ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 921－931
DOI:10. 16337/j. 1004-9037. 2021. 05. 007

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合卷积网络与残差长短时记忆网络的轻量级骨导语音盲增强

邦锦阳[1]，孙　蒙[1]，张雄伟[1]，郑昌艳[2]

(1. 陆军工程大学指挥控制工程学院,南京 210007;2. 火箭军士官学校,青州 262500)

摘　要: 基于深度学习的骨导语音盲增强已经取得了较好的效果,但仍存在模型体积大、计算复杂度高等问题。为此提出一种融合卷积网络和残差长短时记忆网络的轻量级骨导语音增强深度学习模型,该模型在保持语音增强质量的前提下,能有效提升骨导语音盲增强的效率。该模型借助卷积网络参数量小、特征提取能力强等优点,在语谱图频率维度引入卷积结构,从而深入挖掘时频结构的细节和高低频信息间的关联关系以提取新型特征,并将此新型特征输入改进后的长短时记忆网络中,用于恢复高频成分信息并重构语音信号。通过在骨导语音数据库上实验,表明所提模型可以有效改善高频成分的时频结构,在提升增强效果的同时,降低了模型体积和推理的计算复杂度。

关键词: 骨导语音盲增强;卷积网络;长短时记忆网络;轻量级模型

中图分类号: TN912　　**文献标志码:** A

Lightweight Model for Bone-Conducted Speech Enhancement Based on Convolution Network and Residual Long Short-Time Memory Network

BANG Jinyang[1], SUN Meng[1], ZHANG Xiongwei[1], ZHENG Changyan[2]

(1. College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China; 2. Department of Test and Control, High-Tech Institute, Qingzhou 262500, China)

Abstract: Bone-conducted speech enhancement based on deep learning has reached a milestone recently. However, there are still some issues to prevent its real-world applications, such as large models and high computational complexities. In this paper, a lightweight deep learning model is proposed to improve the efficiency of bone-conducted speech enhancement. Inspired by the fact that convolution network has unique advantages in feature extraction with a few of parameters, convolution structures are introduced into the frequency dimensions of the spectrogram in our model. These structures can extract the details of the spectrogram in the time-frequency structures and explore the potential relationship between high and low frequency components. These new features extracted by CNN are fed into the improved long short-term memory network to recover high-frequency components information and reconstruct speech signals. From the experiments on bone conduction speech database, we can draw a conclusion that the proposed model can reconstruct the time-frequency details of the high-frequency components. While improving the enhancement performance, the model size and the computational complexity are reduced.

Key words: bone-conducted speech blind enhancement; convolutional neural network; long short-term memory network; lightweight model

基金项目:国家自然科学基金(62071484)资助项目。
收稿日期:2021-03-01;**修订日期:**2021-07-23

引　言

骨传导麦克风(Bone-conducted microphone,BCM)有别于传统的空气传导麦克风(Air-conducted microphone, ACM),是通过拾取人声带振动采集语音信号的。BCM采集到的语音称为骨导语音,ACM采集到的语音称为气导语音。由于背景噪声的强度不够,无法使BCM产生震动,在声音采集阶段就屏蔽了背景噪声,所以BCM具有较强的抗背景噪声性能,从而在军事行动、抢险救灾、车间工厂等场景中具有非常广阔的应用前景。

骨导语音只能拾取声带振动,根据人体发声的规律研究,采集到的骨导语音缺少了鼻、口腔、嘴唇等器官的辐射效应,因此骨导语音的高频成分衰减十分严重,几乎采集不到2.5 kHz以上的频率成分。图1分别展示了同一句话的气导语音语谱图和骨导语音语谱图。从图1可以看到,骨导语音的低频成分与气导语音非常相似,但高频成分丢失,导致骨导语音的听感沉闷、不清晰。因此,改善骨导语音质量,对于强噪声环境下语音通信具有重要意义。

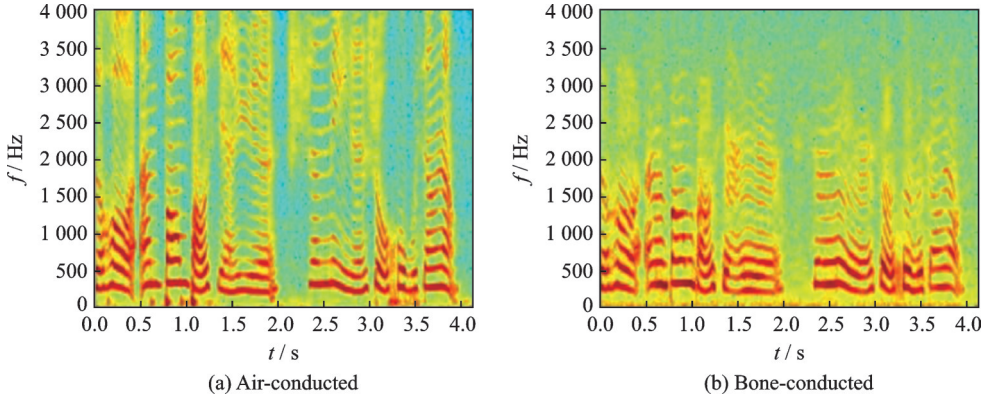


图1　气导语音与骨导语音语谱图

Fig.1　Spectrogram of air-conducted and bone-conducted speeches

当前骨导语音相关的增强方法主要分为两大类:一是融合性的增强方法,结合气导语音的完整性以及骨导语音的抗噪性,实现融合性的语音增强;二是不依赖于气导语音的骨导语音盲增强方法。盲增强方法是指在增强语音时,不需要气导语音作为辅助,只依靠缺失了高频信息的骨导语音信息恢复出原始的气导语音。由于骨导语音缺失大量信息,骨导语音盲增强有别于一般的语音去噪增强,且基于深度学习方法需要大量的数据集进行训练,目前缺少通用的大型骨导语音数据集,因此骨导语音盲增强的难度更大,且相关研究较少,本文对骨导语音盲增强方法进行研究。

传统的盲增强方法有无监督频谱扩展法[1-2]、均衡法和谱包络转换法等。由于骨导语音在声源处缺少了鼻、口腔、嘴唇的辐射模型部分,导致高频信息缺失,这3种方法都尝试寻找一种声道转换模型,实现骨导语音到气导语音的增强。均衡法试图找到一种声道变换函数,建立气导语音与骨导语音在频谱分量上的映射关系,对骨导语音的频谱分量进行增强,此方法能够恢复部分缺失的高频信息,但由于其采用长时谱的平均分量进行计算,容易导致语音信号不连续[3]。谱包络转换法同样基于语音信号的源－滤波器模型,利用谱包络特征表示声道模型的特征,此方法与均衡法相比优势在于增强后的语音信号更为连贯[4]。声道模型包含了复杂精密的人体器官结构,尚无准确的模型能刻画声道特征,受限于计算能力、模型的非线性表达程度,以上增强方法对于语音信号的表征、高频成分的恢复能力有限。

近年来,大量深度学习的方法极大地推动了语音信号处理、图像处理领域的研究,在语音增强、语

音识别、目标检测等各类任务上都取得了不俗的效果。深度学习网络是一种端到端的模型,其优势在于能够拟合非线性特征、处理复杂信息。Xu等[5]设计了一个深度神经网络(Deep neural network, DNN)来学习噪声语音与干净语音间的映射关系,采取全局方差均衡和Dropout策略,提升了增强语音的客观和主观度量指标,同时噪声感知训练技术使其具有良好的泛化能力。Jiang等[6]首先提取语音梅尔倒谱系数(Mel-frequency cepstral coefficient, MFCC),该特征更符合人耳听觉特性,而后输入DNN重构语音幅度谱,结果证明此方法有效提升了语音增强效果,并且减少了模型训练所需的数据量。尽管DNN具有较好的非线性表达能力,但由于语音信号是一种时序性信号,具有上下文关联的特点,而DNN在处理经傅里叶变换后得到的语谱图时,容易忽略相邻帧之间的关联,限制了DNN在语音增强方面的性能。

DNN的隐藏层中节点之间是孤立的,只有不同隐藏层之间的节点间才存在连接,而循环神经网络(Recurrent neural network, RNN)通过在隐藏层节点中建立连接[7],使当前时刻的节点可以保留之前时刻的信息,因此在处理序列问题时,RNN能充分考虑全局信息。然而,若序列长度过长,RNN在反向传播的过程中,梯度持续累积,直到无穷大或无穷小,这种现象称之为梯度爆炸和梯度消失,无法记忆长期的序列。长短时记忆网络(Long short term memory network, LSTM)加入了门控机制,引入输入门、遗忘门、输出门控制不同时刻记忆之前时刻信息的权重,克服了RNN的缺陷,使得网络在处理长序列问题时,依然可以保持"记忆力"。Liang等[8]在LSTM的基础上,结合注意力机制,采用通道间相关性的理想比值掩码作为学习目标,对噪声污染较小的信息进行筛选,有助于重构干净语音。Lee等[9]在双向LSTM(Bi-directional LSTM, BLSTM)的基础上,将语音功率估计和噪声功率估计融合到频谱滤波框架中,并提出一种具有先验信噪比的附加内部约束,有效提升了语音增强质量。

RNN和LSTM的优势在于处理上下文关联信息,但对于语谱图中高低频信息间的关联系利用不足。卷积神经网络(Convolutional neural network, CNN)是图像处理领域的佼佼者,CNN类似于人眼对物体的观察,局部感知特性使其拥有对细节的刻画能力,权值共享结构减少了网络的参数量。CNN对结构特征的表征能力是RNN、LSTM的短板。Kounovsky等[10]利用CNN构造了一个去噪自编码器(Denoisisng autoencoders, DAEs),实验表明对于语音对数功率谱的增强效果中,基于CNN的DAEs比基于全连接(Full connection, FC)的性能提升了8%。此外,Pandey和Wang[11]基于编解码(Encoder-decoder)网络架构,编解码器采用CNN结构,并在编码器和解码器之间增加了一个时域卷积模块,利用当前和之前帧的信息重构增强语音,该模型增强效果强于LSTM,且由于是全卷积的模型,训练参数显著减少。郑昌艳等[12]将LSTM模型应用于骨导语音盲增强的研究中,得到增强语音后,为了解决过平滑问题,采用了非负矩阵分解(Non-negative matrix factorization, NMF),进一步提高了语音质量。LSTM对于骨导语音增强具有不错的效果,但其参数量过大,仍需要一种轻量化的模型用于实现实时性的语音增强。

本文构建了一种卷积网络与残差LSTM联合模型,在浅层LSTM的前端引入卷积网络,以达到简化模型,提升增强效果的目的。首先描述了联合模型的架构;其次介绍了模型中的关键模块及其设计思路;再次进行了实验仿真及结果分析;最后对本文工作进行了总结。

1 骨导语音盲增强的模型架构

1.1 总体架构

考虑深层LSTM模型存在参数量大,计算时间复杂度较高,浅层LSTM增强效果不佳的矛盾,而CNN具有参数量小,对结构特征提取能力强的优势,借助CNN可以在减小模型复杂度的同时,提升浅层LSTM的增强效果。本文提出了一种融合卷积网络与残差LSTM的语音增强模型(Res-convolutional-

recurrent neural network, RCRNN)训练的骨导语音盲增强方法。

RCRNN联合模型的总体结构如图2所示。语谱图作为网络输入,在频域上进行卷积操作,提取频域上的结构特征以及高低频信息间的结构约束,随后将CNN的输出拼接后输入LSTM,得到增强后的语音。同时,为了扩大卷积核的视野,采用了扩张卷积,获取更大的感受野;为了提高网络训练效率,在LSTM中引入了残差连接,进一步减小出现梯度消失和爆炸问题的可能性。损失函数选择均方误差(Mean square error, MSE),将增强后语谱图和原气导语音语谱图进行对比,根据两者的MSE优化模型参数。

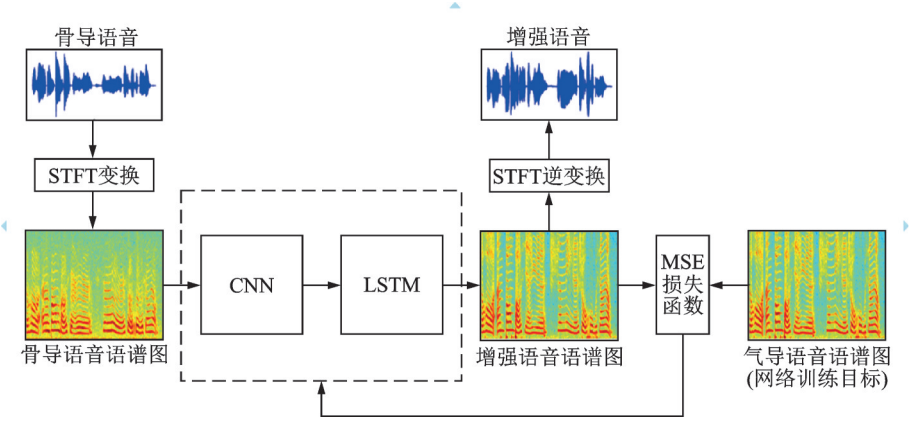


图2 RCRNN联合模型增强方法的结构

Fig.2 Structure of RCRNN joint model enhancement method

1.2 网络结构

本文采用卷积-残差LSTM实现骨导语音盲增强,其网络结构如图3所示。CNN作为残差长短时记忆网络(Residual long short time memory network, RLSTM)的前端特征提取网络,在频率轴方向提取频率特征以及高低频率间的结构相关性特征,不同的卷积核从骨导语音语谱图中提取到不同的高维特征,将卷积网络得到的所有通道的特征按频率方向拼接后输入RLSTM,通过若干LSTM隐藏层的训练,最后添加一个全连接层将高维特征降维映射到低维特征,得到增强后的语音语谱图。

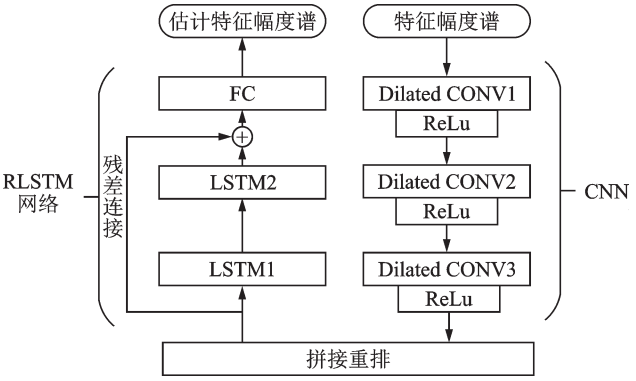


图3 RCRNN网络结构

Fig.3 Network structure of RCRNN

卷积模块中共有3层扩张卷积层,分别为CONV1、CONV2和CONV3,每层卷积后,连接ReLu非线性激活函数。拼接重排层得到所有卷积核提取到的高维特征,将其按频率方向拼接后,作为新的特征矩阵输入RLSTM,残差长短时记忆模块有两个隐藏的LSTM层,对时序上的特征进行建模提取,最后通过一个全连接层FC降维,将高维特征映射到低维特征,得到增强后的语谱图。

网络采用联合训练的方式,引入CNN来弥补LSTM对语音信号频域信息利用不充分的问题,整个网络先后对语谱图的频域、时域信息进行特征提取、训练,达到增强骨导语音的目的。

本文采用了LSTM作为基础模型而不是性能更优的BLSTM,因为BLSTM不仅利用了过去时刻的信息,也利用了未来时刻的信息,以此获得更好的性能,但处理长时间的语音时,BLSTM参数量和预测推理时间无法满足实时语音增强的要求。

1.3 训练目标与优化

在训练时,将预测数据与标签数据的MSE作为损失函数,据此对模型进行优化。骨导语音幅度谱$\boldsymbol{X}$经模型预测后得到的增强幅度谱为$\hat{\boldsymbol{Y}}$,模型训练的输出目标是气导语音幅度谱$\boldsymbol{Y}$,模型利用有监督的方式进行学习,$\hat{\boldsymbol{Y}}$与$\boldsymbol{Y}$的MSE定义为模型的训练误差

$$J_{\mathrm{MSE}}(W,b)=\frac{1}{2}\left\|\hat{\boldsymbol{Y}}(\boldsymbol{X},W,b)-\boldsymbol{Y}\right\|^2 \tag{1}$$

式中J_{MSE}表示模型的训练误差。在语音增强任务中,模型输出的增强后语音应该尽可能接近原始语音,所以$\hat{\boldsymbol{Y}}$与$\boldsymbol{Y}$的MSE越小越好,模型的优化目标可表示为

$$W,b\leftarrow\underset{W,b}{\operatorname{argmin}}\ \frac{1}{2}\left\|\hat{\boldsymbol{Y}}(\boldsymbol{X},W,b)-\boldsymbol{Y}\right\|^2 \tag{2}$$

式中W和b分别为神经元的权值和偏置参数。网络训练时目标为最小化J_{MSE},根据链式法则由后向前逐层更新各层的神经元权值W和偏置b,寻找最优值采用的方法是梯度下降法。

2 算法设计及关键模块

2.1 算法设计

本文工作的算法流程如图4所示。算法流程包括3个步骤:

(1)在数据预处理阶段,首先将骨导语音$\boldsymbol{x}(n)$和气导语音$\boldsymbol{y}(n)$的波形最大最小归一化到[−1,1],而后分别进行分帧加窗、短时傅里叶变换(Short time Fourier transform, STFT),对语音幅度谱取对数得到对数幅度谱并计算其均值方差,最后进行均值方差归一化。

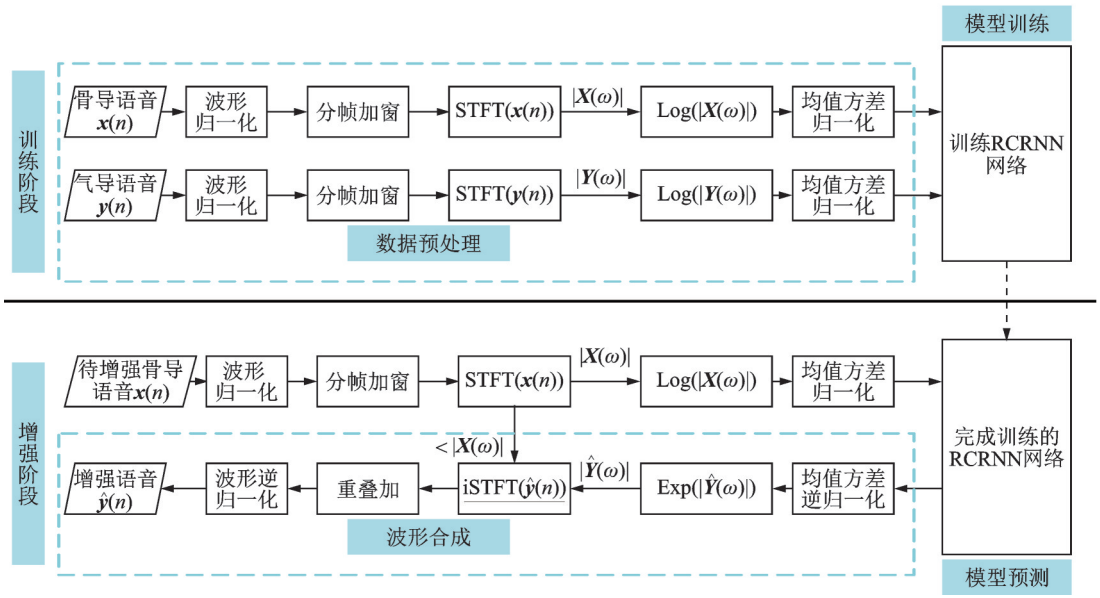


图4 本文算法设计

Fig.4 Design of the proposed algorithm

(2) 在模型训练阶段,首先初始化模型参数记为θ_0,将骨导语音训练数据输入模型得到估计值,以网络训练目标气导语音数据为参照,计算训练误差,并优化模型参数θ_n,直到训练轮次结束或连续5轮误差不再下降。

(3) 在增强阶段,骨导语音经过数据预处理后输入训练好的模型,得到估计的对数幅度谱,最后与原始骨导语音对应的相位谱进行短时傅里叶逆变换和重叠加操作得到增强后的语音波形。

2.2 扩张卷积及其设计思路

扩张卷积(Dilated convolution)[13-14]也被称为空洞卷积或者膨胀卷积,在卷积核大小不变的情况下,卷积计算时跳跃性地选择数据,以此来增加卷积核的感受野,由于未改变卷积核大小,可以在参数量不变的情况下,达到扩大感受野的目的。扩张卷积中引入了扩张率的概念,可以视作在普通卷积核中,每个权值之间填充若干个零后得到一个新的卷积核,由新卷积核完成卷积运算。普通卷积运算可表示为

$$F(s)=(\boldsymbol{x}*k)(s)=\sum_{i=0}^{m-1}\boldsymbol{x}(s-i)\cdot k(i) \tag{3}$$

式中:$\boldsymbol{x}$为输入序列,*表示卷积操作,k为卷积核,m为卷积核尺寸。扩张卷积可表示为

$$F(s)=(\boldsymbol{x}*_d k)(s)=\sum_{i=0}^{m-1}\boldsymbol{x}(s-d\cdot i)\cdot k(i) \tag{4}$$

式中,$*_d$表示扩张率为d的扩张卷积操作,当$d=1$时,扩张卷积等价于普通卷积。

尽管扩张卷积能同时达到扩大感受野和保证特征图信息的目的,但由于卷积核在计算过程中存在空洞,所以输入的语谱图中不是所有时频信息都参与了卷积运算,若连续的卷积层采用相同的扩张率时,便会出现网格效应(Gridding effect)。图5展示的是多次叠加扩张率为2的3×3卷积核出现的结果。此外,尽管扩张卷积扩大了感受野,但会影响卷积核对细节特征的提取,且扩张率越大,细节丢失越严重。

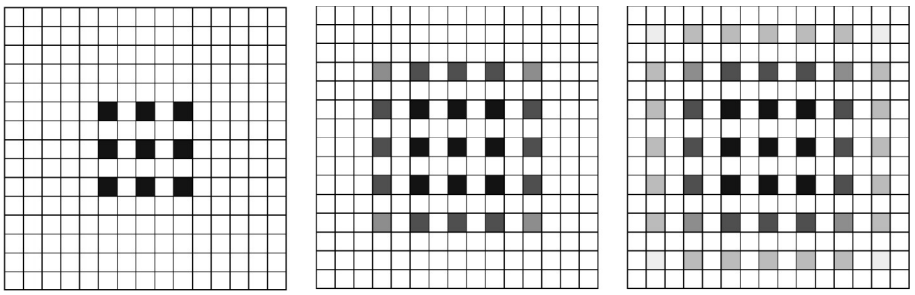


图5 3次扩张率为2的3×3卷积后的结果

Fig.5 Results of three times of 3 × 3 convolution with expansion rate of 2

因此,本文采用了“锯齿状”的扩张率,在3层的卷积网络中,扩张率分别设置为[1,2,5],锯齿状的扩张率可以保证所有输入信息都不会被遗漏。同时,骨导语音的语谱图在增强过程中,既要关注高低频之间的关联性,也要保留时频结构上的细节,不同大小的扩张率恰好可以满足这个需求。在卷积核大小的设置上,借鉴文献[15]的工作,卷积仅在频率轴方向上进行,可以获得较好的增强性能,卷积核在时间轴上尺寸设为1,仅用于提取频域特征。

2.3 残差LSTM及其设计思路

LSTM是一种特殊的RNN,RNN当前时刻t的输入分别是当前时刻输入值$\boldsymbol{x}_t$,上一时刻输出值$\boldsymbol{h}_{t-1}$,以及上一时刻的神经元状态$\boldsymbol{C}_{t-1}$;输出分别是当前时刻输出值$\boldsymbol{h}_t$以及当前时刻的神经元状态$\boldsymbol{C}_t$,

通过节点间建立的连接使网络记住之前的信息。通过增加输入门限、遗忘门限和输出门限,使模型按照一定权重系数将当前时刻输入信息$\boldsymbol{x}_t$、神经元状态信息C_{t-1}、上一时刻输出信息$\boldsymbol{h}_{t-1}$计算出当前时刻输出信息$\boldsymbol{h}_t$,它们之间的关系可表达为

$$\boldsymbol{f}_t=\boldsymbol{\sigma}(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{5}$$

$$i_t=\boldsymbol{\sigma}(\boldsymbol{W}_c\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{6}$$

$$\widetilde{C}_t=\tanh(\boldsymbol{W}_c\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{7}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t*\boldsymbol{C}_{t-1}+\boldsymbol{i}_t*\widetilde{\boldsymbol{C}}_t) \tag{8}$$

$$\boldsymbol{o}_t=\boldsymbol{\sigma}(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{9}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t*\tanh(\boldsymbol{C}_t) \tag{10}$$

式中:$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$分别表示遗忘门、输入门和输出门,$\boldsymbol{C}_t$表示细胞状态。

本文算法中,在隐藏层之间加入残差连接,将上一层的输入和当前层的输出作为下一层的输入,那么当前层网络的训练目标就转换为输入数据与目标之间的残差,随着网络层数的加深,这个残差值会逐渐减小,每层网络只需拟合逼近残差。He等[16]提出的加入残差连接的Resnet,大大加深了神经网络的深度,提升了图像识别的精度。引入残差连接后,可以避免产生梯度消失和梯度爆炸的问题,解决了网络达到一定深度后性能下降的问题,深度网络达到一定深度时,网络会出现退化的现象,在连续的矩阵乘法运算后,权重矩阵的秩会降低,意味着权重矩阵中有效的参数越来越少,特征表达能力越来越弱,把网络浅层的输入连接到深层网络与深层网络的输入融合,网络训练目标就从拟合目标数据变成了拟合目标数据与输入数据的差,随着网络层数加深,拟合会越来越精确,有利于提升深层网络表达的特征质量。

3 实验设置与结果分析

3.1 数据集和评价指标

本文选取了文献[17]中的骨导语音语料库作为训练数据。数据库中有利用喉震式麦克风采集的骨导语音与对应的气导语音,每条语音的时长为3~5 s不等,语音为32 kHz采样率、16 bit量化。本文选取了男1、男2,女1、女2各200条语音作为数据集,对每个人的语音分别进行实验。在实验数据中随机选取单个说话人的140条语音作为训练集,30条语音作为验证集,30条语音作为测试集。本文针对单人的骨导语音进行训练,也用说话人本人的语音作为测试数据,在不同模型上测试增强性能。

感知语音质量评估(Perceptual evaluation of speech quality, PESQ)[18]、短时客观可懂度(Short-time objective intelligibility, STOI)[19]、对数谱距离(Log spectral distance, LSD)[20]是评价语音质量最常用,且具有代表性的客观评价指标。PESQ能预测待测语音的主观MOS值,PESQ将待测语音和原始语音滤波变换后,综合待测语音与原始语音的时频特性,给出一个在[-0.5,4.5]区间的PESQ得分,语音质量与PESQ得分成正比。STOI是衡量语音的重要指标之一,对于语音来说,只有听懂和听不懂两种情况,可以理解为在短时内可懂度是二值的,其范围在[0,1]之间,越接近1质量越好。STOI是将待测语音和原始语音经过移除静音区、STFT变换、归一化后计算短时谱向量的相关系数得到的。LSD衡量待测语音对数谱与原始语音对数谱之间的距离,LSD的值越小,说明待测语音越接近于原始语音,增强质量就越高。

3.2 基线系统及参数设置

本文选取2种不同深度和不同参数的LSTM模型作为对比,分别是:(1)4层隐藏层,每层256个节点(简记为LSTM1);(2)两层隐藏层,每层256个节点(简记为LSTM2);均采用MSE作为损失函数优

化模型。

下面介绍本文所介绍的RCRNN模型参数设置以及实验设定。原始语音采样率为32 kHz,但由于骨导语音的高频成分缺失严重,STFT幅度谱在2.5 kHz以上几乎已没有能量,若要将骨导语音的频率成分恢复到8 kHz甚至16 kHz,难度较大,且耗费的计算资源和参数将大大增加。首先将语音降采样到8 kHz,而后进行分帧加窗操作,利用语音短时平稳性特点使语音具备做傅里叶变换的条件,最后进行256维的STFT,得到频率维度为129维的语音幅度谱。

模型结构和参数以及输入输出数据的维度如表1所示。129维的幅度谱先后通过卷积网络和残差LSTM,两个网络由一个拼接重排层连接,卷积网络的通道数依次是[16,32,64],卷积核大小为3×3,首层填充数为(1,0),其余层的填充数为(1,1),扩张率分别为(1,2,5);残差LSTM共2层,每层都由256个节点组成。模型采用两个网络联合训练的方法,MSE设为损失函数用来优化模型,为了防止模型对于训练集数据出现过拟合问题,所有网络都设置了dropout=0.2。CNN的参数从前至后依次表示卷积的输出通道数(Out channels)、卷积核大小(Kernel size)、填充数(Padding)、扩张率(Dilation rate)。

表1 网络结构参数

Table 1 Parameters of network structure

隐藏层名称	输入数据尺寸	隐藏层参数	输出数据尺寸
Dilated CONV1	$1\times T\times 129$	16,(3×3),(1,0),1	$16\times T\times 64$
Dilated CONV2	$16\times T\times 64$	32,(3×3),(1,1),2	$32\times T\times 31$
Dilated CONV3	$32\times T\times 31$	64,(3×3),(1,1),5	$64\times T\times 12$
Reshape	$64\times T\times 12$	768	$T\times 768$
LSTM1	$T\times 768$	256	$T\times 768$
LSTM2	$T\times 768$	256	$T\times 768$
FC	$T\times 768$	129	$T\times 129$

3.3 实验结果和分析

实验结果的PESQ、STOI、LSD值如表2~4所示,共列出了基线模型与本文所提模型在4个不同说话人数据集上的实验结果。从平均值上看,RCRNN在3个指标上都要优于同样具有2层LSTM隐藏层的LSTM2,同时,在STOI和LSD两个指标上,RCRNN要优于有4层LSTM隐藏层的LSTM1,PESQ指标上,两者的差距很小。

表2 3种模型在不同实验对象下的PESQ值

Table 2 PESQ scores of three models for different speakers

模型	女1	女2	男1	男2	平均值
LSTM1	**3.378**	3.152	3.219	**3.065**	**3.204**
LSTM2	3.317	3.024	3.179	2.978	3.125
RCRNN	3.363	**3.201**	**3.225**	3.014	3.201

表3 3种模型在不同实验对象下的STOI值

Table 3 STOI scores of three models for different speakers

模型	女1	女2	男1	男2	平均值
LSTM1	**0.899**	**0.851**	0.882	0.810	0.861
LSTM2	0.894	0.847	0.879	0.797	0.854
RCRNN	0.894	0.850	**0.889**	**0.817**	**0.863**

表4 3种模型在不同实验对象下的LSD值

Table 4 LSD scores of three models for different speakers

模型	女1	女2	男1	男2	平均值
LSTM1	0.812	0.824	0.953	**0.932**	0.880
LSTM2	0.813	0.831	0.959	0.961	0.891
RCRNN	**0.801**	**0.821**	**0.942**	0.940	**0.876**

RCRNN在LSTM中采用了2层LSTM隐藏层的结构,可以体现出浅层的CNN结构对语音增强性能的提升效果。RCRNN在LSTM2的基础上,PESQ提升了2.5%,STOI提升了1.1%,LSD降低了1.7%,与LSTM1相比,客观指标上几乎相同。

从表2中可以看出,3种基于LSTM模型及其改进模型的骨导语音盲增强方法PESQ得分能达到3分以上,这个分数已经达到了较高水平的增强效果,骨导语音在声源处就已经屏蔽了大部分的背景噪声,增强的目的主要是恢复高频成分,提升语音听感,与当前效果较好的语音去噪方法相比,3分以上的PESQ已经处于较高水平。性能上的提升得益于前端CNN将低维特征扩充为高维特征,利用了高低频信息之间的关联性。

从表3,4可以看出,3种模型对女声的增强性能要好于对男声的增强效果,男女由于身体结构的区别,发声时男声普遍更低沉,而女声更为清脆,体现在频率上就是男声低频厚重,而女声高频更清晰。增强后的男声在STOI值上要低于女声,这可能是由于男声低频信息的权重更大,在恢复高频成分时比女声的难度更大,因此导致增强后语音高频部分的时频结构不够清晰,高频部分不足,语音的主观听感厚重,在一定程度上影响了语音的可懂度。

图6给出的是3种网络结构的参数量大小以及在30条语音的测试集上的预测总用时,与LSTM1相比可以看出,LSTM1的客观指标较好,但参数量最大,耗时最长,而RCRNN达到了和LSTM1几乎相同的性能,但参数量减少了42%,预测耗时降低了46.6%。这里体现了CNN的优势,CNN的参数量小,模型复杂度较低,因此RCRNN虽然加入了CNN网络,但可以使用浅层的LSTM,总体而言在模型复杂度上RCRNN比深层的LSTM降低了近一半,预测时间也大大减少。与LSTM2相比可以看出,由于添加了CNN特征提取模块,RCRNN的参数量和预测时间有所增加,但其增强效果有明显提升,这是因为RCRNN利用了CNN强大的特征提取功能,采用扩张卷积的方式结合小卷积核对于细节的刻画能力和大卷积核对于高低频信息的关联能力,因此,在保证模型复杂度和预测时间不明显增加的情况下,RCRNN比浅层LSTM的性能更好。

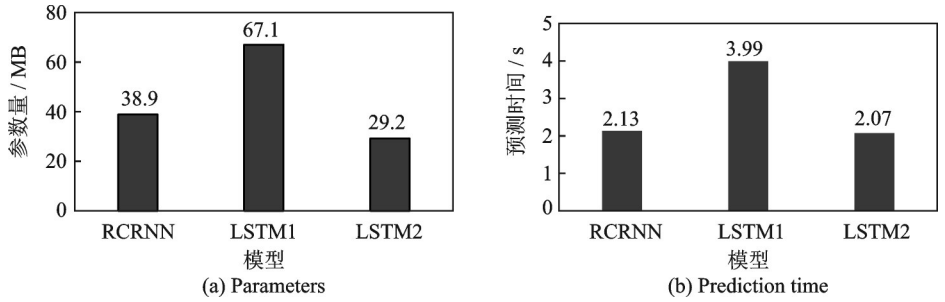


图6 3种模型的参数量和预测时间

Fig.6 Parameters and prediction time of three models

图7展示了不同方法增强后语音的语谱图示例,可以看出,增强后骨导语音的高频成分基本上能够较好地被恢复出来。由图中红色方框标出的位置可以看出,加入CNN特征提取模块的RCRNN,恢复出的语音的语谱图在时频结构上更加清晰,语谱图结构上的细节恢复得更加准确。不过清音和辅音在发声时声带不产生震动,只能依靠上下文信息对其进行恢复,因此RCRNN对于清音和辅音的增强还是存在不足。

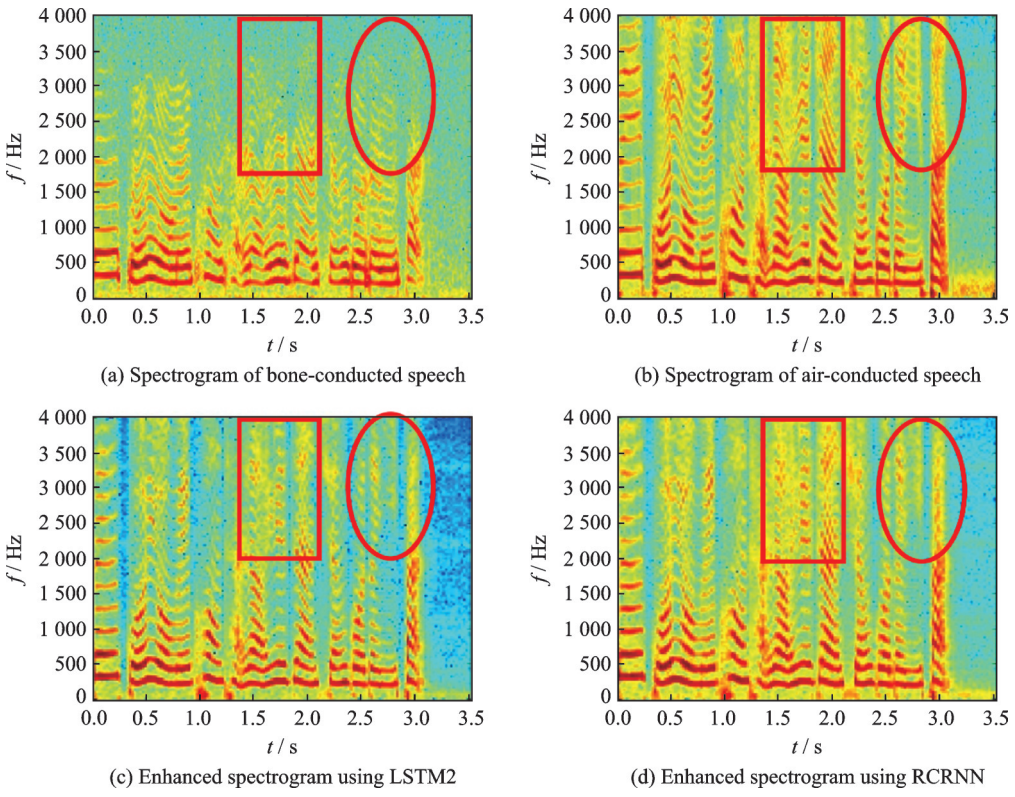


图7 经过不同模型增强的语音语谱图

Fig.7 Speech spectrogram enhanced by different models

4 结束语

本文针对基于LSTM的骨导语音盲增强方法模型复杂度高、预测时延较长，且没有充分利用时频结构信息等问题，提出了一种融合卷积网络和残差LSTM的模型结构(RCRNN)，利用卷积网络参数量小、特征提取能力强的特点，在网络输入的语谱图频率轴方向上进行扩张卷积操作，提取细节信息和高低频关联信息，而后由改进后的残差LSTM在时序上对骨导语音进行处理，以提升增强性能，同时减少模型复杂度和预测时延，为实现语音实时增强提供便利。实验证明，加入特征提取卷积网络后，模型的性能得到了提升，达到了与深层LSTM相同水平的效果，由于使用浅层网络，模型的复杂度大大降低。但由于骨导语音数据库较小，骨导语音与人体发声特点密切相关等原因，该方法对于多说话人的增强效果还有待提高，这也是下一步研究的重点问题。

参考文献：

[1] BOUSERHAL R E, FALK T H, VOIX J. On the potential for artificial bandwidth extension of bone and tissue conducted speech: A mutual information study[C]//Proceedings of IEEE International Conference on Acoustics. [S.l.]: IEEE, 2015: 5108-5112.

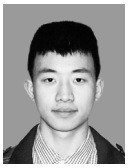
[2] BOUSERHAL R E, FALK T H, VOIX J. In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension[J]. Journal of the Acoustical Society of America, 2017, 141(3): 1321-1331.

[3] KONDO K, FUJITA T, NAKAGAWA K. On equalization of bone conducted speech for improved speech quality[C]//Proceedings of IEEE International Symposium on Signal Processing & Information Technology. [S.l.]: IEEE, 2007: 426-431.

[4] TODA T, NAKAGIRI M, SHIKANO K. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(9): 2505-2517.

[5] XU Y, DU J, DAI L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 23(1): 7-19.
[6] JIANG W, LIU P, WEN F. Speech magnitude spectrum reconstruction from MFCCS using deep neural network[J]. Chinese Journal of Electronics, 2018, 27(2): 393-398.
[7] SALEHINEJAD H, SANKAR S, BARFETT J, et al. Recent advances in recurrent neural networks[EB/OL]. (2017-12-29) [2021-01-20]. https://arxiv.org/abs/1801.01078.
[8] LIANG R, KONG F, XIE Y, et al. Real-time speech enhancement algorithm based on attention LSTM[J]. IEEE Access, 2020(8): 48464-48476.
[9] LEE J, KIM K, TURAJ S. Deep bi-directional long short-term memory based speech enhancement for wind noise reduction [C]//Proceedings of Hands-Free Speech Communications and Microphone Arrays(HSCMA). San Francisco, USA: IEEE, 2017: 41-45.
[10] KOUNOVSKY T, MALEK J. Single channel speech enhancement using convolutional neural network[C]//Proceedings of Electronics, Control, Measurement, Signals & Their Application to Mechatronics(ECMSM). Donostia, Spain: IEEE, 2017: 1-5.
[11] PANDEY A, WANG D L. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 6875-6879.
[12] 郑昌艳, 张雄伟, 曹铁勇, 等. 一种基于LSTM-RNN的喉振传声器语音盲增强算法[J]. 数据采集与处理, 2019, 34(4): 615-624.
ZHENG Changyan, ZHANG Xiongwei, CAO Tieyong, et al. Blind enhancement algorithm for throat microphone speech based on LSTM recurrent neural networks[J]. Journal of Data Acquisition and Processing, 2019, 34(4): 615-624.
[13] ZHANG Z, WANG X, JUNG C. DCSR: Dilated convolutions for single image super-resolution[J]. IEEE Transactions on Image Processing, 2019, 28(4): 1625-1635.
[14] QIN X, JIANG J, YUAN C A, et al. Arbitrary shape natural scene text detection method based on soft attention mechanism and dilated convolution[J]. IEEE Access, 2020(8): 122685-122694.
[15] WANG Z, ZHANG T, SHAO Y, et al. LSTM-convolutional-blstm encoder-decoder network for minimum mean-square error approach to speech enhancement[J]. Applied Acoustics, 2021, 172: 107647.
[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2016: 770-778.
[17] 邢益搏, 张雄伟, 郑昌艳, 等. 骨导语音库的建立与骨气导语音的互信息分析[J]. 声学技术, 2019, 38(3): 312-316.
XING Yibo, ZHANG Xiongwei, ZHENG Changyan, et al. Establishment of bone-conducted speech database and mutual information analysis between bone and airconducted speeches[J]. Technical Acoustics, 2019, 38(3): 312-316.
[18] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs[C]//Proceedings of Acoustics, Speech, and Signal Processing, 2001 Proceedings (ICASSP '01) 2001 IEEE International Conference on. [S.l.]: IEEE, 2001: 749-752.
[19] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]//Proceedings of 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2010: 4214-4217.
[20] GRAY A, MARKEL J. Distance measures for speech processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24(5): 380-391.

作者简介:



邦锦阳(1996-),男,硕士研究生,研究方向:语音处理与网络安全,E-mail:bangjinyang@163.com。

孙蒙(1984-),**通信作者**,男,副教授,硕士生导师,研究方向:智能语音处理、机器学习,E-mail:sunmengccjs@163.com。

张雄伟(1965-),男,教授,博士生导师,研究方向:语音与图像处理、智能信息处理,E-mail:xwzhang9898@163.com。

郑昌艳(1990-),女,讲师,研究方向:智能信息处理、语音信号处理,E-mail:echoaimaomao@163.com。

(编辑:张彤)