文章编号：1003-0530(2019)04-0631-10

利用深度全卷积编解码网络的单通道语音增强

时文华[1,2] 张雄伟[1] 邹 霞[1] 孙 蒙[1]
（1. 陆军工程大学，江苏南京 210007；2. 空军航空大学，吉林长春 130000）

摘 要：针对传统的神经网络未能对时频域的相关性充分利用的问题，提出了一种利用深度全卷积编解码神经网络的单通道语音增强方法。在编码端，通过卷积层的卷积操作对带噪语音的时频表示逐级提取特征，在得到目标语音高级特征表示的同时逐层抑制背景噪声。解码端和编码端在结构上对称，在解码端，对编码端获得的高级特征表示进行反卷积、上采样操作，逐层恢复目标语音。跳跃连接可以很好地解决极深网络中训练时存在的梯度弥散问题，本文在编解码端的对应层之间引入跳跃连接，将编码端特征图信息传递到对应的解码端，有利于更好地恢复目标语音的细节特征。对特征融合和特征拼接两种跳跃连接方式、L_1 和 L_2 两种训练损失函数对语音增强性能的影响进行了研究，通过实验验证所提方法的有效性。

关键词：语音增强；跳跃连接；编解码；卷积神经网络
中图分类号：TN912.3 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2019.04.014

引用格式：时文华，张雄伟，邹霞，等. 利用深度全卷积编解码网络的单通道语音增强[J]. 信号处理，2019，35(4)：631-640. DOI：10.16798/j.issn.1003-0530.2019.04.014.
Reference format：Shi Wenhua，Zhang Xiongwei，Zou Xia，et al. Single Channel Speech Enhancement Based on Deep Fully Convolutional Encoder-Decoder Neural Network[J]. Journal of Signal Processing，2019，35(4)：631-640. DOI：10.16798/j.issn.1003-0530.2019.04.014.

Single Channel Speech Enhancement Based on Deep Fully Convolutional Encoder-Decoder Neural Network

Shi Wenhua[1,2] Zhang Xiongwei[1] Zou Xia[1] Sun Meng[1]
（1. Army Engineering University，Nanjing，Jiangsu 210007，China；2. Air Force Aviation University，Changchun，Jilin 130000，China）

Abstract：Considering the time frequency correlation characteristics of speech is not well utilized in the conventional deep neural network，a single channel speech enhancement method based on deep encoder-decoder neural network is proposed. At the coding end，the time-frequency representation of noisy speech is extracted step by step through convolution and pooling operations of convolution layer to obtain high level feature representation of the target speech. At the same time，the background noise is suppressed. The decoder and the encoder are symmetrical in structure，and the target speech features are reconstructed from the advanced feature representation obtained in the encoder step through de-convolution and up-sampling operations at decoding end. Skip connections are employed to solve the gradient dispersion problem in very deep neural networks. In this paper，low level feature maps which include the detail information of speech are delivered by skip connections from the coding end to the corresponding decoding end feature map in the decoding end. This will help the decoder recover the detailed features of the target speech better. The network is trained in two ways with L_1 loss and L_2 loss，the performance of two forms of connections，feature fusion and feature concatenation are evaluated in the experiments. The

收稿日期：2019-01-24；修回日期：2019-03-09
基金项目：国家自然科学基金项目(61471394)；江苏省优秀青年基金(BK20180080)

results demonstrate the effectiveness of proposed method.

Key words: speech enhancement; skip connection; encoder-decoder; convolutional network

1 引言

从被噪声污染的信号中恢复出目标语音,提高语音的质量和可懂度是语音增强的主要目标。语音增强技术被广泛应用在移动通信、语音识别、听觉辅助设备中,是语音信号处理领域的一个重要分支。基于信号和噪声的分布特性,众多语音增强方法被相继提出,代表性的有谱减法、维纳滤波法、统计模型估计[1]等。这些方法在平稳噪声环境下取得了较好的效果。针对传统方法的局限性,尽管一些新的方法被学者们相继提出[2-3],然而在非平稳噪声和低信噪比环境下的单通道语音增强一直是语音信号处理的一个难点。

近些年来,得益于计算机处理能力的提高和机器学习算法的发展,神经网络突破了早期浅层网络在训练数据和网络规模的限制,在计算机视觉、语音信号处理等多个领域得到了迅速发展。在语音增强方面,作为数据驱动的方法,基于深度学习的方法直接由数据驱动,不需要对信号和噪声的分布做先验假设,在非平稳噪声处理中显示出极大的优势。Xu 等[4]提出了基于深度神经网络(Deep Neural Network,DNN)的自回归语音增强方法,实现带噪语音的对数能量谱到目标语音对数能量谱的直接映射。Wang 等人[5]提出的利用神经网络进行掩蔽估计的方法,将带噪语音的多种声学特征拼接成一个长向量作为网络的输入,听觉 Gammatone 域的理想浮值掩蔽作为网络的目标输出,将语音增强问题转化为时频单元的分类问题。Huang 等人[6]利用深度循环网络(Deep Recurrent Neural Network, DRNN)的动态时序建模能力,提出基于时频掩蔽估计和 DRNN 联合优化的语噪分离方法。传统的基于前馈神经网络(Forward Neural Network,FNN)的增强方法为了充分利用语音的上下文信息,一般是将相邻帧的时频特征拼接成一个长向量作为网络的输入,原时频结构中相邻的时频单元将会位于长向量的不同位置,相对位置的改变会增大网络对语音中的相关性结构进行建模的难度,深度循环神经网络虽然利用循环连接或者存储和门结构单元对语音信号的长短时序相关性进行建模,但同样对时频域的相关性未能很好的利用。卷积神经网络(Conventional Neural Network,CNN)受生物视觉皮层的感知机理的启发,其局部感知和权值共享特性可以对信号局部相关信息进行建模,同时能够极大减少网络模型的参数数量,在计算机视觉和语音识别领域得到了广泛的应用。

传统 CNN 网络在卷积层之后会接上若干个全连接层,将卷积层产生的特征图映射成一个一维的长向量,从而丢失了相邻时间上的相关信息。全卷积网络(Fully Convolutional Network,FCN)将传统卷积网络后面的全连接层换成了卷积层,通过反卷积使输出可以保持和输入相同的大小,保留了原始输入中的结构信息[7]。文献[8]对经典的全卷积网络进行改进,提出了一种基于 U-Net 的全卷积编解码网络结构,通过在上采样层对特征图进行拼接增加特征图数量,将上下文信息传递到更高分辨率层,在细胞边缘检测竞赛中取得了最优的结果。文献[9]提出将全卷积编解码框架用于图像复原,编码层和解码层结构对称,并借鉴了 highway networks 和残差网络的思想,在编解码对应层之间引入跳跃连接(skip connections),将卷积层的特征图信息传递到对应的反卷积层,用于恢复原始图像。

语音信号在经过短时傅里叶变换后(Short Time Fourier Transform,STFT)后的语谱图在时间和频率两个维度上有着和图像类似的二维表示。语音信号浊音段的时域波形呈现出的准周期特性和短时功率谱具有共振峰结构,表明语音信号在时域和频域上存在着局部相关性,因此可以利用 CNN 对语音信号的二维时频表示来建模。但过去几年将 CNN 用于语音增强的研究还比较少。其中,文献[10]提出了 memory efficient 的冗余卷积编解码增强方法,可在参数量减少 68 倍和 12 倍的情况下得到与 FNN 和 RNN 接近甚至较好的增强效果。文献[11]将用于图像风格转换的生成对抗网络结构[12]直接应用于语音增强,将深度编解码结构应用在生成器中,

但全卷积编解码网络在语音增强中的应用研究的还比较缺乏。

受文献[9]工作的启发，本文将深度全卷积编解码框架应用在语音增强中。在网络中引入跳跃连接，可以有效解决深度网络训练中出现的梯度消失问题，本文对两种典型 skip connections 的形式，即特征融合相加和特征拼接对语音增强效果的影响展开了研究。

损失函数是网络的重要组成部分。目前大部分基于神经网络的谱映射语音增强算法都是基于最小均方误差(Minimum Mean Square Error，MMSE)准则，通过反向传播误差更新网络参数。经过 DNN 方法增强后的语音特别是在高频部分会存在噪声冗余，出现过平滑的现象。文献[3]采用全局均衡方差方法来解决经 DNN 增强后语音频谱出现的过平滑问题。近些年来在图像处理任务中，研究者发现基于 L_1 范数的损失函数要优于 L_2 范数因为它引入较小的模糊。文献[11]将这一用于图像处理的结论直接用在 GAN 的生成器中用于语音增强。为了更好地验证结论在语音增强方面的有效性，文献[13]对比了基于 L_1 范数和 L_2 范数的损失函数在传统 FCN 网络中和生成对抗网络中对语音增强性能的影响。本文我们将这一研究扩展到深度全卷积编解码网络中，研究两种不同的跳跃连接方式下，不同损失函数对语音增强性能的影响。

本文内容安排如下：第 2 节描述深度卷积编解码框架，第 3 节是基于深度卷积编解码网络的语音增强，仿真实验和结果分析在第 4 节给出，第 5 节总结全文。

2　深度全卷积编解码网络

2.1　卷积编解码网络

编解码网络是一种灵活的框架模型，由编码器(端、层)和解码器(端、层)组成，可以根据需要灵活地应用到无监督或者有监督任务中，通常用于网络预训练、高维原始数据降维、特征提取及数据压缩、生成等。编码器和解码器可以根据不同任务选取 FCN、CNN、RNN 等不同的网络模型。本文利用 CNN 的局部感知和权值共享特性，将深度全卷积编解码网络用于语音增强。一方面，语音信号浊音段的时域波形具有相似性，呈现出准周期特性，经过 STFT 后的短时功率谱具有共振峰结构，表明语音信号在时域和频域上存在着局部相关性[14]。传统神经网络将相邻帧的时频特征拼接成一个长向量作为网络的输入，相对位置的改变会增大网络对语音中的相关性结构进行建模的难度。同时人耳在嘈杂的背景噪声中选择目标语音时，能对任意声音极快地做出反应，并不需要关注过去的其他人声或噪音，降噪本身应当是一个局部化的处理，这一点和语音识别、机器翻译等任务是不同的。而卷积神经网络具有局部感知特性，网络中的每个神经元只需要和前一层的部分神经元相连，同时，卷积神经网络的权值共享特性使每个滤波器特征图权重相同，这样会使网络的参数量大量减少，提升模型的鲁棒性。

编码端由多个卷积层组成，每个卷积层包括卷积滤波、批量标准化、池化、非线性变换操作组成。解码端和编码端结构对应，由反卷积、上采样、批量标准化、非线性变换操作组成。将该结构应用到语音增强，将带噪语音的时频特征作为网络的输入，在编码端利用卷积网络对带噪语谱中的局部典型性结构进行建模，提取高层语音特征，同时逐层抑制背景噪声的影响。在解码端，通过反卷积层由编码端提取到的高层语音特征信息逐层恢复语音细节成分，重构语音信号。

2.2　跳跃连接

在神经网络中引入跳跃连接，最早是为了解决在训练中随着网络深度的增加，出现梯度消失，导致网络训练困难的问题。本文借鉴文献[9]的思想，在编解码部分对应层之间引入跳跃连接。一方面可以将误差直接传递到低层，使网络的训练更有效，还有一个重要的作用是可以对信号的细节信息进行补偿。随着网络层数的增加，由于逐层的卷积、池化等操作会导致信号的主要结构特征被提取，而一些局部的细节信息丢失，特别是随着网络层数的加深，有可能会产生解码端无法由编码端得到的高层信息恢复目标信号的细节信息的问题。而编码端的特征图包含了信号的大量细节信息，通过引入跳跃连接将编码端的特征图信息由卷积层传递到对应反卷积层，可以有助于恢复目标信号的

细节信息。

本文采用的编解码网络结构如图 1 所示。图 1(a)为无跳跃连接的编解码网络结构,图 1(b)图为残差编解码网络。与文献[9]结构相似,类似于 Resnet [15] 网络中的恒等映射,将使得数据在不同网络层之间传递,将原始的输入 Y 和目标映射 X 的优化拟合问题转换为输入和目标之间残差函数 F(Y)= Y−X 的优化问题。图 1(b)中的'+'号表示直接进行元素级相加而保持特征图层数不变,即将卷积层的特征映射以元素方式传递到对应反卷积层,并与反卷积特征映射按元素求和并传递到下一层。与图 1(b)中将特征图元素求和再传递到下一层不同,图 1(c)是将编码端的特征层传递到对应解码端的特征图并进行拼接,进行通道的合并,从而充分利用特征图信息。

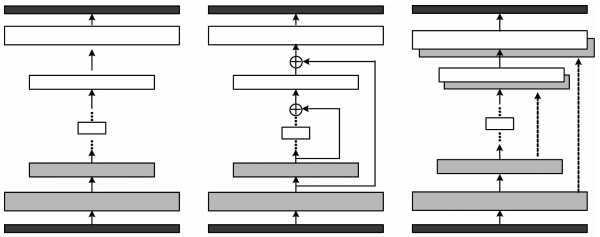

图 1 编解码网络结构图

Fig. 1 Architecture of the encoder-decoder network

3 基于深度全卷积编解码网络的语音增强

设 $Y(m,k)$、$X(m,k)$、$N(m,k)$ 分别为带噪语音、目标语音及加性背景噪声在时刻 m 和频点 k 的 STFT 幅度谱,且满足:

$$Y(m,k)=X(m,k)+N(m,k) \tag{1}$$

其中,$m=\{1,2,\cdots,\Omega\}$ 和 $k=\{1,2,\cdots,K\}$ 分别代表频率和时间,在下文中,我们省略频率和时间索引,简记为 $\boldsymbol{Y}$、$\boldsymbol{X}$ 和 $\boldsymbol{N}$。将第 2 节介绍的深度卷积编解码网络用于语音增强,方法框架如图 2 所示:将带噪语音在时域进行分帧加窗后经 STFT 变换得到带噪语音信号的二维时频表示作为网络的输入,在编码端利用卷积网络对带噪语谱中的典型结构特征进行建模,提取语音特征,同时逐层抑制背景噪声的影响。在解码端通过反卷积层由提取的特征信息逐层恢复语音成分。编解码网络均由卷积层组成,不使用池化层和全连接层。在编解码对应层之间引入跳跃连接,以减少由于卷积操作导致的低层细节信息丢失的问题,从而更好的恢复目标语音的时频特征。反卷积层的输出为带噪语音的时频掩蔽值或者时频增益,由带噪语音的幅度谱联合反卷积层的输出即可得到纯净语音的谱估计值。

损失函数是网络的重要组成部分,目前大部分基于神经网络的谱映射语音增强算法都是基于 MMSE 准则,本文分别考虑了基于 L_1 范数和 L_2 范数的损失函数训练网络对语音增强性能的影响,其对应的损失函数如式(2)和式(3)所示。

$$L_1(\boldsymbol{X},\boldsymbol{Y};\Theta)=\|f(\boldsymbol{Y},\Theta)\otimes\boldsymbol{Y}-\boldsymbol{X}\|_1 \tag{2}$$

$$L_2(\boldsymbol{X},\boldsymbol{Y};\Theta)=\|f(\boldsymbol{Y},\Theta)\otimes\boldsymbol{Y}-\boldsymbol{X}\|_2 \tag{3}$$

式中,$\boldsymbol{Y}$ 代表带噪语音的时频特征输入,Θ 代表网络参数,由权重矩阵和偏置矩阵组成,$f(\boldsymbol{X},\Theta)$ 代表网络的输出,可以理解为掩蔽估计或者增益估计,$f(\boldsymbol{Y},\Theta)\otimes\boldsymbol{Y}$ 代表目标语音的时频特征估计,⊗代表矩阵元素按位相乘。$\|\cdot\|_1$ 表示网络估计和目标语音真值之间距离绝对值之和,$\|\cdot\|_2$ 表示网络估计和目标语音真值之间的距离绝对值的平方和的开方。

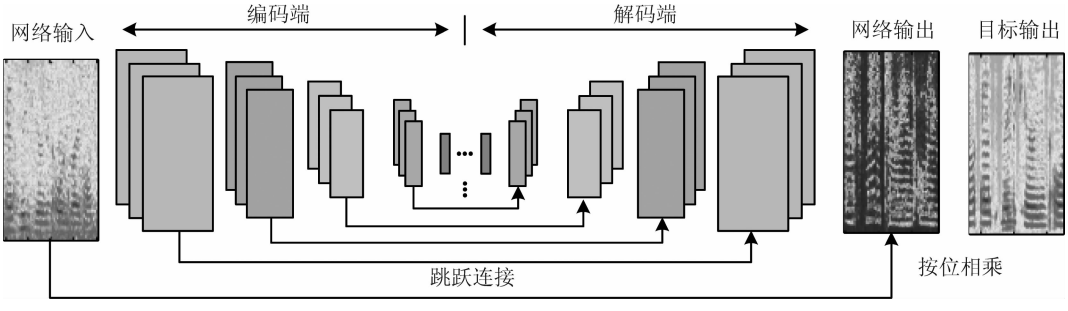

图 2 基于全卷积编解码网络的语音增强

Fig. 2 Fully convolutional encoder-decoder network for speech enhancement

4　实验仿真及性能分析

4.1　数据集和评价指标

本文实验中，从 TIMIT 语音库[16]训练集中随机选取 200 句纯净语音，测试集中选取 20 句纯净语音，从 NOISEX-92 标准噪声库[17]中选取 3 平稳噪声和 3 种非平稳噪声。按照文献[4]的方法对噪声进行分段分别用于训练和测试。按文献[14]的方法对训练和测试用纯净语音按照不同的信噪比添加噪声分别生成 7200 句带噪语音作为训练集，480 句带噪语音作为测试集。

感知语音质量(Perceptual Evaluation of Speech Quality，PESQ)[18]、短时客观可懂度(Short-time objective intelligibility，STOI)[19]和对数谱距离(Log-Spectral Distance，LSD)[20]是被广泛采用的衡量语音增强性能的客观评价指标。PESQ 得分结果与主观听觉测试得分具有很高的相关度。与 PESQ 评估方法侧重于评估处理语音的总体质量不同，STOI 得分主要用于评估语音的可懂度。LSD 是衡量纯净语音和增强语音之间的对数谱失真，其值与语音质量成反比，越小的值表示增强后语音的谱失真越小，其计算公式为：

$$\mathrm{LSD}=\frac{1}{M}\sum_{m=0}^{M-1}\sqrt{\frac{1}{\frac{L}{2}+1}\sum_{l=0}^{L/2}10\log_{10}\frac{|S(m,l)|^2}{|\hat{S}(m,l)|^2}} \quad (4)$$

其中 M 是一段语音总的信号帧数，$S(m,k)$ 和 $\hat{S}(m,k)$ 分别为纯净语音和增强后语音经过短时傅里叶变换后的第 m 帧的第 l 个频谱分量。

4.2　基线方法和参数设置

为了验证本文所提方法的有效性，本文选取基于 DNN 的自回归语音增强方法[4](简记为 DNN-LAS)作为基线方法，对比 Encoder-decoder 全卷积网络(简记为 En-De-Net)和引入两种跳跃连接后的全卷积编解码网络(分别简记为 Res-Net 和 U-Net)的语音增强效果，同时通过实验分析对比利用 L_1 和 L_2 两种损失函数训练网络对语音增强性能的影响。

为了降低网络模型的复杂度，减少训练参数的数量，将训练集和测试集中的语句均下采样至 8000 Hz。帧长取 32 ms，帧移 8 ms，对语句进行 256 点的 STFT 变换，获得语音信号在二维时频域的特征表示。在 DNN-LAS 方法中，网络由输入层、输出层和 3 个隐藏层组成，隐藏层节点数设为 1024。网络的输入为带噪语音当前帧联合相邻 2 帧的对数幅度谱特征，网络的输出目标为当前帧对应的纯净语音的对数幅度谱特征。隐层选取 ReLU 作为激活函数[21]，输出层选取线性激活函数。在实验中 En-De-Net 和 Res-Net 及 U-Net 网络结构相同，均为全连接卷积编解码网络，不包含全连接层和池化层，三种结构的卷积滤波器大小和数量一致。网络的输入和输出均为 128 帧 129 维的时频表示。网络由 11 个卷积层组成，每个卷积层由卷积(反卷积)、批量规范化和非线性激活组成。卷积滤波器大小均为 5×5，数量分别取为 16-32- 64-128-256-512-256-128-64-32-16，卷积步长取 2×2，选取 Leaky ReLUs 非线性激活函数。Dropout 值取为 0.2，在训练时按照 dropout 比率丢弃网络结点，增加网络的泛化性和鲁棒性，在测试时不进行 dropout 操作。其中 DNN-LAS 和 En-De-Net 均是基于 L_2 范数损失函数反向传播误差训练网络参数。为了衡量 L_1 范数用于损失函数对语音增强性能的影响，我们在 U-Net 和 Res-Net 网络增加了基于 L_1 损失函数对网络进行训练的对比实验。

4.3　实验结果和分析

表 1、表 2 和表 3 分别给出了 480 句测试语句经不同网络结构增强后的 PESQ、LSD 和 STOI 得分。由表 1 可以看出，经全连接卷积编解码网络增强后语音的感知质量在各种 SNRs 下均优于 DNN-LAS 方法。在 L_1 范数损失函数下，两种带跳跃连接的 PESQ 均低于 L_2 范数损失函数下不带跳跃连接的全卷积编解码网络，Res-Net 的性能略低于 En-De-Net，高于 U-net 结构。在 L_2 范数损失函数下，带跳跃连接的 U-Net 的增强性能略高于 En-De-Net，而 Res-Net 网络在各个 SNRs 下的语音感知质量得分最优。此结果表明，在相同损失函数下，引入跳跃连接可以有效提升语音的感知质量。由表 2 可以看出，在低 SNRs 下(5 dB 以下)，经 L_1 范数损失函数训练的网络得到的增强语音的对数谱失真要低于经 L_2 范数损失函数训练的网络。在 5 dB 以上，En-De-Net 网络及 U-Net 网络引入的谱失真要高于传统的基于谱映射的全连接神经网络。总的来说，基于 L_1 范数

损失函数训练的 Res-Net 网络的谱失真最小，L_1 范数损失函数下的 U-Net 和 L_2 范数损失函数下的 Res-Net 的谱失真接近。在低 SNRs 下跳跃连接的引入在两种损失函数下均会带来谱失真的减小，随着 SNR 的增加，情况相反。我们认为是跳跃连接在较高 SNR 的情况下，引入的噪声的成分要高于细节信息，导致谱失真的增加。由表 3 给出的不同方法在不同 SNR 下的 STOI 值可以看出，在较低信噪比下，基于 DNN-LAS 的方法对语音可懂度的变化很小，几乎没有改变，而基于全卷积编解码会降低语音的可懂度。这也说明在低信噪比下，基于 DNN-LAS 和全卷积编解码神经网络的方法对可懂度的提升较为明显，但后者的幅度要小于前者。随着信噪比的提高，特别是在 10 dB 下，基于在全卷积编解码网络框架下对可懂度的提升要优于 DNN-LAS 方法，基于 L_1 范数误差准则，以特征图拼接和融合相加的跳跃连接结构反而会降低语音的可懂度。

表 4、表 5 和表 6 分别给出了实验中 6 种噪声

表 1 不同方法在不同 SNR 下的 PESQ 值

Tab. 1 PESQ score of different methods under different SNRs

SNR/dB	Noisy	DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
−5	1.213	1.456	1.598	1.571	1.594	1.601	1.632
0	1.519	1.890	1.957	1.892	1.937	1.956	2.011
5	1.864	2.224	2.270	2.161	2.230	2.277	2.335
10	2.216	2.454	2.551	2.418	2.504	2.579	2.622
Ave.	1.703	2.006	2.094	2.010	2.066	2.103	2.150

表 2 不同方法在不同 SNR 下的 LSD 值

Tab. 2 LSD score of different methods under different SNRs

SNR/dB	Noisy	DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
−5	2.727	1.969	1.892	1.842	1.844	1.913	1.883
0	2.556	1.732	1.719	1.677	1.674	1.727	1.689
5	2.281	1.535	1.566	1.541	1.528	1.561	1.523
10	1.943	1.381	1.413	1.404	1.388	1.398	1.370
Ave.	2.377	1.654	1.648	1.616	1.608	1.650	1.616

表 3 不同方法在不同 SNR 下的 STOI 值

Tab. 3 STOI score of different methods under different SNRs

SNR/dB	Noisy	DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
−5	0.533	0.595	0.560	0.561	0.559	0.576	0.582
0	0.650	0.717	0.675	0.664	0.666	0.687	0.698
5	0.765	0.800	0.776	0.744	0.756	0.781	0.793
10	0.858	0.849	0.855	0.815	0.833	0.857	0.867
Ave.	0.701	0.740	0.717	0.696	0.703	0.725	0.735

在−5、0、5 和 10 dB 经过不同增强方法处理后的 PESQ、LSD 和 STOI 的得分均值。由表 4 可以看出，在平稳噪声和非平稳噪声类型下，基于 En-De-Net 网络的增强方法的 PESQ 得分均高于 DNN-LAS 方法。在 L_1 范数损失函数下，仅在 Babble 噪声和 White 噪声类型下，U-Net 网络的感知质量略高于 DNN-LAS 方法，在其他情况的感知质量得分最低。在 L_2 损失函数下，带跳跃连接的编解码网络的语音感知质量得分要高于 DNN-LAS 和 En-De-Net 方法，以特征图融合相加的跳跃连接方式对语音感知质量的提升要优于特征图拼接的跳跃连接方式。由表 5 可以看出，除了在 Hf channel 噪声类型下，基于卷积编解码网络增强后的语音的谱失真均要高于基于 DNN-LAS 方法，在编解码网络结构中引入跳跃连接会减少语音的谱失真，通过 L_1 范数损失函数训练的网络的谱失真要低于经过 L_2 范数损失函数训练的网络。以特征图融合相加的跳跃连接方式对语音谱失真的减少要优于特征图拼接的跳跃连接方式。由表 6 给出的不同方法在不同 SNR 下的 STOI 值可以看出，在 Babble 噪声下，基于 DNN-LAS 的方法对语音可懂度的变化很小，几乎没有改变，而基于全卷积编解码会降低语音的可懂度。这也说明在 Babble 噪声下，可懂度的提升较为困难。在其他几种平稳和非平稳噪声下，基于 DNN-LAS 的语音增强方法对可懂度的提升要优于全卷积编解码神经网络框架。而在全卷积编解码网络框架下，基于最小均方误差准则，以特征图融合相加的跳跃连接方式对语音可懂度的提升要优于无跳跃连接和以 L_1 范数作为误差准则的方法。

表 4　不同噪声在不同增强方法下的 PESQ 值

Tab. 4　PESQ score of different types of noise under different enhancement methods

噪声	带噪语音	增强方法					
		DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
Babble	1.816	1.863	2.005	1.904	1.953	1.978	2.036
F16	1.752	2.015	2.086	2.007	2.067	2.110	2.151
Factory	1.754	2.021	2.093	2.011	2.047	2.095	2.145
Hf channel	1.591	2.087	2.112	2.053	2.109	2.141	2.181
Pink	1.717	2.042	2.152	2.043	2.119	2.152	2.210
White	1.588	2.010	2.115	2.044	2.102	2.143	2.178

表 5　不同噪声在不同增强方法下的 LSD 值

Tab. 5　LSD score of different types of noise under different enhancement methods

噪声	带噪语音	增强方法					
		DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
Babble	2.036	1.700	1.616	1.593	1.579	1.625	1.582
F16	2.261	1.645	1.606	1.586	1.571	1.607	1.573
Factory	2.321	1.661	1.649	1.608	1.610	1.660	1.622
Hf channel	2.432	1.530	1.643	1.595	1.585	1.619	1.597
Pink	2.453	1.662	1.649	1.626	1.617	1.659	1.626
White	2.758	1.728	1.724	1.690	1.689	1.729	1.697

表6 不同噪声在不同增强方法下的STOI值

Tab. 6 STOI score of different types of noise under different enhancement methods

噪声	带噪语音	增强方法					
		DNN-LAS	En-De-Net	L_1		L_2	
				U-Net	Res-Net	U-Net	Res-Net
Babble	0.704	0.705	0.686	0.672	0.679	0.692	0.701
F16	0.708	0.752	0.730	0.703	0.712	0.737	0.748
Factory	0.683	0.736	0.695	0.682	0.688	0.708	0.718
Hf channel	0.720	0.770	0.756	0.736	0.741	0.768	0.777
Pink	0.705	0.745	0.718	0.692	0.702	0.725	0.734
White	0.688	0.734	0.715	0.691	0.697	0.721	0.731

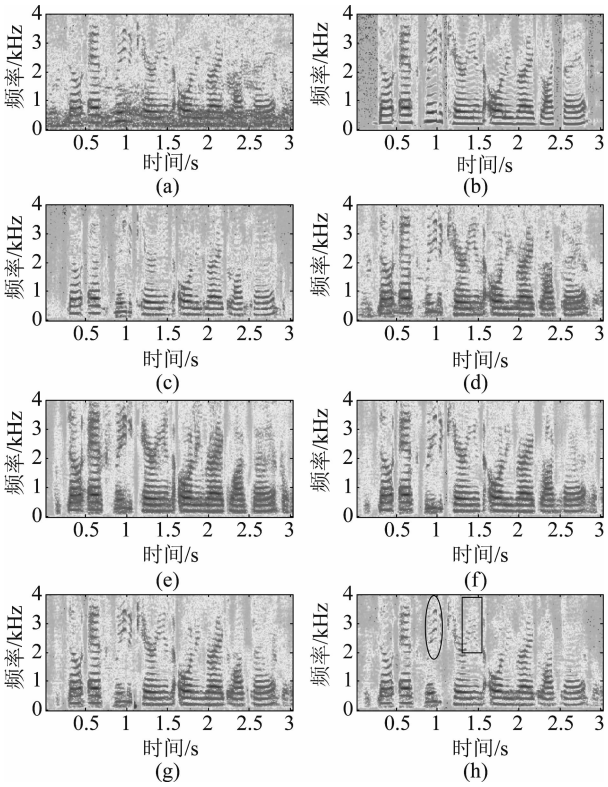

图3 语音语谱图.(a)被5 dB Factory噪声污染的带噪语音;(b)纯净语音;(c)经DNN-LAS方法增强的语音;(d)经En-De-Net方法增强的语音;(e)基于L_1损失函数经U-Net方法增强的语音;(f)基于L_1损失函数经Res-Net方法增强的语音;(g)基于L_2损失函数经U-Net方法增强的语音;(h)基于L_2损失函数经Res-Net方法增强的语音

Fig. 3 Spectrograms of (a) noisy speech corrupted by 5 dB factory noise; (b) clean speech; (c) enhanced by DNN-LAS; (d) enhanced by En-De-Net; (e) enhanced by U-Net with L_1 loss; (f) enhanced by Res-Net with L_1 loss; (g) enhanced by U-Net with L_2 loss; (h) enhanced by Res-Net with L_2 loss

表1至表6从数据上给出了不同网络结构的语音增强性能。为了更直观的理解,图3给出了一段被5 dB Factory噪声污染后的语音经不同的增强方法处理后的语音语谱图。由图3(c)可以看出,经过DNN-LAS增强方法处理后,大部分噪声成分得到了抑制,在低频部分语音成分较好的得到了恢复,但在高频部分存在一些细节成分的缺失。经En-De-Net方法处理后,在高、低频部分语音的结构信息得到了恢复,但是存在频谱模糊的现象,我们分析应该是由高层特征信息通过上采样恢复语音细节信息引入的。图3(e)和图3(f)分在是在L_1损失函数下,在图3(d)结构的基础上引入两种跳跃连接方式得到的,和图3(d)相比,在较好的保留了高低频结构信息的基础上,对噪声的抑制略有提升。图3(g)和图3(h)分别是在L_2损失函数下,在图3(d)结构的基础上引入两种跳跃连接方式得到的,与DNN-LAS方法相比,低频结构信息都得到了较好的恢复,由图中框图部分可以看出,在高频细节成分的恢复要优于DNN-LAS方法,和L_1损失函数相比,尽管噪声成分得到了较好的抑制,但在高频细节结构上还有部分未能很好的恢复。

5 结论

本文将深度全卷积编解码网络用于语音增强,针对编解码过程中出现细节信息缺失的问题,提出一种将特征图融合和特征图拼接两种跳跃连接方

式引入到编解码网络对应层之间的语音增强方法。同时对比了在基于 L_1 和 L_2 两种损失函数的训练方法对语音感知质量和谱失真的影响。实验结果表明，基于特征图融合相加的跳跃连接方法能获得较好的语音感知质量。基于 L_1 损失函数训练的方法在语音高频部分成分的恢复要优于基于 L_2 损失函数的训练方法。

参考文献

[1] Loizou P C. Speech enhancement: Theory and practice [M]. Boca Raton, FL, USA: CRC Press 2007.

[2] 李轶南，张雄伟，曾理，等. 改进的稀疏字典学习单通道语音增强算法[J]. 信号处理，2014，30(1)：44-50.
Li Yinan, Zhang Xiongwei, Zeng Li, et al. An improved monaural speech enhancement algorithm based on sparse dictionary learning [J]. Journal of Signal Processing, 2014, 30(1): 44-50. (in Chinese)

[3] 胡永刚，张雄伟，邹霞，等. 改进的非负矩阵分解语音增强算法[J]. 信号处理，2015，31(9)：1117-1123.
Hu Yonggang, Zhang Xiongwei, Zou Xia, et al. Improved nonnegative matrix factorization based speech enhancement algorithm [J]. Journal of Signal Processing, 2015, 31(9): 1117-1123. (in Chinese)

[4] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.

[5] Wang Y X, Narayanan A, Wang D L. On training targets for supervised speech separation[J]. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

[6] Huang P S, Kim M, Hasegawa-Johnson M, et al. Joint optimization of masks and deep recurrent neural networks for monaural source separation [J]. IEEE/ACM Trans. on Audio Speech and Language Processing, 2015, 23(12): 2136-2147.

[7] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[8] 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. Berlin: Springer, 2015: 234-241.

[9] Mao X J, Shen C, Yang Y B. Image de-noising using very deep fully convolutional encoder-decoder networks with symmetric skip Connections[C]//Proceedings of the 30th Conference on Neural Information Processing Systems(NIPS). NY: Curran Associates, 2016.

[10] Park S R, Lee J. A fully convolutional neural network for speech enhancement[C]//Processing of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, International Speech Communication Association (ISCA) Press, 2017: 1993-1997.

[11] Michelsanti D, Tan Z H. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification[C]//Processing of the 18th Annual Conference of the International Speech Communication Association. Stockholm, Sweden, International Speech Communication Association (ISCA) Press, 2017: 2008-2012.

[12] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). NJ: IEEE, 2017: 5967-5976.

[13] Pandey A, Wang D L. On adversarial training and loss functions for speech enhancement[C]//Proceedings of the 43nd IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2018: 5074-5078.

[14] 时文华，倪永婧，张雄伟，等. 联合稀疏非负矩阵分解和神经网络的语音增强[J]. 计算机研究与发展，2018，55(11)：2430-2438.
Shi Wenhua, Ni Yongjing, Zhang Xiongwei, et al. Deep neural network based monaural speech enhancement with sparse non-negative matrix factorization [J]. Journal of Computer Research and Development, 2018, 55(11): 2430-2438. (in Chinese)

[15] 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), NJ: IEEE, 2016: 770-778.

[16] Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond[J]. Speech Commun., 1990, 9(4): 351-356.

[17] Varga A, Steeneken H J M. Assessment for automatic

speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.

[18] Rix A, Beerends J, Hollier M P, et al. Perceptual evalulation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs[C]// Proceedings of the 26th IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2001: 749-752.

[19] Taal C H, Hendriks R C, Heusdens R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.

[20] Du J, Huo Q. A speech enhancement approach using piecewise linear approximation of an explicit model of environment distortions[C]// Proceedings of the 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2008: 4721-4724.

[21] Andrew L M, Awni Y H, Andrew Y N. Rectifier nonlinearities improve neural network acoustic models[C]// Proceedings of the 30th International Conference on Machine Learning (JMLR: W&CP). Brookline, MA: Microtome Publishing, 2013: 315-323.

作者简介

时文华 女，1982年生，山东人。中国人民解放军陆军工程大学在读博士生，主要研究方向为语音与图像增强。
E-mail: whshi0919@163.com

张雄伟(通讯作者) 男，1965年生，浙江人。中国人民解放军陆军工程大学教授，博士生导师，主要研究方向为多媒体信息处理、数字通信等。
E-mail: xwzhang9898@163.com

邹 霞 男，1979年生，湖北人。中国人民解放军陆军工程大学副教授，硕士生导师，主要研究方向为语音增强、语音编码等。
E-mail: zlc1997@163.com

孙 蒙 男，1984年生，山东人。中国人民解放军陆军工程大学副教授，主要研究方向为语音信号处理、机器学习等。
E-mail: sunmengccjs@163.com