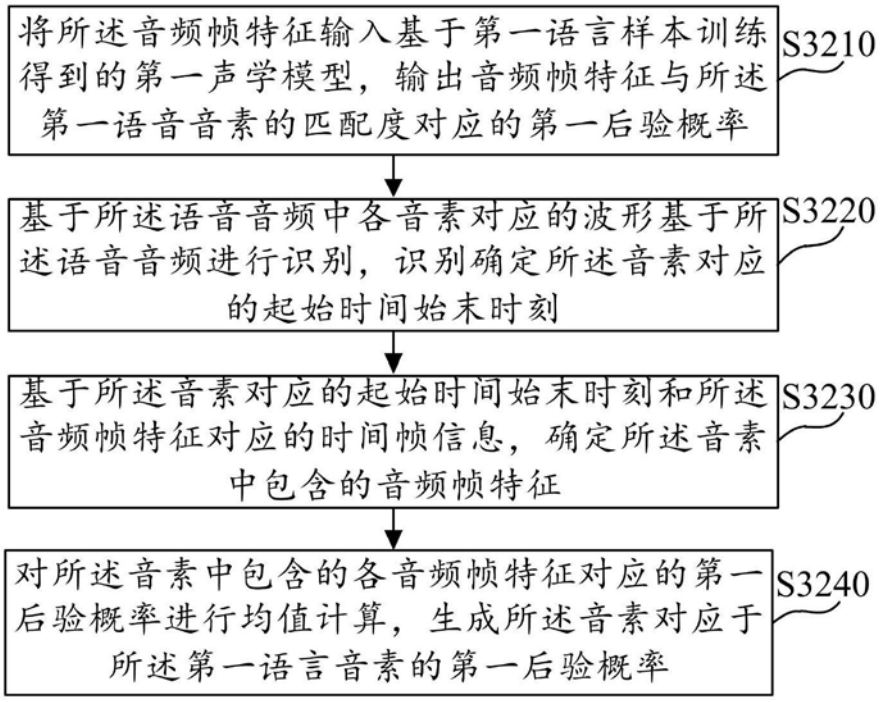

将所述音频帧特征输入基于第一语言样本训练得到的第一声学模型，输出音频帧特征与所述第一语音音素的匹配度对应的第一后验概率
S3210
基于所述语音音频中各音素对应的波形基于所述语音音频进行识别，识别确定所述音素对应的起始时间始末时刻
S3220
基于所述音素对应的起始时间始末时刻和所述音频帧特征对应的时间帧信息，确定所述音素中包含的音频帧特征
S3230
对所述音素中包含的各音频帧特征对应的第一后验概率进行均值计算，生成所述音素对应于所述第一语言音素的第一后验概率
S3240

图12

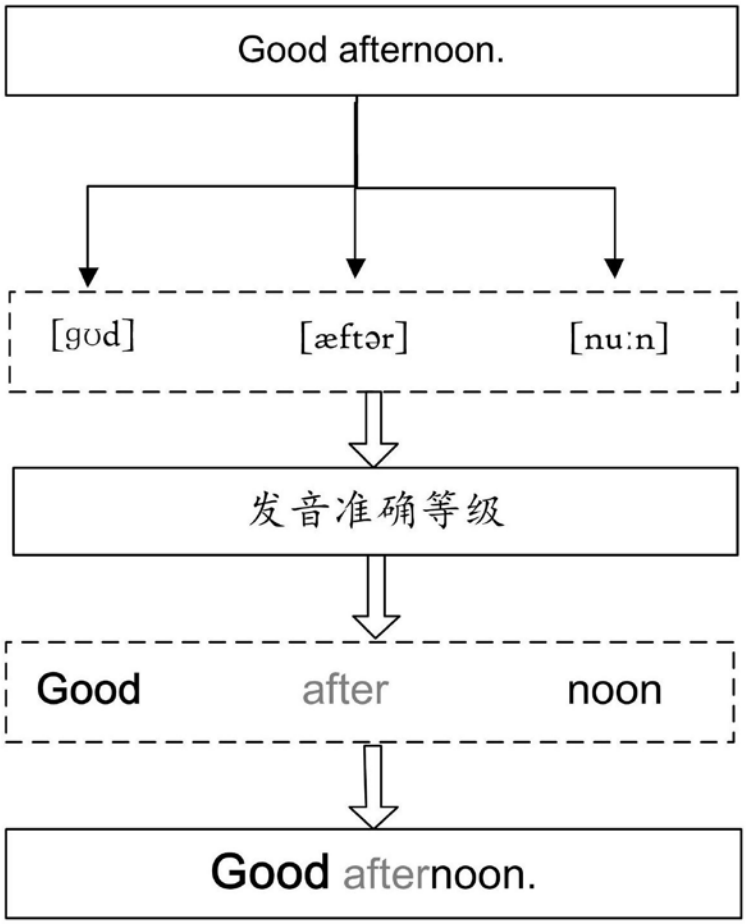

Good afternoon.
[gʊd]
[æftər]
[nu:n]
发音准确等级
Good
after
noon
Good afternoon.

图13


1410
单词跟读1/2
Good afternoon.
下午好。
注意afternoon中[ər]的发音
重做
下一题
1420
发音辅导
afternoon [æftəˈnuːn]
a /ˌæ/
f /f/
t /t/
er /ər/
n /ˈn/
oo /uː/
n /n/
[uː] 音发音时嘴形小而圆，嘴部肌肉要收紧，舌身后缩，舌尖离开下齿，发音与汉语中的"呜"相似，但发音要足够长，要到位。
以下是常发 [uː] 的字母及字母组合，我们巩固练习一下吧
oo
o
u
fool 傻的
[fuːl]

图14

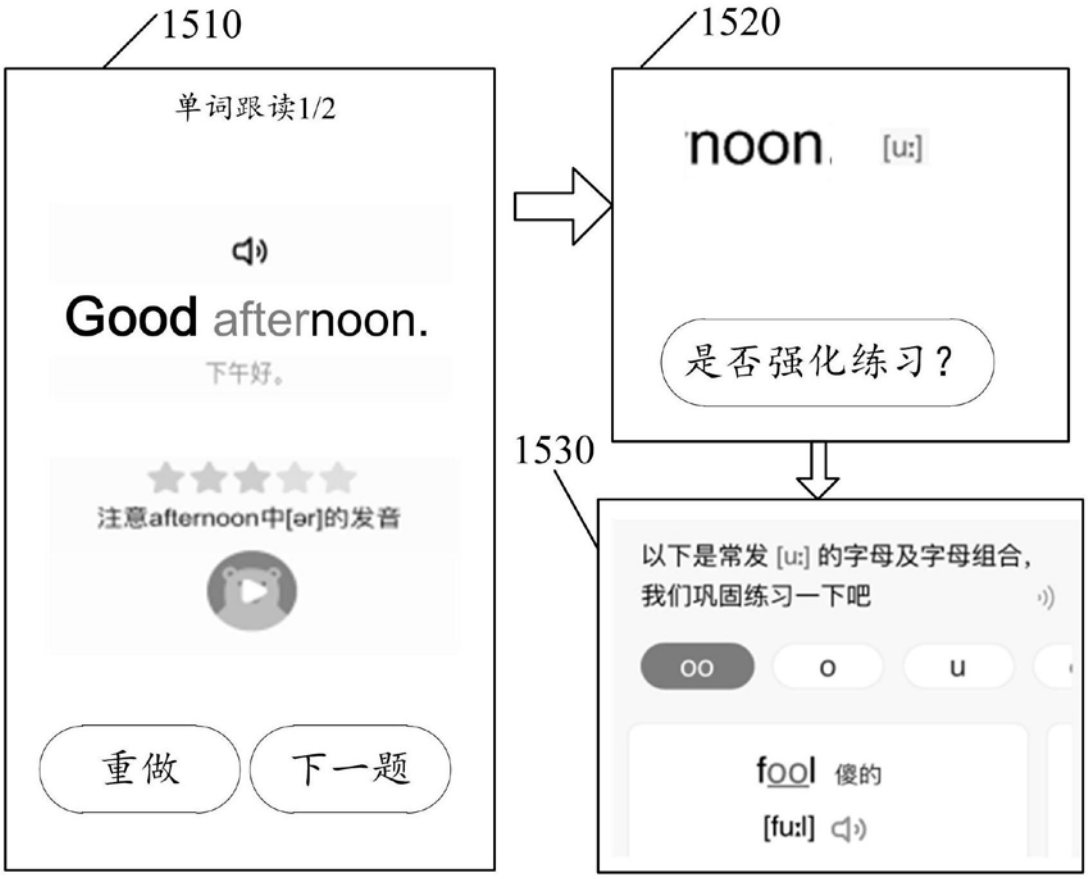

1510
单词跟读1/2
Good afternoon.
下午好。
注意afternoon中[ər]的发音
重做
下一题
1520
noon [u:]
是否强化练习？
1530
以下是常发 [u:] 的字母及字母组合，
我们巩固练习一下吧
oo
o
u
fool 傻的
[fu:l]

图15

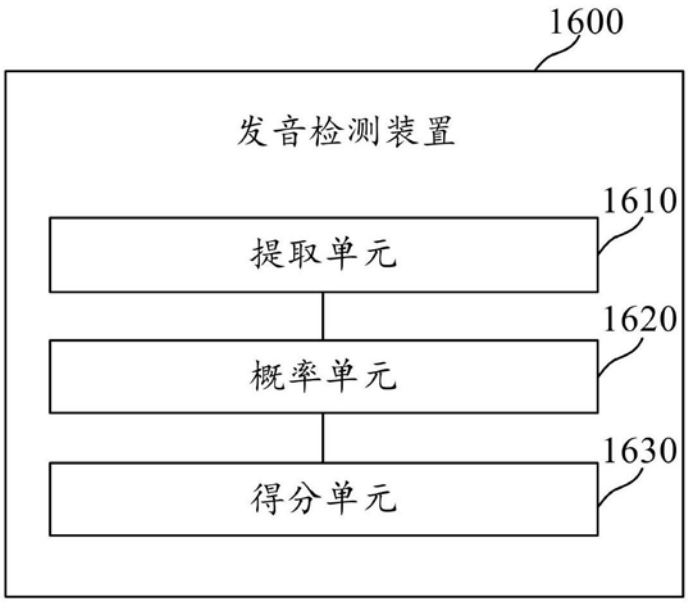

1600
发音检测装置
1610
提取单元
1620
概率单元
1630
得分单元

图16

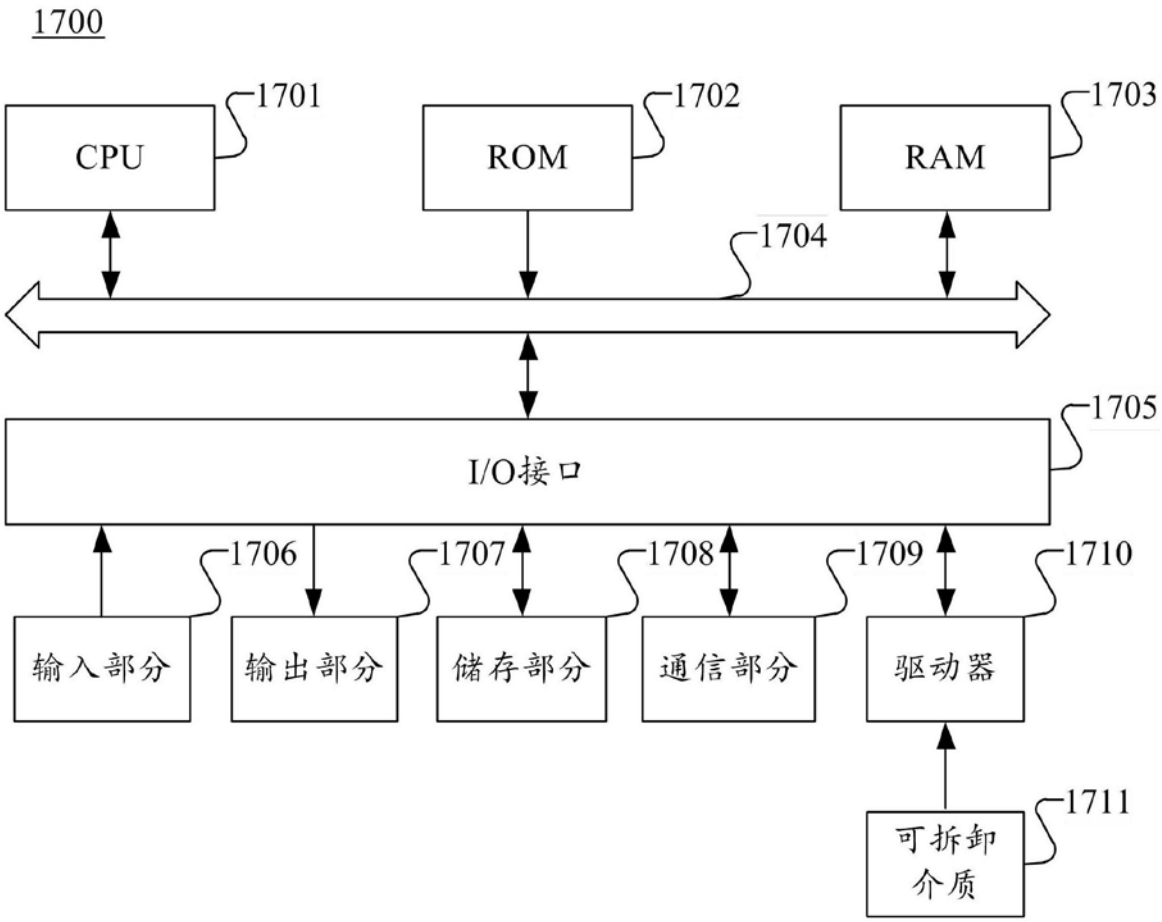

1700
CPU
1701
ROM
1702
RAM
1703
1704
I/O接口
1705
输入部分
1706
输出部分
1707
储存部分
1708
通信部分
1709
驱动器
1710
可拆卸介质
1711

图17