(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 110223706 A
(43)申请公布日 2019.09.10

(21)申请号 201910166373.9

(22)申请日 2019.03.06

(71)申请人 天津大学
地址 300072 天津市南开区卫津路92号

(72)发明人 葛檬 王龙标 党建武

(74)专利代理机构 天津市北洋有限责任专利代理事务所 12201
代理人 程小艳

(51)Int.Cl.
G10L 21/0208(2013.01)
G10L 21/0216(2013.01)
G10L 25/30(2013.01)
G10L 25/03(2013.01)

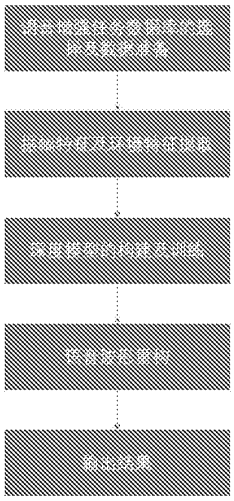
权利要求书2页 说明书4页 附图2页

(54)发明名称

基于注意力驱动循环卷积网络的环境自适应语音增强算法

(57)摘要

本发明公开了一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，包括以下步骤：步骤一，选择语音增强任务数据库，进行输入数据准备；步骤二，提取语音的振幅信息和环境信息，其中语音的环境信息是通过采用权重预测误差方法(WPE)提取，语音的振幅信息主要通过傅里叶变换提取的语谱图信息；步骤三，深度模型的构建和训练；步骤四，语音重构，将步骤三中预测得到的语音振幅转换成语音波形。本发明考虑语音的环境信息，提高了算法在不同环境下的环境自适应性和算法鲁棒性；在真实语音信号保留方面，本发明融入注意力机制构建注意力驱动的循环卷积网络，更加精确地刻画语音的时序上下文信息，有效提高了语音增强的性能。



CN 110223706 A

1.一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，其特征在于，包括以下步骤：

步骤一，语音增强任务数据库的选取及数据准备：

步骤二，振幅特征及环境特征提取：

符号描述：令原始语音信号为s，对语音信号分帧、加窗、短时傅里叶变换，得到的语谱图特征为X；

1)振幅信息提取：直接取语谱图的绝对值，并取log作为语音振幅特征，具体如下：

$x_{振幅}=\log|X|$

2)环境信息提取：基于权重预测误差方法(WPE)来提取语音的环境信息特征；

步骤三，深度模型的构建及训练：

本发明构建环境自适应的端对端深度网络EDANet；

步骤四，语音波形重构：

将步骤三预测得到的语音log振幅特征$\hat{y}$转换成语音波形，转换公式如下：

$$s_{增强}=\exp(\hat{y})\cdot\exp(\angle S)$$

至此，就可以将验证集和测试集的语音进行增强，得到干净的语音波形。

2.根据权利要求1所述的一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，其特征在于，所述步骤二中权重预测误差方法假设原始语音x通过滤波器G能得到想要得到的语音信号是S，其中S满足均值为0，方差为λ的高斯分布，表示为$N_{\mathbb{C}}=(S;0,\lambda)$；

通过最大化log似然函数的方式来求解参数G、λ和语音信号S，求解过程如下：

$$L=\max_{G,\lambda}\prod_{n=1}^{N}N_{\mathbb{C}}(S;0,\lambda)=\min_{G,\lambda}\sum_{n=1}^{N}\frac{|S|^2}{\lambda}+\log\pi\lambda$$

最后求解得到参数G、λ和语音信号S；

环境信息特征提取如下：

$x_{环境}=S$。

3.根据权利要求1所述的一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，其特征在于，所述步骤三中EDANet网络主要分为三个部分：卷积网络，注意力驱动双向循环网络，以及全连接网络，EDANet网络具体的构建细节如下：

1)卷积网络

卷积网络部分，本发明采用了Encoder-DecoderCNN网络，本发明采用了9层卷积层，每层滤波器的数目分别是4,8,16,32,64,32,16,8,4；同时，每个滤波器的大小是3*3；最后经过卷积玩了过部分，总共产生4个2D的特征图，每个特征图的大小是514*7；

2)注意力驱动双向循环网络

将卷积网络产生的所有2D特征图按时间方向拼接在一起，产生特征H(x)，并接着通过注意力驱动的双向循环网络部分；

注意力驱动的双向循环网络细节为给定每个时间步的特征$H_t:=H_t(x)$，计算各帧特征对于目标帧语音的贡献α_t如下：

$$\alpha_t = \frac{\exp(H_t)}{\sum_{i=t-(s-1)/2}^{t+(s-1)/2} \exp(H_i)}$$

然后，将带权的各帧特征$\widehat{H}(\mathrm{x}) := \alpha H(\mathrm{x})$输入到双向循环网络BLSTM，得到融合上下文的时序特征V(x)，具体公式如下：

$$\mathrm{V}(\mathrm{x}) = BLSTM\{\widehat{H}(\mathrm{x})\}$$

其中，本发明中的实验设置的BLSTM层数是2，每层的隐藏单元个数是300；

3)全连接网络

采用全连接网络结合Dropout策略，Dropout是对神经网络进行优化的方法，在学习的过程随机将隐含层的部分权重或者输出归零，降低节点的相互依赖性，从而实现神经网络的正则化，避免模型过拟合。

基于注意力驱动循环卷积网络的环境自适应语音增强算法

技术领域

[0001] 本发明属于语音增强技术领域，尤其是涉及基于注意力驱动循环卷积网络的环境自适应语音增强算法。

背景技术

[0002] 随着智能设备的普及和语音识别技术的快速发展，语音处理技术越来越引起公众关注。在普通的近场(说话人离麦克风比较近)环境下，语音识别的性能已经达到95％以上，许多语音识别和语音合成技术已经商业产品化。然而，在远场(说话人离麦克风距离较远)环境下，往往存在混响及各种背景噪声的影响，语音识别的性能急剧下降。而在远场环境下，由于说话人无须手持麦克风或者佩戴麦克风设备(例如手机设备等)，这种环境更加便利、高效和安全。在当今的物联网人机接口、智能语音交互、智能会议系统等领域有广泛需求。因而，语音增强技术对原始语音降噪并提高语音识别的精度是很有必要的。

[0003] 对于语音增强问题，比较传统的方法是提取语音的振幅特征，通过深度神经网络(DNN) 的方法映射到干净语音。这类方法存在的问题是：仅仅使用语音的振幅信息去增强语音是不够的，这样只能适应当前的噪声环境，往往不能适应其他不同的噪声环境，算法的鲁棒性不够高；另外，利用DNN的方法只是更好的建模时序的语音信号，忽视了语音信号中时间-频率之间的关系，以及难以建模时序语音信号中的动态时序关系，最终会使得增强的语音丢失部分真实语音信号。

发明内容

[0004] 本发明针对现有语音增强模型难以自适应不同噪声环境的问题，提出一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，从而提高了算法在不同环境下的环境自适应性和算法鲁棒性。同时，为了更加精确挖掘语音时序上下文的真实信号关系，本发明融入了注意力机制构建注意力驱动的循环卷积网络，更加精确地刻画语音的时序上下文信息，有效提高了语音增强的性能。

[0005] 为了解决上述技术问题，本发明的技术方案如下：

[0006] 一种基于注意力驱动循环卷积网络的环境自适应语音增强算法，步骤如下：

[0007] 步骤一，语音增强任务数据库的选取及数据准备：

[0008] 本发明选取的语音增强任务数据库是REVERBChallenge2014中的REVERB公开数据集。根据REVERBChallenge2014的要求进行数据准备，划分训练集、验证集、测试集。

[0009] 步骤二，振幅特征及环境特征提取：

[0010] 符号描述：令原始语音信号为s，对语音信号分帧、加窗、短时傅里叶变换，得到的语谱图特征为X。

[0011] 1)振幅信息提取：直接取语谱图的绝对值，并取log作为语音振幅特征，具体如下：

[0012] $x_{振幅}=\log|X|$

[0013] 2)环境信息提取：基于权重预测误差方法(WPE)来提取语音的环境信息特征。权重

预测误差方法假设原始语音x通过滤波器G能得到想要得到的语音信号是S，其中S满足均值为0，方差为λ的高斯分布，表示为$N_{\mathbb{C}}=(S;0,\lambda)$。因此，此时我们就可以通过最大化log似然函数的方式来求解参数G、λ和语音信号S，求解过程如下：

[0014] $$L=\max_{G,\lambda}\prod_{n=1}^{N}N_{\mathbb{C}}\left(S;0,\lambda\right)=\min_{G,\lambda}\sum_{n=1}^{N}\frac{|S|^2}{\lambda}+\log\pi\lambda$$

[0015] 最后求解得到参数G、λ和语音信号S。而本发明中，我们将得到的语音信号S作为环境信息特征，因为S是动态估计不同环境条件下的混响信息从而得到的语音信号，能有效反应不同环境情况下的混响和真实语音信号特点。综上，环境信息特征提取如下：

[0016] $x_{环境}=S$

[0017] 步骤三，深度模型的构建及训练：

[0018] 为了提高算法的鲁棒性，本发明构建环境自适应的端对端深度网络EDANet。。EDANet网络主要分为三个部分：卷积网络，注意力驱动双向循环网络，以及全连接网络。EDANet网络具体的构建细节如下：

[0019] 4)卷积网络

[0020] 卷积网络部分，本发明采用了Encoder-DecoderCNN网络。原因是encoder结构能有效地获取语音语谱图信息的时间-频率的上下文信息，decoder结构能完整地还原语谱图的时间- 频率的结构细节，从而有效地保留原始真实语音信息并去除无关的语音噪声。卷积网络部分的具体设置如图1所示，本发明采用了9层卷积层，每层滤波器的数目分别是4，8，16，32， 64，32，16，8，4。同时，每个滤波器的大小是3*3。最后经过卷积玩了过部分，总共产生4 个2D的特征图，每个特征图的大小是514*7。

[0021] 5)注意力驱动双向循环网络

[0022] 如图1所示，将卷积网络产生的所有2D特征图按时间方向拼接在一起，产生特征H(x)，并接着通过注意力驱动的双向循环网络部分。注意力驱动的双向循环网络细节如图1所示，给定每个时间步的特征$H_t:=H_t(x)$，计算各帧特征对于目标帧语音的贡献α_t如下：

[0023] $$\alpha_i=\frac{\exp(H_t)}{\sum_{i=t-(s-1)/2}^{t+(s-1)/2}\exp(H_i)}$$

[0024] 然后，将带权的各帧特征$\widehat{H}(x):=\alpha H(x)$输入到双向循环网络BLSTM，得到融合上下文的时序特征V(x)，具体公式如下：

[0025] $$V(x)=BLSTM\{\widehat{H}(x)\}$$

[0026] 其中，本发明中的实验设置的BLSTM层数是2，每层的隐藏单元个数是300。

[0027] 6)全连接网络

[0028] 为了避免过拟合的问题，本发明采用全连接网络结合Dropout策略来提高模型的泛化能力并减缓此问题。Dropout是对神经网络进行优化的方法，在学习的过程随机将隐含层的部分权重或者输出归零，降低节点的相互依赖性，从而实现神经网络的正则化，避免模型过拟合。实验中，Dropout省略了20％的网络节点连接。具体地，针对每个时间步t，网络的计算公式如下：

[0029] $\hat{y} = \max\{0, W_V(x) + b_w\}$

[0030] 其中，W和bw都是模型的参数。

[0031] 模型训练过程：如图1所示，模型的输入是步骤二中提取的语音振幅和语音环境特征 $x = [x_{振幅}, x_{环境}]$，经过构建的环境自适应的端对端深度网络EDANet，预测语音的log振幅特征。其中，预测的语音log振幅特征$\hat{y}$和原始干净语音的log振幅特征y的最小平方误差作为目标函数，模型的优化方法采用AdaDelta方法。

[0032] 步骤四，语音波形重构：

[0033] 将步骤三预测得到的语音log振幅特征$\hat{y}$转换成语音波形，转换公式如下：

[0034] $s_{增强} = \exp(\hat{y}) \cdot \exp(\angle S)$

[0035] 至此，就可以将验证集和测试集的语音进行增强，得到干净的语音波形。

[0036] 与现有技术相比，本发明的有益效果为：如图2所示，本发明考虑了不同噪声环境的影响，动态估计不同的环境信息，大大的提高了模型的鲁棒性，获得更好语音增强效果。同时，通过融入注意力机制，更精确地挖掘语音时序上下文之间的关系，丰富了语音增强过程中的信息获取，有效提高了语音增强性能。

附图说明

[0037] 图1是本发明提出的基于注意力驱动循环卷积网络的环境自适应语音增强算法框架图；

[0038] 图2是本发明方法和现有语音增强技术(DNN)的对比图：

[0039] (a)DNN基线方法增强语音过程图

[0040] (b)本发明增强语音过程图

[0041] 图3是本发明的方法流程图。

具体实施方式

[0042] 为了更好地理解本发明的技术方案，现结合附图及具体实施方式来对本发明进行更进一步详细的描述

[0043] 图1是本发明的基于注意力驱动循环卷积网络的环境自适应语音增强算法框架图，主要包含以下步骤：

[0044] 步骤一，输入数据准备：为了验证本发明的效果，在REVERBChallenge2014数据库进行语音增强实验。REVERBChallenge2014中所有句子采样频率为16KHz。

[0045] 步骤二，振幅特征和环境特征提取：

[0046] 1)振幅特征提取：把每一段语音信号经过预加重、分帧、加窗、快速傅里叶变换，FFT 点数设为512，窗长512，窗移256，特征维数设为257维。

[0047] 2)环境特征提取：本发明采用WPE算法进行环境信息提取，其中参数FFT点数设为512，窗长512，窗移256，特征维数也是257维。

[0048] 步骤三，模型构建及训练：

[0049] EDANet网络的设置如下：卷积层设置9层，每层的滤波器数量分别是4，8，16，32，

64，32,16,8和4。其中每个滤波器的大小都是3*3。注意力驱动的双向循环网络,实验中设置了2层,每层都是300个隐藏单元。将卷积网络产生的所有2D特征图按时间方向拼接在一起,产生特征大小是2056*7。全连接网络部分是一层Dropout层和一层全连接层。最后的目标函数是使用的最小平方误差,然后回传误差,使用AdaDelta算法进行优化。模型优化收敛后,输入验证集或者测试集的语音,预测干净语音的log振幅。

[0050] 步骤四,语音波形重构:

[0051] 将步骤三预测得到的语音log振幅特征转换成语音波形。

[0052] 表1是在REVERBChallenge2014数据库上语音增强的结果对比

[0053]

方法		验证集模拟数据 PESQ			验证集模拟数据 SRMR			验证集真实数据 SRMR		
		远场	近场	平均	远场	近场	平均	远场	近场	平均
基线方法	未处理	2.16	2.59	2.38	3.46	3.96	3.71	3.19	3.17	3.18
	MSLP	2.14	2.53	2.34	3.04	3.32	3.18	3.04	3.19	3.07
	WPE	2.26	2.72	2.49	3.76	4.27	4.02	3.58	3.42	3.50
	DNN	2.42	2.69	2.56	4.34	4.87	4.61	4.97	4.92	4.95
	BLSTM	2.53	2.89	2.71	4.58	4.95	4.77	5.28	5.07	5.18
提出方法	CNN-BLSTM	2.57	2.95	2.76	4.70	4.97	4.84	5.46	5.22	5.34
	EDANet (无环境特征)	2.60	2.96	2.78	4.74	4.97	4.86	5.58	5.28	5.43
	EDANet	**2.66**	**3.08**	**2.87**	**4.81**	**5.07**	**4.94**	**5.59**	**5.40**	**5.50**

[0054] 表1是在REVERBChallenge2014数据库上进行语音增强的结果对比,评价指标为验证集上的PESQ(越高越好)和SRMR(越高越好)。首先,从CNN-BLSTM和基线方法对比发现,本方法构建的网络中,Encoder-DecoderCNN对语音真实信号的时间-频率特征刻画是有效的。其次,对比EDANet(无环境信息)方法,证明了本发明融入的注意力机制在语音增强任务中表现良好,有助于语音信号的时序上下文信息的更精细挖掘。最后,相比于没有融入环境信息的EDANet方法,环境信息的融入提高了语音增强的性能,这表明融入环境特征对语音增强是有效的,证明了环境信息能提高模型的鲁棒性,能使得模型在不同环境中有自适应性。

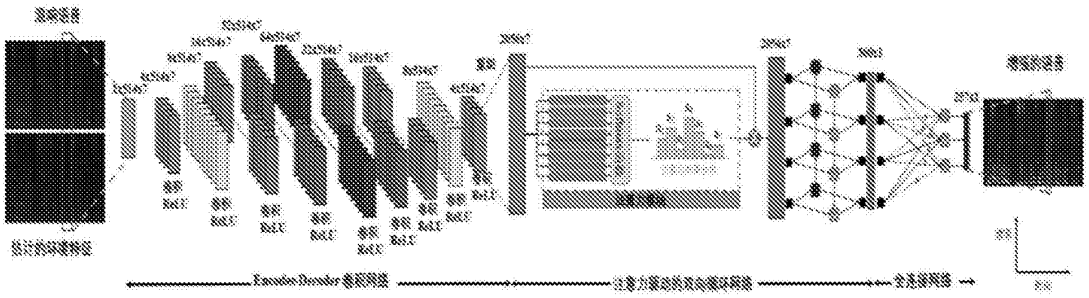
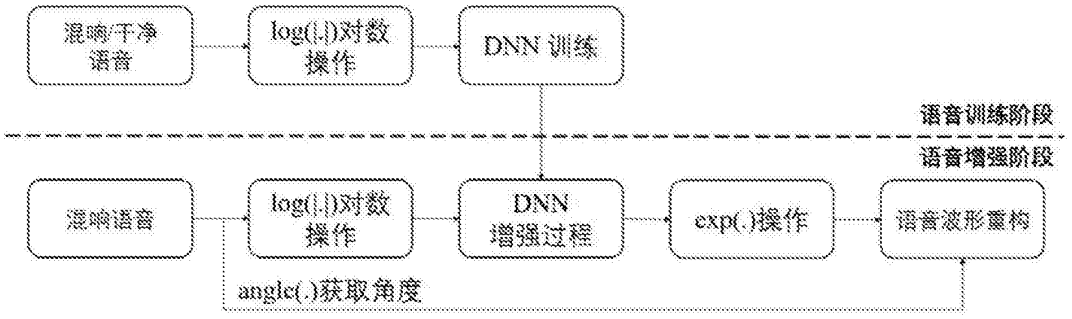
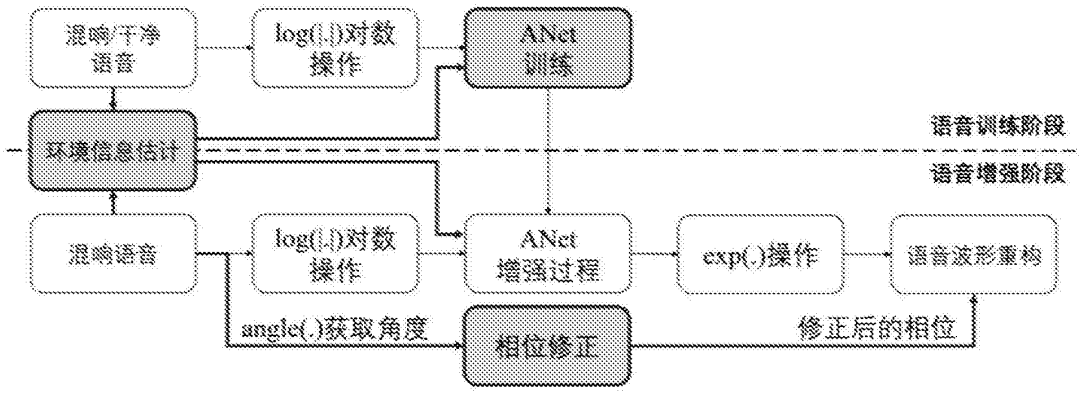


图1



（a）DNN 基线方法增强语音过程图



（b）本发明增强语音过程图

图2

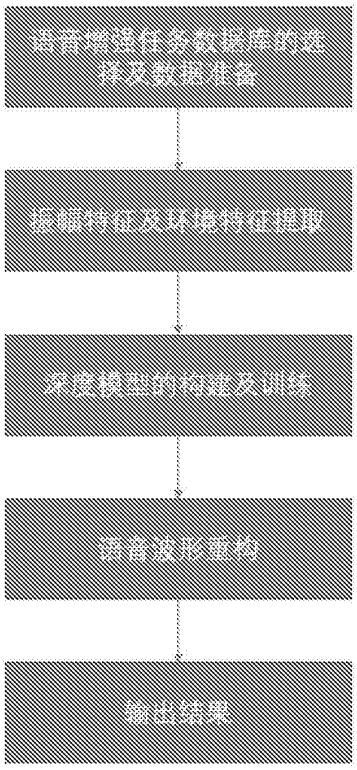


图3