(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115910093 A
(43)申请公布日 2023.04.04

(21)申请号 202211573044.4

(22)申请日 2022.12.08

(71)申请人 思必驰科技股份有限公司
地址 215123 江苏省苏州市苏州工业园区新平街388号腾飞创新园14栋

(72)发明人 钱彦旻 李晨达 吴逸飞

(74)专利代理机构 北京商专永信知识产权代理事务所(普通合伙) 11400
专利代理师 黄谦 侯晓艳

(51)Int.Cl.
G10L 21/0272(2013.01)
G10L 17/04(2013.01)
G10L 15/06(2013.01)
G10L 15/18(2013.01)

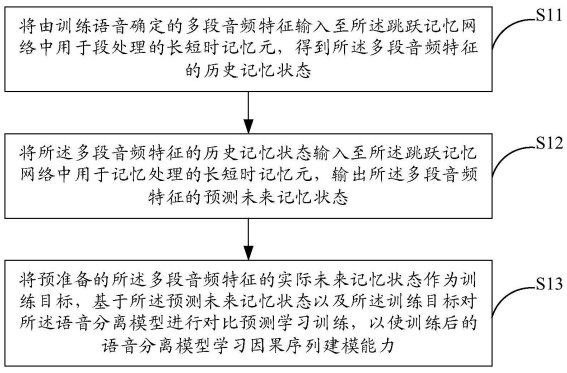
权利要求书2页 说明书8页 附图4页

(54)发明名称

基于跳跃记忆网络的语音分离模型训练方法及系统

(57)摘要

本发明实施例提供一种基于跳跃记忆网络的语音分离模型训练方法及系统。该方法包括：将由训练语音确定的多段音频特征输入至跳跃记忆网络中用于段处理的长短时记忆元，得到多段音频特征的历史记忆状态；将多段音频特征的历史记忆状态输入至跳跃记忆网络中用于记忆处理的长短时记忆元，输出多段音频特征的预测未来记忆状态；将预准备的多段音频特征的实际未来记忆状态作为训练目标，基于预测未来记忆状态以及训练目标对语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力。本发明实施例训练了模型的因果建模能力，使其在不增加延迟的基础上提高了语音分离性能，同时降低了处理延迟。



CN 115910093 A

1.一种基于跳跃记忆网络的语音分离模型训练方法,包括:

将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元,得到所述多段音频特征的历史记忆状态;

将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元,输出所述多段音频特征的预测未来记忆状态;

将预准备的所述多段音频特征的实际未来记忆状态作为训练目标,基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练,以使训练后的语音分离模型学习因果序列建模能力。

2.根据权利要求1所述的方法,其中,所述用于段处理的长短时记忆元的结构包括:编码器网络、连续长短时记忆元构成的中间网络、解码器网络。

3.根据权利要求2所述的方法,其中,所述将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元,得到所述多段音频特征的历史记忆状态包括:

基于编码器网络对所述多段音频特征进行特征压缩,得到第一帧率的多段音频特征;

将所述第一帧率的多段音频特征输入至所述连续长短时记忆元构成的中间网络,得到所述第一帧率的多段音频特征的历史记忆状态;

通过所述解码器网络对所述第一帧率的多段音频特征的历史记忆状态进行解码,得到第二帧率的多段音频特征的历史记忆状态,其中,所述第二帧率高于所述第一帧率,以降低所述中间网络计算量。

4.根据权利要求3所述的方法,其中,所述将所述第一帧率的多段音频特征输入至所述连续长短时记忆元构成的中间网络包括:

将所述多段音频特征中的每一段音频特征分别依次的输入至所述连续长短时记忆元,通过所述连续长短时记忆元为对应段的音频特征进行局部上下文编解码,得到各段音频特征的历史记忆状态。

5.一种基于跳跃记忆网络的语音分离模型训练系统,包括:

历史记忆确定程序模块,用于将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元,得到所述多段音频特征的历史记忆状态;

预测程序模块,用于将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元,输出所述多段音频特征的预测未来记忆状态;

训练程序模块,用于将预准备的所述多段音频特征的实际未来记忆状态作为训练目标,基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练,以使训练后的语音分离模型学习因果序列建模能力。

6.根据权利要求5所述的系统,其中,所述用于段处理的长短时记忆元的结构包括:编码器网络、连续长短时记忆元构成的中间网络、解码器网络。

7.根据权利要求6所述的系统,其中,所述历史记忆确定程序模块用于:

基于编码器网络对所述多段音频特征进行特征压缩,得到第一帧率的多段音频特征;

将所述第一帧率的多段音频特征输入至所述连续长短时记忆元构成的中间网络,得到所述第一帧率的多段音频特征的历史记忆状态;

通过所述解码器网络对所述第一帧率的多段音频特征的历史记忆状态进行解码,得到

第二帧率的多段音频特征的历史记忆状态，其中，所述第二帧率高于所述第一帧率，以降低所述中间网络计算量。

8.根据权利要求7所述的系统，其中，所述历史记忆确定程序模块还用于：

将所述多段音频特征中的每一段音频特征分别依次的输入至所述连续长短时记忆元，通过所述连续长短时记忆元为对应段的音频特征进行局部上下文编解码，得到各段音频特征的历史记忆状态。

9.一种电子设备，其包括：至少一个处理器，以及与所述至少一个处理器通信连接的存储器，其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-4中任一项所述方法的步骤。

10.一种存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求1-4中任一项所述方法的步骤。

基于跳跃记忆网络的语音分离模型训练方法及系统

技术领域

[0001] 本发明涉及智能语音领域，尤其涉及一种基于跳跃记忆网络的语音分离模型训练方法及系统。

背景技术

[0002] 语音分离技术可以用作于复杂声学环境下语音处理系统的前端，来分离多个说话人混合语音中每个说话人的语音。然而许多对话场景要求语音处理系统具有低延迟，例如远程会议、同声传译、助听器等。语音分离前端的延迟决定了整个系统的延迟下限。因此，低延迟的语音分离系统具有重要意义。

[0003] 通常可以使用SkiM模型(Skipping Memory，跳跃记忆网络)来进行语音分离，其中，SkiM模型是一种用于在线的语音分离的神经网络模型，上述语音分离系统的实际延迟来自两个部分。第一个是理想延迟，或称为算法延迟。在特征编码过程中，可以利用小窗口可以将系统的理想延迟压缩到很小的程度。另一个导致实际延迟的因素是处理延迟，这与模型的计算量和硬件速度有关。但是，当使用可以利用小窗口将系统的理想延迟压缩到很小的程度时，就会使得需要处理的特征帧变得更多，这增加了模型的总计算量。

[0004] 在实现本发明过程中，发明人发现相关技术中至少存在如下问题：

[0005] SkiM模型为了达到极低的理论延迟，对特征的建模粒度相对很细。细粒度的特征使得SkiM模型每秒钟需要处理的特征数量大大增加。从而对计算性能的需求较大，存在一定的处理延迟。为了进一步提升语音分离效果，会在语音分离时引入更多的“未来”信息(例如采集后几秒的语音来辅助当前语音片段的语音分离)，上述操作都会引入更多的延迟。延迟和语音分离系统的性能通常是一种互相取舍的关系。

发明内容

[0006] 为了至少解决现有技术中语音分离系统的延迟和性能互相取舍的问题。第一方面，本发明实施例提供一种基于跳跃记忆网络的语音分离模型训练方法，包括：

[0007] 将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元，得到所述多段音频特征的历史记忆状态；

[0008] 将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元，输出所述多段音频特征的预测未来记忆状态；

[0009] 将预准备的所述多段音频特征的实际未来记忆状态作为训练目标，基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力。

[0010] 第二方面，本发明实施例提供一种基于跳跃记忆网络的语音分离模型训练系统，包括：

[0011] 历史记忆确定程序模块，用于将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元，得到所述多段音频特征的历史记忆状态；

[0012] 预测程序模块，用于将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元，输出所述多段音频特征的预测未来记忆状态；

[0013] 训练程序模块，用于将预准备的所述多段音频特征的实际未来记忆状态作为训练目标，基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力。

[0014] 第三方面，提供一种电子设备，其包括：至少一个处理器，以及与所述至少一个处理器通信连接的存储器，其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本发明任一实施例的基于跳跃记忆网络的语音分离模型训练方法的步骤。

[0015] 第四方面，本发明实施例提供一种存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现本发明任一实施例的基于跳跃记忆网络的语音分离模型训练方法的步骤。

[0016] 本发明实施例的有益效果在于：对用于语音分离的SkiM模型的两个长短时记忆元分别进行了改进，对于Mem-LSTM用于记忆处理的长短时记忆元训练了因果建模能力，使其在不增加延迟的基础上提高了语音分离性能；改进了Seg-LSTM用于段处理的长短时记忆元的结构，压缩了局部特征，降低SkiM模型的处理延迟。改进后的SkiM模型可以在更加低功耗的设备上进行部署，提升了SkiM模型的使用广度以及处理性能。

附图说明

[0017] 为了更清楚地说明本发明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一简单地介绍，显而易见地，下面描述中的附图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。

[0018] 图1是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的流程图；

[0019] 图2是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的跳跃记忆网络示意图；

[0020] 图3是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的对比预测编码示意图；

[0021] 图4是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的局部特征可视化示意图；

[0022] 图5是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的模型之间结果比较示意图；

[0023] 图6是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的数据对比示意图；

[0024] 图7是本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练系统的结构示意图；

[0025] 图8为本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练的电子设备的实施例的结构示意图。

具体实施方式

[0026] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作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。

[0027] 如图1所示为本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练方法的流程图，包括如下步骤：

[0028] S11：将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元，得到所述多段音频特征的历史记忆状态；

[0029] S12：将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元，输出所述多段音频特征的预测未来记忆状态；

[0030] S13：将预准备的所述多段音频特征的实际未来记忆状态作为训练目标，基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力。

[0031] 在本实施方式中，考虑到用于语音分离的SkiM模型由两个LSTM组成，一个是用于长跨度建模的Mem-LSTM(memory Long short-term memory，用于长跨度记忆处理的长短时记忆元)，一个是用于本地局部建模的Seg-LSTM(segmentation Long short-term memory，用于段处理的长短时记忆元)。为了降低模型的延迟，分别对这两个LSTM进行改进训练。

[0032] 本方法对原有SkiM模型的扩展。在Mem-LSTM训练阶段采用了CPC(contrastive predictive coding，对比预测编码)；改进后的Mem-LSTM用于从历史记忆状态(由“historical memory states”直译)中估计未来记忆(由“future memory”直译)，使训练后的SkiM具有更好的因果建模能力。训练后的SkiM模型在推理阶段不再使用未来信息(例如，也就是不在分离当前语音时等待后续语音的输入，降低了理想延迟)。

[0033] 对于步骤S11，在低延迟在线语音分离目标下，需要处理语音中的长特征序列，如图2所示，识别输入的训练语音的音频特征W，将特征W分成更小的段$\{\mathbf{W}^s \in \mathbb{R}^{K\times N}\}$，其中s＝1，…，s。s代表段数，K代表段长，N代表特征大小。整体用于语音分离的SkiM模型结构如图3所示，将其输入至Seg-LSTM中，第l块Seg-LSTM的映射函数为：

[0034] $$\mathbf{W}_{l+1}^{s}, \hat{\mathbf{r}}_{l}^{s} = \text{Seg-LSTM}\left(\mathbf{W}_{l}^{s}, \mathbf{r}_{l}^{s}\right)$$

[0035] 其中，$\mathbf{W}_l^s$是第l块中第s段的输入特征，$\mathbf{r}_l^s$表示Seg-LSTM中LSTM层的初始隐藏和单元存储状态。第一个SkiM块中的$\mathbf{r}_1^s$为0，$\hat{\mathbf{r}}_l^s$为输出的编码第s段的局部信息的记忆状态，然后Mem-LSTM所有记忆状态，用于大跨度建模：

[0036] $$\hat{\mathbf{R}}_l = \begin{cases} [\mathbf{0}, \hat{\mathbf{r}}_l^1, \cdots, \hat{\mathbf{r}}_l^{S-1}], & \text{如果有因果关系} \\ [\hat{\mathbf{r}}_l^1, \cdots, \hat{\mathbf{r}}_l^S], & \text{如果没有因果关系} \end{cases}$$

[0037] $$\mathbf{R}_{l+1} = \text{Mem-LSTM}\left(\hat{\mathbf{R}}_l\right)$$

[0038] 其中，$\mathbf{R}_{l+1} = [\mathbf{r}_{l+1}^1, \cdots, \mathbf{r}_{l+1}^S]$是全局同步的记忆状态，它将被用作下一个SkiM块的Seg-LSTM的初始状态。在最后一个SkiM块之后，最后的输出段$\{\mathbf{W}_{L+1}^s\}$被合并，得到长度

为t的连续输出记忆状态，将其确定为多段音频特征的历史记忆状态。

[0039] 对于步骤S12，将多段音频特征的历史记忆状态输入至SkiM模型中的Mem-LSTM，通过交替使用Seg-LSTMs和Mem-LSTM，训练后的SkiM模型可以在因果语音分离模型中对一个非常长的序列的历史信息进行建模。继续如图3所示，Mem-LSTM的输入是从每个Seg-LSTMs确定的历史记忆状态$\{\hat{\mathbf{r}}_l^s\}$，Mem-LSTM输出长跨度的预测未来记忆状态：

[0040] $$\mathbf{r}_{l+1}^s = \text{Mem-LSTM}(\{\hat{\mathbf{r}}_l^{\leq s}\})$$

[0041] 其中，长跨度的预测未来记忆状态$\mathbf{r}_{l+1}^s$可视为对历史输入的上下文进行编码$\{\mathbf{W}_l^{\leq s}\}$。在SkiM的预测中，希望Mem-LSTM具有从$\mathbf{r}_{l+1}^s$预测特征$\mathbf{W}_l^{s+d}$的能力，即$p(\mathbf{W}_l^{s+d}|\mathbf{r}_{l+1}^s)$，其中d为未来多个段的数量。

[0042] 对于步骤S13，本方法使用预准备的所述多段音频特征的实际未来记忆状态作为训练目标作为$I(\mathbf{W}_l^{s+d}, \mathbf{r}_{l+1}^s)$呈正相关的函数$f_l^d$：

[0043] $$f_l^d(\mathbf{W}_l^{s+d}, \mathbf{r}_{l+1}^s) \propto \frac{p(\mathbf{W}_l^{s+d}|\mathbf{r}_{l+1}^s)}{p(\mathbf{W}_l^{s+d})}$$

[0044] 通过最大化f_l^d，可以最大化$\mathbf{W}_l^{s+d}$与$\mathbf{r}_{l+1}^s$之间的互信息，这意味着从$\mathbf{r}_{l+1}^s$重构$\mathbf{W}_l^{s+d}$更容易。采用对数双线性模型f_d：

[0045] $$f_l^d(\mathbf{W}_l^{s+d}, \mathbf{r}_{l+1}^s) = \exp(\hat{\mathbf{r}}_l^{sT} \mathbf{P}_l^d \mathbf{r}_{l+1}^s)$$

[0046] 其中，$\mathbf{P}_l^d$是一个线性变换的参数，从$\mathbf{r}_{l+1}^s$估计$\hat{\mathbf{r}}_l^s$。为了使$f_l^d(\mathbf{W}_l^{s+d}, \mathbf{r}_{l+1}^s)$最大化，使用对比预测学习损失来对语音分离模型进行训练：

[0047] $$\mathcal{L}_{l,d}^M = -\mathbb{E}_{\mathcal{W}}\left[\log \frac{f_l^d(\mathbf{W}_l^{s+d}, \mathbf{r}_{l+1}^s)}{\sum_{\mathbf{W}_m \in \mathcal{W}} f_l^d(\mathbf{W}_m, \mathbf{r}_{l+1}^s)}\right]$$

[0048] 其中W＝{W_1,…,W_M}是包含$\mathbf{W}_l^{s+d}$和M-1个随机负样本的集合。本方法训练的SkiM模型的最终训练目标可以写成：

[0049] $$\mathcal{L} = \mathcal{L}_{PIT} + \lambda \sum_{l=1}^{L} \sum_{d=1}^{D} \mathcal{L}_{l,d}^M$$

[0050] 其中，D为预测的最大轮数，L_{PIT}是一种基于置换不变训练(PIT)方法的不变源噪声比(SI-SNR)语音分离损失，λ是一个加权因子，最终训练后的语音分离模型学习因果序列建模能力，能够直接从输入的语音特征推理出未来记忆，进而在确保提升语音分离效率的基础上不增加语音分离模型的延迟。

[0051] 作为一种实施方式，所述用于段处理的长短时记忆元的结构包括：编码器网络、连续长短时记忆元构成的中间网络、解码器网络。

[0052] 基于编码器网络对所述多段音频特征进行特征压缩，得到第一帧率的多段音频特征；

[0053] 将所述多段音频特征中的每一段音频特征分别依次的输入至所述连续长短时记

忆元,通过所述连续长短时记忆元为对应段的音频特征进行局部上下文编解码,得到各段音频特征的历史记忆状态。

[0054] 通过所述解码器网络对所述第一帧率的多段音频特征的历史记忆状态进行解码,得到第二帧率的多段音频特征的历史记忆状态,其中,所述第二帧率高于所述第一帧率,以降低所述中间网络计算量。

[0055] 在本实施方式中,考虑到现有的语音分离模型的基本层被分为SIMO(Single Input and Multiple Outputs,单输入多输出)和SISO单输入单输出(Single Input and Single Outputs)模块。SIMO模块处理多源混合语音的深度特征。它或其后续SIMO模块将深度特征分离为多个流,这些流对应于单说话人语音。大多数传统的盲源分离系统都是纯SIMO设计的,其中深度特征被分离在分离器的最后一层。SIMO-SISO模型是一个分离前和增强后的管道。深度特征在早期用编码器网络SIMO层分离,在后期用解码器网络SISO层增强。研究表明,在相同的参数数量下,这种编码-解码(SIMO-SISO)设计可以提高分离性能。也可以将本方法SIMO-SISO结构的模型称为后增强(PE)模型。

[0056] 本方法训练了一个现有技术的SIMO-only SkiM模型和本方法所述的编码器网络、连续长短时记忆元构成的中间网络、解码器网络结构的PE SkiM模型。然后在两种模型中分别对混合语音进行前向传播。图4给出了局部段特征$\mathbf{W}_s \in \mathbb{R}^{K\times N}$在不同中间SkiM块中的可视化结果。通过观察,可以发现混合音频SIMO模块(在SIMO-only和PE模型中)的特征没有显示出独特的模式。在PE SkiM模型中,从两个扬声器分支的SISO模块中分离出的特征可以观察到周期性。周期性可能来自于小段分离语音的短时固定。局部SISO特征具有较好的周期性,意味着特征具有较强的冗余性和可压缩性。基于这一发现,本方法继续在训练后的SkiM模型中的用于段处理的长短时记忆元集成了一个局部上下文编解码器(LCC),以减少实时应用中的计算成本。在第一个SISO之前,插入一个编码器将$\mathbf{W}_s \in \mathbb{R}^{K\times N}$映射到$\bar{\mathbf{W}}_s \in \mathbb{R}^{\bar{K}\times N}$,其中,$\bar{K} \ll K$,在最后一个SISO块之后,解码器映射长度从$\bar{K}$到K。因此,可以降低中间块的计算成本。

[0057] 通过该实施方式可以看出,本方法对用于语音分离的SkiM模型的两个长短时记忆元分别进行了改进,对于Mem-LSTM用于记忆处理的长短时记忆元训练了因果建模能力,使其在不增加延迟的基础上提高了语音分离性能;改进了Seg-LSTM用于段处理的长短时记忆元的结构,压缩了局部特征,降低SkiM模型的处理延迟。改进后的SkiM模型可以在更加低功耗的设备上进行部署,提升了SkiM模型的使用广度以及处理性能。

[0058] 对本方法进行实验说明,本方法在WSJ0-2mix数据集上进行了实验,其是一个广泛使用的语音分离基准数据集。基线模型是一个因果SkiM模型,它包含6个SkiM块,所有的SkiM块都是SIMO模块。Seg-LSTM和Mem-LSTM都是单向LSTM,有256个隐藏单元。卷积编码器和转置卷积解码器有128个通道。

[0059] 在4种不同的设置上比较了提出的本方法训练的SkiM模型,为了方便说明将本方法训练的SkiM模型表示为pSkiM模型和现有SkiM模型。它们的步幅大小为{4,8,16,32},卷积核大小为{8,16,32,64},SkiM块中相应的分割大小K分别为{64,48,32,24}。

[0060] 在pSkiM中,估计的最大步幅数是10,λ为1.0。训练过程中批次大小设置为32,CPC损失的负样本从其他小批次的样本中选择。

[0061] 在本地上下文编解码器(LCC)实验中,在PE SkiM中,前两个SkiM块用作SIMO层,而其余4个用作SISO层。PE SkiM中的参数数量与SIMO-only基线相当。上下文编码器和解码器都是128单元的单层单向LSTM。编解码器步幅设置为4。编码器每4步输出一个编码特征,而解码输出自回归地为每个输入运行4步。为了避免算法延迟的增加,解码器的当前输入是最后4步的特征,因此不需要等待编码器4步解码当前步骤。

[0062] 所有模型都是用ESPNet-SE工具包实现的。Adam优化器用于训练。初始学习率(LR)设置为10^{-3}。训练了150个epoch的模型,LR在每个周期减少0.98。

[0063] 首先检验了CPC训练对Mem-LSTM的影响。如图5列出了不同系统下分离语音的SI-SNRi改善情况。还报告了每秒乘数-累加运算(mac)的次数。在实际应用程序中,具有较小mac/s的系统将具有较小的处理延迟。根据上述实验结果,发现卷积编码器的内核尺寸越小,语音分离性能越好,理想延迟越小,mac/s越大。与所有基线模型相比,本方法的SkiM模型具有较好的分离性能。在4ms和8ms设置中,改进更加明显,而在2ms和16ms设置中,性能的提高是最小的。特征上下文的预测只在训练阶段进行,因此推理阶段的总体计算成本与基线模型相同。

[0064] PE SkiM的性能如图6的第二行。配备CPC训练的pSkiM与基线模型的差距也不大。本方法将PE SkiM训练与CPC训练相结合时,可以观察到明显的改善。通过预测分离的单一说话者流的记忆状态,可以获得比混合流更好的上下文依赖性。

[0065] 比较了三种不同的本地上下文编解码器(LCC)策略。第一种是将LCC应用于PE SkiM的SISO模块上。第二种是在SIMO模块上进行LCC。最后一种是在整个网络上应用LCC。如图6的结果表明,在SISO模块上应用LCC在所有模型中具有优势。它比基线模型提高了1.2dB,同时需要更少的计算成本。在SIMO模块中增加LCC时,性能提升较小。如果整个模型都配备了LCC,计算成本可以降低60%,但会有轻微的性能下降。可以得出SISO分支中的特征更适合使用局部上下文编码进行压缩。编码和解码处理不会损伤深度特征内的信息,同时节省了计算成本。

[0066] 如图7所示为本发明一实施例提供的一种基于跳跃记忆网络的语音分离模型训练系统的结构示意图,该系统可执行上述任意实施例所述的基于跳跃记忆网络的语音分离模型训练方法,并配置在终端中。

[0067] 本实施例提供的一种基于跳跃记忆网络的语音分离模型训练系统10包括:历史记忆确定程序模块11,预测程序模块12和训练程序模块13。

[0068] 其中,历史记忆确定程序模块11用于将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元,得到所述多段音频特征的历史记忆状态;预测程序模块12用于将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元,输出所述多段音频特征的预测未来记忆状态;训练程序模块13用于将预准备的所述多段音频特征的实际未来记忆状态作为训练目标,基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练,以使训练后的语音分离模型学习因果序列建模能力。

[0069] 本发明实施例还提供了一种非易失性计算机存储介质,计算机存储介质存储有计算机可执行指令,该计算机可执行指令可执行上述任意方法实施例中的基于跳跃记忆网络的语音分离模型训练方法;

[0070] 作为一种实施方式，本发明的非易失性计算机存储介质存储有计算机可执行指令，计算机可执行指令设置为：

[0071] 将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元，得到所述多段音频特征的历史记忆状态；

[0072] 将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元，输出所述多段音频特征的预测未来记忆状态；

[0073] 将预准备的所述多段音频特征的实际未来记忆状态作为训练目标，基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力。

[0074] 作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块，如本发明实施例中的方法对应的程序指令/模块。一个或者多个程序指令存储在非易失性计算机可读存储介质中，当被处理器执行时，执行上述任意方法实施例中的基于跳跃记忆网络的语音分离模型训练方法。

[0075] 图8是本申请另一实施例提供的基于跳跃记忆网络的语音分离模型训练方法的电子设备的硬件结构示意图，如图8所示，该设备包括：

[0076] 一个或多个处理器810以及存储器820，图8中以一个处理器810为例。基于跳跃记忆网络的语音分离模型训练方法的设备还可以包括：输入装置830和输出装置840。

[0077] 处理器810、存储器820、输入装置830和输出装置840可以通过总线或者其他方式连接，图8中以通过总线连接为例。

[0078] 存储器820作为一种非易失性计算机可读存储介质，可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块，如本申请实施例中的基于跳跃记忆网络的语音分离模型训练方法对应的程序指令/模块。处理器810通过运行存储在存储器820中的非易失性软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例基于跳跃记忆网络的语音分离模型训练方法。

[0079] 存储器820可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储数据等。此外，存储器820可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实施例中，存储器820可选包括相对于处理器810远程设置的存储器，这些远程存储器可以通过网络连接至移动装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0080] 输入装置830可接收输入的数字或字符信息。输出装置840可包括显示屏等显示设备。

[0081] 所述一个或者多个模块存储在所述存储器820中，当被所述一个或者多个处理器810执行时，执行上述任意方法实施例中的基于跳跃记忆网络的语音分离模型训练方法。

[0082] 上述产品可执行本申请实施例所提供的方法，具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节，可参见本申请实施例所提供的方法。

[0083] 非易失性计算机可读存储介质可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据装置的使用所创建的数据等。此外，非易失性计算机可读存储介质可以包括高速随机存取存储器，还

可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实施例中，非易失性计算机可读存储介质可选包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0084] 本发明实施例还提供一种电子设备，其包括：至少一个处理器，以及与所述至少一个处理器通信连接的存储器，其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本发明任一实施例的基于跳跃记忆网络的语音分离模型训练方法的步骤。

[0085] 本申请实施例的电子设备以多种形式存在，包括但不限于：

[0086] (1)移动通信设备：这类设备的特点是具备移动通信功能，并且以提供话音、数据通信为主要目标。这类终端包括：智能手机、多媒体手机、功能性手机，以及低端手机等。

[0087] (2)超移动个人计算机设备：这类设备属于个人计算机的范畴，有计算和处理功能，一般也具备移动上网特性。这类终端包括：PDA、MID和UMPC设备等，例如平板电脑。

[0088] (3)便携式娱乐设备：这类设备可以显示和播放多媒体内容。该类设备包括：音频、视频播放器，掌上游戏机，电子书，以及智能玩具和便携式车载导航设备。

[0089] (4)其他具有数据处理功能的电子装置。

[0090] 在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”，不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0091] 以上所描述的装置实施例仅仅是示意性的，其中所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的劳动的情况下，即可以理解并实施。

[0092] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0093] 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。

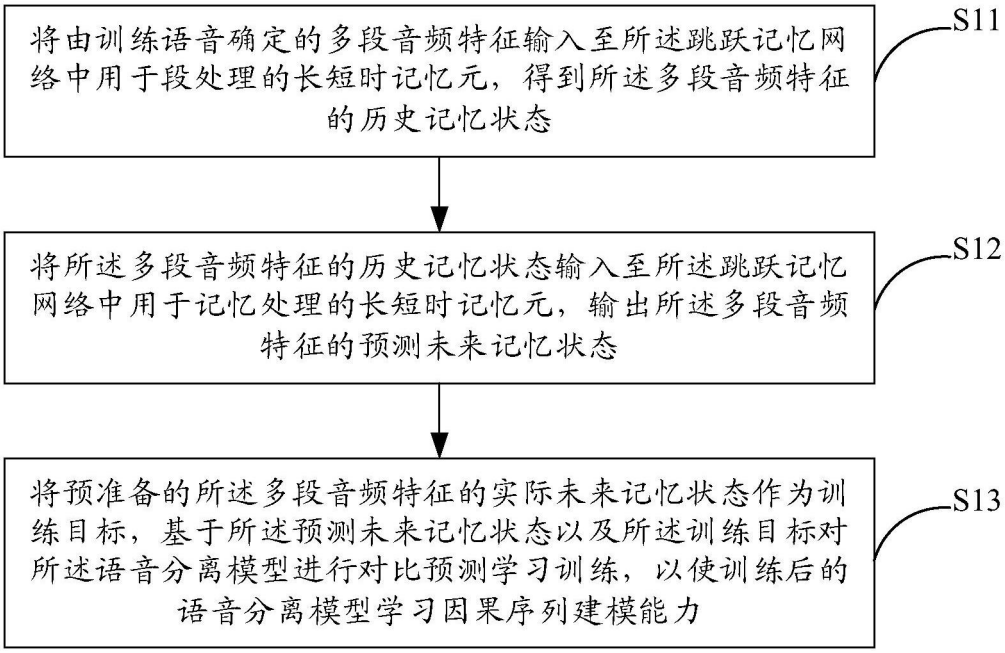

将由训练语音确定的多段音频特征输入至所述跳跃记忆网络中用于段处理的长短时记忆元，得到所述多段音频特征的历史记忆状态
S11
将所述多段音频特征的历史记忆状态输入至所述跳跃记忆网络中用于记忆处理的长短时记忆元，输出所述多段音频特征的预测未来记忆状态
S12
将预准备的所述多段音频特征的实际未来记忆状态作为训练目标，基于所述预测未来记忆状态以及所述训练目标对所述语音分离模型进行对比预测学习训练，以使训练后的语音分离模型学习因果序列建模能力
S13

图1

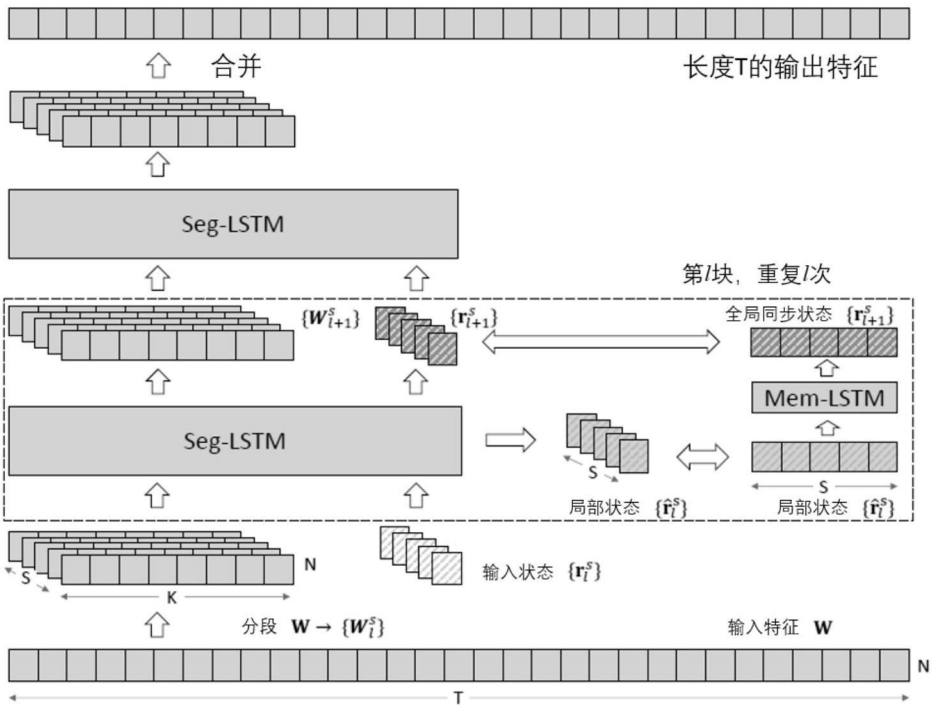

合并
长度T的输出特征
Seg-LSTM
第l块，重复l次
全局同步状态
Mem-LSTM
Seg-LSTM
局部状态
局部状态
S
K
N
输入状态
分段 W → {W_l^s}
输入特征 W
T
N

图2

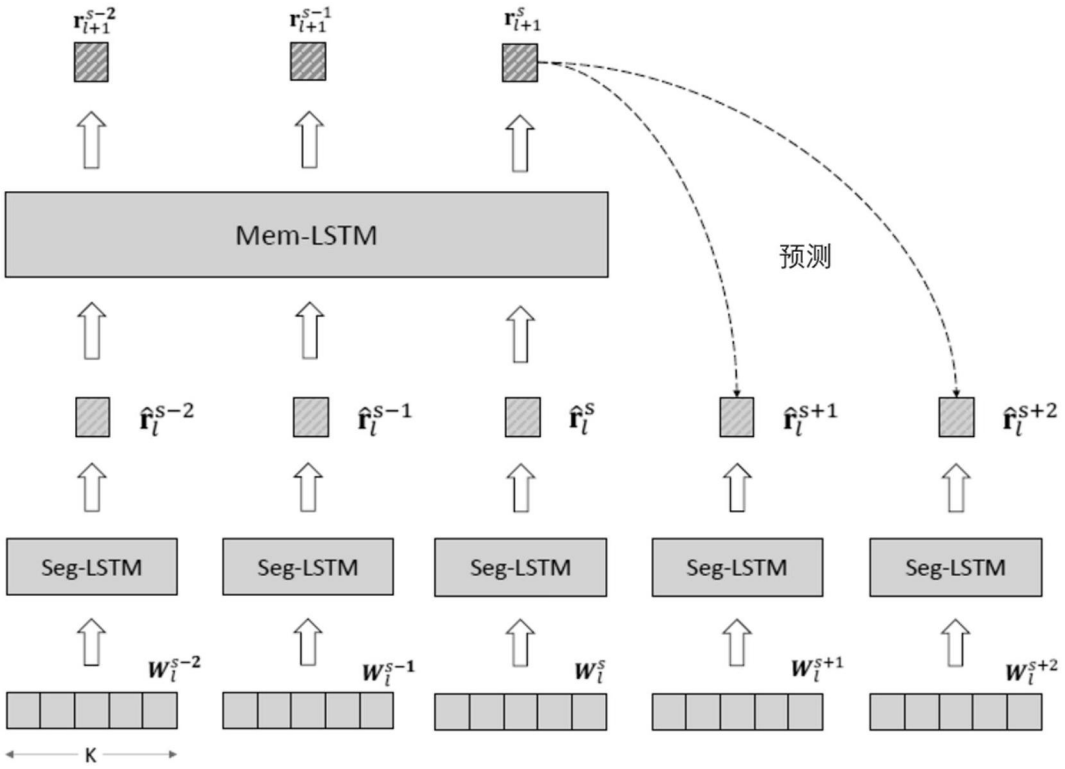

Mem-LSTM
预测
Seg-LSTM
Seg-LSTM
Seg-LSTM
Seg-LSTM
Seg-LSTM
K

图3

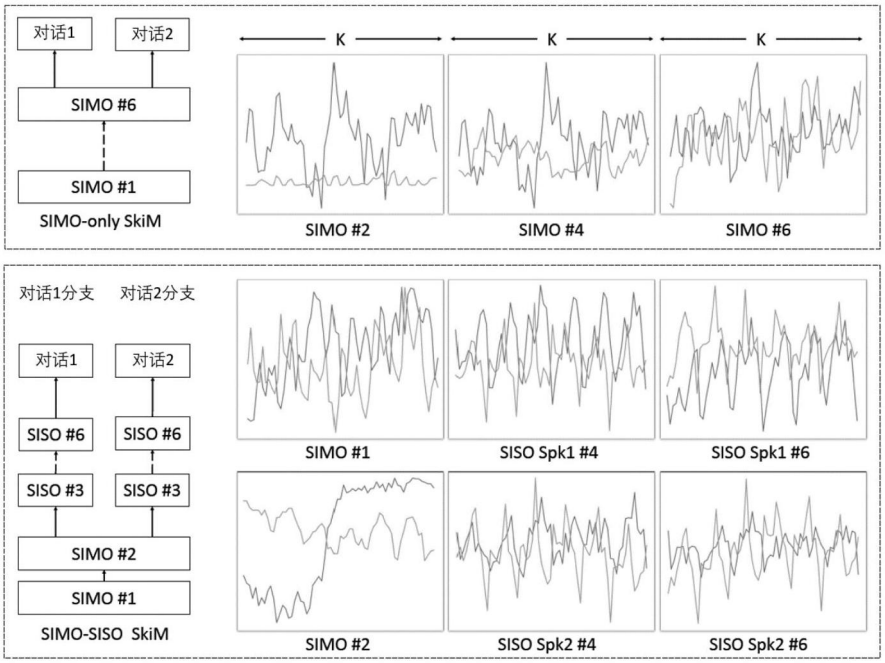

对话1
对话2
SIMO #6
SIMO #1
SIMO-only SkiM
K
K
K
SIMO #2
SIMO #4
SIMO #6
对话1分支
对话2分支
对话1
对话2
SISO #6
SISO #6
SISO #3
SISO #3
SIMO #2
SIMO #1
SIMO-SISO SkiM
SIMO #1
SISO Spk1 #4
SISO Spk1 #6
SIMO #2
SISO Spk2 #4
SISO Spk2 #6

图4

模型	内核大小	模型大小	MACs (G/s)	理想等待时间	SI-SNRi (dB) 训练/测试
SkiM	8	8.5	10.0	2ms	15.1/14.3
+ CPC	8	8.5	10.0	2ms	**15.3/14.5**
SkiM	16	8.5	5.1	4ms	14.7/13.6
+ CPC	16	8.5	5.1	4ms	**15.1/14.0**
SkiM	32	8.5	2.6	8ms	14.0/12.6
+ CPC	32	8.5	2.6	8ms	**14.1/13.2**
SkiM	64	8.5	1.9	16ms	13.0/11.9
+ CPC	64	8.5	1.9	16ms	**13.1/12.0**

图5

模型	内核大小	模型大小	MACs (G/s)	理想等待时间	SI-SNRi (dB) 训练/测试
SkiM	8	8.5	10.0	2ms	15.1/14.3
+ PE	8	8.9	18.3	2ms	15.2/14.3
+ CPC	8	8.5	10.0	2ms	15.3/14.5
+ PE & CPC	8	8.9	18.3	2ms	**15.7/15.1**
++ LCC SISO	8	9.3	9.6	2ms	**16.0/15.5**
++ LCC SIMO	8	9.3	15.2	2ms	15.9/15.0
++ LCC ALL	8	9.3	6.5	2ms	15.3/14.5

图6

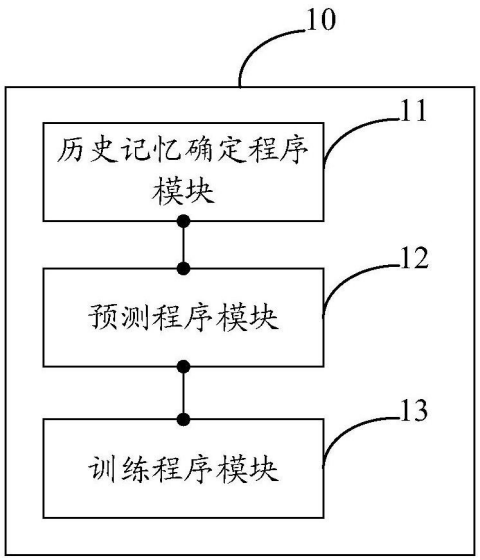

10
11
历史记忆确定程序模块
12
预测程序模块
13
训练程序模块

图7

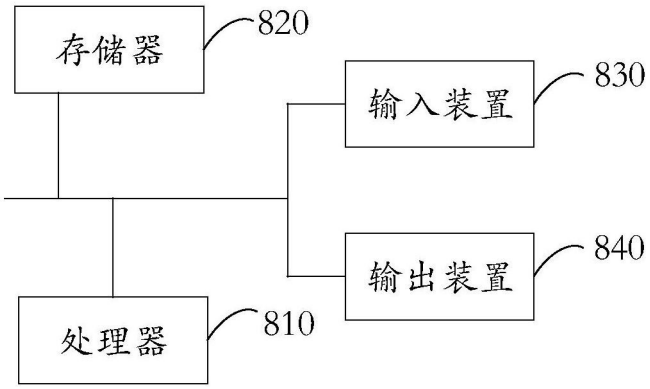

存储器
820
输入装置
830
输出装置
840
处理器
810

图8