DOI:10.19651/j.cnki.emt.2208788

基于 UNet 自适应特征融合的语音增强*

任 健 李鸿燕 张 昱 邢 璐
(太原理工大学信息与计算机学院 榆次 030600)

摘 要: 针对传统的语音增强网络对未知噪声增强效果不理想的问题,本文从语谱图增强,网络结构,特征融合机制 3 方面提出改进方法。首先为了提取语谱图深层特征信息,使用 VGG19 结构来代替 UNet 结构中编码器部分,同时在解码器部分加入残差网络以加深网络深度,防止训练退化;其次,为了更好地结合语谱图中特征信息,在 UNet 结构跳跃连接部分加入自适应特征融合机制来融合深浅层特征。此外,为增强说话人信息,通过直方图均衡算法对语谱图进行特征优化,得到直方图均衡化增强后的语谱图。在不同的噪声环境中,本文所提方法在质量和可理解性度量方面评分都优于其他增强方法。

关键词: 语音增强;卷积神经网络;自适应特征融合;VGG19;直方图均衡化增强
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 UNet adaptive feature fusion

Ren Jian Li Hongyan Zhang Yu Xing Lu
(College of Information and Computer Science, Taiyuan University of Technology, Yuci 030600, China)

Abstract: Aiming at the problem that the traditional speech enhancement network is not ideal for unknown noise enhancement, this paper proposes an improved method from the aspects of spectral enhancement, network structure and feature fusion mechanism. Firstly, in order to extract the deep feature information of the spectrum, VGG19 structure was used to replace the encoder part of UNet structure, and residual network was added to the decoder part to deepen the network depth and prevent the training degradation. Secondly, in order to better combine the feature information in the spectrogram, an adaptive feature fusion mechanism is added to the jump connection part of the UNet structure to fuse the deep and shallow features. In addition, in order to enhance the speaker information, the histogram equalization algorithm is used to optimize the feature of the spectrogram, and the histogram equalization enhancement spectrogram is obtained. In different noise environments, the proposed method outperforms other enhancement methods in terms of quality and comprehensibility.

Keywords: speech enhancement;convolutional neural networks;adaptive feature fusion;VGG19;histogram equalization is enhanced

0 引 言

语音增强算法作为语音信号处理中非常重要的一部分,多年来吸引了大量学者继续研究。语音增强算法包含了信号检测[1]、波形重筑等信号处理理论,而且与语音来源的本身特性、人耳对其感知特性[2-3]等生理学紧密相关。

过去传统的语音增强方法多是针对平稳噪声设计,其中最著名和最有效的方法为最优改进对数谱幅度估计(OMLSA)。但是这类方法在外界环境中比较容易受到非平稳噪声的影响。近年来基于深度神经网络(deep neural network,DNN)的语音增强快速发展,受到学者的广泛关注。此类方法在非平稳噪声环境下语音增强性能[4-6]更好。徐勇等人以纯净语音的对数功率谱为训练目标,通过训练网络构造带噪语音对数功率谱[7](logarithmic power spectra, LPS)与纯净语音对数功率谱之间的映射函数,从而提升在非平稳噪声环境中语音增强效果。

DNN 存在大量的参数非常不便于训练。研究发现,卷积神经网络[8-10](convolutional neural-networks, CNN)比等效的 DNN 产生更少的参数。这使得 CNN 在低资源的情况下具有吸引力,如助听器中的语音增强[11]和医学图像处理[12]。另一方面,CNN 固有的特征提取特性鼓励了研

收稿日期:2022-01-06
* 基金项目:山西省自然科学基金(201701D121058)、山西省回国留学科研项目(2020-042)资助

究人员实现时域和波形端到端语音增强网络[13],但是单一的 CNN 结构语音增强效果并不理想。在此基础上,本文提出了一种基于 UNet[14] 的新结构,(VGG19-residual-adaptive f-eature fusion-unet, VGG19-RAFFUNet)。首先深度全卷积网络 VG-G19 嵌入在 UNet 编码器-解码器网络架构的编码器部分;其次在跳跃连接[15]之间加入自适应特征融合[16](adaptive feature fusion, AFF)融合有效特征信息,最后在解码器部分加入残差网络[17]以增加网络深度。

1　语音增强算法

VGG19-RAFFUNet 进行语音增强任务分为训练和增强阶段,语音增强框图如图 1 所示。

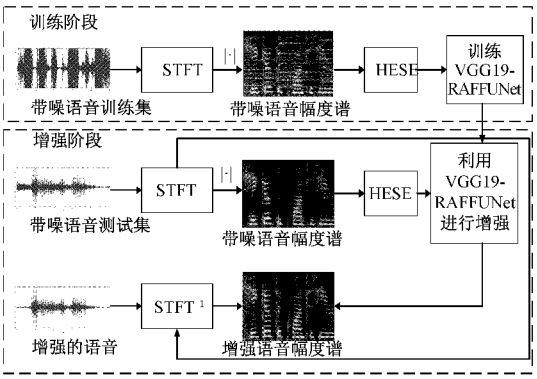


图 1　语音增强框图

在训练阶段,首先进行短时傅里叶变换(short time Fourier transform, STFT),得到带噪语音的语谱图,对语谱图进行直方图均衡化增强(histogram equalization spectrogram enhancement, HESE),将增强后的语谱图作为 VGG19-RAFFUNet 的输入,对 VGG19-RA-FFUNet 进行训练。

在增强阶段,将带噪语音语谱图最为网络结构的输入,输出估计的纯净语音谱,使用带噪语音的相位信息进行短时傅里叶逆变换 $STFT^{-1}$ 最后输出增强的语音信号。

1.1　语谱图增强

本文提出将语谱图使用直方图均衡化算法(HESE)进行增强,通过映射函数,将输入的语谱图不均匀的像素点转化为在整个灰度区间呈现均匀分布的像素点,再计算每个像素点上的新像素实现图像增强,将灰度动态范围扩展,算法如下:

1)首先根据原图像的灰度计算灰度概率密度函数$P(r_k)$。

$$P(r_k)=N_k/N,k=0,1,2,\cdots,L-1 \tag{1}$$

式中:N_k 是第 k 个灰度级出现总次数;N 是灰度图像素总数;L 是灰度级数量。

2)由 $P(r_k)$计算累计分布函数 S_k,并将 S_k 归一到[0,255]。

经过算法增强的语谱图更能显示出说话人信息,如图 2 为原始语谱图,图 3 为增强后的语谱图。

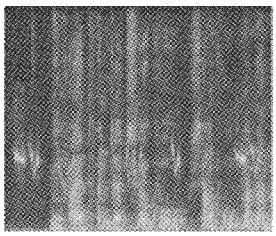


图 2　原始语谱图

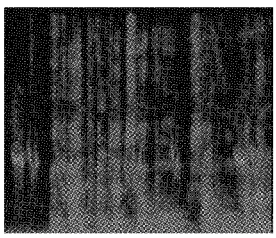


图 3　增强后的语谱图

每个语谱图都是 256×256 的图像。经过语谱图增强,提供图像数据集,包括噪声信号的(HESE)增强语谱图和相应的干净语音信号的图像数据集。

1.2　语谱图特征提取

语谱图能够结合语音的时域和频域特性进行分析,可以看出语音信号的频谱在时间序列上的变化情况。语谱图的纵轴和横轴分别是频率序列和时间序列,颜色的深浅代表具有的能量。分析语谱图可以获得大量语音的语义信息,可以使用 CNN 来提取增强后的语谱图中相关的特征信息。本文采用 VGG19 结构来代替经典 UNet 网络编码器部分,它的组成有 3 部分:输入层、卷积层和池化层以及全连接层。卷积层输出特征图,全连接层将这些特征图中的特征,得到深层特征的含义。VGG19 特征提取过程如图 4 所示,输入是 256×256 的语谱图,卷积部分采用大小为(3×3)卷积核,最大池化尺寸为(2×2);卷积层 N_1,卷积核个数为 64,将卷积后输出的特征图填充,保持尺寸不变,接着连接 ReLU 函数引入非线性单元。在 N_1 后连接池化层 M_1 进行最大池化,得到压缩特征图,可以减少运算并筛选出有效特征,输出尺寸 128×128 的特征图。第 2 个卷积层 N_2,由 128 个卷积核组成,通过卷积输出 128 个 128×128 的特征图,连接 ReLU 激活函数。在 N_2 后连接池化层 M_2 进行最大池化,输出特征图大小为 64×64,经过卷积层 N_3 和池化层 M_3,输出 32×32 的特征图,再经过两次卷积和池化得到 512 个 8×8 的特征图,最后全连接层将特征图映射成长度不变的特征向量。

2　网络模型

UNet 是一个经典的全卷积网络,用于精准快速的图像到图像的转换任务中,比如图像分割和图像降噪。有些研究中,通过使用 VGG 代替前一代的模型,已经证明了网络深度有利于提高系统的性能。在此基础上,本文对经典

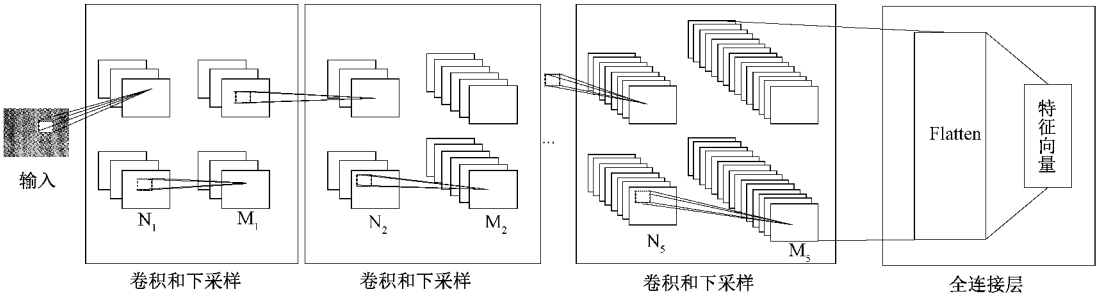


图 4　VGG19 特征提取图

UNet 做 3 点优化，提出一种新的用于语音增强任务的编解码器式网络。1)编码器部分，使用 VGG19 的 5 个卷积块作为一个强大的深度特征提取器，先将每层下采样得到的浅层特征通过跳跃连接部分，然后通过 5 次下采样得到深层特征传递到解码器部分。2)在解码器中加入残差网络，加深网络深度，防止训练退化现象发生，应用了一个 2×2 上采样和两个具有非线性的 3×3CNN 层以及残差结构，然后将上一步重复 5 次。与编码器路径相比，每个序列中的信道数减半，最终达到 32 个，通过一个具有线性激活函数的 1×1CNN 转换为频谱图像。3)在跳跃连接中加入自适应特征融合(AFF)融合 VGG19 每层下采样部分得到的浅层特征和经过下采样和上采样得到的深层特征。VGG19-RAFFUNet 结构如图 5 所示。

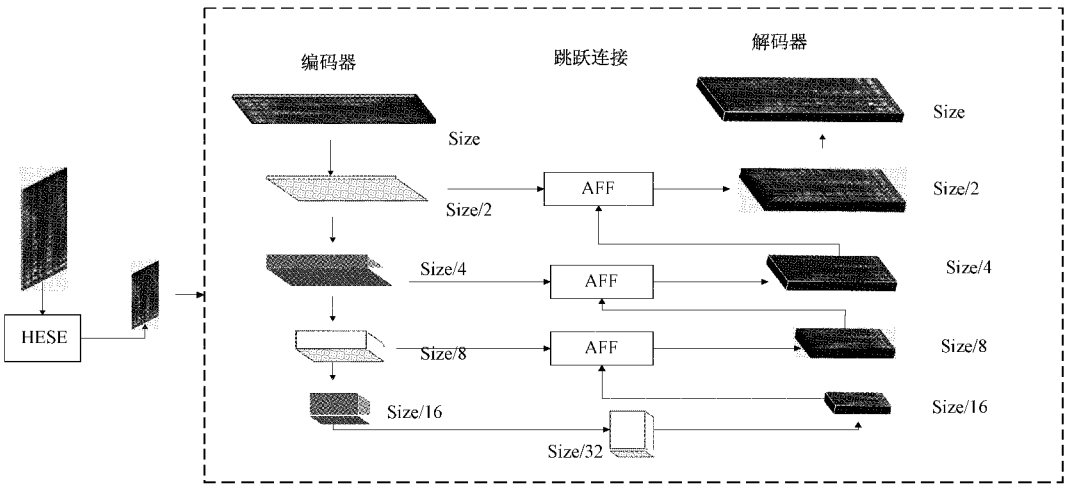


图 5　VGG19-RAFFUNet 结构

2.1　解码器

网络深度增加，会出现训练退化，准确率也会趋于平缓，但误差会增加，为了防止出现这个问题，在 AFF-UNet 中加入了残差(Residual)结构，得到 RAFFUNet 结构。残差结构有一定的正则化作用，有助于深层模型的训练，提升容错率。

2.2　基于 AFF 的跳跃连接

在 UNet 结构跳跃连接部分加入 AFF，对输入的浅层特征图和深层特征图赋予不同的权重，进行加权融合语音信号的位置信息和语义信息。AFF 网络结构如图 6 所示。

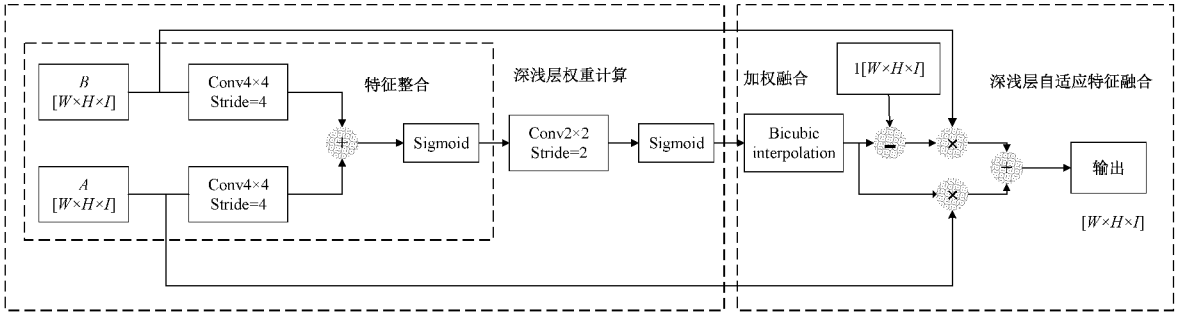


图 6　AFF 的网络结构

AFF-UNet 与标准的 UNet 相比,整体结构相似,不同的是增加了自适应特征融合。创建权值网络,通过训练该网络,计算出深浅层特征图的权重,对不同卷积层进行自适应分配。图 7 表示两种融合对比,图 7(a)为本文提出的 AFF 方法,图 7(b)为普通跳跃连接的融合方法。

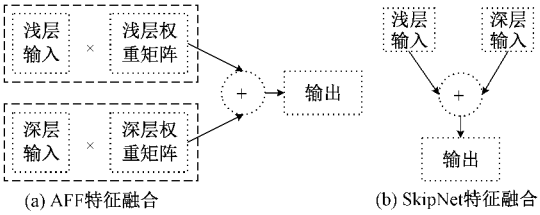


图 7　深浅层特征融合与 SkipNet 叠加融合

1)权重矩阵计算

两个卷积层分别有 I 个大小为 $W\times H$ 的特征图,将其作为权重计算层的输入,融合相对应的两个特征图信息,假设输入矩阵分别为 $\boldsymbol{I}_a$ 和 $\boldsymbol{I}_b$,大小是 $p\times q$,$\boldsymbol{o}$ 表示输出矩阵,则整合过程公式表达为:

$$\begin{cases}\boldsymbol{K}_1=\boldsymbol{I}_a*\omega_1+c_1\\ \boldsymbol{K}_2=\boldsymbol{I}_b*\omega_2+c_2\\ \boldsymbol{o}=\alpha(\boldsymbol{K}_1+\boldsymbol{K}_2)\end{cases}\tag{2}$$

式中:ω_1,c_1 与 ω_2,c_2 分别表示相对于输入 $\boldsymbol{I}_a$ 与 $\boldsymbol{I}_b$ 的权重与偏置,$*$ 表示卷积操作。

采用的激活函数 sigmoid 计算公式如下:

$$\alpha(x)=1/(1+\mathrm{e}^{-x})\tag{3}$$

2)深浅层特征加权融合

权值矩阵的维度一般小于特征图,因此在加权时要提升维度。本文采用双三次插值(bicubic in-terpolation)对维度进行放大,该方法效果最好,但速度最慢。假设输入矩阵分别为 $\boldsymbol{I}_a$ 和 $\boldsymbol{I}_b$,大小是 $p\times q$,$\boldsymbol{o}$ 表示输出矩阵,深浅层特征加权融合计算公式为:

$$\boldsymbol{o}=\boldsymbol{\delta}\circ\boldsymbol{I}_a+(\boldsymbol{I}_{p\times q}-\boldsymbol{\delta})\circ\boldsymbol{I}_b\tag{4}$$

式中:$\circ$表示哈达玛积,即矩阵对应元素相乘;$\boldsymbol{\delta}$ 为维度增加后权重矩阵,大小是 $p\times q$;$\boldsymbol{I}_{p\times q}$ 表示一个常数矩阵,元素都为 1,大小是 $p\times q$。

3　仿真实验与结果分析

3.1　数据集

首先,建立一个由带噪语音和相应纯语音组成的训练集。干净语音来自 TIMIT 库的训练集,噪声数据来自 100 类真实噪声。合成信噪比为−10、−5、0、5 和 10 dB 的 25 000 条训练数据;接着,使用 TIMIT 库测试集中的纯语音数据和 Noisex92 库中的噪声数据创建测试集。第一步,从噪声库中选择和训练集噪声不一样的 4 类未知噪声,即 Factory2、Buccaneer1、Dest-royer engine 和 HF channel;然后合成信噪比为−9、−4.5、0、4.5、9、13.5 dB 的 4 800 条测试数据。

3.2　学习 VGG19-RAFFUNet 架构

基于 tensorflow 和 keras 深度学习框架学习 VGG19-RAFFUNet 网络。batch size 设置为 20,在每轮训练中,从训练谱图中随机选择 20 幅,输入网络来进行学习。在训练过程中,平均绝对误差(MAE)作为代价函数。最后选择产生最小 MAE 的学习模型来生成增强的语音信号。本文采用 MAE 作为客观评价指标,计算公式如下:

$$\mathrm{MAE}(a,b)=\frac{1}{n}\left(\sum_{i=1}^{n}|a-b|\right)\tag{5}$$

式中:a 是真实值 b 是预测值,n 代表个数,MAE 的结果越小,代表模型更精确。

3.3　评价指标

评价指标采用语音质量感知评价(perceptual-evaluation of speech quality, PESQ)和短时客观可懂度(short-time objective intelligibility, STOI)。

PESQ(语音质量的感知评估)是一种客观的、全参考的语音质量评估方法,通过带噪信号和原始信号,PESQ 算法能够对客观语音质量评估提供一个主观 MOS 的预测值,而且可以映射到 MOS 刻度范围,计算过程包含预处理、时间对齐、感知滤波、掩蔽效果等,PESQ 范围在 [−0.5～4.5]之间。得分越高表示语音质量越好。

STOI 是衡量语音可懂度的重要指标之一,它的计算相对简单,主要包括 5 个步骤:1)移除静音区,因为静音区没有内容;2)STFT 变换,得到与人耳听觉系统相似的时域特征;3)1/3 倍频分析,划分时频点,得到短时谱向量;4)归一化和裁剪,通过归一化补偿差异性,再经过裁剪约束损坏语音的敏感度;5)可懂度计算,计算待测试语音和干净语音之间短时谱向量的相关系数。最后通过对所有帧频带的相关系数求均值得到 STOI 的值,范围是[0,1],它的值越高说明语音增强效果越好。

3.4　对比实验

为了验证本文提出的语音增强方法效果,选择如下方法进行对比实验。

1) OMLSA:采用传统的基于统计的增强器,利用统计模型中 log-MMSE 的估计器。

2) UNet:将传统 UNet 结构在线性频率域和对数幅度域进行训练。

3) UNet-HESE:将经过(HESE)增强的语谱图输入 UNet 中进行训练。

4) RAFFUNet-HESE:将自适应特征融合机制和残差网络加入 UNet 结构中并输入经过(HESE)增强的语谱图进行训练。

5) VGG19-RAFFUNet-HESE:将 UNet 编码器部分用 VGG19 代替并在跳跃连接中加入自适应特征融合机制以及解码器部分加入残差网络,然后输入经过(HESE)增强的语谱图进行训练。

表 1 和 2 分别是不同算法的 PESQ 和 STOI 对比结

果。表中计算了不同信噪比水平的平均性能，从非常低的　　信噪比(−9 dB)到较高的信噪比(13.5 dB)。

表 1　不同算法的 PESQ 对比结果

方法	信噪比/dB						
	−9	−4.5	0	4.5	9	13.5	平均
带噪语音	1.03	1.06	1.13	1.17	1.23	1.57	1.20
OMLSA	1.20	1.27	1.55	1.80	2.19	2.57	1.77
UNet	1.33	1.42	1.78	2.17	2.69	2.93	2.05
UNet-HESE	1.56	1.71	2.29	2.46	2.80	3.12	2.38
RAFFUNet-HESE	1.63	2.13	2.16	2.50	2.98	3.51	2.48
VGG19-RAFFUNet-HESE	**1.75**	**2.15**	**2.32**	**2.72**	**3.14**	**3.65**	**2.61**

表 2　不同算法的 STOI 对比结果

方法	信噪比/dB						
	−9	−4.5	0	4.5	9	13.5	平均
带噪语音	0.31	0.42	0.53	0.67	0.79	0.85	0.60
OMLSA	0.32	0.45	0.58	0.71	0.82	0.88	0.63
UNet	0.50	0.57	0.71	0.84	0.89	0.91	0.73
UNet-HESE	0.46	0.61	0.74	0.86	0.91	0.92	0.75
RAFFUNet-HESE	0.47	0.63	**0.84**	0.89	0.92	0.93	0.78
VGG19-RAFFUNet-HESE	**0.51**	**0.67**	0.80	**0.93**	**0.93**	**0.94**	**0.80**

这几种深度神经网络显著优于基于统计的增强器，OMLSA 和深度学习方法在信噪比水平较高的时候能获得更好的语音质量，在低信噪比情况下语音可懂度相对得分较高。

比较 UNet 和 UNet-HESE 方法，第 2 种方法基本在所有的信噪比下效果都更好。结果表明，与输入未作增强的语谱图训练相比，输入经过 HESE 优化的语谱图进行训练，语音增强效果更好。

从实验结果可以看出，本文提出的方法在不同信噪比和未知噪声情况下都显示出最佳的性能。平均 PESQ 提升到 2.61，STOI 均值也有所提升。

3.5　不同噪声环境下实验对比

图 8 和 9 是在测试集中不同噪声环境下，采用不同方法的平均 PESQ 和 STOI 统计图。从实验结果可以看出，未知噪声情况下，本文提出的基于 VGG19-RA-FFUNet-HESE 网络增强被证明是最优的。

3.6　语谱图分析

为了更直观比较本文提出的方法与其他方法的语音增强性能，使用之前提到的其中 3 种语音增强方法：OMLSA 方法、基于 RAFFUNct-HESE 的语音增强方法和本文提出的基于 VGG19-RAFFUNet-HESE 的语音增强方法，对一段含有 Buccaneer1 噪声，信噪比为−6 dB 的含噪语音增强。

图 10 是在信噪比为−6 dB 下各方法增强后的语谱图：包含图 10(a)干净语音语谱图、图 10(b)带噪语音语谱

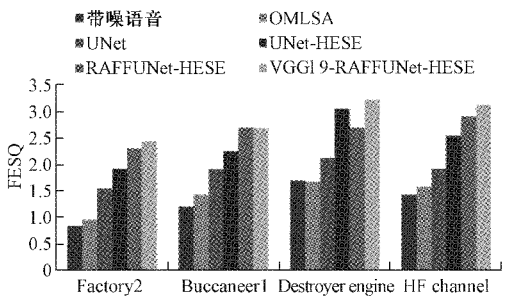


图 8　各方法不同噪声环境下 PESQ 均值

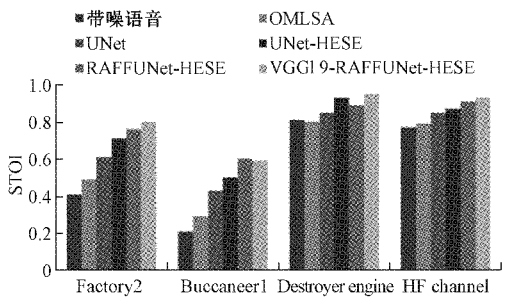
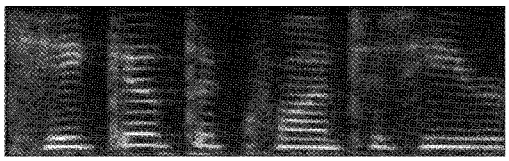
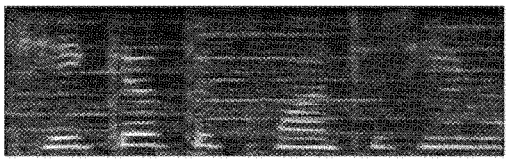


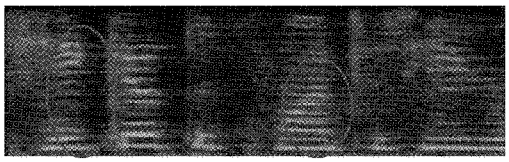
图 9　各方法不同噪声环境下 STOI 均值

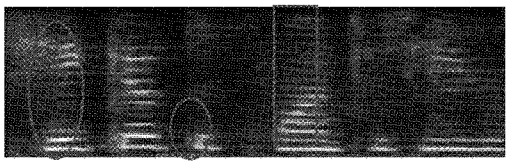
图、图 10(c) OMLSA 方法增强语音语谱图、图 10(d) RAFFUNet-HESE 方法语音增强语谱图和图 10(e) VGG19-RAFFUNet-HESE 方法语音增强语谱图。

相比图 10(a)的纯净语音，图 10(c)OMLSA 的语音增强方法虽然减少了噪声残留，但会丢失一些语义信息，如


(a) 干净语音语谱图


(b) 带噪语言语谱图


(c) OMLSA方法增强语言语谱图


(d) RAFFUNet-HESE方法语音增强语谱图

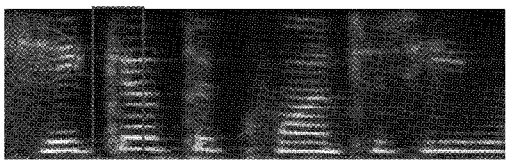

(e) VGG19-RAFFUNet-HESE-E方法语音增强语谱图

图 10　不同方法增强后的语谱图

图 10(c)中椭圆形区域,会降低可懂度得分。图 10(d)可以看出 RAFFUNet-HESE 语音增强的方法,存在如图 10(d)中椭圆形区域相比图 10(c)中有了一定改善,但矩形部分与干净语音对比缺少明显谐波结构。图 10(e)是本文提出的 VGG19-RAFFUNet-HES-E 方法,从椭圆形区域可以看出谐波可以清晰显示,有明显的横纹,说明这种方法能更好降低噪声影响。

4　结　　论

本文提出基于 VGG19-RAFFUNet-HESE 的语音增强网络以更好地恢复语音信息,首先将语谱图进行直方图均衡化增强(HESE),输入到深度特征提取器-VGG19 中进行特征提取,其次在网络结构中加入残差网络和自适应特征融合机制能够更有效训练网络和选择有效特征,得到干净的语音谱,再逆变换输出纯净语音。实验结果表明,在未知噪声情况下,该算法能够减小增强语音的失真现象,可以获得质量和可懂度更好的语音,增强效果明显。

参考文献

[1] 王生霄,侯兴松,黑夏萌. 嵌入 CBAM 结构的改进 YOLOV3 超宽带雷达生命信号检测算法[J]. 国外电子测量技术,2020,39(3):1-6.

[2] 房慧保,马建芬,田玉玲,等. 基于感知相关代价函数的深度学习语音增强[J]. 计算机工程与设计,2020,41(11):3212-3217.

[3] 陆生礼,时龙兴,余崇智,等. 听觉模拟的语音增强方法[J]. 声学学报,1996(6):879-883.

[4] 徐勇. 基于深层神经网络的语音增强方法研究[D]. 合肥:中国科学技术大学,2015.

[5] 刘文举,聂帅,梁山,等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报,2016,42(6):819-833.

[6] 张晓艳,张天骐,葛宛营,等. 联合深度神经网络和凸优化的单通道语音增强算法[J]. 声学学报,2021,46(3):471-480.

[7] 龙华,张林濮,邵玉斌,等. 说话人特征约束的多任务卷积网络语音增强[J]. 小型微型计算机系统,2021,42(10):2178-2183.

[8] ZHAO Z P,LI Q F,ZHANG Z X, et al. Combining a parallel 2D CNN with a self-attention dilated residual network for CTC based discrete speech emotion recognit-ion[J]. Neural Networks,2021,141:52-60.

[9] 陈思佳,罗志增. 基于长短时记忆和卷积神经网络的手势肌电识别研究[J]. 仪器仪表学报,2021,42(2):162-170.

[10] 王志杰,张学良. 基于双路径循环神经网络的单通道语音增强[J]. 信号处理,2021,37(10):1872-1879.

[11] 朱亚涛,陈霏,张雨晨,等. 基于循环神经网络的双耳助听器语音增强算法[J]. 传感技术学报,2021,34(9):1165-1172.

[12] KUMAR S,BHANDARI A K,RAJ A, et al. Triple clipped histogram based medical image enhancement using spatial frequency[J]. IEEE Transactions on Nanboscience,2021:278-286.

[13] 蓝天,彭川,李森,等. 基于 RefineNet 的端到端语音增强方法[J/OL]. 自动化学报:1-10[2021-09-17].

[14] 李梅梅,胡春海,龙平,等. 基于 MultiRes+UNet 网络的车道线检测算法[J]. 电子测量与仪器学报,2020,34(9):117-122.

[15] ZHANG L,ZHANG J M,SHEN P Y, et al. Block level skip connections across cascaded V-Net for multi organ segmentation[J]. IEEE Transactions on Medical Imaging,2020,39(9):2782-2793.

[16] 黄庭鸿,聂卓赟,王庆国,等. 基于区块自适应特征融合的图像实时语义分割[J]. 自动化学报,2021,47(5):1137-1148.

[17] 谢小红,李文韬,孙晓燕. 深度残差网络综述[J]. 信息与电脑(理论版),2021,33(16):85-87.

作者简介

任健,硕士研究生,主要研究方向为语音增强。
E-mail:tylgrj@163.com

李鸿燕(通信作者),博士,教授,主要研究方向为信号与信息处理。
E-mail:tylihy@163.com

张昱,硕士研究生,主要研究方向为语音识别。
E-mail:871589401@qq.com

邢璐,硕士研究生,主要研究方向为语音增强。
E-mail:1024530967@qq.com