DOI: 10.13382/j.jemi.B2205308

基于 A-DResUnet 的语音增强方法*

李吉祥 倪旭昇 颜上取 邹 孝 钱盛友
(湖南师范大学物理与电子科学学院 长沙 410081)

摘 要:为了更精确地从语谱图中提取特征信息,提出了一种基于 A-DResUnet 的语音增强方法。A-DResUnet 模型在 ResUnet 模型的基础上融合了空洞卷积,提升捕获语音上下文信息的能力;同时在编码器中加入卷积注意力模块(CBAM),提高对噪声谱图特征的关注。实验结果表明,与模型输出目标为干净语音语谱图相比,用噪声谱图作为模型输出目标时,该模型对未知噪声具有更强的分离能力;相较 ResUnet 模型,提出的 A-DResUnet 模型减少了语音细节信息的损失;对比基于 DNN、GAN 的语音增强方法,PESQ 平均提升了 22.81%、33.11%,STOI 平均提升了 9.62%、15.33%,为复杂环境下的语音增强提供了一种更有效的方法。

关键词: 语音增强;语谱图;模型输出目标;空洞卷积;卷积注意力模块
中图分类号: TN912.35 **文献标识码**: A **国家标准学科分类代码**: 510.4040

Speech enhancement method based on A-DResUnet

Li Jixiang Ni Xusheng Yan Shangqu Zou Xiao Qian Shengyou
(School of Physics and Electronics, Hunan Normal University, Changsha 410081, China)

Abstract: In order to extract feature information from spectrogram more accurately, this paper proposes a speech enhancement method based on A-DResUnet (attention-dilated ResUnet). The A-DResUnet model incorporates dilated convolution on the basis of ResUnet model to improve the ability to capture the contextual information of speech; at the same time, the convolution block attention module (CBAM) is added into the ResUnet encoder to improve the attention to the features of the noise spectrogram. The experimental results show that when the noise spectrum is used as the output target of the model, the model has a stronger ability to separate unknown noise than when the output target of the model is clean speech spectrum; compared with the ResUnet model, the proposed A-DResUnet model reduces the loss of speech detail information; compared with the speech enhancement methods based on DNN and GAN, PESQ increased by an average of 22.81%, 33.11%, STOI increased by an average of 9.62%, 15.33%, which is a more effective method for speech enhancement in complex environments.

Keywords: speech enhancement; spectrogram; the output target of the model; dilated convolution; convolution block attention module

0 引 言

随着深度学习的发展,有监督的语音增强方法受到了广泛的关注[1]。基于深度学习的语音增强方法主要分为基于时频掩蔽的方法和基于特征映射的方法。基于特征映射的方法按处理域的类型又可分为两种:一种是直接对语音的时域波形进行映射[2-5],保留了更多的原始波形信息且简化了处理流程,由于时域语音波形本身并不能表达语音的特征信息,直接对时域波形建模比较困难,而且在低信噪比且非平稳噪声的情况下波形十分复杂,增强难度进一步提升;另一种方法是在时频域中处理,大多使用神经网络模型学习从带噪语音语谱图到干净语音语谱图的特征映射,同时利用带噪语音的相位来重构时域信号。要获得精确的语谱图估计,语音上下文信息是十分重要的,而传统的深度神经网络(deep neural network,DNN)模型[6]无法捕获语音上下文信息。因此,有学者将循环神经网络(recurrent neural network,RNN)

收稿日期: 2022-03-25 Received Date: 2022-03-25
*基金项目:国家自然科学基金(11774088)项目资助

和卷积神经网络(convolutional neural network,CNN)用于语音增强[7-8]。U-Net是CNN的一种,其结构以及改进算法广泛用于图像分割任务[9-12],也已被用于语音增强[13-14],但是当层数过深时会影响模型性能。He等[15]在U-Net的基础上,融合残差学习(Residual learning)提出了用于音乐源分离的ResUnet[16],解决了深层网络性能退化的问题,也通过残差模块中的残差连接进行特征融合取得了比U-Net更好的分离效果,但仍然没有充分考虑语音上下文信息,容易造成语音细节信息的损失。针对这些问题,本文在ResUnet的基础上提出了A-DResUnet语音增强模型,用噪声谱图作为该模型的输出目标,并通过实验研究其对未知噪声的增强效果。

1　算法原理

1.1　A-DResUnet模型

ResUnet模型由编码器、解码器组成,本文对其编码器进行改进:在原有的残差模块(Res)中融合空洞卷积得到空洞残差模块(DRes);并在每一个DRes之后加卷积注意力模块(convolutional block attention module,CBAM),最终得到A-DResUnet(attention-dilated ResUnet)模型,结构如图1所示。A-DResUnet模型的编码器由4个DRes以及3个CBAM组成,编码器从输入的语谱图中提取高层次的特征,得到尺寸更小的特征图;解码器由3个Res组成,不使用DRes以防止空洞卷积降低从编码器中获得的特征精细度,解码器将特征图还原至原图大小。考虑到解码过程中信息的损失,采用跳层连接将低层次特征与高层次特征结合进行信息的补充。网络的卷积层使用批量标准化(batch normalization,BN)[17]加快网络的训练、防止过拟合,使用带泄露整流线性单元(leaky rectified linear units,LeakyReLU)[18]解决某些神经元梯度永远为0的问题。

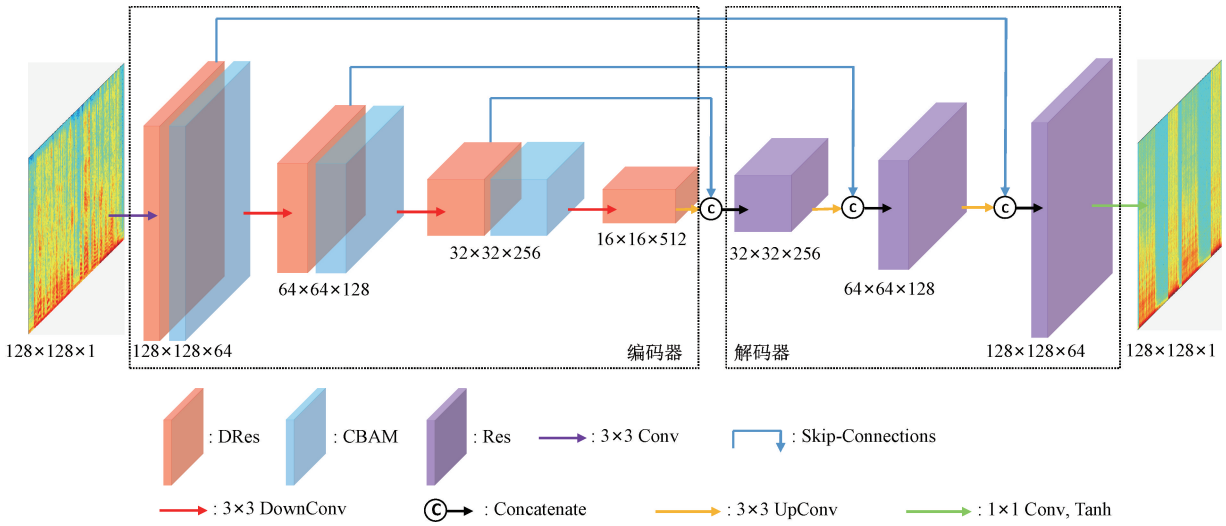


图1　A-DResUnet模型结构图

Fig. 1　The structure diagram of A-DResUnet

1.2　空洞残差模块(DRes)

语谱图中横轴代表时间、纵轴代表频率,相邻帧包含着语音的上下文信息。ResUnet模型通过池化下采样以扩大感受野来捕获语音上下文信息,但是池化操作会损失语谱图的局部特征,导致语音细节信息的丢失。因此,为了弥补下采样过程中的信息损失,本文在其Res中融合空洞卷积得到DRes。空洞卷积可以在不改变分辨率、不增加计算量的情况下扩大感受野,捕获多尺度上下文信息[19],其原理是在普通卷积的卷积核中插入多个零值。对于扩张率为d,卷积核大小为k的空洞卷积,其感受野大小n的计算公式为:

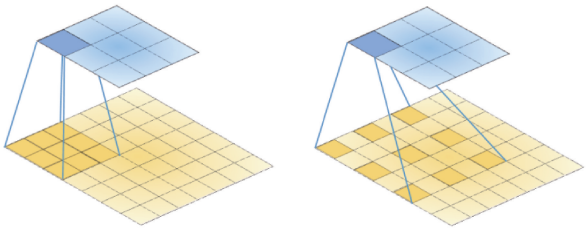
$$n = k + (k - 1) \times (d - 1) \tag{1}$$

图2(a)是一个卷积核为3×3的普通卷积,其感受野为3,图2(b)是一个卷积核为3×3、扩张率为2的空洞卷积,其感受野被扩大到5。

DRes包含空洞卷积部分和残差连接,其结构如图3所示。空洞卷积部分由两个卷积核为3×3、扩张率分别为2、3的空洞卷积以及BN和LeakyReLU组成;利用空洞卷积对时间、频率方向的扩张充分捕获语音上下文信息,以减少原语音信息的损失。

1.3　卷积注意力模块(CBAM)

由于本文用噪声谱图作为模型的输出目标,提取噪声特征的过程会受到说话人特征的影响,导致估计的噪声谱图不够精确。因此,本文在ResUnet模型的编码器



(a) 普通卷积
(a) Normal convolution

(b) 扩张率为2的空洞卷积
(b) Dilated convolution with dilation rate 2

图 2　普通卷积与空洞卷积示意图

Fig. 2　Schematic diagram of convolution and dilated convolution

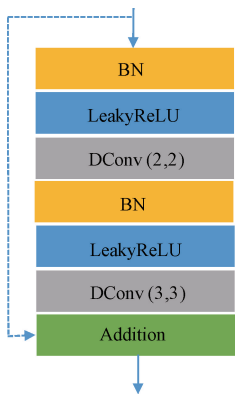


图 3　DRes 结构图

Fig. 3　The structure diagram of DRes

中加入 CBAM，增加对语谱图中噪声特征的关注。CBAM 是一种融合通道维度以及空间维度的注意力模块，用于学习关键信息并抑制无效的特征信息[20]，该模块计算量小，不会增加太多的网络参数，其结构如图 4 所示。

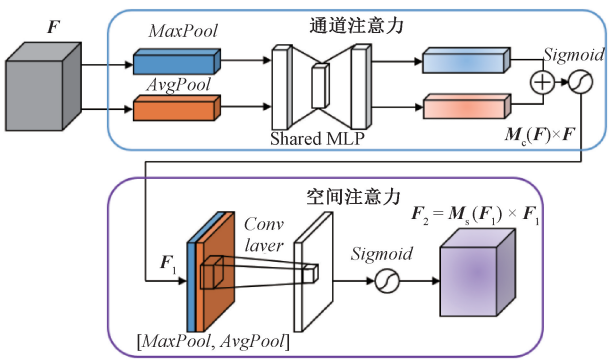


图 4　CBAM 结构图

Fig. 4　The structure diagram of CBAM

首先，通道注意力模块为目标通道特征赋予更大的权重，加强该通道的影响。设输入特征图 F 维度为 $H\times W\times C$，分别经过最大池化和平均池化得到两个 $1\times1\times C$ 的特征图；使用一个包含两个全连接层的多层感知器（multilayer perceptron，MLP），将两个 $1\times C$ 的特征图合并；最后经过 Sigmoid 激活后得到各通道的注意力权重。计算公式为：

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (2)$$

$$F_1 = M_c(F) \times F \quad (3)$$

其中，F 是输入特征，$AvgPool(F)$ 是平均池化，$MaxPool(F)$ 是最大池化，MLP 是多层感知器，$\sigma(\cdot)$ 是 Sigmoid 函数；$M_c(F)$ 是通道注意力模块输出的权重系数，F_1 为经过通道注意力模块得到的特征。

其次，将通道注意力模块的输出 F_1 作为空间注意力模块的输入，关注输入特征空间维度上的重要信息。对 F_1 同时做基于通道的平均池化和最大池化得到两个 $H\times W\times1$ 的特征图；将两个特征图沿通道拼接成 $H\times W\times2$ 的特征图，再通过 3×3 卷积的降维成 $H\times W\times1$；最后经过 Sigmoid 激活后得到空间的注意力权重。计算公式为：

$$M_s(F_1) = \sigma(f([AvgPool(F_1);MaxPool(F_1)])) = \sigma(f([F^s_{avg};F^s_{max}])) \quad (4)$$

$$F_2 = M_s(F_1) \times F_1 \quad (5)$$

其中，F_1 是经过通道注意力模块得到的特征，$AvgPool(F_1)$ 是平均池化，$MaxPool(F_1)$ 是最大池化，$\sigma(\cdot)$ 是 Sigmoid 函数，f 是 3×3 卷积运算；$M_s(F_1)$ 是空间注意力模块输出的权重系数，F_2 为经过空间注意力模块得到的特征。

2　实验结果与分析

2.1　实验方法

基于 A-DResUnet 的语音增强方法分训练阶段和增强阶段。在训练阶段，带噪语音和对应的噪声分别通过短时傅里叶变换（short-time Fourier transform，STFT）转换成带噪语音语谱图和噪声谱图，其中带噪语音语谱图作为 A-DResUnet 模型的输入、对应的噪声谱图作为训练的标签，通过最小化损失函数来更新网络参数得到最优模型；增强阶段的处理流程如图 5 所示，带噪语音通过 STFT 转换成带噪语音语谱图输入到训练好的 A-DResUnet 模型中估计出噪声谱图，再用带噪语音的语谱图减去输出的噪声谱图得到增强后的语音语谱图，最后结合带噪语音的相位信息通过逆短时傅里叶变换（inverse short-time Fourier transform，ISTFT）还原成时域信号。

本文通过实验验证所提方法的有效性，在实验中干净语音来自 LibriSpeech、噪声来自 ESC-50。从 ESC-50 中选取 10 类环境噪声、从 LibriSpeech 中选取 2 702 条干净语音以随机信噪比、随机说话人混合成约 6 h 带噪语音与对应的噪声语音作为训练集；从 LibriSpeech 中选取 40 条不同说话人的干净语音、从 ESC-50 中选取 4 类环境噪

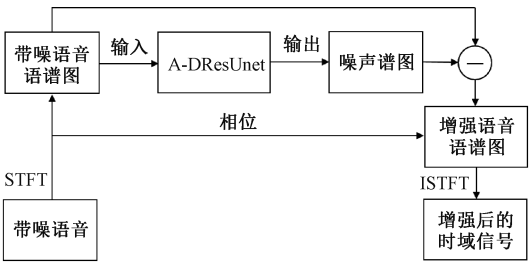


图5　增强阶段的处理流程图

Fig. 5　Process flow diagram for the enhancement phase

声(carhorn、engine、knock、wind),40 条干净语音分别与其中一类噪声以-5、0、5 dB 的信噪比混合成 120 条带噪语音,4 类噪声共得到 480 条带噪语音作为测试集,测试集中的语音以及噪声类型都不同于训练集。

训练集、测试集中的语音采样率为 8 KHz。将训练集中约 6 h 的带噪语音以及对应的噪声语音分别截取成 21 000 段,每段约 1 s;每个截取片段做短时傅里叶变换,使用海宁窗,窗长为 256、窗移 63;取 128 个频率段、128 个时间段得到 128×128 对称大小的单通道语谱图,并且将数据归一化至[-1,1]。最终训练集转换成 21 000 张带噪语音语谱图以及 21 000 张对应噪声谱图,21 000 张带噪语音语谱图为模型的输入,21 000 张噪声谱图为模型的标签。

神经网络模型基于 Keras2.3.1 和 GPU 版 Tensorflow 2.6.1 搭建,使用 RTX3060GPU、Adam 优化器进行训练,损失函数为 Huber,批处理大小为 16,初始学习率设为 0.001,当损失连续 3 次不下降就将学习率减半,当损失连续 10 次不下降代表模型基本训练完毕停止训练。

本文采用的评价指标为短时客观可懂度(short-time objective intelligibility,STOI)取值范围为 0~1,值越高代表可懂度越高;以及语音质量的感知评估(perceptual evaluation of speech quality,PESQ)取值范围为-0.5~4.5,值越大代表语音质量越好。

2.2　输出目标的影响

在以往基于时频域处理的增强方法中,通常是以干净语音的语谱图作为模型的输出目标,而本文用噪声谱图作为模型的输出目标以排除说话人特征的影响、加强模型对于未知噪声的分离能力。为了探究模型相同时不同输出目标对增强效果的影响,同时验证空洞残差模块(DRes)、卷积注意力模块(CBAM)的有效性,本文使用测试集进行实验。每种模型都分别以干净语音语谱图和噪声谱图作为输出目标进行对比,实验结果如表 1 所示,表 1 中结果为同类噪声下 120 条不同带噪语音增强结果的平均值。表 1 中第 1 行为带噪语音未处理的结果,共列出了 ResUnet、DResUnet、A-DResUnet 这 3 种模型的增强效果;同一模型中上行代表输出目标为干净语音语谱图的结果,下行代表输出目标为噪声谱图的结果。

表 1　输出目标对不同模型增强效果的影响

Table 1　The effect of output target on the enhancement effect of different models

模型	PESQ				STOI/%			
	carhorn	engine	knock	wind	carhorn	engine	knock	wind
带噪语音	1.75	2.10	2.63	2.42	70.96	83.78	92.15	85.96
ResUnet	2.57	2.69	2.98	2.69	83.82	86.73	92.67	86.26
	2.63	**2.73**	**3.01**	**2.76**	**84.91**	**87.07**	**92.76**	**87.10**
DResUnet	2.60	2.66	3.02	2.73	85.01	87.32	93.25	86.83
	2.65	**2.80**	**3.10**	**2.85**	**85.70**	**87.77**	**93.83**	**87.70**
A-DResUnet	2.64	2.77	3.10	2.80	84.94	87.89	93.37	87.58
	2.70	**2.82**	**3.12**	**2.86**	**86.60**	**88.30**	**94.10**	**88.24**

由表 1 可知,当模型相同时,用噪声谱图作为输出目标的增强效果要优于用干净语音语谱图作为输出目标的增强效果,其 PESQ 平均提升 2.37%、STOI 平均提升 0.80%。以噪声谱图为目标时相比于 ResUnet 模型,使用融合空洞卷积的 DResUnet 模型增强后 PESQ 平均提升 2.39%、STOI 平均提升 0.89%;使用同时融合空洞卷积以及 CBAM 的 A-DResUnet 模型增强后 PESQ 平均提升 3.30%、STOI 平均提升 1.54%。从结果中发现 STOI 的提升相比于 PESQ 的提升要小,其原因是 STOI 的计算过程中,混合噪声后静音区域的噪声会被直接消除再进行后续的计算,这就导致带噪语音本身的 STOI 分数较高、提升相对较小。

为了进一步比较模型相同、输出目标不同时的增强效果,随机取一段含 carhorn 噪声且信噪比为-5 dB 的语音,分别使用以干净语音语谱图作为输出目标的 A-DResUnet 模型、以噪声谱图作为输出目标的 A-DResUnet 模型对其增强,语谱图结果对比如图 6 所示。可以看出用噪声谱图作为模型输出目标时增强后的语音噪声残留和信息失真更少,因为用干净语音语谱图作为输出目标时模型会学习到说话人的部分特征,当测试语音为未知

说话人以及未知噪声时会将一些特征信息误处理，从而影响增强效果；而用噪声谱图作为输出目标时模型能更好地排除说话人特征的影响、分离未知噪声。

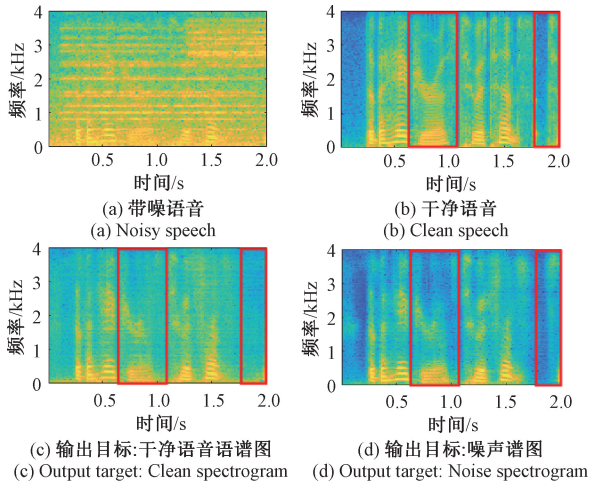


图 6　模型输出目标不同所获得的增强语音语谱图对比

Fig. 6　Comparison of enhanced speech spectrograms obtained with different output targets of model

2.3　不同增强方法的对比

基于 DNN 的语音增强方法[6]、基于 GAN 的语音增强方法[2]分别为时频域和时域中比较经典的方法。为了验证本文方法的优越性，将本文提出的基于 A-DResUnet 语音增强方法与 DNN、GAN 进行对比。DNN、GAN 使用与本文方法相同的训练集、测试集。

图 7 展示了对测试集中 4 类未知噪声、3 种不同信噪比的语音使用不同方法的增强效果比较，结果表明本文方法在各种噪声环境下的增强性能均优于 DNN、GAN。在较平稳的 engine、wind 噪声条件下，GAN 能取得不错的增强效果，而在非平稳的 carhorn、knock 噪声环境下通过 GAN 增强后语音会失真，导致 PESQ、STOI 指标要低于未处理的带噪语音；DNN 和 A-DResUnet 在 4 类噪声环境下均能有效增强语音，且 A-DResUnet 的增强效果要显著优于 DNN。从结果中还发现在 carhorn 噪声环境下，本文提出的方法对信噪比为−5 dB 的语音增强后，其 PESQ 从平均 1.54 提升至平均 2.37；而 DNN、GAN 对信噪比为 5 dB 的语音增强后，其 PESQ 从平均 2.01 分别只提升至平均 2.36 和 2.07，说明本文方法在低信噪比环境下也具有极强的噪声分离能力。

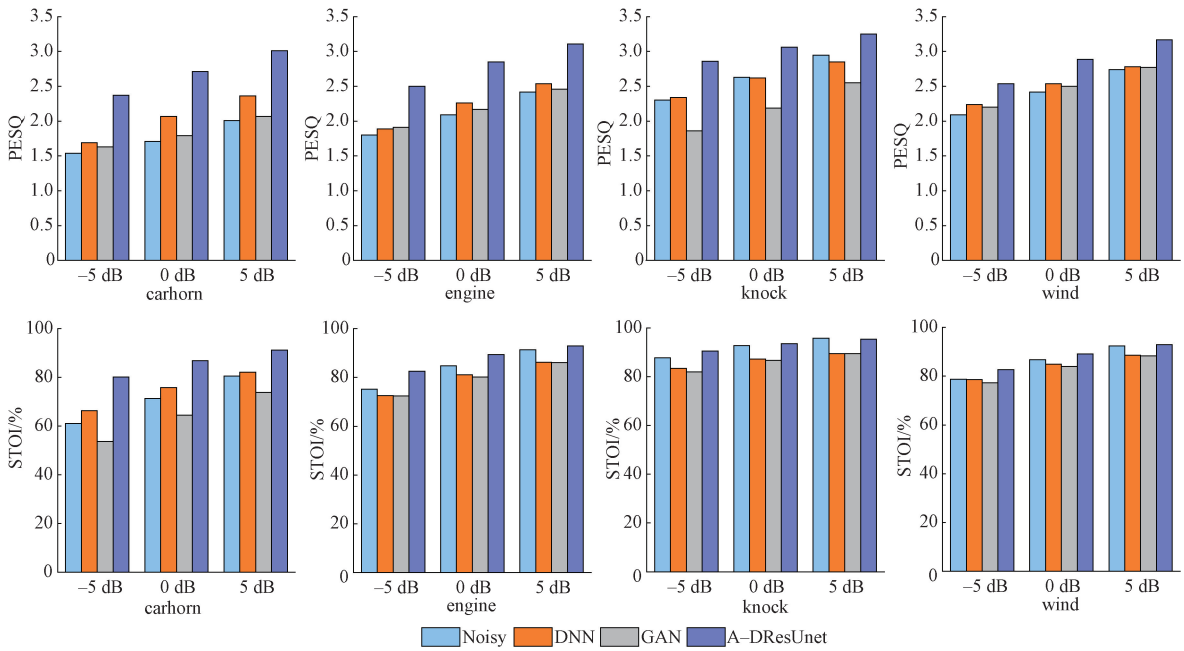


图 7　不同方法对含不同噪声的语音增强效果比较

Fig. 7　Comparison of speech enhancement effects with different methods for different noises

图 8 为一段带噪语音使用不同方法进行增强后的语谱图结果对比，该语音含 carhorn 噪声且信噪比为−5 dB。可以看出 GAN 的增强效果不佳，残留了许多噪声且原语音信息失真严重；DNN 效果稍好，但与 ResUnet 以及 A-DResUnet 相比，在图中红框内 DNN 失真严重且部分区域噪声残留较多；A-DResUnet 相比于 ResUnet 模型，由图中两处红框可以看出 A-DResUnet 模型增强后语音的细节信息损失较少、更接近原始的干净语音。

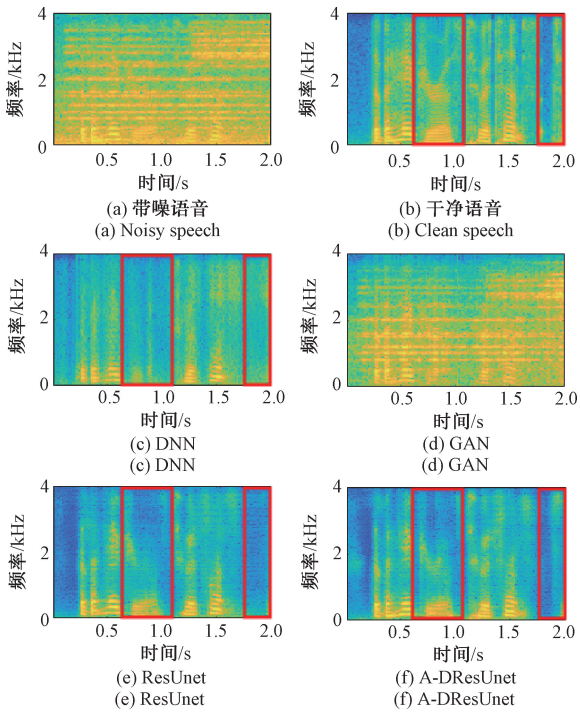


图 8　使用不同方法所获得的增强语音语谱图对比

Fig. 8　Comparison of enhanced speech spectrograms obtained by different methods

3　结　论

本文提出一种基于 A-DResUnet 的语音增强方法，多种噪声条件下的实验结果表明本文方法在复杂环境下具有良好的增强效果，体现在如下 3 点：1）用噪声谱图作为 A-DResUnet 模型的输出目标，加强了该模型对于未知噪声的分离能力、提高了模型的泛化性；2）提出的 A-DResUnet 模型提高了捕获语音上下文信息的能力以及提取噪声谱图特征的精度，从语谱图结果可以看出相比于 ResUnet 模型，增强后噪声残留更少、更好地保留了原语音信息，有效提升了增强效果；3）本文方法在各类噪声环境下的增强效果要显著优于 DNN、GAN，且在 carhorn 噪声环境下，本文方法对信噪比为−5 dB 的语音增强后，其 PESQ、STOI 指标要高于 DNN、GAN 对信噪比为 5 dB 语音的增强结果，进一步说明了本文方法的优越性。

参考文献

[1] 鲍长春，项扬. 基于深度神经网络的单通道语音增强方法回顾[J]. 信号处理，2019，35(12)：1931-1941.
BAO CH CH, XIANG Y. Review of monaural speech enhancement based on deep neural networks[J]. Journal of Signal Processing, 2019, 35(12): 1931-1941.

[2] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: Speech enhancement generative adversarial network[C]. Interspeech, 2017: 3642-3646.

[3] 韩鑫怡，张洪德，柳林，等. 基于 WDGAN-div 的语音增强方法[J]. 电子测量技术，2021，44(21)：64-70.
HAN X Y, ZHANG H D, LIU L, et al. Speech enhancement method based on WDGAN-div [J]. Electronic Measurement Technology, 2021, 44(21): 64-70.

[4] GIRI R, ISIK U, KRISHNASWAMY A. Attention wave-U-Net for speech enhancement [C]. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2019: 249-253.

[5] ZHU Y Y, XU X, YE Z F. FLGCNN: A novel fully convolutional neural network for end-to-end monaural speech enhancement with utterance-based objective functions[J]. Applied Acoustics, 2020, 170: 107511.

[6] 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, and Language Processing, 2014, 23(1): 7-19.

[7] CHEN J, WANG D L. Long short-term memory for speaker generalization in supervised speech separation [J]. The Journal of the Acoustical Society of America, 2017, 141(6): 4705-4714.

[8] TAN K, CHEN J, WANG D L. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 27(1): 189-198.

[9] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical Image Computing and Computer Assisted Intervention, 2015: 234-241.

[10] SHI M Y, GAO J C. Research on high altitude remote sensing building segmentation based on improved U-Net algorithm[J]. Instrumentation, 2021, 8(4): 47-54.

[11] 叶明，李晓丞，刘凯，等. 一种基于 U^2-Net 模型的电阻抗成像方法[J]. 仪器仪表学报，2021，42(2)：235-243.
YE M, LI X CH, LIU K, et al. Image reconstruction method for electrical impedance tomography using U^2-Net [J]. Chinese Journal of Scientific Instrument, 2021, 42(2): 235-243.

[12] 何晓云，许江淳，陈文绪. 基于改进 U-Net 网络的眼底血管图像分割研究[J]. 电子测量与仪器学报，2021，35(10)：202-208.
HE X Y, XU J CH, CHEN W X. Research on fundus blood vessel image segmentation based on improved U-

Net network[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(10): 202-208.

[13] JANSSON A, HUMPHREY E, MONTECCHIO N, et al. Singing voice separation with deep U-Net convolutional networks[C]. Proceedings of the International Society for Music Information Retrieval Conference, 2017: 323-332.

[14] ERNST O, CHAZAN S E, GANNOT S, et al. Speech derverberation using fully convolutional networks[C]. 2018 26th European Signal Processing Conference, 2018: 390-394.

[15] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[16] KONG Q, CAO Y, LIU H, et al. Decoupling magnitude and phase estimation with deep ResUNet for music source separation[C]. Proceedings of the International Society for Music Information Retrieval Conference, 2021: 342-349.

[17] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International Conference on Machine Learning, 2015: 448-456.

[18] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]. International Conference on Machine Learning, 2013: 1152-1160.

[19] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]. International Conference on Learning Representations, 2016: 1-13.

[20] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision, 2018: 3-19.

作者简介

李吉祥，2020 年于湖南涉外经济学院获得学士学位，现为湖南师范大学硕士研究生，主要研究方向为语音分离和语音增强。
E-mail：ljx1544@foxmail.com

Li Jixiang received B. Sc. degree from Hunan International Economics University in 2020. Now he is a M. Sc. candidate in Hunan Normal University. His main research interests include speech separation and speech enhancement.

钱盛友（通信作者），1997 年于上海交通大学获得博士学位，现为湖南师范大学教授、博士生导师，主要研究方向为生物医学电子学和智能仪器。
E-mail：syqian@foxmail.com

Qian Shengyou received Ph. D. degree from Shanghai Jiaotong University in 1997. Now he is a professor and doctoral supervisor in Hunan Normal University. His main research interests include biomedical electronics and intelligent instruments.