(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116863920 B
(45)授权公告日 2024.06.11

(21)申请号 202310874348.2
(22)申请日 2023.07.17
(65)同一申请的已公布的文献号
申请公布号 CN 116863920 A
(43)申请公布日 2023.10.10
(73)专利权人 北京邮电大学
地址 100876 北京市海淀区西土城路10号
(72)发明人 明悦 范春晓 吕柏阳 胡楠楠 周江琬
(74)专利代理机构 北京市商泰律师事务所 11255
专利代理师 毛燕生
(51)Int.Cl.
G10L 15/06(2013.01)
G10L 15/16(2006.01)
G10L 15/02(2006.01)
G10L 15/26(2006.01)
G06N 3/09(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 115310461 A,2022.11.08
CN 115810351 A,2023.03.17
US 11551668 B1,2023.01.10
US 2023096805 A1,2023.03.30
Junwen Bai等.Joint Unsupervised and Supervised Training for Multilingual ASR.ICASSP 2022.2022,全文.
张文林等.基于正样本对比与掩蔽重建的自监督语音表示学习.通信学报.2022,全文.
Zhehuai Chen等.Tts4pretrain 2.0: Advancing the use of text and speech in ASR pretraining with consistency and contrastive losses.ICASSP 2022.2022,全文.
刘娟宏等.端到端的深度卷积神经网络语音识别.计算机应用与软件.2020,(第04期),全文.
(续)

审查员 可杨

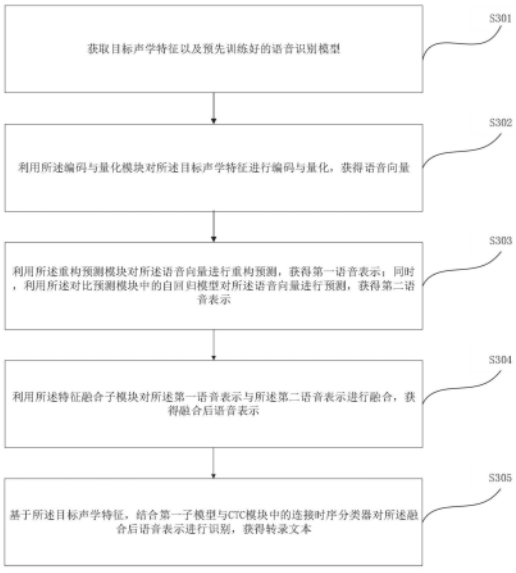
权利要求书2页 说明书14页 附图4页

(54)发明名称

基于双流自监督网络的语音识别方法、装置、设备及介质

(57)摘要

本发明提供了一种基于双流自监督网络的语音识别方法、装置、设备及介质,包括:利用编码与量化模块对目标声学特征进行编码与量化获得语音向量;利用重构预测模块对语音向量进行重构预测获得第一语音表示;同时,利用对比预测模块中的自回归模型对语音向量进行预测获得第二语音表示;利用特征融合子模块对第一语音表示与第二语音表示进行融合获得融合后语音表示;基于目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对融合后语音表示进行识别获得转录文本。本发明能够关注语音详细的上下文信息及语音不同特征之间的差异信息,提高自监督学习的鲁棒性,有效结合生成式和判别式自监督学习的互补优势。



CN 116863920 B

[转续页]

[接上页]

(56)对比文件

唐振韬等.深度强化学习进展:从AlphaGo到AlphaGo Zero.控制理论与应用.2017,(第12期),全文.

1.一种基于双流自监督网络的语音识别方法,其特征在于,包括:

获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

2.根据权利要求1所述的基于双流自监督网络的语音识别方法,其特征在于,所述特征融合子模块包括门控循环单元和自适应融合层;

相应地,所述利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示,包括:

利用所述门控循环单元分别对所述第一语音表示与所述第二语音表示进行特征选择,对应获得第一选择后特征与第二选择后特征;

利用所述自适应融合层对所述第一选择后特征与第二选择后特征进行自适应融合。

3.根据权利要求1所述的基于双流自监督网络的语音识别方法,其特征在于,所述预先训练好的语音识别模型通过如下方式训练得到:

获取声学特征样本以及预先构建的语音识别模型;

将所述声学特征样本输入至所述预先构建的语音识别模型;

基于所述重构预测模块输出的第一语音表示与所述声学特征样本计算获得重建损失;

基于所述特征融合子模块输出的融合后语音表示与所述声学特征样本计算获得对比损失;

基于所述声学特征样本的码本信息计算得到多样性损失;

根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块以及对比预测模块中的初始网络参数进行迭代更新,获得所述编码与量化模块、重构预测模块以及对比预测模块中的更新后网络参数;

将所述更新后网络参数作为所述CTC模块的特征提取器提取的语音表征,并基于所述声学特征样本以及标注数据对所述CTC模块进行训练解码,从而获得训练好的语音识别模型;

或者,根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块、对比预测模块以及CTC模块中的随机初始化的网络参数进行迭代更新,从而获得训练好的语音识别模型。

4.根据权利要求1所述的基于双流自监督网络的语音识别方法,其特征在于,所述编码与量化模块包括编码器以及向量量化层,所述编码器基于Conformer网络获得;

相应地,所述利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量,包括:

利用所述编码器对所述目标声学特征进行编码,获得潜在语音表示;

通过所述向量量化层对所述潜在语音表示进行离散化处理,以获得所述语音向量。

5.根据权利要求4所述的基于双流自监督网络的语音识别方法,其特征在于,所述编码器包括多层Conformer,每一层Conformer包括:

依次连接的第一前馈层、第一残差与标准化模块、多头自注意层、第二残差与标准化模块、卷积模块、第三残差与标准化模块、第二前馈层、第四残差与标准化模块以及Layernorm层;其中,所述第一残差与标准化模块与第二残差与标准化模块、第二残差与标准化模块与第三残差与标准化模块、第三残差与标准化模块与第四残差与标准化模块之间进行残差连接。

6.根据权利要求1-5任一所述的基于双流自监督网络的语音识别方法,其特征在于,所述预先训练好的语音识别模型还包括随机掩码模块;

相应地,在所述获取目标声学特征之后,方法还包括:

利用所述随机掩码模块对所述目标声学特征进行时间随机掩码与频率随机掩码处理,获得目标掩码声学特征;

所述利用所述编码与量化模块对所述目标掩码声学特征进行编码与量化,获得语音向量。

7.一种基于双流自监督网络的语音识别装置,其特征在于,包括:

声学特征与模型获取模块,用于获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

编码与量化模块,用于利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

重构与对比模块,用于利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

融合模块,用于利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

分类模块,用于基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

8.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1-6任一项所述基于双流自监督网络的语音识别方法。

9.一种计算机可读存储介质,其特征在于,其存储有计算机程序,所述计算机程序被处理器执行时实现如权利要求1-6任一项所述基于双流自监督网络的语音识别方法。

基于双流自监督网络的语音识别方法、装置、设备及介质

技术领域

[0001] 本发明涉及语音识别技术领域,尤其涉及一种基于双流自监督网络的语音识别方法、装置、设备及介质。

背景技术

[0002] 语音作为信息传递最直接有效的方式,是人们彼此感情交流和思想传递最主要的途径。自动语音识别(Automatic Speech Recognition,ASR)技术是指将语音信号正确地识别为对应的文本内容或命令,让机器听懂人类语言并执行相关操作。ASR技术融合多学科知识的前沿技术,覆盖了数学与统计学、声学与语音学、计算机与人工智能等基础学科与前沿学科,是人机语言通信以及信息交流的关键环节,有很强的实用价值。随着计算机的广泛应用,ASR技术成为实现简单便捷的人机智能交互的关键技术,被广泛应用到检索查询、自动导航、自助服务、机器翻译、自动驾驶等许多真实场景,涉及到工业、文化、商业等领域。

[0003] ASR经历了传统方法和深度学习方法两个发展时段。传统方法主要是将声学模型、发音模型和语言模型三个模块整合,用以发现给定语音观测时最可能出现的词语序列。随着深度学习技术的迅速发展,使用深度学习的语音任务性能逐渐超过传统算法。其中,基于深度神经网络的端到端语音识别(End-to-End Automatic Speech Recognition,E2E ASR)模型解决了需要对标注语音数据做对齐预处理的问题,并且可以直接得到输入语音波形或特征和输出文本内容之间的映射关系。E2E ASR简化了模型训练流程的同时凭借强大的建模和学习能力相较于传统语音识别技术显著提高了语音识别准确率。值得注意的是,不同于传统的ASR系统,E2E模型的性能很大程度上取决于可用的目标标注语料数量。然而语音数据收集及人工标注工作量巨大,并且小语种或者方言等因素均会导致标注语料数量不足的低资源应用场景出现。这为开展有效的E2EASR带来了严峻挑战。目前面向标注数据有限的端到端语音识别的方案主要通过预训练策略在大量无标注数据上学习语音基础结构信息,然后在有限的标注数据进行监督训练。在监督训练过程根据监督学习的方式不同,可以具体分为如下问题:

[0004] (1)无监督学习的问题。由于数据收集和标注的巨大工作量会导致标注语料数量不足的应用场景出现,这将会显著降低模型建模能力。而无监督学习不依赖于标注数据,通过对数据本身蕴含的结构或特征,找到数据样本间的关系,能够一定程度缓解因标注数据不足导致的性能下降。然而,由于无监督学习使用未标注的数据来捕获数据本身的分布或结构,会使得模型预测过程中监督信息缺失,造成了模型预测的偏差增加,限制了标注数据有限的实际场景应用。

[0005] (2)半监督学习的问题。半监督学习是监督学习与无监督学习相结合的一种学习方法。与无监督学习不同,为了缓解无监督学习过程中监督信息不足而导致的预测偏差问题,半监督学习考虑对无标注数据进行部分标注的思路。即在标注数据上训练模型,使用经过训练的模型来预测无标注数据的标签,从而创建伪标签。然后将标签数据和新生成的伪标签数据结合起来作为新的训练据以此缓解无监督学习中监督信息不足的问题,但是半监

督训练的性能严重依赖于模型预测伪标签的准确率。

[0006] (3)自监督学习的问题。自监督学习主要是利用辅助任务从大规模的无标注数据中挖掘自身的监督信息,使用构建的监督信息训练模型。与无监督和半监督相比可以学习到更多的语义关系和对下游任务有价值的表征。但是语音信号具有复杂的潜在结构(包含音素、音节、单词、韵律特征、句子上下文信息等),包含不同时间尺度的相关信息。而当前的自监督学习方案不能够兼顾不同特征之间的差异信息和数据自身分布的上下文信息导致预测的准确性和鲁棒性较差。

[0007] 综上,为了推动端到端语音识别在标注数据有限的实际场景应用,提高自监督学习对语音基础结构信息捕获的完整性,需要对上述问题进行深入研究,提出合理的解决方案。

发明内容

[0008] 本发明的实施例提供了一种基于双流自监督网络的语音识别方法、装置、设备及介质,以克服现有技术的缺陷。

[0009] 为了实现上述目的,本发明采取了如下技术方案。

[0010] 第一方面,本发明提供一种基于双流自监督网络的语音识别方法,包括:

[0011] 获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

[0012] 利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

[0013] 利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

[0014] 利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

[0015] 基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0016] 可选地,所述特征融合子模块包括门控循环单元和自适应融合层;

[0017] 相应地,所述利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示,包括:

[0018] 利用所述门控循环单元分别对所述第一语音表示与所述第二语音表示进行特征选择,对应获得第一选择后特征与第二选择后特征;

[0019] 利用所述自适应融合层对所述第一选择后特征与第二选择后特征进行自适应融合。

[0020] 可选地,所述预先训练好的语音识别模型通过如下方式训练得到:

[0021] 获取声学特征样本以及预先构建的语音识别模型;

[0022] 将所述声学特征样本输入至所述预先构建的语音识别模型;

[0023] 基于所述重构预测模块输出的第一语音表示与所述声学特征样本计算获得重建损失;

[0024] 基于所述特征融合子模块输出的融合后语音表示与所述声学特征样本计算获得对比损失；

[0025] 基于所述声学特征样本的码本信息计算得到多样性损失；

[0026] 根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块以及对比预测模块中的初始网络参数进行迭代更新，获得所述编码与量化模块、重构预测模块以及对比预测模块中的更新后网络参数；

[0027] 将所述更新后网络参数作为所述CTC模块的特征提取器提取的语音表征，并基于所述声学特征样本以及标注数据对所述CTC模块进行训练解码，从而获得训练好的语音识别模型；

[0028] 或者，根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块、对比预测模块以及CTC模块中的随机初始化的网络参数进行迭代更新，从而获得训练好的语音识别模型。

[0029] 可选地，所述编码与量化模块包括编码器以及向量量化层，所述编码器基于Conformer网络获得；

[0030] 相应地，所述利用所述编码与量化模块对所述目标声学特征进行编码与量化，获得语音向量，包括：

[0031] 利用所述编码器对所述目标声学特征进行编码，获得潜在语音表示；

[0032] 通过所述向量量化层对所述潜在语音表示进行离散化处理，以获得所述语音向量。

[0033] 可选地，所述编码器包括多层Conformer，每一层Conformer包括：

[0034] 依次连接的第一前馈层、第一残差与标准化模块、多头自注意层、第二残差与标准化模块、卷积模块、第三残差与标准化模块、第二前馈层、第四残差与标准化模块以及Layernorm层；其中，所述第一残差与标准化模块与第二残差与标准化模块、第二残差与标准化模块与第三残差与标准化模块、第三残差与标准化模块与第四残差与标准化模块之间进行残差连接。

[0035] 可选地，所述预先训练好的语音识别模型还包括随机掩码模块；

[0036] 相应地，在所述获取目标声学特征之后，方法还包括：

[0037] 利用所述随机掩码模块对所述目标声学特征进行时间随机掩码与频率随机掩码处理，获得目标掩码声学特征；

[0038] 所述利用所述编码与量化模块对所述目标声学特征进行编码与量化，获得语音向量，包括：

[0039] 利用所述编码与量化模块对所述目标掩码声学特征进行编码与量化，获得语音向量。

[0040] 第二方面，本发明还提供一种基于双流自监督网络的语音识别装置，包括：

[0041] 声学特征与模型获取模块，用于获取目标声学特征以及预先训练好的语音识别模型；所述预先训练好的语音识别模型包括第一子模型与第二子模型，所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块，所述对比预测模块包括特征融合子模块，所述第二子模型包括CTC模块；

[0042] 编码与量化模块，用于利用所述编码与量化模块对所述目标声学特征进行编码与

量化,获得语音向量;

[0043] 重构与对比模块,用于利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

[0044] 融合模块,用于利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

[0045] 分类模块,用于基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0046] 第三方面,本发明还提供一种电子设备,包括存储器和处理器,处理器和存储器相互通信,存储器存储有可被处理器执行的程序指令,处理器调用程序指令执行如上的基于双流自监督网络的语音识别方法。

[0047] 第四方面,本发明还提供一种计算机可读存储介质,其存储有计算机程序,计算机程序被处理器执行时实现如上的基于双流自监督网络的语音识别方法。

[0048] 本发明有益效果:本发明提供的基于双流自监督网络的语音识别方法、装置、设备及介质,在编码与量化模块之后并行结合重构预测模块(Reconstruction Prediction Module,RPM)和对比预测模块(Contrastive Prediction Module,CPM)设计了一个双通道结构。其中,将重建预测作为对比预测的辅助任务分别对语音向量进行预测语音帧,从而在建模不同语音表示之间的归属关系捕获语音不同特征差异信息的同时,关注详细的语音上下文信息。此外,为了有效地利用双通道语音表示,还通过特征融合子模块来融合两个分支的语音表示,该特征融合子模块通过参数可学习策略自适应融合两个分支的语音表示,并利用权值来控制各种语音特征的暴露。最后,本发明提供的双流自监督学习网络可以很好地初始化ASR模型的权重。与其他自监督学习方法相比,本发明提供的语音识别方法可以达到具有竞争力的预测精度。此外,在有限的标记数据场景下,与最先进的自监督学习方法相当。

[0049] 本发明附加的方面和优点将在下面的描述中部分给出,这些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

[0050] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例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发明的一些实施例,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。

[0051] 图1是现有技术中的基于掩码重建的语音识别方法的流程示意图;

[0052] 图2是现有技术中的基于对比预测的语音识别方法的流程示意图;

[0053] 图3为本发明实施例提供的一种基于双流自监督网络的语音识别方法的流程示意图之一;

[0054] 图4为本发明实施例提供的一种基于双流自监督网络的语音识别方法的流程示意图之二;

[0055] 图5为本发明实施例提供的特征融合子模块的结构示意图;

[0056] 图6为本发明实施例提供的编码器的结构示意图。

具体实施方式

[0057] 下面详细描述本发明的实施方式,实施方式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方式是示例性的,仅用于解释本发明,而不能解释为对本发明的限制。

[0058] 本技术领域技术人员可以理解,除非特意声明,这里使用的单数形式“一”、“一个”、“”和“该”也可包括复数形式。应该进一步理解的是,本发明的说明书中使用的措辞“包括”是指存在特征、整数、步骤、操作、元件和/或组件,但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、组件和/或它们的组。应该理解,当我们称元件被“连接”或“耦接”到另一元件时,它可以直接连接或耦接到其他元件,或者也可以存在中间元件。此外,这里使用的“连接”或“耦接”可以包括无线连接或耦接。这里使用的措辞“和/或”包括一个或更多个相关联的列出项的任一单元和全部组合。

[0059] 本技术领域技术人员可以理解,除非另外定义,这里使用的所有术语(包括技术术语和科学术语)具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应该理解的是,诸如通用字典中定义的那些术语应该被理解为具有与现有技术的上下文中的意义一致的意义,并且除非像这里一样定义,不会用理想化或过于正式的含义来解释。

[0060] 现有技术中,自监督学习不仅可以学习到对下游任务有价值的语音表征,还可以将训练的模型用于初始化ASR任务,且自监督预训练学习到的模型参数被发现是用于初始化ASR模型的有效方法。目前,自监督学习方法主要有两个主要分支,即生成式自监督学习和判别式自监督学习。

[0061] 其中,生成式自监督学习基于一些有限的语音帧生成或重建输入数据这包括从过去的输入预测未来的输入,从未掩码中预测掩码,或者从其他损坏的语音正中预测原始语音。下面分为自回归预测语音帧和掩码预测语音帧的方法进行介绍。

[0062] 其中,自回归预测的灵感主要来自于文本的语言模型(LM),并将此扩展到语音领域。与传统的线性回归不同,自回归模型对过去的声波序列时间信息进行编码。然后,该模型在预测未来的语音帧的同时对过去语音帧进行调节。但是它只对先前时间步长的信息进行编码,而不是对整个输入进行编码,因此会导致全局上下文信息的缺失。

[0063] 掩码重建则很大程度上受到掩码语言模型的启发后扩展到语音领域。将输入句子中的某些令牌进行随机替换为掩码令牌,然后该模型依靠未屏蔽令牌从损坏或者屏蔽令牌特征中恢复隐藏语音特征。其中掩码策略与BERT相似,一般可以沿着时间和频率两个维度进行掩码运算。这允许模型对整个输入中的信息进行编码以学习语音基础结构信息。但是语音信号具有复杂的潜在结构(包含音素、音节、单词、韵律特征、句子上下文信息等),掩码重建模型会对语音信号中的所有信息进行编码重建,但是会为特定的ASR任务编码冗余信息。

[0064] 基于上述原因,学习重建原始语音信号可能不是发现语音潜在结构的最佳方式。对比模型通过最大化给定语音和正样本之间的相似性,同时最小化给定语音和负样本之间的相似性,以此区分给定语音的目标样本(正)和干扰样本(负)来学习语音表示。下面分为基于对比预测编码(Contrastive Predictive Coding,CPC)和基于wav2vec2.0的方法。

[0065] 基于对比预测的方法。基于对比预测编码采用了特征空间中的一种单向建模形式,首先使用一个非线性编码器将输入语音序列映射到隐藏空间,此时语音表示只具有较低的时间分辨率。然后使用自回归模型对语音潜在表示进行编码得到语音上下文表征,并结合语音的历史上下文表征经过预测网络来预测未来几帧的潜在特征。最后通过最大化未来几帧音频片段与其上下文表征之间的互信息来判断预测结果跟真实特征的接近程度。这不仅是得模型能够学习语音编码(高维)信号不同部分之间的基本共享信息表征,同时它摒弃了低层次信息和更局部的噪音。wav2vec2.0在对比预测编码的基础上结合掩码操作使用InfoNCE损失来最大化上下文化表示和原始语音表示之间的相似性来学习语音表征。其专注于学习输入和输出之间的映射关系,导致对对训练数据本身的特性捕获不足,造成上下文信息的缺失。

[0066] 基于掩码重建的语音识别是利用掩码重建自监督预训练策略对声学模型进行训练后,将得到的声学模型用于语音表示提取或进行微调来进行语音识别。其主要算法流程如图1所示,基于掩码重建的语音识别的具体步骤为:

[0067] 首先,将每个输入语音特征视为维度为T×F的图像,其中T是帧数,F是频率区间数。通过使用时间和频率两种随机掩码方案策略沿这两个维度进行掩码。对于时间掩码,每个序列中从T_I开始随后的T_n个连续时间步被随机屏蔽(T_I,T_I+T_n),其中共有15%的语音帧没有重叠的被屏蔽。上述过程中80%的帧被零向量替换,10%的帧被来自随机位置的帧替换,并且在其余时间保持相同。与时间掩码类似,频率在整个输入序列的所有时间步中将连续频率区间块的值随机屏蔽为零。对f个连续的mel频率信道$[F_I,F_I+f]$进行屏蔽,其中f从{0,1,...,f}均匀采样f的宽度来选择掩蔽频率块,F_I从[0,F-f]中随机选择。

[0068] 另外,在掩码重建过程中,随机掩蔽了一部分特征,通过单独或者混合使用时间和频率两种掩蔽码策略鼓励RNN/LSTM/Transforemer等自回归模型网络充分学习输入特征中的时空信息,即语音的全局上下文信息和空间信息。

[0069] 最后,通过表示提取或者微调的方式将自监督学习到的语音知识纳入到语音识别网络中。对于表示提取,通过冻结自监督训练自回归模型的网络参数作为ASR网络的特征提取器,将提取的语音表示作为输入特征馈送到ASR网络进行监督训练得到文本输出。对于微调,将自监督训练自回归模型与随机初始化的ASR一起进行监督训练,更新网络参数,得到最终的文本输出。

[0070] 上述基于掩码重建的语音识别方法,由于语音信号包含杂的潜在结构(包含音素、音节、韵律特征、句子上下文信息等),依靠上下文信息预测掩码特征导致模型对影响ASR性能的韵律特征等信息捕获不足。

[0071] 为了充分利用上下文预测掩码特征,须要对语音信号中的所有信息进行编码来学习语音数据的数据本身特性,这将导致学习成本比判别式自监督学习更高,需要更多的计算资源。

[0072] 此外,重建预测会对语音信号中的所有信息进行编码,还会为特定的ASR任务编码冗余信息,导致其预测的鲁棒性较低。

[0073] 除了上述的基于掩码重建的语音识别之外,还有一种基于对比预测的语音识别的方法,其结合wav2vec2.0与下游语音识别网络通过语音表示提取或进行微调来进行语音识别,其中wav2vec2.0包含特征编码器,量化模块和Transformer的上下文表示三部分组成。

其主要算法流程如图2所示,基于对比预测的语音识别方法具体步骤为:

[0074] 首先,采用七层卷积网络的特征提取器将原始音频编码为帧特征序列,通过向量量化模块把每帧特征转变为离散特征,并作为自监督目标。

[0075] 进而,向量量化模块被用来离散特征编码器的输出,它包含G组码本,每个码本包含V个变量。对于特征编码器输出的每一个连续型变量,会在每一组码本中找到一个变量,并把G个变量拼接起来,再做一次线性变化得到最后的离散特征。

[0076] 随后,Transformer被用来获取语音上下文表征。特征编码器的输出在输入transformer之前会进行一些掩码操作,使用掩码的可训练嵌入令牌代替。在进行向量量化时不会进行掩码操作。通过语音表征的上下文和离散特征计算对比损失函数,以使得掩码令牌在transformer输出能够在包含干扰项的候选离散特征中被识别出来,其中干扰项是从其它掩码时刻采样得到。最后通过表示提取或微调的方式得到最终语音识别任务的文本输出。

[0077] 然而,基于对比预测的语音识别方法存在如下缺陷:判别式学习通过对比目标样本(正)和干扰样本(负)相似性度量专注于学习输入和输出之间的映射关系,对数据的内在结构考虑不充足,因此对于缺失掩码数据的处理能力较弱。

[0078] 综上所述,现有技术中,基于自监督学习语音识别算法存在如下问题:

[0079] 1.对于语音基础结构信息捕获不完整。生成式重建预测通过语音上下文信息重建掩码数据关注数据自身分布,但是由于语音信号的复杂特性,依然存在对影响ASR性能的韵律特征等信息捕获不足。判别式模型通过对比目标样本(正)和干扰样本(负)相似性度关注的是数据的差异信息,寻找的是分类面,其专注于学习输入和输出之间的映射关系,对语音信号的自相关特性考虑不充分,因此对于缺失掩码数据的处理能力较弱。现有的自监督学习方案均存在对语音基础结构信息捕获不完整的问题。

[0080] 2.生成式和判别式自监督的优势没被有效结合。不同类型的自监督模型在不同的下游任务上表现出不同的优势。现有的技术方案中缺少有效的融合两种自监督学习的方案。因此,为了更好地利用两种模型的潜力,有效融合策略是十分必要的。

[0081] 3.重建预测为特定的ASR任务编码冗余信息。现有的技术方案中,重建预测利用上下文信息预测掩码特征,这须要对语音信号中的所有信息进行编码来学习语音数据的本身特性,这将导致其为ASR任务编码冗余信息,使得预测的鲁棒性较低。

[0082] 本发明专注于解决自监督学习中对语音基础结构信息捕获不完整,和模型预测结果鲁棒性较差的问题,对现有自监督学习方案的缺陷与不足进行改进,促进其在实际中更广泛的应用。

[0083] 下面结合附图对本发明提到的基于双流自监督网络的语音识别方法进行说明。

[0084] 术语解释:

[0085] 自监督学习(Self-Supervised Learning,SSL):通过设计辅助任务挖掘无标签数据自身的表征特性作为监督信息,从而提升模型的特征提取能力的学习方式。

[0086] 端到端语音识别(End-to-End Automatic Speech Recognition,E2E ASR):基于端到端网络模型的语音识别系统通过一个神经网络模型直接将输入语音波形列映射到输出文本,不再需要像传统语音识别算法中对系统中各个模块单独训练,简化了语音识别流程,并且都很好的解决了输入序列和输出序列的自动对齐问题,不需要对输入序列进行强

制对齐处理。

[0087] 注意力机制(Attention Mechanism):注意力机制指计算深度学习中模仿人类时间系统可以在复杂的场景中自然有效地找到突出区域的这一特点,所设计的一种方法。深度学习中的注意力机制包含空间注意力、通道注意力、自注意力等。

[0088] 实施例1

[0089] 图3为本发明实施例提供的一种基于双流自监督网络的语音识别方法的流程示意图之一;图4为本发明实施例提供的一种基于双流自监督网络的语音识别方法的流程示意图之二;如图3以及图4所示,一种基于双流自监督网络的语音识别方法,包括如下步骤:

[0090] S301,获取目标声学特征以及预先训练好的语音识别模型。

[0091] 其中,所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块。目标声学特征即为待识别的语音数据。

[0092] S302,利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量。

[0093] 在本步骤中,利用编码与量化模块对所述目标声学特征进行编码与量化,从而学习到更有意义的语音单元信息来丰富语音表示。并将语音向量输入至双通道结构的重构预测模块和对比预测模块中。

[0094] S303,利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示。

[0095] 在本步骤中,重构预测作为对比预测的辅助任务进行联合训练。

[0096] S304,利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示。

[0097] 在本步骤中,利用特征融合子模块进行语音表示融合后,该特征融合子模块可以通过参数可学习策略自适应融合两种语音表示在关注上下文信息的同时探索不同语音表征之间的归属关系以捕获不同特征差异信息。

[0098] S305,基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0099] 在本步骤中,通过连接时序分类器(Connectionist Temporal Classification,简称CTC)进行最终的文本输出。

[0100] 需要说明的是,本发明提供的语音识别方法可以应用于人机交互、机器翻译、自动驾驶、智能家居等许多真实场景,涉及到工业、文化、商业等领域。例如Google,亚马逊,百度,阿里巴巴,科大讯飞等各大互联网公司相继发布的智能音箱便是ASR技术成功落地的一个产品。

[0101] 根据本发明实施例提供的基于双流自监督网络的语音识别方法,在编码与量化模块之后并行结合重构预测模块(Reconstruction Prediction Module,RPM)和对比预测模块(Contrastive Prediction Module,CPM)设计了一个双通道结构。其中,将重建预测作为对比预测的辅助任务分别对语音向量进行预测语音帧,从而在建模不同语音表示之间的归属关系捕获不同特征差异信息的同时,关注详细的语音上下文信息。此外,为了有效地利用

双通道语音表示,还通过特征融合子模块来融合两个分支的语音表示,该特征融合子模块通过参数可学习策略自适应融合两个分支的语音表示,并利用权值来控制各种语音特征的暴露。最后,本发明提供的双流自监督学习网络可以很好地初始化ASR模型的权重。与其他自监督学习方法相比,本发明提供的语音识别方法可以达到具有竞争力的预测精度。此外,在有限的标记数据场景下,与最先进的自监督学习方法相当。

[0102] 可选地,所述预先训练好的语音识别模型还包括随机掩码模块;

[0103] 相应地,在所述获取目标声学特征之后,方法还包括:

[0104] 利用所述随机掩码模块对所述目标声学特征进行时间随机掩码与频率随机掩码处理,获得目标掩码声学特征。即,对于目标声学特征x,使用时间和频率两种随机掩码策略来获得掩码声学特征$\hat{x}$。

[0105] 具体地,对于时间掩码,随机选择起始索引T_I来屏蔽具有最大宽度为T_n的语音,其中每个序列被随机屏蔽为(T_I, T_I+T_n),占整个序列的15\%。在上述过程中,80%的语音帧被零向量取代,10%被从同一语音中随机采样的其他语音帧取代。类似地,频率掩码在所有时间步长上随机地将连续的频率(F_I, F_I+F)的值掩蔽为零,其中F从{0,1,…,F}中均匀地采样F的宽度以选择掩蔽频率。

[0106] 所述利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量,包括:

[0107] 利用所述编码与量化模块对所述目标掩码声学特征进行编码与量化,获得语音向量。

[0108] 可选地,所述编码器包括多层Conformer,每一层Conformer包括:

[0109] 依次连接的第一前馈层、第一残差与标准化模块、多头自注意层、第二残差与标准化模块、卷积模块、第三残差与标准化模块、第二前馈层、第四残差与标准化模块以及Layernorm层;其中,所述第一残差与标准化模块与第二残差与标准化模块、第二残差与标准化模块与第三残差与标准化模块、第三残差与标准化模块与第四残差与标准化模块之间进行残差连接。

[0110] 具体地,本发明使用基于Conformer的编码器结构,该编码器由N层组成,每个层由多头自注意层(Multi-Head Self-Attention,MHSA)、卷积模块(Convolution module,Conv)和前馈层(Feed forward module,FFN)、残差与标准化层(Add&Norm)构成,如图6所示,整体Conformer结构将原始前馈层替换为两个half-step前馈层,一个在多头注意力层之前,第二个在卷积模块其后。第二个前馈模块之后紧接一个Layernorm层。因此,给定一个输入$\hat{x}$经过Conformer得到输出H^X定义如下:

[0111] $$H = \hat{x} + \frac{1}{2}\mathrm{FFN}(\hat{x}) \qquad (1)$$

[0112] $$H' = H + \mathrm{MHSA}(H) \qquad (2)$$

[0113] $$H'' = H' + \mathrm{Conv}(H) \qquad (3)$$

[0114] $$H^X = \mathrm{Layernorm}(H'' + \frac{1}{2}\mathrm{FFN}(H'')) \qquad (4)$$

[0115] 多头自注意实际上是一种多通道并行自注意机制。对于自注意机制而言,先从输入的掩码语谱特征表示$\hat{x}$通过线性计算得到查询、键和值(Q,K,V)后进行点积计算:

[0116] $$Q=\widehat{X}W^{Q}, K=\widehat{X}W^{K}, K=\widehat{X}W^{V} \quad (5)$$

[0117] 其中,W^{Q}、W^{K}、W^{V}分别为可学习的参数矩阵。然后通过softmax函数进行点积注意计算:

$$head_i = Att(Q, K, V) = softmax(QK^{T}/\sqrt{d_k})V \quad (6)$$

[0118] 随后多头自注意机制即将注意输入等分为h个不同注意通道并行计算,并对将所有通道注意结果进行连接:

[0119] $$Multihead(Q,K,V)=concat(head_1,\cdots,head_h)W^{o} \quad (7)$$

[0120] 其中,W^{Q},W^{K},W^{V},W^{o}分别为可学习的参数矩阵,$1/\sqrt{d_k}$为缩放系数。一般来说,它通常使用h=8个平行注意空间或头部。在实际应用中,总是设定$d_k=d_{model}/h$使多头注意的计算复杂度与单个自注意相同,d_{model}表示输入向量的维度。其中,卷积模块由Pointwise卷积、Depthwise卷积和GLU激活层、Swish激活层组成。前馈层由两个线性转换组成,在中间有一个ReLU激活,

[0121] $$FFN(x)=max(0,xW_1+b_1)W_2+b_2 \quad (8)$$

[0122] 其中x表示前馈层的输入,W_1,W_2表示可学习参数矩阵,b_1,b_2为保持线性引入的常数介质。

[0123] 尽管线性变换在不同位置上是相同的,但它们在不同层之间使用不同的参数。另外,还在每个两个子层周围使用残差连接,进行层归一化,以实现更稳定和更快的收敛。

[0124] 为了更多地关注语言/语音单元信息,本发明在Conformer后接入量化层。首先通过线性层将Conformer输出的潜在语音表示H^{X}映射到logitsI$\in R^{G\times V}$,其中G是码本数量,V是码本的大小。然后通过从一个固定大小的码本$C=\{C_1,\cdots,C_V\}$中选择一个变量并将结果向量叠加,并应用线性变换来获得语音离散化表示v_t。在第g个码本中选择第v个码的概率定义如下:

[0125] $$p_{g,v}=\frac{\exp\ (I_{g,v}+n_v)/\tau}{\sum_{k=1}^{V}\exp\ (I_{g,k}+n_k)/\tau} \quad (9)$$

[0126] 可选地,所述预先训练好的语音识别模型通过如下方式训练得到:

[0127] 获取声学特征样本以及预先构建的语音识别模型;

[0128] 将所述声学特征样本输入至所述预先构建的语音识别模型。

[0129] 基于所述重构预测模块输出的第一语音表示与所述声学特征样本计算获得重建损失。

[0130] 基于所述特征融合子模块输出的融合后语音表示与所述声学特征样本计算获得对比损失。

[0131] 基于所述声学特征样本的码本信息计算得到多样性损失。

[0132] 根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块以及对比预测模块中的初始网络参数进行迭代更新,获得所述编码与量化模块、重构预测模块以及对比预测模块中的更新后网络参数。

[0133] 将所述更新后网络参数作为所述CTC模块的特征提取器提取的语音表征,并基于所述声学特征样本以及标注数据对所述CTC模块进行训练解码,从而获得训练好的语音识

别模型。

[0134] 或者,根据所述重建损失、所述对比损失以及所述多样性损失对所述编码与量化模块、重构预测模块、对比预测模块以及CTC模块中的随机初始化的网络参数进行迭代更新,从而获得训练好的语音识别模型。

[0135] 在本实施例中,本发明在所述编码与量化模块之后构造了基于重构预测模块和对比预测模块的双流结构。重构预测模块主要由预测网络P_{net}组成,其目的是从掩码特征$\hat{x}_t$中重建声学特征x_t。本发明中预测网络由位置前馈网络(Position-wise Feed Forward Network,FFN)组成。然后在输入x和P_{net}的网络输出之间计算L1重建损失,以更新网络参数$\theta_{C_{encoder}}$和$\theta_{P_{net}}$。

[0136] $$L_{Reconstruction} = \sum_t |x_t - P_{net}(C_{encoder}(\hat{x}_t))| \qquad (10)$$

[0137] 式中,x_t表示原始语音特征输入(即掩码声学特征),$\hat{x}_t$表示经过掩码操作后的语音特征(目标掩码声学特征),$\theta_{C_{encoder}}$为编码与量化模块中的Conformer编码器的参数,该网络参数保留用于ASR任务,而预测网络P_{net}被丢弃。重构预测模块通过从对以前和未来内容的上下文化理解重建的掩码的语音帧,有效地提高了语音识别预测的准确性。

[0138] 对比预测模块利用自回归模型将离散表示总结为新的上下文向量c_t。但是,本发明没有直接使用上下文向量c_t来计算对比预测,而是使用GFF模块将RPM的输出与CPM的自回归网络输出融合得到的语音表示c_{GFF},来提高预测语音表示的准确性。然后利用融合得到的语音表示c_{GFF}计算对比损失,有利于学习更全面的语音结构信息。模型利用对比损失在一组K+1个候选表示中识别真正的上下文向量语音表示c_t,其中包括x_t和K个干扰项,其中包括K个干扰项。干扰项是从同一话语的其他掩蔽时间步中统一采样的。对比损失的定义为:

[0139] $$L_{Contrastive} = -\log \frac{\exp\ (\mathrm{sim}(c_{GFF}, x_t)/\kappa)}{\sum_{\tilde{x}\sim X_t} \exp\ (\mathrm{sim}(c_{GFF}, \tilde{x}_t)/\kappa)} \qquad (11)$$

[0140] 在$L_{Contrastive}$中,sim表示两个向量之间的余弦相似性,κ是temperature超参数。此外,还利用了多样性损失来增加量化码本表示,并通过最大化一批音频中每个码本p_g的码本条目上的平均softmax分布的熵来平衡每个码本中使用所有entries的概率,其中$p_{g,v}$表示在第g个码本中选择第v个码的概率。

[0141] $$L_{Diversity} = \frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V} p_{g,v}\log p_{g,v} \qquad (12)$$

[0142] 本发明最终训练目标L_{Total}由重建损失$L_{Reconstruction}$、对比损失$L_{Contrastive}$和多样性损失$L_{Diversity}$三部分组成,可以同时解决这两个自监督任务。最终要最小化的训练损失为:

[0143] $$L_{Total} = L_{Contrastive} + \alpha L_{Diversity} + \beta L_{Reconstruction} \qquad (13)$$

[0144] 其中α和β是可学习的超参数。$L_{Contrastive}$通过语音表示和声学特征计算,其中声学特征的噪声样本从同一语音的其他掩模中均匀采样。对于$L_{Diversity}$,设置α为0.1以平衡$L_{Diversity}$的权重。$L_{Reconstruction}$由声学特征X和重建输出$P_{net}(C_{encoder}(\hat{x}_t))$计算得到。

[0145] 在确定了最终的训练目标函数L_{Total}之后,对于各个模块的参数训练有两种不同的方式,可以通过表示提取和微调两种方式将双流自监督网络学习到语音知识纳入到ASR任务中进行训练和解码,以实现标注数据有限的端到端语音识别。

[0146] 其中,表示提取是指与下游ASR训练时,通过冻结DSSLNet的参数作为训练CTC模块的特征提取器提取语音表征,这本质上是DSSLNet编码器最后一层的隐藏状态。提取的表示作为输入替换FBANK/MFCC等特征馈送到CTC模块进行训练解码,得到文本输出。

[0147] 微调则是用下游CTC模块对DSSLNet进行微调。这里DSSLNet的输出连接到CTC模块,其中DSSLNet的参数并未冻结。然后,将训练的DSSLNet与随机初始化的CTC模块一起更新进行训练解码,得到文本输出。

[0148] 可选地,所述编码与量化模块包括编码器以及向量量化层,所述编码器基于Conformer网络获得;

[0149] 相应地,所述利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量,包括:

[0150] 利用所述编码器对所述目标声学特征进行编码,获得潜在语音表示;

[0151] 通过所述向量量化层对所述潜在语音表示进行离散化处理,以获得所述语音向量。

[0152] 根据本发明实施例提供的基于双流自监督网络的语音识别方法,本发明提出由门控循环单元(GRU)和自适应融合层组成的GUR特征融合模块,通过控制不同特征的曝露来达到自适应特征融合的目的,以降低重建预测为特定的ASR任务产生的冗余信息。

[0153] 可选地,所述特征融合子模块包括门控循环单元和自适应融合层。

[0154] 相应地,所述利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示,包括:

[0155] 利用所述门控循环单元分别对所述第一语音表示与所述第二语音表示进行特征选择,对应获得第一选择后特征与第二选择后特征。

[0156] 利用所述自适应融合层对所述第一选择后特征与第二选择后特征进行自适应融合。

[0157] 具体地,本发明设计的特征融合子模块(GRUfeaturefusion,简称GFF)能够避免融合特征中存在大量冗余信息,该模块包括门控循环单元(GRU)和自适应融合层,如图5所示。GFF模块的工作流程分为两个步骤。

[0158] 首先,将第一语音表示与第二语音表示输入GRU,其中,GRU由重置门r_t和更新门z_t组成。利用GRU的门控机制,从大量的特征映射中选择最有用的信息,然后根据得到的结果选择性地聚合信息。此步骤的输出使用门控机制让信息选择性地通过。第二步通过自适应融合层对GRU的输出进行特征融合处理。

[0159] 具体地,在第一语音表示的处理过程中,通过当前RPM的输出O_{Recon}(即第一语音表示)和上一个节点传递下来的隐状态h_{t-1}来获取两个门控的信息:

[0160] $$r_t=\sigma(W_r\cdot[h_{t-1},O_{Recon}]) \quad (14)$$

[0161] $$z_t=\sigma(W_z\cdot[h_{t-1},O_{Recon}]) \quad (15)$$

[0162] 式中,σ为sigmoid型函数,W_r和W_z分别为重置门和更新门的权值。

[0163] 在获得门控信息后,将O_{Recon}和复位数据拼接在一起,其中复位门决定了过去需要记忆多少信息。然后,通过激活函数tanh得到当前隐藏节点的输出。

[0164] 最后是“更新内存”阶段,更新后的表达式为:

[0165] $$\tilde{h}_t=\tanh(W\cdot[r_t h_{t-1},O_{Recon}]) \quad (16)$$

[0166] $$h_q = z_t h_{t-1} + (1 - z_t)\tilde{h}_t \qquad (17)$$

[0167] 式中,W表示GRU的可学习参数,tanh表示激活函数,h_{t-1}表示上一时刻的输入的隐藏状态。如果忽略前面权重为z_t的信息,则选择当前权重为(1-z_t)的输入信息。

[0168] 同样地,针对第二语音表示进行相同地计算,从而最后获得h_p,在此不再赘述。

[0169] 在获得h_q和h_p之后,利用自适应融合层进行自适应融合。具体步骤如下:

[0170] $O_{GFF}=\eta h_p+\mu h_q$ (18)

[0171] 式中,η和μ表示可学习的超参数,h_p和h_q分别代表RPM的输出O_{Recon}和CPM的输出O_{Con}经过GRU处理得到的结果。

[0172] 实施例2

[0173] 在实施例1的基础上,本实施例2提供一种基于双流自监督网络的语音识别装置,该基于双流自监督网络的语音识别装置与上述基于双流自监督网络的语音识别对应,基于双流自监督网络的语音识别装置包括:

[0174] 声学特征与模型获取模块,用于获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

[0175] 编码与量化模块,用于利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

[0176] 重构与对比模块,用于利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

[0177] 融合模块,用于利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

[0178] 分类模块,用于基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0179] 具体细节参见基于双流自监督网络的语音识别方法部分的描述,在此不再赘述。

[0180] 实施例3

[0181] 本发明实施例3提供一种电子设备,包括存储器和处理器,处理器和存储器相互通信,存储器存储有可被处理器执行的程序指令,处理器调用程序指令执行基于双流自监督网络的语音识别方法,该方法包括如下流程步骤:

[0182] 获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

[0183] 利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

[0184] 利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

[0185] 利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

[0186] 基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0187] 实施例4

[0188] 本发明实施例4提供一种计算机可读存储介质,其存储有计算机程序,计算机程序被处理器执行时实现基于双流自监督网络的语音识别方法,该方法包括如下流程步骤:

[0189] 获取目标声学特征以及预先训练好的语音识别模型;所述预先训练好的语音识别模型包括第一子模型与第二子模型,所述第一子模型包括编码与量化模块、重构预测模块以及对比预测模块,所述对比预测模块包括特征融合子模块,所述第二子模型包括CTC模块;

[0190] 利用所述编码与量化模块对所述目标声学特征进行编码与量化,获得语音向量;

[0191] 利用所述重构预测模块对所述语音向量进行重构预测,获得第一语音表示;同时,利用所述对比预测模块中的自回归模型对所述语音向量进行预测,获得第二语音表示;

[0192] 利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合,获得融合后语音表示;

[0193] 基于所述目标声学特征,结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别,获得转录文本。

[0194] 综上所述,本发明实施例提供的基于双流自监督网络的语音识别方法,在编码与量化模块之后并行结合重构预测模块(Reconstruction Prediction Module,RPM)和对比预测模块(Contrastive Prediction Module,CPM)设计了一个双通道结构。其中,将重建预测作为对比预测的辅助任务分别对语音向量进行预测语音帧,从而在建模不同语音表示之间的归属关系捕获不同特征差异信息的同时,关注详细的上下文信息。此外,为了有效地利用双通道语音表示,还通过特征融合子模块来融合两个分支的语音表示,该特征融合子模块通过参数可学习策略自适应融合两个分支的语音表示,并利用权值来控制各种语音特征的暴露。最后,本发明提供的双流自监督学习网络可以很好地初始化ASR模型的权重。与其他自监督学习方法相比,本发明提供的语音识别方法可以达到具有竞争力的预测精度。此外,在有限的标记数据场景下,与最先进的自监督学习方法相当。

[0195] 本领域普通技术人员可以理解:附图只是一个实施例的示意图,附图中的模块或流程并不一定是实施本发明所必须的。

[0196] 本说明书中的各个实施例均采用递进的方式描述,各个实施例之间相同相似的部分互相参见即可,每个实施例重点说明的都是与其他实施例的不同之处。尤其,对于方法或装置实施例而言,由于其基本相似于方法实施例,所以描述得比较简单,相关之处参见方法实施例的部分说明即可。以上所描述的方法及装置实施例仅仅是示意性的,其中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模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下,即可以理解并实施。

[0197] 以上,仅为本发明较佳的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到的变化或替换,都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应该以权利要求的保护范围为准。

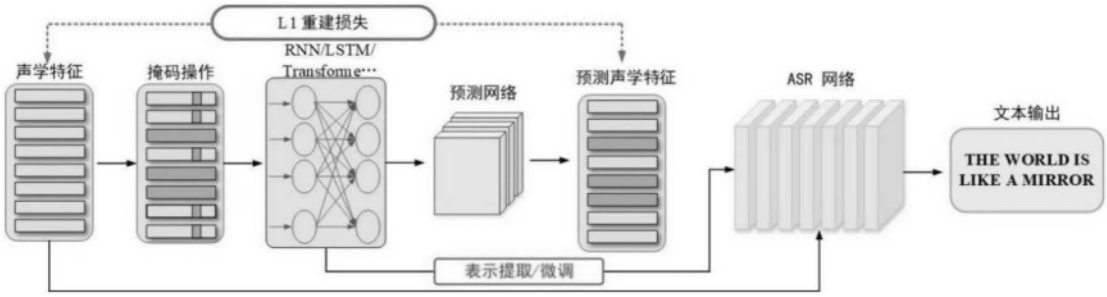

L1重建损失
声学特征
掩码操作
RNN/LSTM/
Transforme…
预测网络
预测声学特征
ASR 网络
文本输出
THE WORLD IS
LIKE A MIRROR
表示提取/微调

图1

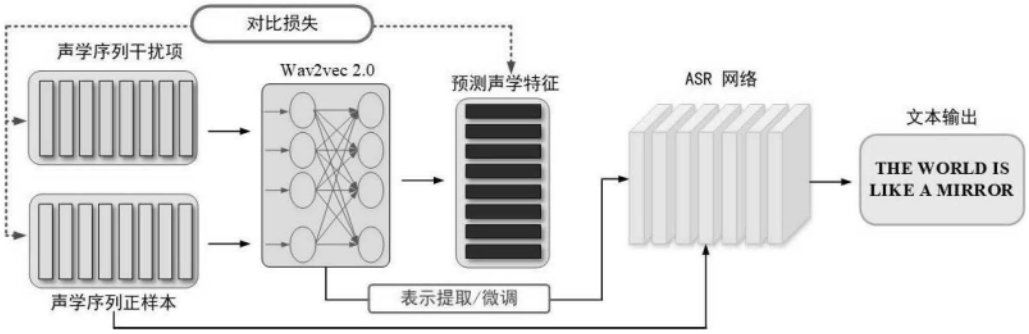

对比损失
声学序列干扰项
Wav2vec 2.0
预测声学特征
ASR 网络
文本输出
THE WORLD IS
LIKE A MIRROR
表示提取/微调
声学序列正样本

图2

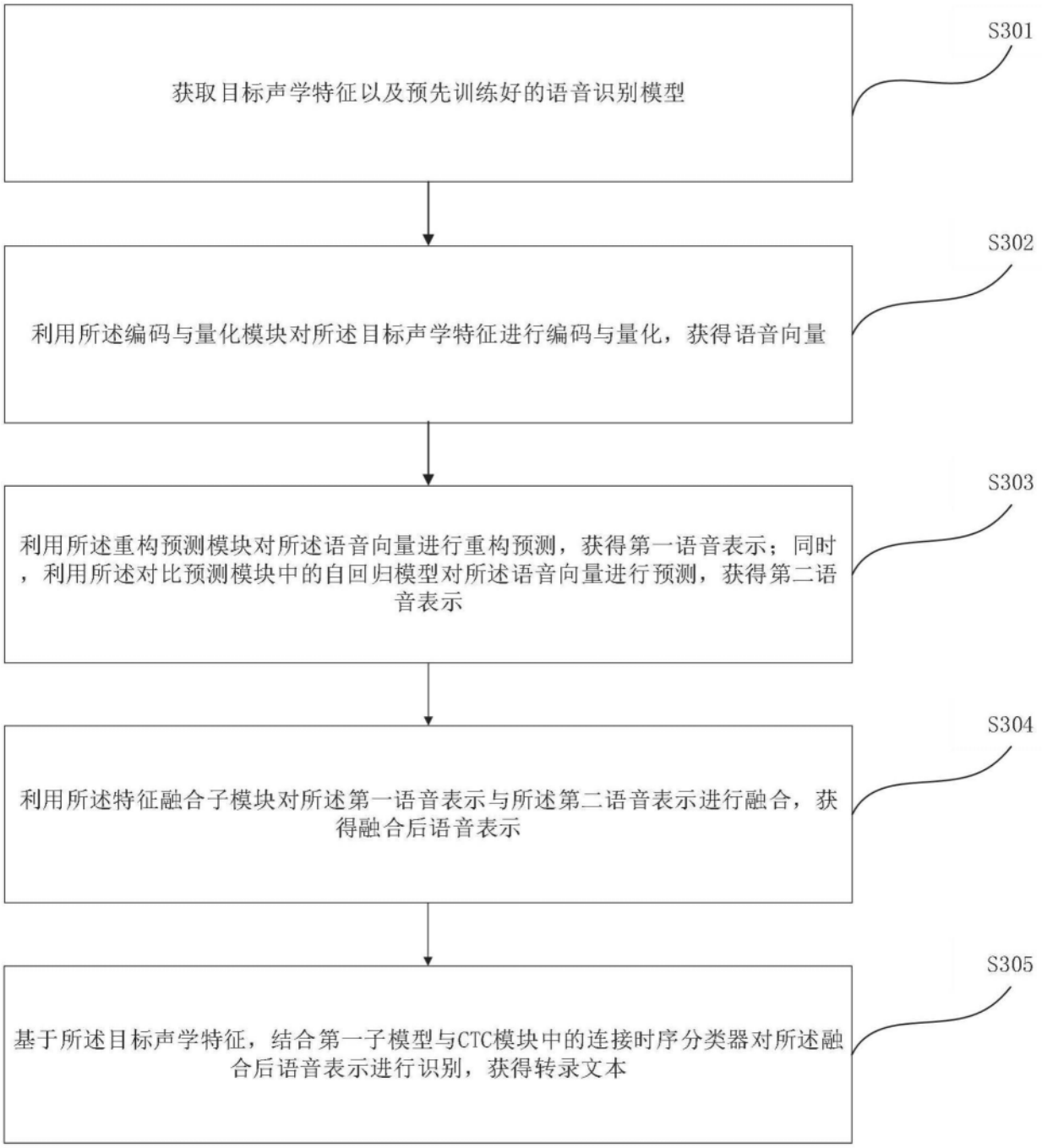

S301
获取目标声学特征以及预先训练好的语音识别模型
S302
利用所述编码与量化模块对所述目标声学特征进行编码与量化，获得语音向量
S303
利用所述重构预测模块对所述语音向量进行重构预测，获得第一语音表示；同时，利用所述对比预测模块中的自回归模型对所述语音向量进行预测，获得第二语音表示
S304
利用所述特征融合子模块对所述第一语音表示与所述第二语音表示进行融合，获得融合后语音表示
S305
基于所述目标声学特征，结合第一子模型与CTC模块中的连接时序分类器对所述融合后语音表示进行识别，获得转录文本

图3

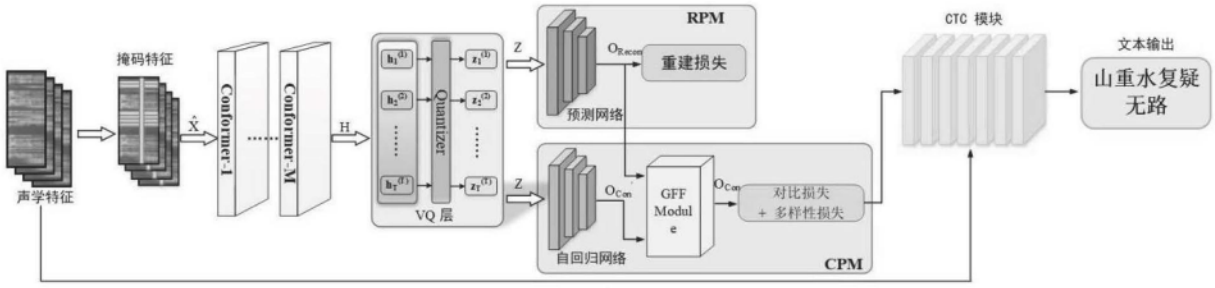

声学特征
掩码特征
Conformer-1
Conformer-M
H
Quantizer
VQ 层
Z
RPM
重建损失
预测网络
自回归网络
GFF Module
对比损失 + 多样性损失
CPM
CTC 模块
文本输出
山重水复疑无路

图4

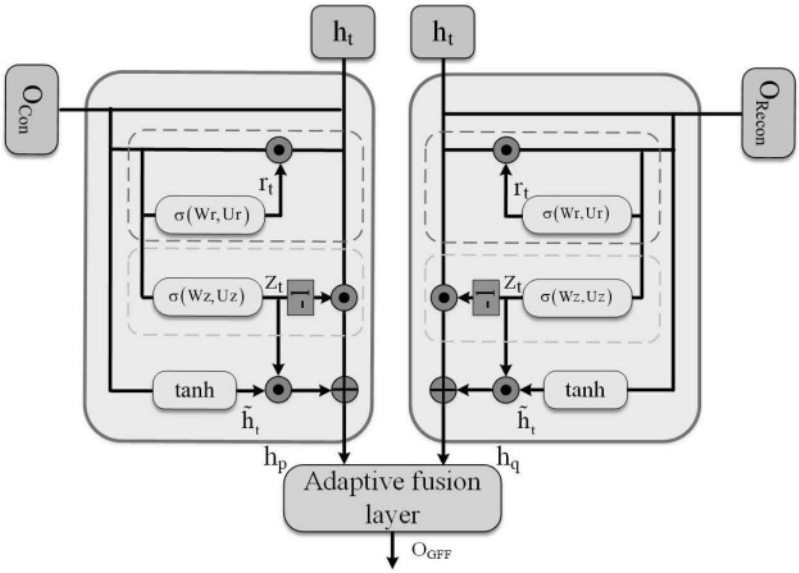

Adaptive fusion layer

图5

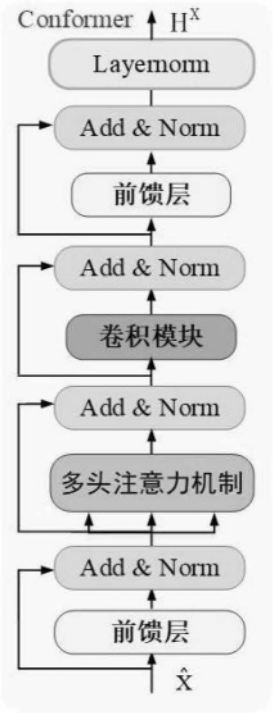

Conformer
H^X
Layernorm
Add & Norm
前馈层
Add & Norm
卷积模块
Add & Norm
多头注意力机制
Add & Norm
前馈层
$\hat{X}$

图6