DOI: 10.13382/j.jemi.B2407824

基于残差膨胀卷积与门控编解码网络的语音增强*

李 珂[1] 王雅静[2] 昝志辉[1] 齐瑞洁[1]
(1. 山东理工大学计算机科学与技术学院 淄博 255049;2. 山东理工大学电气与电子工程学院 淄博 255049)

摘 要:语音信号的时序依赖性特征和上下文信息在语音增强任务中至关重要,针对编解码网络对其捕获不充分导致增强效果差的问题,构建了一种非对称的残差膨胀卷积与门控编解码网络(RD-EGN),该网络包含编码器、中间层和解码器 3 部分。编码器设计了一种因果卷积层结构,以时序特征建模,捕获语音序列中不同层的特征,并保持语音信号的因果性;中间层设计了残差膨胀卷积网络(RDCN),融合膨胀卷积、残差连接和级联的扩张块使网络拥有更高的感受野,以跨层的方式传递信息并提取语音长时依赖性特征,在此基础上将 RDCN 与长短时记忆网络相结合,捕获更广泛的上下文信息;解码器引入门控机制,动态调整信息流的门控程度,获得更丰富的全局特征并重建增强语音。分别在 TIMIT、UrbanSound8k、VoiceBank 及 NOISE92 数据集上进行消融及性能对照,实验结果表明,RD-EGN 相较于卷积循环网络(CRN)、自编码器卷积神经网络(AECNN)、膨胀-密集自动编码器(DDAEC)等具有较少的训练参数和较高的 SSNR 得分、主观评价指标(CSIG, CBAK 和 COVL)得分,并且在客观评价指标方面,语音质量客观评价指标(PESQ)提高了 2.5%~7.1%,短时客观可懂度(STOI)提高了 1%~5.3%,具有较为突出的增强性能与泛化能力。

关键词: 语音增强;深度学习;编解码网络;膨胀卷积;门控机制
中图分类号: TN912.35 **文献标识码**: A **国家标准学科分类代码**: 510.40

Speech enhancement based on residual dilatation convolutional and gated codec networks

Li Ke[1] Wang Yajing[2] Zan Zhihui[1] Qi Ruijie[1]
(1. School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China;
2. School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255049, China)

Abstract: The time-dependent features and context information of speech signals are crucial in speech enhancement tasks. Aiming at the problem that codec networks insufficiently capture these features, resulting in poor enhancement performance, an asymmetric residual dilatation convolutional and gated codec network (RD-EGN) is constructed. The network comprised three parts: the encoder, intermediate layer and decoder. The encoder designed a causal convolution layer structure to model the temporal feature, capture the features of different layers in the speech sequence and maintain the speech signal's causality. The intermediate layer incorporated a residual dilated convolutional network (RDCN), which integrated dilated convolution, residual connections, and cascaded expansion blocks to endow the network with a larger receptive field. It facilitated cross-layer information transfer and extracted long-term dependency features in speech. The RDCN is combined with the long short-term memory network to capture broader context information. The decoder introduced a gating mechanism to adjust the gating degree of information flow dynamically, obtain richer global features and reconstruct enhanced speech. Ablation and performance comparison experiments were conducted on the TIMIT, UrbanSound8k, VoiceBank, and NOISE92 datasets. The results show that, RD-EGN has fewer training parameters and higher scores in SSNR and subjective evaluation metrics (CSIG, CBAK, and COVL) than CRN, AECNN and DDAEC. In objective evaluation metrics, the PESQ is increased by 2.5% to 7.1%, and the STOI is increased by1% to 5.3%. RD-EGN demonstrates outstanding enhancement performance and generalization ability.

Keywords: speech enhancement; deep learning; codec network; dilatational convolution; gating mechanism

收稿日期: 2024-09-13 Received Date: 2024-09-13
*基金项目:山东省自然科学基金(ZR2024MD031)项目资助

0　引　言

语音增强(speech enhancement, SE)是指当语音信号受到各种噪声干扰后,从噪声背景中提取有用信息,降低噪声干扰的技术。随着通信技术的发展,语音增强技术可以应用于语音通信、语音识别、语音合成等多个领域。传统语音增强方法大多是无监督方法,在复杂环境下需要基于一定的条件假设,例如噪声平稳假设的谱减法[1],语音和噪声服从特定分布的统计模型法[2]。但是现实生活中的噪声大多是非平稳的,传统方法的增强效果较差。近年来,基于深度学习的有监督语音增强方法发展迅速,通过训练大量语音和噪声样本数据[3],使模型能够学习到有效的特征表示。Xu 等[4]提出基于深层神经网络(deep neural network, DNN)的语音增强方法,利用其非线性建模能力,估计含噪语音到干净语音的非线性映射,相较于传统方法,增强效果明显提升,但 DNN 在捕获长时序列依赖关系方面存在局限性。Takeuchi 等[5]将循环神经网络(recurrent neural network, RNN)用于语音增强,取得了优于 DNN 的增强效果,但 RNN 仅能够处理短期依赖问题,且在训练时容易引起梯度消失。Saleem 等[6]在 RNN 的基础上引入门控机制,提出长短时记忆网络(long short-term memory, LSTM)选择性传递信息,缓解梯度消失等问题。卷积神经网络(convolutional neural networks, CNN)凭借其模型尺寸小、计算快及自动学习特征能力强等优点[7],受到了研究人员的广泛关注,并用于语音增强任务中。Tan 等[8]提出卷积循环网络(convolutional recurrent network, CRN),联合 CNN 和 LSTM 获取上下文信息,但不能更好地保留输入特征结构。Pandey 等[9]提出基于自动编码器的全卷积神经网络,在时域进行语音增强,但存在特征训练困难的问题。Tan 等[10]提出门控卷积循环网络(gated convolutional recurrent networks, GCRN),捕获语音时频特征,使用分组策略提高模型效率。随后,深度复卷积循环网络(deep complex convolution recurrent network, DCCRN)[11]结合 DCUNET(deep complex u-network)[12]和 CRN 的优点,利用 LSTM 对时间上下文进行建模,取得了优于 GCRN 的性能。Zhao 等[13]在 DCCRN 的基础上结合前馈顺序记忆网络(feedforward sequential memory network, FSMN)[14]提出频率递归卷积循环网络(frequency recurrence convolutional recurrent network, FRCRN),使用循环模块加强语音的输入特征表示,并在跳跃连接中加入注意模块促进信息流动,但整体计算成本非常高。Pandey 等[15]提出膨胀-密集自动编码器网络(dilated and dense autoencoder, DDAEC),引入密集连接网络(dense connected network, DCN)进行远程上下文聚合,但是较宽的 DCN 会降低网络的参数效率。尽管上述方法在语音增强方面已经取得了一些进展,但其难以处理语音信号中的时序依赖性特征[16]和上下文信息,无法更好地捕捉语音序列中的全局特征,且训练时间较长,导致语音增强效果不理想。

基于上述问题,构建一种残差膨胀卷积与因果-门控编解码网络((residual dilation-encoder gated network, RD-EGN)。该网络在编码器和解码器中分别增加因果卷积、门控线性单元,以时序特征建模,捕获语音序列中不同层的时序数据,选择性地学习和传递有用信息,抑制噪声和无关特征,并引入跳跃连接实现编码层多层次特征的重用,提升对细节特征的保留能力。中间层采用 LSTM 和残差膨胀卷积网络(residual dilated convolutional network, RDCN),有效提取语音长时依赖性特征并在不增加参数量的情况下显著扩大感受野,捕捉更广泛的上下文信息,提升对全局特征的捕获能力。在此基础上,提出了一种 RD-EGN 语音增强算法,以较少的训练参数,并通过 TIMIT、UrbanSound8k、VoiceBank 及 NOISE92 数据集验证了算法的有效性与泛化能力。

1　RD-EGN 网络模型

1.1　网络结构

网络采用编码器-解码器结构,由因果卷积编码器、LSTM、RDCN 和门控反卷积解码器组成,网络参数和结构分别如表 1、图 1 所示。

表 1　模型参数设置

Table 1　Parameter setting of network model

模块	输入维度	层超参数	输出维度
Conv2d_1	1×T ×638	3×5,(1,2)	16×T×319
Conv2d_2	16×T ×319	3×5,(1,2)	16×T×159
Conv2d_3	16×T×159	3×5,(1,2)	16×T×79
Conv2d_4	16×T×79	3×5,(1,2)	32×T×39
Conv2d_5	32×T×39	3×5,(1,2)	32×T×19
Conv2d_6	32×T×19	3×5,(1,2)	64×T×9
Conv2d_7	64×T×9	3×5,(1,2)	64×T×4
Reshape_1	64×T×4		T×256
LSTM	T×256	256	T×256
RDCN	T×256	256	T×256
Reshape_2	T×256		64×T×4
DeConv2d_7	128×T×4	3×5,(1,2)	64×T×9
DeConv2d_6	64×T×9	3×5,(1,2)	32×T×19
DeConv2d_5	32×T×19	3×5,(1,2)	32×T×39
DeConv2d_4	32×T×39	3×5,(1,2)	16×T×79
DeConv2d_3	16×T×79	3×5,(1,2)	16×T×159
DeConv2d_2	16×T×159	3×5,(1,2)	16×T×319
DeConv2d_1	16×T×319	3×5,(1,2)	1×T×638

注:输入、输出维度格式为通道数×时间帧×帧维度(C×T×F),编解码器的层超参数为卷积核、步长

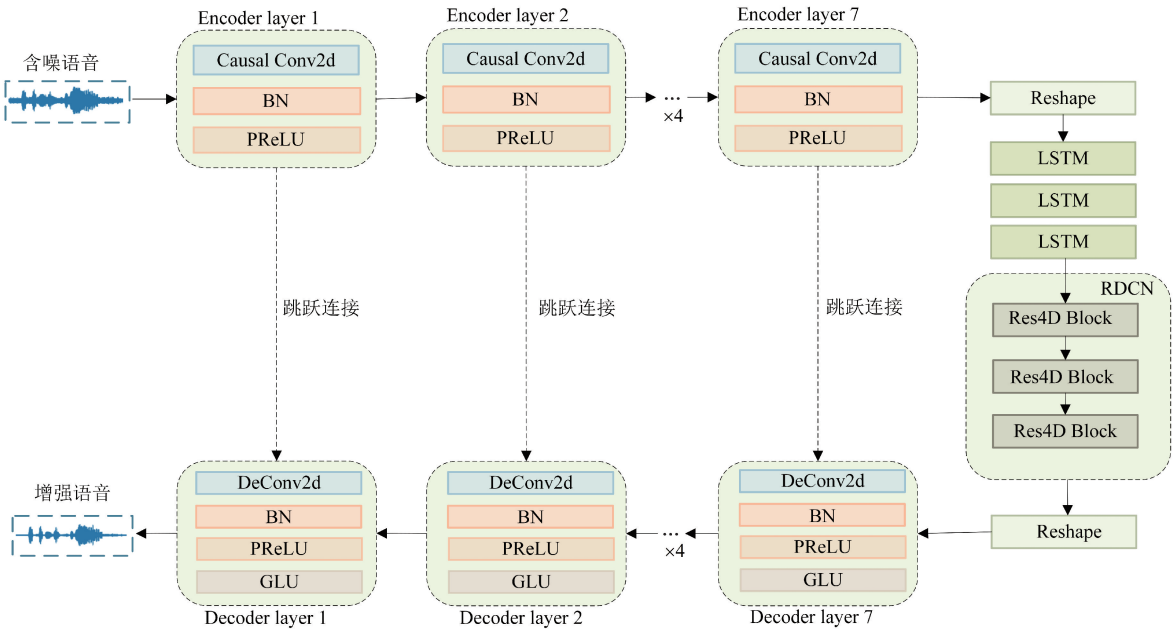


图 1 RD-EGN 网络结构

Fig. 1 The network structure of RD-EGN

编码器由 7 个二维因果卷积层组成,对含噪语音进行时间维度上的卷积操作,每层都重用前面层的输出,逐层捕捉语音时序依赖特征,输出一个包含整个噪声输入帧序列的低维特征向量。第 1 层将通道数量从 1 增加到 16,而后沿着帧维度使用以跨度为 2 依次减小的卷积,最终编码器的输出尺寸为 64×T×4。

中间层由 LSTM 网络和 RDCN 网络组成,其中 RDCN 由因果膨胀卷积、残差连接组成残差膨胀块(Res4D Block)级联在一起。卷积前先对编码器输出的张量进行降维操作,生成大小为 $T\times 256$ 的一维信号。中间层通过全连接层对编码器提取的特征数据进行传递,产生相同大小的输出。

解码器由 7 个二维门控反卷积层组成,将输入的低维语音特征恢复为高维特征表示,通过门控机制调整中间层输出语音特征的重要性,并结合待增强语音特征重建干净语音。解码层不修改沿时间维度的大小,使得输出与输入信息具有相同的帧数,保持一致的上下文信息。此外,编码器与解码器的对应层采用跳跃连接,允许低层次的信息与高层次的信息相结合,补偿编码过程中的信息丢失。

1.2 因果卷积层

在长序列建模时,为了有效地提取语音信号特征,在编码层设计了一种二维因果卷积层结构,每层都包括因果卷积、批归一化层(batch normalization,BN)和非线性激活函数(parametric rectified linear unit,PReLU),高效地通过多层卷积操作提取不同层次的特征,捕获各个时间点之间的相关性和依赖性,并保持语音信号的因果性,因果卷积层结构如图 2 所示。

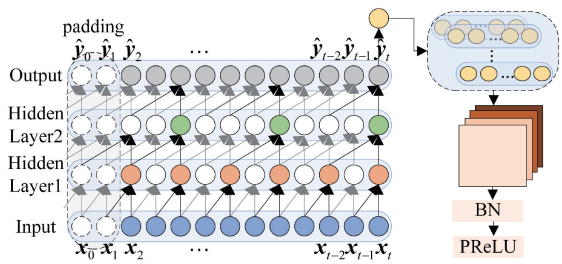


图 2 因果卷积层结构

Fig. 2 The structure of causal convolutional layer

因果卷积使用步长为(1,2)、卷积核大小为(3,5)的滤波器,以卷积核在时间维度上滑动提取特征。卷积前先在输入序列张量左侧进行 padding 零填充(填充数量等于卷积核-1),使输出序列张量与输入语音序列张量具有相同的长度。设语音序列 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$,滤波器 $\boldsymbol{F}=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_K)$,输入节点 $\boldsymbol{x}_t$ 处的因果卷积定义为:

$$(\boldsymbol{F}\times\boldsymbol{X})(\boldsymbol{x}_t)=\sum_{k=1}^{K}\boldsymbol{f}_K\boldsymbol{x}_{t-K+k} \tag{1}$$

当以第 1 层隐藏层作为输出时,输出层节点 $\hat{\boldsymbol{y}}_t$ 表达式为:

$$\hat{\boldsymbol{y}}_t=\boldsymbol{f}_1\boldsymbol{x}_{t-1}+\boldsymbol{f}_2\boldsymbol{x}_t \tag{2}$$

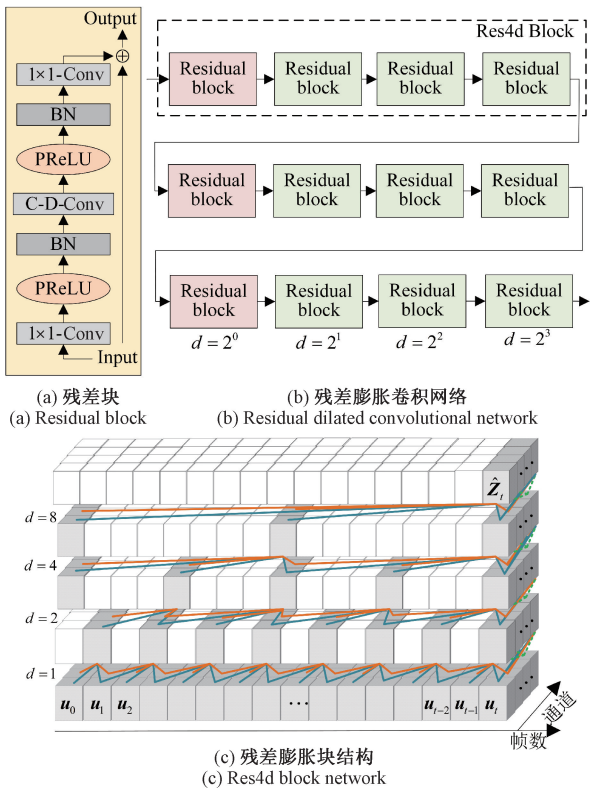
式中:$\boldsymbol{x}_{t-1}$、$\boldsymbol{x}_t$ 分别为输入层最后两个节点;$\boldsymbol{F}=(\boldsymbol{f}_1,\boldsymbol{f}_2)$ 为滤波器。隐藏层的最后一个节点关联了输入的最后两个节点即 $\boldsymbol{x}_{t-1}$、$\boldsymbol{x}_t$,卷积操作只与当前和过去的

语音序列有关,确保在计算时不会将未来的语音信息泄露到当前时刻[17]。

因果卷积逐个时间步长地提取语音特征,生成一系列特征图,表示输入信号在不同时间尺度的特征,对应于卷积核在不同位置上的响应,特征图经过批量归一化处理和 PReLU 激活函数后送入下一层。

1.3　残差膨胀卷积网络

语音信号特征分布在不同时间尺度上,为了增加模型的感受野[18]并捕捉到长序列语音的上下文关系,进一步提取和整合语音序列中的特征。在中间层设计了 RDCN,使用残差连接有效地处理语音上下文特征,并采用膨胀卷积扩大感受野而不增加计算复杂度。RDCN 由 3 个残差膨胀块(Res4d Block)级联在一起,每个 Res4d Block 由 4 个膨胀率呈指数增长的残差块叠加而成,膨胀率分别为 2^0、2^1、2^2、2^3,RDCN 模型结构如图 3(a)所示。



(a) 残差块
(a) Residual block

(b) 残差膨胀卷积网络
(b) Residual dilated convolutional network

(c) 残差膨胀块结构
(c) Res4d block network

图 3　RDCN 结构

Fig. 3　The structure of RDCN

残差块包括输入卷积、因果膨胀卷积和输出卷积,结构如图 3(b)所示,输入和输出都通过一维卷积来控制大小。输入卷积接收来自前一层的输出特征作为输入,并将输入特征的通道数量增加 1 倍。因果膨胀卷积使用不同扩张率的卷积核和深度可分离卷积操作来覆盖更大的输入区域,以较少的计算代价实现更大的感受野,其表达式为:

$$\boldsymbol{Z}(p)=(\boldsymbol{u}\times\boldsymbol{f}_d)(p)=\sum_{i=0}^{k-1}\boldsymbol{f}_d(i)\boldsymbol{u}_{p-d\cdot i} \tag{3}$$

式中:$\boldsymbol{f}_d$ 为卷积核大小,d 是膨胀因子系数,网络 i 级的 $d=O(2^i)$;$\boldsymbol{u}$ 和 $\boldsymbol{Z}$ 分别是输入特征与输出特征;k 是滤波器大小;$p-d\cdot i$ 为感受野的边界。

在输入卷积和因果膨胀卷积之后,采用 PReLU 和 BN 来正则化网络。输出卷积用于将输入特征通道数恢复到原始通道数,使输入卷积和输出卷积能够相加。在输出端使用残差连接,将增强后的输出特征与原始输入特征相加作为残差块的最终输出,传递给下一层。残差连接使网络以跨层的方式传递信息,保留输入语音的重要特征。经过残差块运算的第 N 层输出表示为:

$$\boldsymbol{Z}_N=\boldsymbol{F}_N(\boldsymbol{Z}_{N-1}) \tag{4}$$

式中:$\boldsymbol{Z}_N$ 与 $\boldsymbol{Z}_{N-1}$ 分别为一个残差块在第 N 层与第 $N-1$ 层输出的一维特征;$\boldsymbol{F}_N$ 表示一系列操作的组合。将 4 个残差块以呈指数增长的膨胀率堆叠在一起形成 Res4d Block,在时序上进行卷积操作,结构如图 3(c)所示。多个级联的 Res4d Block,帮助模型结合相邻时间点的信息,实现局部特征的融合和整体特征的提取。

1.4　门控反卷积层

由于普通一维卷积对语音序列的处理能力有限,门控结构已被证明在构建分层表示和捕获语音序列远程依赖关系方面是有效的[19]。因此在解码层中引入门控机制,构造了一个二维门控解卷积层,包括二维反卷积、批量归一化层、PReLU 激活函数和门控线性单元(gated linear unit, GLU),有选择地输出重要特征,实现对特征的动态控制,有针对地重建和增强语音序列。

GLU 通过控制输入特征来优化语音特征提取,包括两个激活函数,分别为 Linear 线性激活函数和 Sigmoid 激活函数。Linear 用来缓解反向传播产生的梯度消失问题,Sigmoid 用来维持网络的非线性特性,其控制 $\boldsymbol{X}_L*\boldsymbol{V}_L+\boldsymbol{c}_L$ 中哪些语音特征信息可以传入下一层。GLU 层中蕴含的单元数取决于输入特征序列的数量。计算公式如式(5)所示。

$$\boldsymbol{X}_{L+1}=(\boldsymbol{X}_L*\boldsymbol{W}_L+\boldsymbol{b}_L)\otimes\sigma(\boldsymbol{X}_L*\boldsymbol{V}_L+\boldsymbol{c}_L)=\boldsymbol{v}_1\otimes\sigma(\boldsymbol{v}_2) \tag{5}$$

式中:$\boldsymbol{X}_{L+1}$、$\boldsymbol{X}_L\in\boldsymbol{R}^{N\times m}$ 分别表示第 $L+1$ 层和第 L 层的输出语音特征;m、n 分别为输入和输出语音序列特征映射的个数;k 为 patch 大小;$\boldsymbol{W}$、$\boldsymbol{V}\in\boldsymbol{R}^{k\times m\times n}$,$\boldsymbol{b}$、$\boldsymbol{c}\in\boldsymbol{R}^n$ 为学习参数;σ 为 Sigmoid 激活函数;$*$ 表示卷积算子;$\otimes$为矩阵元素之间的哈达玛积。

门控反卷积层将中间层输出的语音特征作为输入,通过转置卷积进行上采样,逐步恢复至原始信号维度。采用 GLU 门控机制对转置卷积输出的语音特征向量进行调节,使其经过激活函数后,通过哈达玛积逐位相乘得到相应的隐层向量,采用 Softmax 函数,将神经元的输出

映射到[0,1]的区间,得到解码层的最终输出特征向量。由 GLU 构建的门控解卷积层结构如图 4 所示。

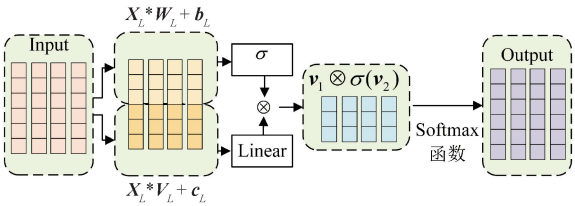


图 4 门控反卷积层结构

Fig. 4 The structure of gated deconvolution layer

1.5 RD-EGN 网络语音增强算法

在上述网络基础上,搭建一个 RD-EGN 语音增强算法,算法框架如图 5 所示。包括训练阶段、测试阶段两部分。

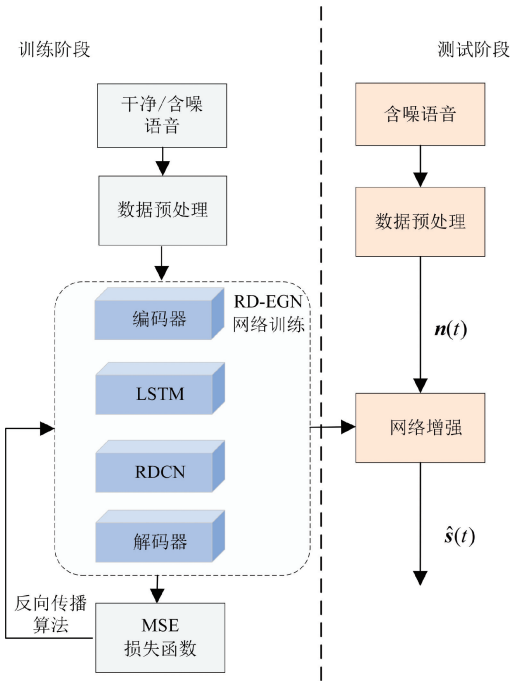


图 5 RD-EGN 语音增强算法框架

Fig. 5 Framework of RD-EGN speech enhancement algorithm

在训练过程中,首先进行数据预处理,即对含噪语音、干净语音进行分帧处理,生成带噪语音帧序列 $\boldsymbol{n}(t)$:$\boldsymbol{n}_0,\boldsymbol{n}_1,\cdots,\boldsymbol{n}_{t-1},\boldsymbol{n}_t$ 作为输入序列和干净语音帧序列 $\boldsymbol{s}(t)$:$\boldsymbol{s}_0,\boldsymbol{s}_1,\cdots,\boldsymbol{s}_{t-1},\boldsymbol{s}_t$ 作为输出序列,将其进行监督学习。形式上对语音序列进行时间建模任务就是学习一个 $M:\boldsymbol{n}_0\cdots\boldsymbol{n}_t\rightarrow\hat{\boldsymbol{s}}_0\cdots\hat{\boldsymbol{s}}_t$ 的映射关系:

$$\hat{\boldsymbol{s}}_0\cdots\hat{\boldsymbol{s}}_t=M(\boldsymbol{n}_0\cdots\boldsymbol{n}_t) \tag{6}$$

对 RD-EGN 网络模型进行训练,当映射关系 M 满足输入语音序列与输出语音序列之间的因果限制时,通过计算均方误差损失函数(mean-square error, MSE)最小化含噪语音和干净语音之间的差异,经过网络多轮次的训练和学习,调整网络参数后得到估计的输出语音序列即增强语音序列 $\hat{\boldsymbol{s}}(t)=\hat{\boldsymbol{s}}_0,\hat{\boldsymbol{s}}_1,\cdots,\hat{\boldsymbol{s}}_{t-1},\hat{\boldsymbol{s}}_t$。

2 实验数据与配置

2.1 数据集

本文以一种说话人和噪声无关的方式评估了 RD-EGN。纯净语音选自 TIMIT 语料库[20],分别选取 80%、10%、10% 作为训练集、测试集和验证集。噪音选自 UrbanSound8k 数据集,在-5、0 和 5 dB 的信噪比(SNR)下生成训练语音。

实验的测试集是由 TIMIT 数据集与 NOISE92 数据集中 Babble、Factory2、Destroyerengine、Hfchannel 噪音分别以-5、0 和 5 dB 混合得到。为了评估模型的增强效果,创建两个测试集,评估受过训练和未受过训练说话人的表现。第 1 个测试集使用来自 TIMIT 训练集中 20 位说话者的语音,第 2 个测试集使用 20 位不包括在训练集中说话者的语音。为了研究模型的泛化性,在 VoiceBank 数据集中随机选择 200 条语音与 NOISE92 数据集中 Babble、Factory2 分别以-5 和 0 dB 混合得到。验证集是与 NOISE92 数据集中的 Factory1 噪音以-5 dB 混合得到的。所有训练的含噪语音由随机选取的干净语音与噪声语音在随机信噪比下产生的。

2.2 实验设置

本文实验所有语音采用 16 kHz 的采样频率,使用大小为 20 ms 的汉明窗提取帧,相邻帧之间重叠 50%,得到每帧 161 维的频率特征向量。训练时使用 8 个话语的 batch size 进行训练,对于短语音采用同 batch 中最长语音大小的零填充匹配,对于时长大于 4 s 的语音取 4 s 的随机片段。所有模型都使用(MSE 作为损失函数,采用 Adam 优化器用于基于随机梯度下降(stochastic gradient descent,SGD)的优化,学习速率为 0.000 2。

2.3 评价指标及对比模型

在实验中,使用语音质量感知评价(perceptual evaluation of speech quality, PESQ)[21]、短时客观可懂度(short-Time objective intelligibility, STOI)[22]分数和分段信噪比(segmented signal-to-noise ratio,SSNR)来客观评估模型。PESQ 的取值范围为-0.5~4.5,STOI 的取值范围为 0~1,SSNR 越大,表示语音所含噪声和失真越少。

同时采用 3 种主观平均意见得分(mean opinion score, MOS):信号失真测度(CSIG)、噪声失真测度(CBAK)和综合质量测度(COVL),取值范围都为 1~5,得分越高,表示模型性能越好[23]。本文选用 5 种不同的语音增强网络作为对比模型,包括 LSTM、CRN[8]、自编码

卷积神经网络(AECNN)[9]、DCCRN[11]、DDAEC[15]。其中 LSTM 从输入层到输出层分别有 161、1 024、1 024、1 024、1 024、161 个单元。

3　实验结果与分析

3.1　模型参数实验

为了探究不同参数设置对模型性能的影响,设置不同的编解码层数及中间层 Res4d Block 中残差块个数,计算 PESQ 和 STOI 得分,如图 6 所示。当残差块个数固定时,随着编解码层数从 5 增加到 8,PESQ 和 STOI 得分均有所提升,7 和 8 层编解码层结构性能较好且均优于 6 层。在 7 层结构下,随着残差块个数增加,PESQ 得分先增后减,在个数为 4 时表现出最佳值,说明适当增加残差块个数可以增大感受野,使模型能够捕捉到更全局的特征。虽然编解码层数为 7 和 8 层的性能相当,但随着层数增加模型复杂度也会提高,可能导致过拟合。综上,使用 7 层编解码层结构和 4 个残差块。

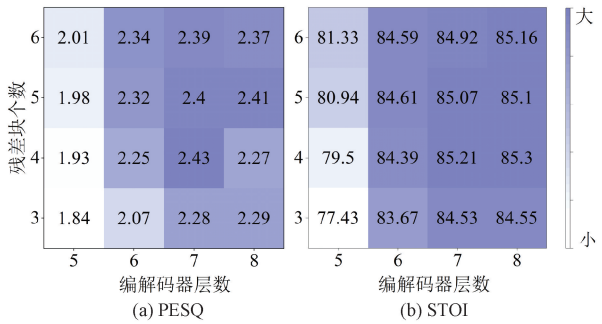


图 6　模型在不同参数下的 PESQ 和 STOI

Fig. 6　PESQ and STOI of models at different parameters

3.2　模型性能评估

为了探究 RD-EGN 模型的优越性,与 LSTM、CRN、AECNN、DDAEC 和 DCCRN 语音增强模型进行比较,其 PESQ 和 STOI 得分、3 种主观平均意见得分和 SSNR 得分分别如表 2、图 7 和 8 所示。以 TIMIT 数据集为例,可以看出,RD-EGN 在-5、0 dB 信噪比下均取得了较好的增强效果。相较于 LSTM,RD-EGN 在处理长序列语音时受梯度消失影响较小,因此增强效果明显优于 LSTM,例如在-5 dB 时 PESQ 提高 18.3%,STOI 提高 11.1%。相较于 CRN,虽然二者都采用编码器-解码器结构,但 CRN 采用递归网络,利用增强的幅度信息与噪声相位重构得到增强语音;而 RD-EGN 采用改进的卷积网络以时域波形建模,保留了更多的原始语音信息,捕获更广泛的上下文特征,取得了优于递归网络的效果,因此得分显著高于 CRN。另外,相较于 AECNN、DDAEC 和 DCCRN 模型,RD-EGN 增强效果最好,PESQ 平均提高了 2.5% ~ 7.1%,STOI 有 1% ~ 5.3% 的平均改善,说明此模型网络能够更好地控制层次结构中传递的信息,使模型学习到更准确的特征。

另外,本文还在 VoiceBank 数据集上进行对比实验,可以看出 RD-EGN 获得了更高 PESQ 和 STOI,说明 RD-EGN 对不同数据集有较好的泛化能力。

由图 7 可以看出,RD-EGN 的主观指标得分均优于其他模型,其中 CSIG 和 CBAK 指标的提升幅度较大,充分说明 RD-EGN 模型在增强过程中提取的特征能更好的还原语音信号,并有效抑制背景噪声,使得语音信号质量和听觉效果较好。

4 种测试噪声下各模型在 3 种信噪比下的 SSNR 如图 8 所示,可以看出,RD-EGN 模型在 0 和 5 dB 下均获得

表 2　不同数据集上各模型的 PESQ 和 STOI

Table 2　PESQ and STOI of each model

数据集	模型	Babble				Factory2			
		PESQ		STOI		PESQ		STOI	
		-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
TIMIT	Noisy	1.54	1.78	0.546 8	0.682 2	1.47	1.63	0.526 1	0.670 9
	LSTM	1.93	2.09	0.715 4	0.786 7	1.89	1.98	0.703 5	0.768 4
	CRN	2.09	2.17	0.743 5	0.815 9	2.05	2.13	0.736 9	0.803 0
	AECNN	2.16	2.29	0.754 1	0.817 2	2.11	2.24	0.743 1	0.810 5
	DDAEC	2.21	2.34	0.779 5	0.827 6	2.14	2.30	0.756 3	0.818 9
	DCCRN	2.23	2.39	0.782 4	0.845 2	2.18	2.33	0.777 4	0.838 2
	RD-EGN	2.28	2.44	0.790 3	0.853 6	2.24	2.41	0.786 2	0.846 7
VoiceBank	LSTM	1.91	2.05	0.689 4	0.753 1	1.79	1.92	0.675 8	0.739 6
	CRN	2.04	2.11	0.727 5	0.789 3	1.96	2.04	0.713 4	0.772 1
	AECNN	2.12	2.23	0.735 9	0.804 7	2.08	2.20	0.726 5	0.800 3
	DDAEC	2.18	2.28	0.753 8	0.826 6	2.10	2.26	0.738 6	0.807 8
	DCCRN	2.20	2.31	0.779 1	0.840 9	2.15	2.37	0.753 3	0.815 7
	RD-EGN	2.24	2.37	0.788 2	0.837 2	2.21	2.35	0.774 8	0.830 7

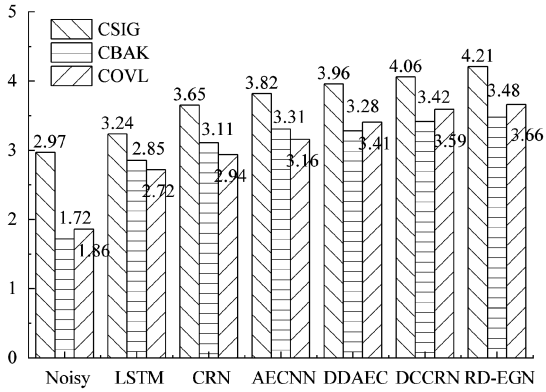


图 7 各模型的 CSIG、CBAK 和 COVL 得分

Fig. 7 CSIG, CBAK and COVL scores of each model

了较高的 SSNR 值,并且 SSNR 得分普遍优于另外 5 种模型,在 0 dB 时,RD-EGN 模型的 SSNR 值已经达到了其他模型在更高信噪比(如 3.5 或 5 dB)时方能达到的水平,这充分说明经 RD-EGN 模型增强后的语音信号残留噪声较少,并且随着信噪比的增加,该模型在噪声抑制方面更为出色。

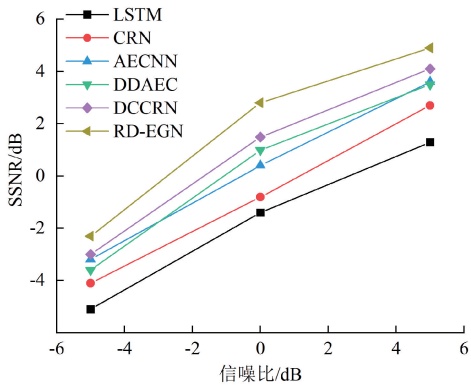


图 8 4 种测试噪声下各模型在 3 种信噪比下的 SSNR

Fig. 8 The average SSNR of each model in three SNR under four test noises

不同模型的可训练参数数量如表 3 所示。RD-EGN 使用卷积网络和门控机制,使模型选择重要的特征信息来建模,与其他模型相比,具有较少的参数,大大减少整个模型的训练时间,同时保持了更好的性能,更适合用于需要高效实现的应用。

表 3 各模型的可训练参数个数

Table 3 The number of trainable parameters for each model

模型	LSTM	CRN	AECNN	DCCRN	DDAEC	RD-EGN
参数量/($\times 10^6$)	36.8	17.6	19.2	14.3	15.7	10.4

3.3 语音增强实验

不同说话人的语音特征不同,在语音增强过程中,模型需要能够适应各种语音特征,既不能过度拟合受过训练的说话人语音,同时对未受训练的说话人语音有一定的泛化能力。为了研究模型的增强效果,比较受过训练和未受过训练说话人语音的 PESQ 和 STOI,如表 4 所示。可以看出,受过训练的说话人在 PESQ 和 STOI 得分方面优于未受训练的说话人,例如,在信噪比为-5 dB 时,受过训练的说话人比未受过训练的说话人语音的 STOI 提高 0.4%,在 0 dB 时提高 1.3%,以及 6.2%的 PESQ。受过训练的说话人模型使用大量说话人的语音数据进行训练,这些语音信号包含各种声学特征和说话场景,RD-EGN 网络允许模型学习到更广泛的语音信号特征并捕获关键特征,其中门控机制动态控制特征的传输,防止网络过度依赖特定说话人特征,提升网络对带噪语音的建模能力和语音质量。

另外,未受过训练的说话人语音的 PESQ 和 STOI 得分相较于原始含噪语音分别提高了 42.4%、35.5%,说明 RD-EGN 网络具有一定的泛化能力,能够很好的概括说话人语音,即使说话人未经过专门训练,模型仍然能够从语音信号中提取出一些有用的信息。

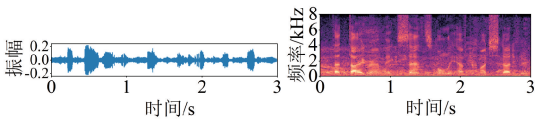
表 4 受过训练的说话人和未受过训练的说话人语音的 PESQ 和 STOI(%)

Table 4 PESQ and STOI of trained and untrained speakers

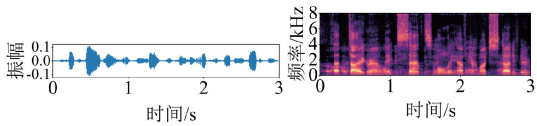
说话人语音	模型	Babble				Factory2			
		PESQ		STOI		PESQ		STOI	
		-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
受过训练	Noisy	1.54	1.78	0.546 8	0.682 2	1.47	1.63	0.526 1	0.670 9
	RD-EGN	2.28	2.44	0.790 3	0.853 6	2.21	2.41	0.786 2	0.846 7
未受过训练	Noisy	1.49	1.73	0.543 1	0.689 1	1.42	1.60	0.518 8	0.667 2
	RD-EGN	2.21	2.31	0.779 4	0.847 5	2.19	2.22	0.776 9	0.839 7

为了更直观的展示 RD-EGN 网络的语音增强效果,以-5、0 dB 下混合了 Babble 噪音的语音为例,含噪语音、纯净语言和经 RD-EGN 增强后语音的波形图和语谱图如图 9、10 所示,可以看出,增强后的语音接近纯净语音,实现了有效的语音增强。波形图都呈现出清晰的语音特征,噪声能量被有效减少;语谱图呈现出清晰的谱时结

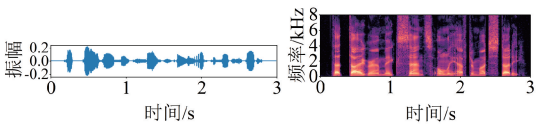
构,意味着语音信号的细节信息得到更好的保留。



(a) 含噪语音
(a) Noisy speech



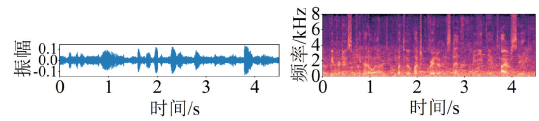
(b) 纯净语音
(b) Clean speech



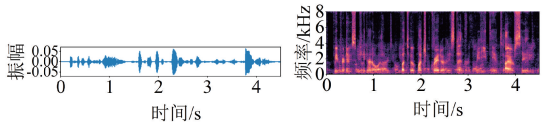
(c) RD-EGN增强语音
(c) RD-EGN enhanced speech

图 9　在−5 dB 下混合 Babble 噪音的波形图和语谱图

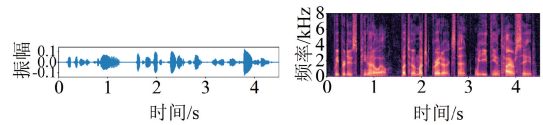
Fig. 9　Waveform and spectrogram of mixed babble noise at −5 dB SNR



(a) 含噪语音
(a) Noisy speech



(b) 纯净语音
(b) Clean speech



(c) RD-EGN增强语音
(c) RD-EGN enhanced speech

图 10　在 0 dB 下混合 Babble 噪音的波形图和语谱图

Fig. 10　Waveform and spectrogram of mixed babble noise at 0 dB SNR

3.4　消融实验

为了探究不同模块对 RD-EGN 网络整体性能的影响,在 TIMIT 数据集上以 PESQ 和 STOI 为评价指标进行消融实验,结果如表 5 所示。Noisy 表示含噪语音测试样本,EGN-A、EGN-B 分别为用普通卷积替换编码器中因果卷积层、解码器中门控反卷积层,EGN 为编码器-解码器结构,EGN-LSTM 和 EGN-RDCN 分别为在编解码器结构中添加 LSTM、残差膨胀卷积网络,EGN-LSTM-RDCN 即为 RD-EGN 网络。

表 5　不同模块消融实验的 PESQ 和 STOI

Table 5　PESQ and STOI of different module ablation experiments

模型	Babble				Factory2			
	PESQ		STOI		PESQ		STOI/%	
	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB	−5 dB	0 dB
Noisy	1.53	1.75	0.549 5	0.667 2	1.49	1.66	0.526 4	0.674 9
EGN-A	1.84	1.92	0.683 2	0.721 5	1.78	1.90	0.642 8	0.710 3
EGN-B	1.90	1.98	0.707 6	0.735 8	1.85	1.93	0.674 5	0.723 6
EGN	1.99	2.05	0.732 4	0.785 5	1.90	2.01	0.713 8	0.779 1
EGN-LSTM	2.08	2.17	0.748 9	0.804 9	2.06	2.12	0.727 6	0.793 5
EGN-RDCN	2.25	2.29	0.763 1	0.837 6	2.11	2.24	0.752 9	0.816 5
EGN-LSTM-RDCN	2.31	2.43	0.797 2	0.852 1	2.25	2.32	0.786 2	0.837 1

由表 5 发现,使用普通卷积替换编解码器中的卷积层后,性能明显下降,可见因果卷积层的时序性利于网络提取更准确的特征,门控反卷积层能够减少冗余信息的传递,提高网络性能。另外,只添加 LSTM 或 RDCN 时,与单一的编解码器(EGN)结构相比,PESQ 和 STOI 有所提高,但是当同时添加 LSTM 和 RDCN 时,PESQ 和 STOI 有大幅度提升。在信噪比为−5 dB 时,EGN-LSTM-RDCN 比 EGN-LSTM 模型提高了约 10.1%的 PESQ 和约 2.7%的 STOI。即 RDCN 能够帮助获取更丰富的原始语音特征,且中间层使用 LSTM 和 RDCN 处理编码器提取的序列特征,优于仅使用 EGN 结构,与 LSTM 处理序列信息的能力相结合,使得网络具有更好的建模能力及增强效果。

4　结　论

本文针对编解码网络对时序依赖性特征和上下文信息捕获不充分,导致增强效果差的问题,在分别设计编码器、中间层和解码器基础上,提出一种非对称的 RD-EGN 网络。该网络编码器通过因果卷积层捕获语音序列中不

同层次的时序特征;中间层将设计的 RDCN 与 LSTM 相结合,获得更大的感受野,提高对语音长时依赖性特征和更广泛上下文信息的提取能力;解码器引入门控机制自适应地学习特征表示,有效地捕获语言序列中的重要特征,进一步提升增强效果。利用该网络在波形域通过语音序列建模实现语音增强,实验结果表明,RD-EGN 在 4 个不同数据集上的增强语音质量与可懂度明显优于其他模型,展现出卓越的泛化能力,实现了模型参数与增强效果之间的平衡。未来的工作包括提出一些轻量级模型结构,以满足实时应用或计算资源受限场景的需求,并探索不同领域下的潜在应用,如语音识别、语音合成等。

参考文献

[1] HAO L, CAO S, ZHOU P, et al. Denoising method based on spectral subtraction in time-frequency domain [J]. Advances in Civil Engineering, 2021(1):6621596.

[2] FARAJI N, KOHANSAL A. MMSE and maximum a posteriori estimators for speech enhancement in additive noise assuming at-location-scale clean speech prior[J]. IET Signal Processing,2018,12(4):532-543.

[3] 袁文浩,屈庆洋,梁春燕,等. 基于感知条件网络的可控语音增强模型[J]. 仪器仪表学报,2023,44(5):53-60.
YUAN W H,QU Q Y,LIANG CH Y,et al. Controllable speech enhancement model based on perceptual conditional network [J]. Chinese Journal of Scientific Instrument, 2023,44(5):53-60.

[4] XU Y, DU J, HUANG Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[C]. Interspeech,2015: 1508-1512.

[5] TAKEUCHI D, YATABE K, KOIZUMI Y, et al. Real-time speech enhancement using equilibriated RNN[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020:851-855.

[6] SALEEM N, KHATTAK M I, AL-HASAN M, et al. On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks[J]. IEEE Access,2020,8:160581-160595.

[7] 罗庆予,张天骐,方蓉,等. 联合频谱映射与掩蔽估计的协作式语音增强方法[J]. 电子测量与仪器学报,2023,37(10):14-23.
LUO Q Y,ZHANG T Q,FANG R,et al. Collaborative speech enhancement method combining spectral mapping and masking estimation [J]. Journal of Electronic Measurement and Instrumentation,2023,37(10):14-23.

[8] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement [C]. Interspeech, 2018:3229-3233.

[9] PANDEY A, WANG D L. A new framework for CNN-based speech enhancement in the time domain [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.

[10] TAN K, WANG D L. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 380-390.

[11] HU Y, LIU Y, LYU S, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]. Interspeech,2020:2472-2476.

[12] CHOI H S, KIM J H, HUH J, et al. Phase-aware speech enhancement with deep complex u-Net [C]. International Conference on Learning Representations, 2019.

[13] ZHAO S, MA B, WATCHARASUPAT K N, et al. FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement [C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022:9281-9285.

[14] WANG Z, NA Y, LIU Z, et al. Weighted recursive least square filter and neural network based residual echo suppression for the aec-challenge [C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 141-145.

[15] PANDEY A, WANG D L. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain [C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6629-6633.

[16] 喻永振,刘大明. 基于幅值滤波与分层特征融合策略的语音情感识别[J]. 国外电子测量技术,2024,43(3):35-42.
YU Y ZH,LIU D M. Speech emotion recognition based on amplitude filtering and hierarchical feature fusion strategy[J]. Foreign Electronic Measurement Technology, 2024,43(3):35-42.

[17] 孙坤,尹晓红. 基于数据去噪和 CNN-BiGRU 的 SO_2 排放预测[J]. 电子测量技术,2023,46(13):66-72.
SUN K,YIN X H. SO_2 emission prediction based on data denoising and CNN-BiGRU[J]. Electronic Measurement Technology,2023,46(13):66-72.

[18] LYU M N, ZHOU X C, DU Z T, et al. Image denoising using dual convolutional neural network with skip connection[J]. Instrumentation, 2024, 11(3): 74-85.
[19] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks [C]. International Conference on Machine Learning, 2017.
[20] LYONS J W. DARPA TIMIT acoustic-phonetic continuous speech corpus [R]. Gaithersburg: National Institute of Standards and Technology, 1993.
[21] OLATUBOSUN A, OLABISI O P. An improved logistic function for mapping raw scores of perceptual evaluation of speech quality (PESQ) [J]. Journal of Engineering Research and Reports, 2018, 3(1): 1-10.
[22] NOGUCHI K, KOBAYASHI Y, KISHIGAMI J, et al. Listening difficulty estimation model using short-time objective intelligibility measure for outdoor public address systems[J]. Acoustical Science and Technology, 2020, 41(1): 420-422.
[23] LIN Z, ZHOU L, QIU X. A composite objective measure on subjective evaluation of speech enhancement algorithms [J]. Applied Acoustics, 2019, 145 (1): 144-148.

作者简介

李珂，2021 年于郑州工程技术学院获得学士学位，现为山东理工大学硕士研究生，主要研究方向为语音信号处理。
E-mail：like_0619@163.com

Li Ke received her B. Sc. degree from Zhengzhou Institute of Technology in 2021. Now she is a M. Sc. candidate at Shandong University of Technology. Her main research interest includes speech signal processing.

王雅静（通信作者），2011 年于上海理工大学获得博士学位，现为山东理工大学教授、博士生导师，主要研究方向为信号处理技术。
E-mail：wangyajing@sdut.edu.cn

Wang Yajing (Corresponding author) received her Ph. D. degree from University of Shanghai for Science and Technology in 2011. Now she is a professor and Ph. D. supervisor at Shandong University of Technology. Her main research interest includes signal processing technology.