(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114360552 B
(45)授权公告日 2025.05.02

(21)申请号 202111495680.5
(22)申请日 2021.12.08
(65)同一申请的已公布的文献号
申请公布号 CN 114360552 A
(43)申请公布日 2022.04.15
(73)专利权人 深圳大学
地址 518060 广东省深圳市南山区南海大道3688号
(72)发明人 王佳 兰天浩 林秋镇 李坚强
(74)专利代理机构 深圳市君胜知识产权代理事务所(普通合伙) 44268
专利代理师 王娅洁
(51)Int.Cl.
G10L 17/04(2013.01)
G10L 17/02(2013.01)
G10L 17/14(2013.01)
G10L 15/02(2006.01)
G06F 18/214(2023.01)
G06F 18/20(2023.01)
G06F 18/241(2023.01)
G06N 3/047(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 112259104 A,2021.01.22
CN 112259105 A,2021.01.22

审查员 孟令鹏

权利要求书3页 说明书10页 附图2页

(54)发明名称

用于说话人识别的网络模型训练方法、装置及存储介质

(57)摘要

本发明涉及语音识别技术领域,具体是涉及用于说话人识别的网络模型训练方法、装置及存储介质。本发明首先将跨域的说话人样本数据集和音素样本数据集分别输入到一个多任务网络模型中,根据说话人分类子网和音素分类子网输出的结果,采用最大均值差异算法计算这两个结果之间的差异损失值,并作为总损失的一部分,通过增加了差异损失值的总损失不断去训练多任务网络模型,最终得到训练之后的模型,而训练之后的说话人子网模型对不同域的音素样本数据集具有较高的泛化能力,即训练之后的模型能够弱化跨域音素样本数据集与说话人样本数据集所具有的差异给网络模型识别说话人准确性所带来的影响。

CN 114360552 B

1.一种用于说话人识别的网络模型训练方法,其特征在于,包括:

将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练;

将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练;

对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值;

依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型;

所述依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

获取所述说话人样本数据集所对应的说话人样本身份标签;

依据所述说话人分类子网模型,得到训练的说话人嵌入模型;

依据所述说话人样本数据集和所述说话人嵌入模型,得到预测的说话人身份标签;

计算所述说话人样本身份标签和预测的所述说话人身份标签之间的身份标签差异;

获取所述音素样本数据集所对应的音素样本标签;

依据所述音素样本数据集和所述音素分类子网模型,得到预测的音素标签;

计算所述音素样本标签和预测的所述音素标签之间的音素差异;

将所述身份标签差异、所述差异损失值、所述音素差异作为新的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

2.如权利要求1所述的用于说话人识别的网络模型训练方法,其特征在于,所述依据所述身份标签差异、所述差异损失值、所述音素差异,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

将所述身份标签差异、所述差异损失值、所述音素差异进行加权计算,得到所述网络模型所对应的损失总值;

依据所述损失总值对所述网络模型进行训练,得到训练之后的所述网络模型。

3.如权利要求1所述的用于说话人识别的网络模型训练方法,其特征在于,所述依据所述身份标签差异、所述差异损失值、所述音素差异,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

依据两个所述音素分类子网模型,得到所述音素分类子网模型中的帧级音素分类子网络和段级音素分类子网络;

依据所述音素样本数据集和所述帧级音素分类子网络,得到预测的所述音素标签中的音素第一标签;

依据所述音素样本数据集和所述段级音素分类子网络,得到预测的所述音素标签中的音素第二标签;

计算所述音素差异中的所述音素样本标签和预测的所述音素第一标签之间的音素第一差异；

计算所述音素差异中的所述音素样本标签和预测的所述音素第二标签之间的音素第二差异；

依据所述身份标签差异、所述差异损失值、所述音素第一差异、所述音素第二差异，对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练，得到训练之后的所述网络模型。

4.如权利要求1所述的用于说话人识别的网络模型训练方法，其特征在于，还包括：

获取所述说话人样本数据集所对应的说话人样本身份标签；

获取所述音素样本数据集所对应的音素样本标签；

依据所述说话人样本数据集和训练之后的所述说话人分类子网模型，得到预测的说话人身份标签；

依据所述音素样本数据集和训练之后的所述音素分类子网模型，得到预测的所述音素标签；

将所述说话人样本身份标签相异于预测的所述说话人身份标签所对应的说话人样本，记为说话人误分类样本；

对所述说话人样本数据集应用误分类向量指导损失函数，得到新的说话人的损失函数值；

依据新的说话人的所述损失函数值，更新训练之后的多任务网络总损失。

5.如权利要求1所述的用于说话人识别的网络模型训练方法，其特征在于，所述音素分类子网模型为用于识别音素标签的音素神经网络模型，所述说话人分类子网模型为用于识别说话人身份的说话人神经网络模型，所述音素分类子网模型中的设定层为位于所述音素神经网络模型用于输出音素标签所在层的上一层，所述说话人分类子网模型中的设定层为位于所述说话人神经网络模型用于输出说话人身份标签所在层的上一层。

6.一种用于说话人识别的网络模型训练方法的装置，其特征在于，所述装置包括如下组成部分：

第一结果计算模块，用于将说话人样本数据集输入到多任务网络模型中，提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果，所述说话人分类子网模型用于说话人分类训练；

第二结果计算模块，用于将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中，提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果，所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应，所述音素分类子网模型用于音素分类训练；

差异损失值计算模块，用于对所述第一结果和所述第二结果应用最大均值差异算法，得到所述第一结果和所述第二结果所对应的差异损失值；

模型训练模块，用于依据添加了所述差异损失值的总损失，对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练，得到训练之后的所述网络模型；

所述依据添加了所述差异损失值的总损失，对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练，得到训练之后的所述网络模型，包括：

获取所述说话人样本数据集所对应的说话人样本身份标签；

依据所述说话人分类子网模型，得到训练的说话人嵌入模型；

依据所述说话人样本数据集和所述说话人嵌入模型，得到预测的说话人身份标签；

计算所述说话人样本身份标签和预测的所述说话人身份标签之间的身份标签差异；

获取所述音素样本数据集所对应的音素样本标签；

依据所述音素样本数据集和所述音素分类子网模型，得到预测的音素标签；

计算所述音素样本标签和预测的所述音素标签之间的音素差异；

将所述身份标签差异、所述差异损失值、所述音素差异作为新的总损失，对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练，得到训练之后的所述网络模型。

7.一种终端设备，其特征在于，所述终端设备包括存储器、处理器及存储在所述存储器中并可在所述处理器上运行的用于说话人识别的网络模型训练程序，所述处理器执行所述用于说话人识别的网络模型训练程序时，实现如权利要求1-5任一项所述的用于说话人识别的网络模型训练方法的步骤。

8.一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上存储有用于说话人识别的网络模型训练程序，所述用于说话人识别的网络模型训练程序被处理器执行时，实现如权利要求1-5任一项所述的用于说话人识别的网络模型训练方法的步骤。

用于说话人识别的网络模型训练方法、装置及存储介质

技术领域

[0001] 本发明涉及语音识别技术领域,具体是涉及用于说话人识别的网络模型训练方法、装置及存储介质。

背景技术

[0002] 说话人识别是验证输入话语(语音信号)是否属于特定说话人的任务。语音信号由说话人情感、口音(音素)和语言等多种内在成分组成。这些内在因素的不确定性,尤其是语音内容的不确定性,会影响系统的识别性能。因此,说话人嵌入的提取不能只考虑说话人标签。受说话人自适应技术在自动语音识别(ASR)中应用的启发,多任务学习(MTL)策略被提出来学习包含在多个相关任务中的语音信息,以帮助提高主任务(说话人识别)的泛化能力。近些年,对抗学习等策略被引入到多任务框架中,可以在语音信息使用方面发挥两者的优势。许多研究表明在帧级鼓励音素信息与段级抑制音素信息都是有效的。然而在许多实际情况下,获得同时具有说话人标签和音素标签的域内理想数据集是非常昂贵且不灵活的。而当在音素辨别子网(音素辨别子网用于辅助说话人网络模型对说话人进行分类训练)中引入跨域ASR数据集或跨语言ASR数据集时,这些方法通常不会为说话人辨别子网(说话人网络模型)提供更多有用的信息。尤其是进行小语种说话人识别时,生成新的人工转录音素标签会导致更长的训练时间。因此,由于用于识别说话人的音素感知网络模型不适用于来源不同的音素数据集和说话人数据集所带来的差异,即用于进行说话人分类训练的音素感知的网络模型的泛化能力较差。

[0003] 综上所述,现有的用于进行说话人分类训练的音素感知网络模型的泛化能力较差。

[0004] 因此,现有技术还有待改进和提高。

发明内容

[0005] 为解决上述技术问题,本发明提供了用于说话人识别的网络模型训练方法、装置及存储介质,解决了现有的用于进行说话人分类训练的音素感知网络模型的泛化能力较差的问题。

[0006] 为实现上述目的,本发明采用了以下技术方案:

[0007] 第一方面,本发明提供一种用于说话人识别的网络模型训练方法,其中,包括:

[0008] 将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练;

[0009] 将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练;

[0010] 对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值;

[0011] 依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0012] 在一种实现方式中,所述依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

[0013] 获取所述说话人样本数据集所对应的说话人样本身份标签;

[0014] 依据所述说话人子网模型,得到训练的说话人嵌入模型;

[0015] 依据所述说话人样本数据集和所述说话人嵌入模型,得到预测的说话人身份标签;

[0016] 计算所述说话人样本身份标签和所述说话人身份标签之间的身份标签差异;

[0017] 获取所述音素样本数据集所对应的音素样本标签;

[0018] 依据所述音素样本数据集和所述音素分类子网模型,得到预测的所述音素标签;

[0019] 计算所述音素样本标签和预测的所述音素标签之间的音素差异;

[0020] 将所述身份标签差异、所述差异损失值、所述音素差异作为新的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0021] 在一种实现方式中,所述依据所述身份标签差异、所述差异损失值、所述音素差异,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

[0022] 将所述身份标签差异、所述差异损失值、所述音素差异进行加权计算,得到所述网络模型所对应的损失总值;

[0023] 依据所述损失总值对所述网络模型进行训练,得到训练之后的所述网络模型。

[0024] 在一种实现方式中,所述依据所述身份标签差异、所述差异损失值、所述音素差异,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型,包括:

[0025] 依据所述音素分类子网模型,得到所述音素分类子网模型中的帧级音素分类子网络和段级音素分类子网络;

[0026] 依据所述音素样本数据集和所述帧级音素分类子网络,得到预测的所述音素标签中的音素第一标签;

[0027] 依据所述音素样本数据集和所述段级音素分类子网络,得到预测的所述音素标签中的音素第二标签;

[0028] 计算所述音素差异中的所述音素样本标签和所述音素第一标签之间的音素第一差异;

[0029] 计算所述音素差异中的所述音素样本标签和所述音素第二标签之间的音素第二差异;

[0030] 依据所述身份标签差异、所述差异损失值、所述音素第一差异、所述音素第二差异,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得

到训练之后的所述网络模型。

[0031] 在一种实现方式中,还包括:

[0032] 获取所述说话人样本数据集所对应的说话人样本身份标签;

[0033] 获取所述音素样本数据集所对应的音素样本标签;

[0034] 依据所述说话人样本数据集和训练之后的所述说话人分类子网模型,得到预测的说话人身份标签;

[0035] 依据所述音素样本数据集和训练之后的所述音素分类子网模型,得到所述音素标签;

[0036] 将所述说话人样本身份标签相异于所述说话人身份标签所对应的说话人样本,记为说话人误分类样本;

[0037] 对所述说话人样本数据集应用误分类向量指导损失函数,得到新的说话人的损失函数值;

[0038] 依据新的说话人的所述损失函数值,更新训练之后的多任务网络总损失。

[0039] 在一种实现方式中,所述音素分类子网模型为用于识别音素标签的音素神经网络模型,所述说话人分类子网模型为用于识别说话人身份的说话人神经网络模型,所述音素分类子网模型中的设定层为位于所述音素神经网络模型用于输出音素标签所在层的上一层,所述说话人分类子网模型中的设定层为位于所述说话人神经网络模型用于输出说话人身份标签所在层的上一层。

[0040] 第二方面,本发明实施例还提供一种用于说话人识别的网络模型训练方法的装置,其中,所述装置包括如下组成部分:

[0041] 第一结果计算模块,用于将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练;

[0042] 第二结果计算模块,用于将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练;

[0043] 差异损失值计算模块,用于对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值;

[0044] 模型训练模块,用于依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0045] 第三方面,本发明实施例还提供一种终端设备,其中,所述终端设备包括存储器、处理器及存储在所述存储器中并可在所述处理器上运行的用于说话人识别的网络模型训练程序,所述处理器执行所述用于说话人识别的网络模型训练程序时,实现上述所述的用于说话人识别的网络模型训练方法的步骤。

[0046] 第四方面,本发明实施例还提供一种计算机可读存储介质,所述计算机可读存储介质上存储有用于说话人识别的网络模型训练程序,所述用于说话人识别的网络模型训练程序被处理器执行时,实现上述所述的用于说话人识别的网络模型训练方法的步骤。

[0047] 有益效果:本发明首先将说话人样本数据集和音素样本数据集分别输入到多任务音素网络模型中,根据说话人分类子网模型和音素分类子网模型输出的结果,采用最大均值差异算法计算这两个结果之间的差异损失值,通过添加了的差异损失值的总损失不断去训练音素感知模型,最终得到训练之后的模型,而训练之后的音素感知模型对不同域的音素样本数据集具有较高的泛化能力,即训练之后的模型能够弱化不同域的音素样本数据集与说话人样本数据集所具有的差异给网络模型识别说话人准确性所带来的影响。

附图说明

[0048] 图1为本发明的整体流程图;

[0049] 图2为本发明的包括三个子模块的网络模型;

[0050] 图3为本发明的包括四个子模块的网络模型;

[0051] 图4为本发明实施例提供的终端设备的内部结构原理框图。

具体实施方式

[0052] 以下结合实施例和说明书附图,对本发明中的技术方案进行清楚、完整地描述。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0053] 经研究发现,说话人识别是验证输入话语(语音信号)是否属于特定说话人的任务。语音信号由说话人情感、口音(音素)和语言等多种内在成分组成。这些内在因素的不确定性,尤其是语音内容的不确定性,会影响系统的识别性能。因此,说话人嵌入的提取不能只考虑说话人标签。受说话人自适应技术在自动语音识别(ASR)中应用的启发,多任务学习(MTL)策略被提出来学习包含在多个相关任务中的语音信息,以帮助提高主任务(说话人识别)的泛化能力。近些年,对抗学习等策略被引入到多任务框架中,可以在语音信息使用方面发挥两者的优势。许多研究表明在帧级鼓励音素信息与段级抑制音素信息都是有效的。然而在许多实际情况下,获得同时具有说话人标签和音素标签的理想数据集是非常昂贵且不灵活的。而当在音素辨别子网(音素辨别子网用于辅助说话人分类子网模型对说话人分类训练)中引入跨域ASR数据集或跨语言ASR数据集时,这些方法通常不会为说话人辨别子网(说话人子网模型)提供更多有用的信息。尤其是进行小语种说话人识别时,生成新的人工转录音素标签会导致更长的训练时间。因此,由于用于说话人分类训练的音素感知网络模型不适用于来源不同的音素数据集和说话人数据集所带来的差异,即用于进行说话人分类训练的音素感知网络模型的泛化能力较差。

[0054] 为解决上述技术问题,本发明提供了用于说话人识别的网络模型训练方法、装置及存储介质,解决了现有的用于进行说话人分类训练的音素感知网络模型的泛化能力较差的问题。具体实施时,本发明首先将说话人样本数据集和音素样本数据集分别输入到多任务音素网络模型中,根据说话人分类子网模型和音素分类子网模型输出的结果,采用最大均值差异算法计算这两个结果之间的差异损失值,通过添加了差异损失值的总损失不断去训练音素感知模型,最终得到训练之后的模型。本发明能够提高训练之后的模型的泛化能力。

[0055] 举例说明,说话人样本数据集来源于A域,音素样本数据集来源于B域。说话人样本

数据集和音素样本数据集来源于不同的域(比如这两个数据集来源于不同的数据库,不同的数据库的数据分布是有差异的,或者数据编码方式也是有差异的),使得说话人样本数据集和音素样本数据集存在差异,而这种差异的存在会导致由说话人分类子网模型和音素分类子网模型所在的网络模型不能利用音素信息去提高说话人分类训练性能,而现有的网络模型不能适应于这种差异。本实施例为了使得网络模型能够适应于这种差异,采用最大均值差异算法计算两种子网模型所输出的结果所对应的差异损失值,依据添加了差异损失值的总损失去训练网络模型,使得训练之后的网络模型能够适应于来源于不同域的说话人样本数据集和音素样本数据集。

[0056] 示例性方法

[0057] 本实施例的一种用于说话人识别的网络模型训练方法可应用于终端设备中,所述终端设备可为具有计算的终端产品,比如电脑等。在本实施例中,如图1中所示,所述用于说话人识别的网络模型训练方法具体包括如下步骤:

[0058] S100,将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练。

[0059] 本实施例中,说话人分类子网模型为说话人嵌入模型,是一种神经网络模型M_s。神经网络模型M_s如图2所示,本实施例中M_s的设定层位于用于输出说话人标签所在层的上一层,本实施例中设定层为M_s中的第七层。

[0060] S200,将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练。

[0061] 本实施例中,音素分类子网模型是一种神经网络模型M_p。神经网络模型M_p如图2所示,本实施中,M_p的设定层为位于用于输出音素标签所在层的上一层,本实施例的设定层为M_p的第七层。

[0062] S300,对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值。

[0063] 本实施例,利用最大均值差异算法(MMD)计算差异损失值L_{mmd}的原理如下:

[0064] 假设 $D_s=\{x_i^s\}_{i=1}^n$ 和 $D_p=\{x_j^p\}_{j=1}^m$ 是分布S和P的样本集,采用如下的公式计算S和P的差异损失值:

[0065] $$MMD\left(D_s,D_p\right)=\left\|\frac{1}{n}\sum_{i=1}^{n}\phi\left(x_i^s\right)-\frac{1}{m}\sum_{j=1}^{m}\phi\left(x_j^p\right)\right\|_H^2 \quad (1)$$

[0066] 其中H表示特征空间即再生希尔伯特空间,φ(•)表示映射函数。因为高斯核函数可以映射无限维空间,引用高斯核函数 $k\left(x,y\right)=\exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$ 来表示映射函数内积,其中σ是带宽参数,用于控制径向作用范围。将引入高斯核函数后的原公式拆开为:

[0067] $$MMD\left(D_s,D_p\right)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k\left(x_i^s,x_j^s\right)-\frac{2}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}k\left(x_i^s,x_j^p\right)+\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k\left(x_i^p,x_j^p\right) \quad (2)$$

[0068] 当采用公式(2)计算本实施例的说话人分类子网模型第七层输出的第一结果与音素分类子网模型第七层输出的第二结果所对应的差异损失值L_{mmd}时,只需要将公式(2)中的$D_s=\{x_i^s\}_{i=1}^n$替换成第一结果,将$D_p=\{x_j^p\}_{j=1}^m$替换成第二结果即可。

[0069] S400,依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0070] 依据差异损失值L_{mmd}、身份标签差异L_s、音素差异L_p这三者共同训练网络模型。

[0071] 本实施例依据这三者的加权和所对应的损失总值L_{total}(损失总值L_{total}即总损失)进行反向传播来训练网络模型。

[0072] $L_{total}=L_s+\alpha\cdot L_p+\beta\cdot L_{mmd}$ (3)

[0073] 式中,α表示音素子网权重,β表示最大均值差异损失权重。

[0074] 计算身份标签差异L_s时,包括如下步骤S401、S402、S403、S404:

[0075] S401,获取所述说话人样本数据集所对应的说话人样本身份标签。

[0076] S402,依据所述说话人子网模型,得到训练的说话人嵌入模型。

[0077] S403,依据所述说话人样本数据集和所述说话人嵌入模型,得到预测的说话人身份标签。

[0078] S404,计算所述说话人样本身份标签和预测的所述说话人身份标签之间的身份标签差异。

[0079] 音素差异L_p时,包括如下步骤S405、S406、S407:

[0080] S405,获取所述音素样本数据集所对应的音素样本标签。

[0081] S406,依据所述音素样本数据集和所述音素分类子网模型,得到预测的所述音素标签。

[0082] S407,计算所述音素样本标签和预测的所述音素标签之间的音素差异。

[0083] 下面介绍L_s和L_p的详细计算过程:

[0084] 本实施例的网络模型除了包括说话人分类子网模型和音素分类子网模型,还包括共享帧级网络模块M_f,其中共享帧级网络模块M_f用于挖掘说话人信息与音素信息的共同特征。

[0085] 如图2所示,给定一个训练段对[X^s,X^p]和相对应的说话人身份标签y^s和音素样本标签y^p,s用于表示说话人,p用于表示音素,其中$X^s=[x_1^s,x_2^s,\cdots,x_m^s]$由M个帧组成,$x_i^s$为$X^s$中第i帧,$X^p=[x_1^p,x_2^p,\cdots,x_n^p]$由N个帧组成。

[0086] $L_s=CE(M_s(M_f(X^s)),y^s)$ (4)

[0087] $$L_p=\frac{1}{N}\sum_{i=1}^{N}CE\left(M_p\left(M_f\left(x_i^p\right)\right),y_i^p\right) \quad (5)$$

[0088] 式中,CE代表交叉熵损失函数(Coss Entropy loss)。

[0089] 为了提升整个网络模型的泛化能力,本实施例将音素分类子网模型划分为如图3所示的帧级音素分类子网络M_{ps}和段级音素分类子网络M_{pf},则公式(3)变换为公式(6):

[0090] $L_{total}=L_s+\alpha\cdot L_{pf}+\beta\cdot L_{ps}+\gamma\cdot L_{mmd}$ (6)

[0091] 式中,α表示帧级音素子网权重,β表示段级音素子网权重,γ表示说话人子网与帧

级音素子网的最大均值差异损失权重,L_{pf}为音素第二差异,L_{ps}为音素第一差异,计算L_{pf}和L_{ps}的具体过程包括:依据两个所述音素分类子网模型,得到所述音素分类子网模型中的帧级音素分类子网络M_{ps}和段级音素分类子网络M_{pf};依据所述音素样本数据集和所述帧级音素分类子网络,得到预测的所述音素标签中的音素第一标签;依据所述音素样本数据集和所述段级音素分类子网络,得到预测的所述音素标签中的音素第二标签;计算所述音素差异中的所述音素样本标签和所述音素第一标签之间的音素第一差异L_{ps};计算所述音素差异中的所述音素样本标签和所述音素第二标签之间的音素第二差异L_{pf}。

[0092] 通过步骤S100、S200、S300、S400得到训练之后的网络模型,本实施例还对训练之后的网络模型进行更新,更新训练之后的网络模型包括如下步骤S501、S502、S503、S504、S505、S506、S507:

[0093] S501,获取所述说话人样本数据集所对应的说话人样本身份标签。

[0094] 说话人样本身份标签就是预先给说话人设定的身份标签,本实施例中的身份标签可以类比说话人身份证号,因此本实施例中的身份标签是唯一的标签。

[0095] S502,获取所述音素样本数据集所对应的音素样本标签。

[0096] 音素样本标签就是预先给音素设定的标签,本实施例的音素标签可以类比每个单词的音标。

[0097] S503,依据所述说话人样本数据集和训练之后的所述说话人分类子网模型,得到预测的说话人身份标签。

[0098] 说话人身份标签就是训练之后的说话人分类子网模型根据说话人样本数据集而得到的标签。

[0099] S504,依据所述音素样本数据集和训练之后的所述音素分类子网模型,得到所述音素标签。

[0100] S505,将所述说话人样本身份标签相异于所述说话人身份标签所对应的说话人样本,记为说话人误分类样本。

[0101] 如果一个说话人样本数据集所对应的说话人样本身份标签为A,但是在训练过程的说话人分类子网模型对说话人样本数据集进行识别后,得到预测的说话人身份标签为B,A和B不同,就将该说话人样本数据集记为说话人误分类样本。

[0102] S506,对所述说话人误分类样本应用误分类向量指导损失函数,得到损失函数值L_{mv}。

[0103] S507,依据新的说话人所述损失函数值,更新总损失,训练所述网络模型。

[0104] 本实施例中,计算损失函数值L_{mv}的原理如下:

[0105] 在说话人领域应用广泛的margin-based softmax存在两个问题:1)这些损失函数没有考虑训练中误分类带来的困难样本的重要性;2)类之间的间隔在训练中不能自适应改变。本实施例引入误分类向量指导的损失函数(mv-softmax),公式如下:

[0106]
$$L_{mv}=-\log\frac{e^{sf(m,\theta_{\omega_y,x})}}{e^{sf(m,\theta_{\omega_y,x})}+\sum_{k\neq y}^{K}h\left(t,\theta_{\omega k,x},I_k\right)e^{s\cos(\theta_{\omega k,x})}} \tag{7}$$

[0107] 式中I_k的公式如下:

[0108] $$I_k=\begin{cases}0, f(m,\theta_{\omega y,x})-\cos(\theta_{\omega k,x})\geq 0\\ 1, f(m,\theta_{\omega y,x})-\cos(\theta_{\omega k,x})\leq 0\end{cases} \quad (8)$$

[0109] 当I_k=1时,表示该样本当前是困难样本,我们会重点强调此错误分类向量:

[0110] $$h(t,\theta_{\omega k,x},I_k)=e^{st(\cos(\theta_{\omega i,x})+1)I_k} \quad (9)$$

[0111] 其中t≥0,是一个超参,f(•)不同的margin-based softmax的距离公式,m表示设置的参数间隔margin,ω_y为对应y类的权重,x为学习到的特征向量。在加入噪声和混响数据扩充后的数据集Voxceleb2上的实验表明,在说话人识别领域,同时重视误分类样本权重和间隔自适应的mv-softmax较am-softmax的EER降低了5.5%。

[0112] 综上,本发明首先将说话人样本数据集和音素样本数据集分别输入到多任务音素感知网络模型中,根据说话人分类子网模型和音素分类子网模型输出的结果,采用最大均值差异算法计算这两个结果之间的差异损失值,通过添加了差异损失值总损失不断去训练音素感知模型,最终得到训练之后的模型,而训练之后的说话人分类子网模型对不同域的音素样本数据集具有较高的泛化能力,即训练之后的模型能够弱化跨域的音素样本数据集与说话人样本数据集所具有的差异给网络模型识别说话人准确性所带来的影响。

[0113] 另外,本发明首次对音素感知网络应用最大均值差异(MMD)的差异最小化,提高说话人子网在不同域的音素信息情况下产生的说话人嵌入的泛化能力。最大均值差异(MMD)用于在再生希尔伯特空间中度量两个分布的距离。

[0114] 误分类向量指导的损失函数mv-softmax首次将样本间隔损失(margin-based softmax)和困难样本挖掘损失(mining-base softmax)的优点结合在一起,本发明首次将它应用到说话人识别网络中,能够充分利用训练阶段的误分类样本信息,并且使样本间的间隔能够在训练阶段自适应改变。

[0115] 示例性装置

[0116] 本实施例还提供一种用于说话人识别的网络模型训练方法的装置,所述装置包括如下组成部分:

[0117] 第一结果计算模块,用于将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练;

[0118] 第二结果计算模块,用于将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练;

[0119] 差异损失值计算模块,用于对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值;

[0120] 模型训练模块,用于依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0121] 基于上述实施例,本发明还提供了一种终端设备,其原理框图可以如图4所示。该终端设备包括通过系统总线连接的处理器、存储器、网络接口、显示屏、温度传感器。其中,

该终端设备的处理器用于提供计算和控制能力。该终端设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统和计算机程序。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该终端设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种用于说话人识别的网络模型训练方法。该终端设备的显示屏可以是液晶显示屏或者电子墨水显示屏,该终端设备的温度传感器是预先在终端设备内部设置,用于检测内部设备的运行温度。

[0122] 本领域技术人员可以理解,图4中示出的原理框图,仅仅是与本发明方案相关的部分结构的框图,并不构成对本发明方案所应用于其上的终端设备的限定,具体的终端设备以包括比图中所示更多或更少的部件,或者组合某些部件,或者具有不同的部件布置。

[0123] 在一个实施例中,提供了一种终端设备,终端设备包括存储器、处理器及存储在存储器中并可在处理器上运行的用于说话人识别的网络模型训练程序,处理器执行用于说话人识别的网络模型训练程序时,实现如下操作指令:

[0124] 将说话人样本数据集输入到多任务网络模型中,提取所述多任务网络模型中的说话人分类子网模型中设定层所输出的第一结果,所述说话人分类子网模型用于说话人分类训练;

[0125] 将与所述说话人样本数据集所对应的跨域的音素样本数据集输入到所述多任务网络模型中,提取所述多任务网络中的音素分类子网模型中设定层所输出的第二结果,所述音素分类子网模型中的设定层与所述说话人分类子网模型中的设定层相对应,所述音素分类子网模型用于音素分类训练;

[0126] 对所述第一结果和所述第二结果应用最大均值差异算法,得到所述第一结果和所述第二结果所对应的差异损失值;

[0127] 依据添加了所述差异损失值的总损失,对说话人分类子网模型和音素分类子网模型所在的所述多任务网络模型进行训练,得到训练之后的所述网络模型。

[0128] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一非易失性计算机可读取存储介质中,该计算机程序在执行时,可包括如上述各方法的实施例的流程。其中,本发明所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用,均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0129] 综上,本发明公开了用于说话人识别的网络模型训练方法、装置及存储介质,所述方法包括:首先将说话人样本数据集和音素样本数据集分别输入到多任务音素网络模型中,根据说话人分类子网模型和音素分类子网模型输出的结果,采用最大均值差异算法计算这两个结果之间的差异损失值,通过添加了差异损失值的总损失不断去训练音素感知模型,最终得到训练之后的模型。本发明训练之后的说话人子网模型对跨域的音素样本数据集具有较高的泛化能力,即训练之后的模型能够弱化不同域的音素样本数据集与说话人样

本数据集所具有的差异给网络模型识别说话人准确性所带来的影响。

[0130] 最后应说明的是:以上实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

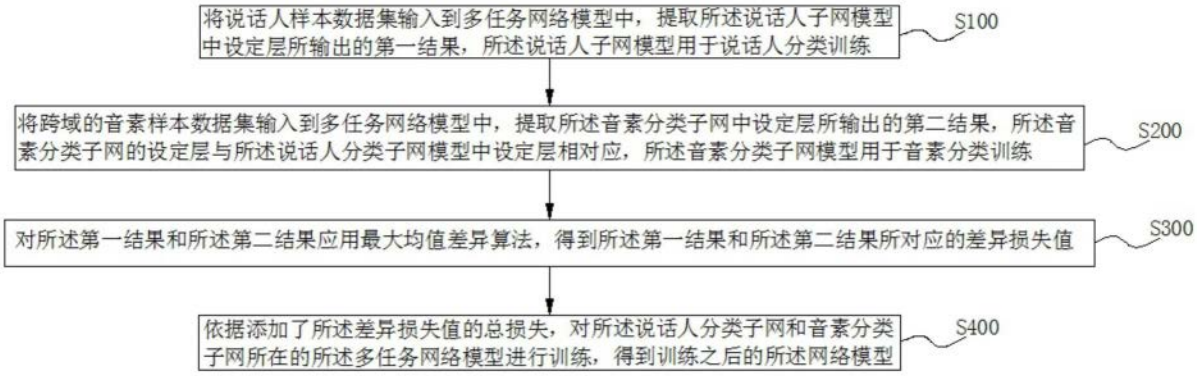
将说话人样本数据集输入到多任务网络模型中，提取所述说话人子网模型中设定层所输出的第一结果，所述说话人子网模型用于说话人分类训练
S100
将跨域的音素样本数据集输入到多任务网络模型中，提取所述音素分类子网中设定层所输出的第二结果，所述音素分类子网的设定层与所述说话人分类子网模型中设定层相对应，所述音素分类子网模型用于音素分类训练
S200
对所述第一结果和所述第二结果应用最大均值差异算法，得到所述第一结果和所述第二结果所对应的差异损失值
S300
依据添加了所述差异损失值的总损失，对所述说话人分类子网和音素分类子网所在的所述多任务网络模型进行训练，得到训练之后的所述网络模型
S400

图1

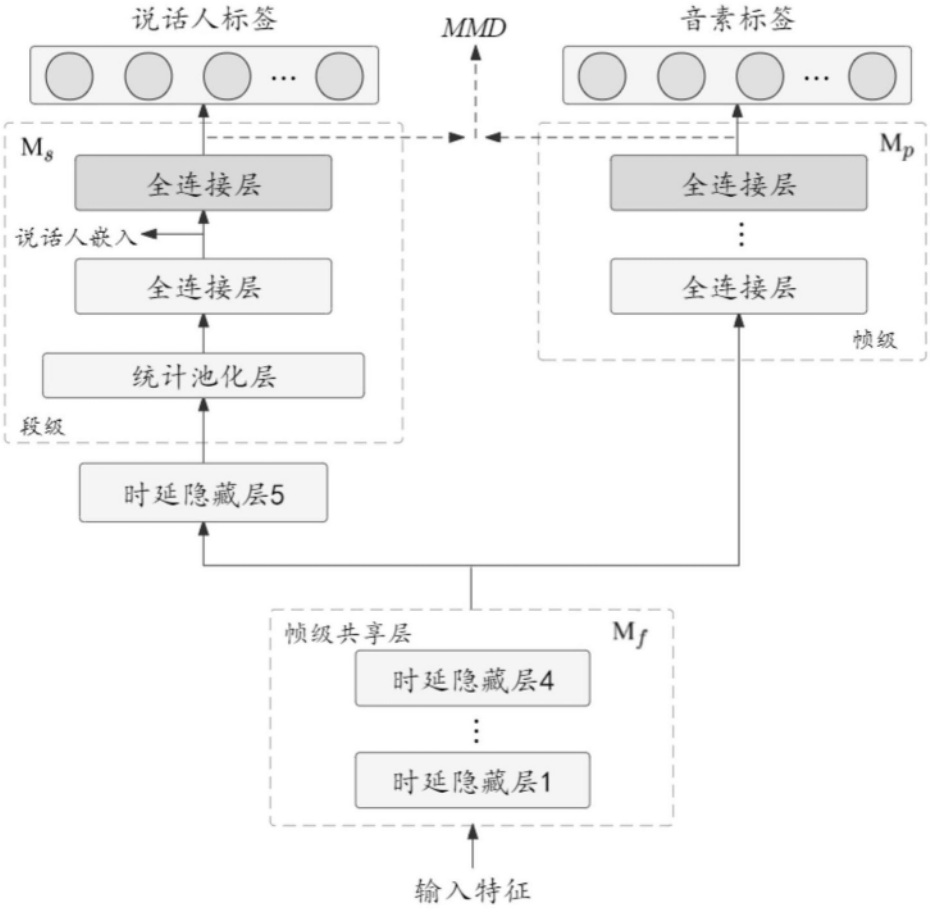
说话人标签
MMD
音素标签
M_s
M_p
全连接层
说话人嵌入
全连接层
统计池化层
段级
全连接层
全连接层
帧级
时延隐藏层5
帧级共享层
M_f
时延隐藏层4
时延隐藏层1
输入特征

图2

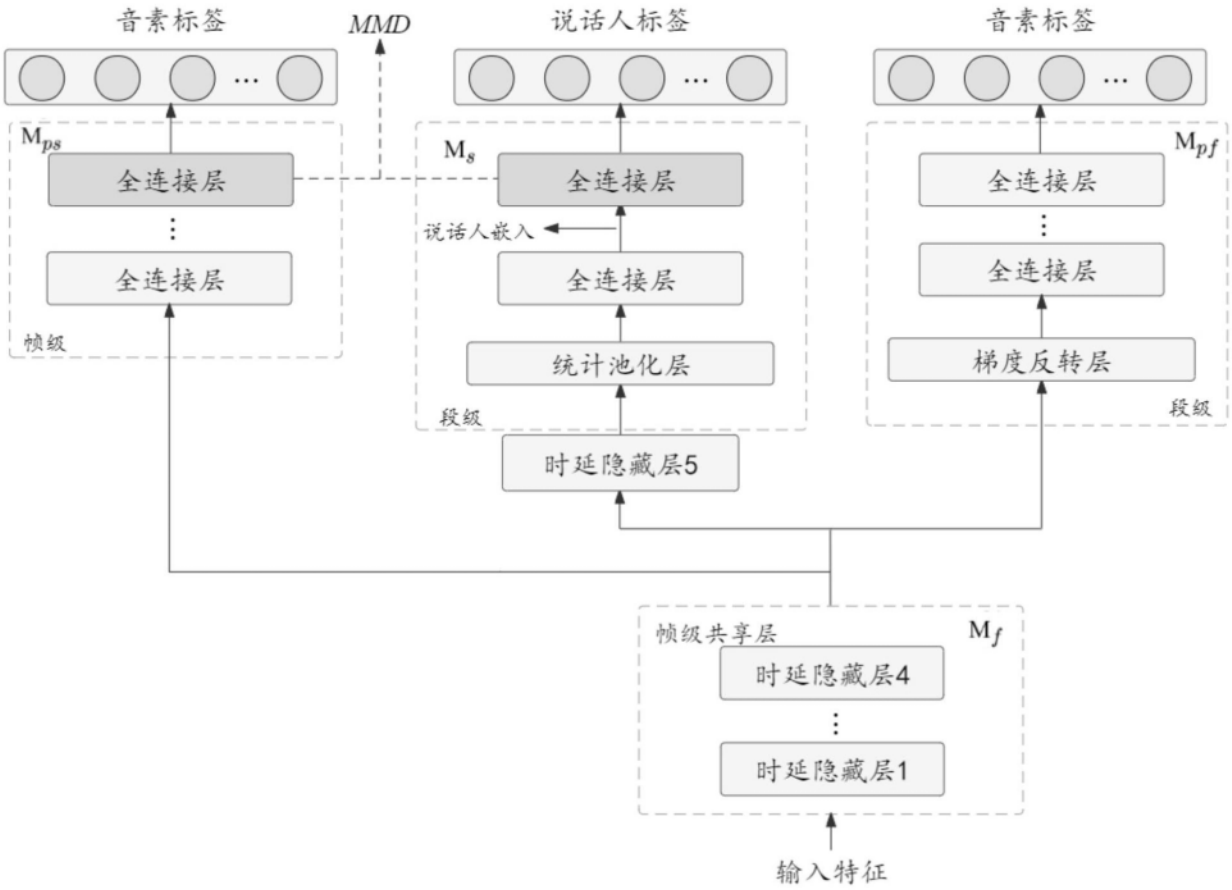
音素标签
MMD
说话人标签
音素标签
M_{ps}
全连接层
全连接层
帧级
M_s
全连接层
说话人嵌入
全连接层
统计池化层
段级
M_{pf}
全连接层
全连接层
梯度反转层
段级
时延隐藏层5
帧级共享层
M_f
时延隐藏层4
时延隐藏层1
输入特征

图3

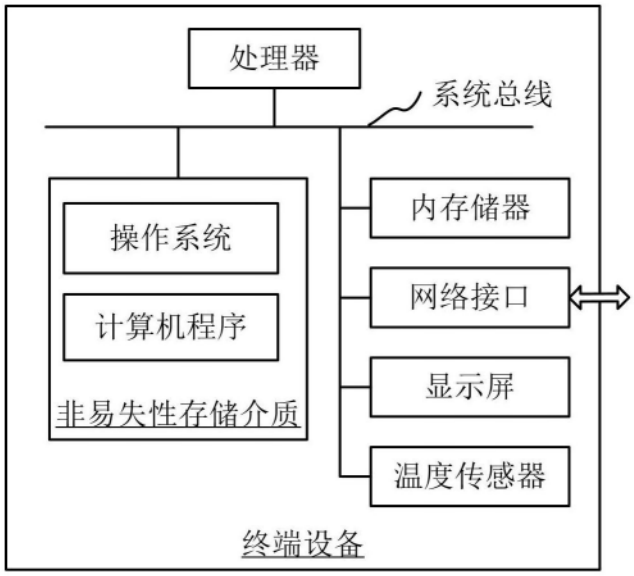
处理器
系统总线
操作系统
计算机程序
非易失性存储介质
内存储器
网络接口
显示屏
温度传感器
终端设备

图4