文章编号：1003-0530(2021)10-1986-10

# 基于门控残差卷积编解码网络的单通道语音增强方法

张天骐 柏浩钧 叶绍鹏 刘鉴兴
（重庆邮电大学通信与信息工程学院，信号与信息处理重庆市重点实验室，重庆 400065）

**摘 要**：针对卷积编解码网络（CED，Convolution encoder-and-decoder）对语音时序相关信息捕获困难的问题，本文提出了一种基于门控残差卷积编解码网络的语音增强方法。该方法在卷积编解码网络的基础上引入了门控机制、膨胀卷积与残差连接：门控机制能够很好地处理序列前后相关信息；膨胀卷积使得卷积过程获得更大的感受野，提取更加丰富的全局信息；残差连接能够防止梯度消失与梯度爆炸，提升网络精度。此外，采用频域损失函数与时域评价指标联合优化的策略对网络进行训练，以进一步提升网络增强效果。实验表明，在匹配噪声和不匹配噪声下，相比于基线 CED 与其他对比方法，本文方法取得了更高的 PESQ、STOI 与 SI-SDR，对语音的清浊音都有较好恢复效果，且具有较强的泛化能力。

**关键词**：语音增强；门控机制；卷积编解码网络；残差连接
**中图分类号**：TN912.35 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2021.10.023

**引用格式**：张天骐，柏浩钧，叶绍鹏，等. 基于门控残差卷积编解码网络的单通道语音增强方法[J]. 信号处理，2021，37(10)：1986-1995. DOI：10.16798/j.issn.1003-0530.2021.10.023.
**Reference format**：ZHANG Tianqi，BAI Haojun，YE Shaopeng，et al. Single-channel speech enhancement method based on gated residual convolution encoder-and-decoder network[J]. Journal of Signal Processing，2021，37(10)：1986-1995. DOI：10.16798/j.issn.1003-0530.2021.10.023.

# Single-channel Speech Enhancement Method Based on Gated Residual Convolution Encoder-and-Decoder Network

ZHANG Tianqi BAI Haojun YE Shaopeng LIU Jianxing
(School of Communication and Information Engineering, Chongqing Key Laboratory of Signal and Information Processing (CQKLS&IP), Chongqing University of Posts and Telecommunications (CQUPT), Chongqing 400065, China)

**Abstract**: In order to solve the problem that it is difficult for Convolution Encoder-and-Decoder (CED) network to capture temporal related contexts of speech, a speech enhancement method based on gated residuals convolution encoder-and-decoder network is proposed. Based on CED, this proposed method introduces the gating mechanism, dilated convolution and residual connection to the network: The gating mechanism can well handle the relevant contexts of sequence; Dilated convolution makes the convolution process obtain larger receptive field and extract more abundant global information; Residual connection can prevent vanishing gradient and exploding gradient and improve network accuracy. In addition, the combined optimization strategy of frequency-domain loss function and time-domain evaluation index is adopted to train the network to further improve the enhancement effect of propose network. Experimental results show that, compared with the baseline CED and other comparison methods, the proposed method achieves higher PESQ, STOI and SI-SDR under matched noise and mismatched noise, and it has a good recovery effect on the voiceless and voiced sounds of speech and has strong generalization ability.

**Key words**: speech enhancement; gating mechanism; convolution encoder-and-decoder network; residual connection

收稿日期：2021-04-08；修回日期：2021-05-11
基金项目：国家自然科学基金项目(61671095，61702065，61701067，61771085)；信号与信息处理重庆市市级重点实验室建设项目(CSTC2009CA2003)；重庆市研究生科研创新项目(CYS19248)；重庆市教育委员会科研项目(KJ1600427，KJ1600429)

## 1　引言

语音增强的任务为从被噪声污染的语音信号中滤除噪声，提取干净语音，提升语音的质量与可懂度。根据麦克风数量划分，语音增强可分为多通道语音增强（多个麦克风）与单通道语音增强（单个麦克风），由于单通道语音增强方法具有结构简单、成本低廉、应用范围广、时频域能有效建模等优点，仍是国内外语音增强的重点研究方向。

早期研究一般采用传统信号处理方法来进行语音增强，如谱减法[1]、滤波法[2]以及基于最小均方误差（MMSE，Minimum mean square error）的谱估计算法[3]。谱减法即直接从含噪语音谱中减去噪声谱，估计得到增强语音谱，在平稳噪声下降噪效果较好，但面对非平稳噪声，谱减法容易将谱成分过多或过少的减去，从而造成语音失真并产生音乐噪声。滤波法能够在一定程度上处理非平稳噪声，克服了谱减法的部分问题，但其仍需对模型进行假设。不同于滤波法的是，基于 MMSE 的方法能够得到增强语音的非线性估计，在特定条件下有较好的处理能力。上述方法统称为基于模型的无监督语音增强方法，这类方法一般需要条件假设，对平稳和慢变噪声处理效果较好，但针对非平稳和变化明显的噪声，传统方法的增强效果会明显减弱，特别是在低信噪比下，跟踪噪声特征非常困难。

针对上述问题，有监督语音增强方法迅速发展起来。有监督语音增强采用有监督学习的方式，从大量的语音和噪声数据下学习一个数学模型，并通过该模型对含噪语音进行预测，得到增强语音。由于深度神经网络（DNN，Deep neural network）具有较强的非线性建模能力，WANG 等人将其引入了语音分离和增强领域[4]，相比于传统方法，分离与增强效果提升明显。2014 年，XU Yong 等人将含噪语音的对数功率谱（LPS，Logarithmic power spectrum）作为 DNN 的输入特征[5]，利用 DNN 的非线性建模能力，构造了含噪语音 LPS 到纯净语音 LPS 的非线性映射函数，进一步提升了 DNN 的增强效果，值得一提的是，该方法未假设语音和噪声需满足任何前提条件，具有较强的泛化能力。在上述网络的基础上，衍生出了多种基于 DNN 的语音增强方法，文献[6]提出一种多目标学习方法，联合 LPS 以及理想比值掩蔽（IRM，Ideal ratio masking）作为 DNN 的学习目标，并利用语音子带特征来进行噪声感知训练，增强后的语音在质量与可懂度上都有较大提升。文献[7]结合稀疏非负矩阵分解（SNMF，Sparse non-negative matrix factorization）与 DNN 来进行语音增强，利用谱分解特性，能够保证清音和无结构语音部分不会引入额外失真，取得了较好的增强效果。尽管 DNN 在语音增强上展现出了许多优势，但其仍存在一定不足，随着神经元数量以及网络层数增加，网络参数量也会明显增加，计算开销加大，且网络训练容易陷入局部最优解和过拟合的情况，使得模型精度降低。

而近年来，卷积神经网络（CNN，Convolution neural network）迅速发展，其具有训练参数小、平移不变和对多维度数据处理能力好等优点，在图像和语音领域得到广泛应用。KYONG 等人使用 CNN 来进行语音情感识别[8]，取得了不错的效果。2015 年，RONNEBERGER 等人提出一种新颖的卷积神经网络（U-net）来进行图像分割[9]，U-net 包含编码层、中间层与解码层，其主要思想在于解码层的池化（pooling）操作用上采样来替代，为了融合高层与底层的特征信息，防止梯度消失，将高分辨率的输入（编码层）和经过上采样的输出（解码层）进行跳跃连接（Skip connection），在较小的数据集下，U-net也能取得较好的效果，这引起了语音领域的高度关注。2017 年，PARK S R 等人在 U-net 的基础上提出一种全卷积网络（FCN，Fully convolution network）[10]，又称冗余卷积编解码网络（CED），该网络取消了 U-net 的上下采样操作与 CNN 的全连接层，通过调整卷积核的大小与步长来实现特征维度变化，相较于 DNN 与 CNN，在较小网络参数的情况下取得了较多的指标提升。2018 年，STOLLER D 等人提出一种基于 FCN 的 Wave-U-Net 结构[11]，该网络直接以语音的一维时域信号作为网络输入，并采用一维卷积的方式对输入语音进行处理，网络直接输出增强语音波形，具有较强的特征提取及处理能力，实时性较好。在前面网络的基础上，有学者对网络进行了进一步创新，PANDEY A 等人搭建了一种深层的复数域卷积编解码网络[12]，该网络直接估计语音的实部与虚部，重构阶段可以直接估计得到相位信息，但其存在特征训练困难的问题。

在上述的卷积编解码网络中，中间层一般只起到特征传递的作用，即将编码层提取到的抽象特征传递给解码层，对语音序列前后相关特征的处理能力有限。TAN Ke 等人针对该问题提出一种卷积循环神经网络（CRN，Convolution recurrent network）[13]，该网

络以 LSTM 作为 CED 网络的中间层，能够更好关注语音的时序特征，对上下文关键信息进行捕获，但 LSTM 存在计算量大、并行性差的问题，在一定程度上影响了模型的增强效果。

针对卷积编解码网络对无法很好处理语音序列信息的问题，我们提出一种基于门控残差卷积编解码网络的语音增强方法。该方法借鉴自然语言处理[14]中的门控机制，将全卷积门控线性单元引入 CED 网络的中间层，同时采用膨胀卷积以及残差连接等操作来提升卷积过程的感受野与模型精度，所提方法能够更好关注语音的时序相关信息，有利于提升语音的整体质量与可懂度。同时在模型的目标函数上，采用时域评价指标与频域损失函数联合优化的策略，以进一步提升模型的增强效果。

## 2 系统结构

假设语音模型为：

$$\boldsymbol{y}=\boldsymbol{s}+\boldsymbol{d} \tag{1}$$

其中 $\boldsymbol{y}$、$\boldsymbol{s}$ 和 $\boldsymbol{d}$ 分别为含噪语音、干净语音和噪声的波形，表示一维向量，语音增强的任务是从 $\boldsymbol{y}$ 中滤除 $\boldsymbol{d}$ 得到尽可能干净的 $\boldsymbol{s}$。对式(1)进行短时傅里叶变换(STFT, Short time Fourier transform)到频域为：

$$\boldsymbol{Y}_n(k)\exp(\angle\boldsymbol{Y}_n(k))=\boldsymbol{S}_n(k)\exp(\angle\boldsymbol{S}_n(k))+\boldsymbol{D}_n(k)\exp(\angle\boldsymbol{D}_n(k)) \tag{2}$$

其中 $\boldsymbol{Y}_n(k)$、$\boldsymbol{S}_n(k)$、$\boldsymbol{D}_n(k)$ 和 $\angle\boldsymbol{Y}_n(k)$、$\angle\boldsymbol{S}_n(k)$、$\angle\boldsymbol{D}_n(k)$ 分别为含噪语音、纯净语音和噪声在第 $n$ 帧的幅度谱和相位谱向量，$n$ 为帧索引 $n=(1,2,\cdots,N)$，$k$ 为频率索引 $k=(1,2,\cdots,K)$，由于每一帧的频率数相同，式(2)可简写为：

$$\boldsymbol{Y}_n\exp(\angle\boldsymbol{Y}_n)=\boldsymbol{S}_n\exp(\angle\boldsymbol{S}_n)+\boldsymbol{D}_n\exp(\angle\boldsymbol{D}_n) \tag{3}$$

考虑到人耳对相位信息不敏感，网络训练过程忽略相位信息，则有：

$$\boldsymbol{Y}_n=\boldsymbol{S}_n+\boldsymbol{D}_n \tag{4}$$

神经网络在语音增强任务中的作用可概括为：通过训练网络参数集 $\theta$ 构造一个含噪语音到干净语音特征空间的复杂的非线性映射函数 $f_\theta$，使得其逼近真实语音，谱映射的形式为：

$$\min \left\| f_\theta(\boldsymbol{Y}_n)-\boldsymbol{S}_n \right\|_2^2 \tag{5}$$

从而得到目标输出(增强语音幅度谱)：

$$\hat{\boldsymbol{S}}_n=f_\theta(\boldsymbol{Y}_n) \tag{6}$$

联合带噪音相位进行短时傅里叶逆变换(ISTFT, Inverse short time Fourier transform)得到时域增强语音：

$$\hat{\boldsymbol{s}}_n=\mathrm{ISTFT}(\hat{\boldsymbol{S}}_n\cdot\exp(\angle\boldsymbol{Y}_n)) \tag{7}$$

### 2.1 卷积编解码网络

我们采用的基线网络为卷积编解码网络(CED)，该网络采用全卷积操作，取消了上下采样以及全连接层，具体结构如图 1 所示。其中，网络的输入特征为含噪语音幅度谱，即时间(帧)和频率两个维度的特征图，输出为增强语音幅度谱，结合含噪语音相位可重构得到增强语音波形。总体上，网络大致可划分为三层，分别为编码层、中间层和解码层，每一层的结构及功能如下所述：

1)编码层：编码层由 5 个二维卷积层组成，每个二维卷积层都包含二维卷积(2D-Conv, Two-dimension convolution)、批次归一化(BN, Batch Normalization)层以及泄露修正线性单元(LRelu, Leaky rectified linear unit)激活函数，首先对特征图进行二维卷积，然后进行批次归一化，最后通过激活函数得到每一层的输出。BN 层对数据进行减均值和去相关等操作，使得卷积后的数据特征满足独立同分布假设，经研究发现，BN 层有利于加快网络的收敛，防止梯度爆炸；激活函数采用 LRelu，其能给负值

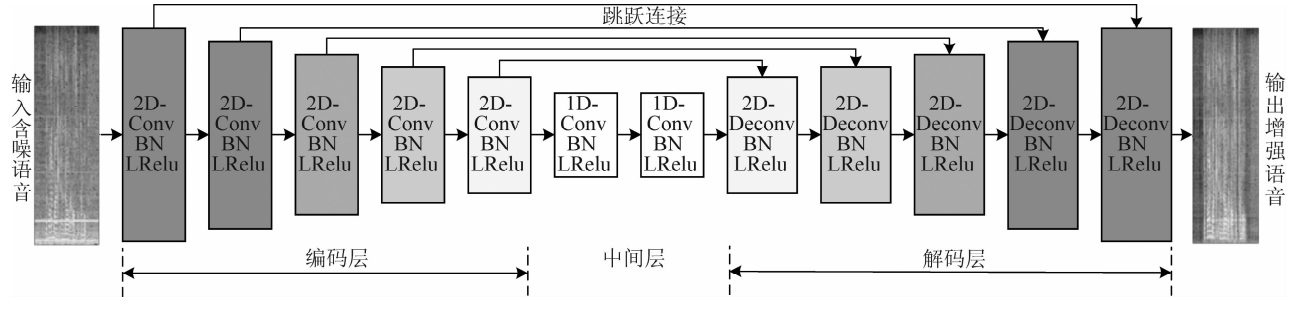


图 1 基于 CED 网络的语音增强流程图

Fig. 1 Speech enhancement flow chart based on CED network

特征添加一个斜率，保证训练过程负值特征不丢失，减少未训练单元数量；5 层二维卷积（2D-Conv）逐层提取语音的幅度谱特征，每次卷积操作后，特征图在时间上保持不变、频率维度上减半、通道数翻倍，通过调整卷积核的大小、步长与数量来实现这一操作。值得注意的是，通过编码层后的特征图在时间维度上与输入特征图保持一致，这使得模型能够处理任意长度（帧长）的语音，具有较好的实时特性。

2）中间层：中间层由 2 个一维卷积层组成，每层都包含一维卷积（1D-Conv，One-dimension convolution）、BN 层以及激活函数 LRelu，一维卷积的卷积核与卷积步长都为 1。卷积之前需要对编码器输出的二维张量进行调整（Reshape）降维，用以满足一维卷积的需求，同样，中间层的输出需要经过调整还原维度。整体上，中间层主要作用为特征传递。

3）解码层：解码层由 5 个二维反卷积层组成，每层都包含二维反维卷积（2D-Deconv，Two-dimension deconvolution）、BN 层以及激活函数 LRelu。2D-Deconv 可以看作 2D-Conv 的逆过程，通过调整卷积步长即可还原特征图位置信息。同时，我们将编码层特征图输入到相同维度的解码层，通过通道拼接使得相应解码层特征图通道数扩大一倍，该操作有利于在解码过程中恢复细粒特征信息。

### 2.2　门控线性单元

由于中间的一维卷积层的主要作用为特征传递，对序列信息的处理能力有限。因此，我们采用一种门控机制来处理一维信息流，在早期研究中，一般采用基于 RNN 的门控机制来处理序列信息，如长短时记忆单元（LSTM，Long short-term memory）和门控循环单元（GRU，Gated recurrent unit），但其存在并行性差和计算量大的问题。针对上述问题，我们采用了一种全卷积门控线性单元（GLU，Gate linear unit）[14]，如图 2 所示，该门控单元能够减小网络训练参数，并行性好，且能够选择性传递信息。值得注意的是，GLU 中间层的两个激活函数分别为线性激活函数 linear 与 Sigmoid（用符号 $\sigma$ 表示，计算公式如式（9）所示），linear 为梯度反向传播提供了线性路径来缓解梯度消失的问题，而 Sigmoid 用以维持网络的非线性特性，其取值为 0 到 1，通过该激活函数可以关注我们所需要的语音特征同时忽略不相关的特征，GLU 的表达式如式（8）所示：

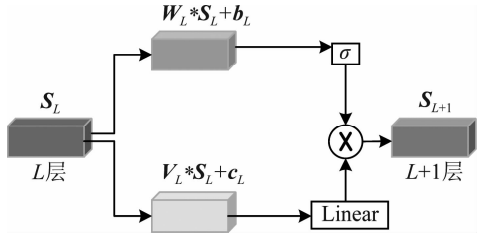


图 2　门控线性单元

Fig. 2　Gate linear unit

$$S_{L+1}=\sigma(W_L*S_L+b_L)\odot(V_L*S_L+c_L) \tag{8}$$

$$\sigma(x)=\frac{1}{1+e^{x}} \tag{9}$$

其中 $S_L$ 和 $S_{L+1}$ 分别表示第 $L$ 层和第 $L+1$ 层的输出特征，$W_L$、$V_L$、$b_L$ 和 $c_L$ 分别表示第 $L$ 层的权重和偏执，$\sigma$、$*$ 和 $\odot$ 分别表示 Sigmoid 激活函数、卷积操作和阿达玛乘积（逐点相乘）。

### 2.3　一维膨胀卷积

在中间层中，我们使用了一维膨胀卷积（Dilated convolution），相比一维普通卷积，一维膨胀卷积能够获得更大的感受野（Receptive field）[23]，随着膨胀率的提升，感受野往往呈指数级增长，这意味卷积过程能够获取更加丰富的语音上下文特征信息，可以更好挖掘序列中的信息依赖关系，其表达式如式（10）所示。

$$Y=(X*_{r}F)(p)=\sum_{s+rt=p}X(s)F(t) \tag{10}$$

其中 $X$ 和 $Y$ 分别表示输入特征与输出特征，$F$ 和 $r$ 分别表示卷积核与膨胀率，$p$、$s$、$t\in Z$，$Z$ 表示整数集。每层网络感受野的计算公式为：

$$R_{L+1}=R_L+(F-1)r \tag{11}$$

$R_L$ 表示第 $L$ 层卷积层的感受野，$R_{L+1}$ 表示第 $L+1$ 层卷积层的感受野。

### 2.4　门控残差模块

残差网络能够解决深层神经网络的过拟合问题，同时还能防止梯度消失与梯度爆炸，提升模型精度，利用其优点，我们将残差网络与上述门控线性单元 GLU 相结合，并引入一维膨胀卷积，得到一种门控残差模块，其结构如图 3 所示。门控残差模块一共包含四个卷积层，上半部分相当于把 GLU 中的两个一维卷积替换为一维膨胀卷积，其卷积核的大小、步长与输出通道数分别为 5、1 和 128，下半部分为两个并行的一维普通卷积层，卷积核的大小、

步长与输出通道数为 1、1 和 128，分别得到残差输出以及跳跃连接输出，由于模块输入输出的通道数一致，假设 $s$ 为网络输入，$F(s)$ 为经过多个隐藏层的输出，则残差输出为 $o=F(s)+s$。在两个一维普通卷层后，同样加入了 BN 层以及激活函数 LRelu，用以保证模块输出特征仍然满足独立同分布假设以及保持网络的非线性特性。值得注意的是，模块中两个并行的膨胀卷积层采用同样的膨胀率，通过堆叠门控残差模块，并逐渐增大膨胀率，可以达到扩大感受野的目的。

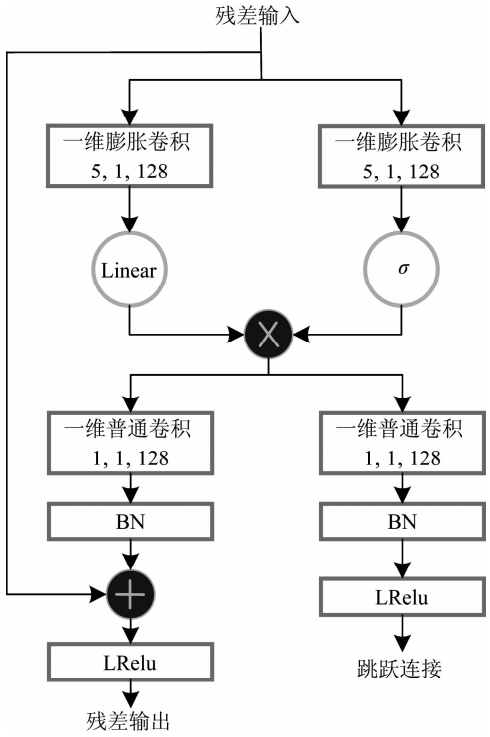


图 3 门控残差模块

Fig. 3 Gate residual block

## 2.5 门控残差卷积编解码网络

在基线卷积编解码网络的基础上，我们将上述搭建的门控残差模块引入网络中间层，得到了一种门控残差卷积编解码网络，其网络结构如图 4 所示。在网络中间层中，通过将一组膨胀率 $r$ 膨胀到某个最大系数 $2^N$ 的门控残差模块前后堆叠而形成了门控残差网络，该结构能够在较小参数量下显著提升感受野，并同时提升模型对序列信息的处理能力，更好地关注时序相关特征[15]。此外，门控残差网络中使用了跳跃连接，这允许网络将对应层级提取的特征合并(Add)到最终预测之中。值得注意的是，受文献[16]启发，我们通过卷积层来实现 ISTFT，从而使得时域增强语音能够参与网络训练，且原始时域语音 $\boldsymbol{s}_n$ 的相位能够补偿含噪语音相位并重建更加精确的时域增强语音 $\hat{\boldsymbol{s}}_n$。

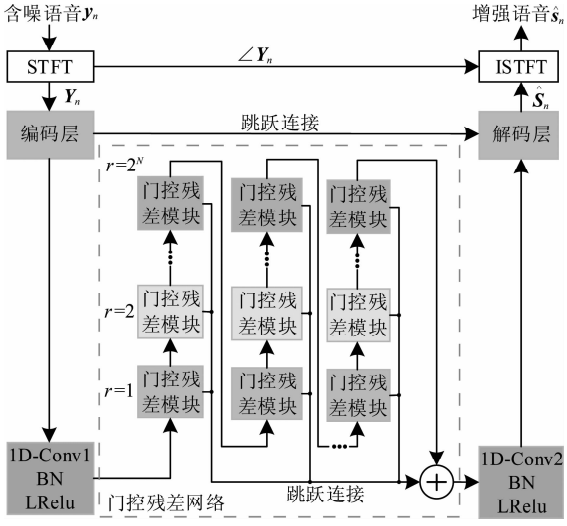


图 4 基于门控残差卷积编解码网络的语音增强流程图

Fig. 4 Speech enhancement flow chart based on gated residual CED network

表 1 网络模型参数

Tab. 1 Network model parameters

| 模块 | 输入维度 | 输出维度 | 超参数 |
|---|---|---|---|
| Reshape1 | $T\times128$ | $T\times128\times1$ | – |
| 编码层 | $T\times128\times1$ | $T\times4\times64$ | $k=3\times3$, $s=1\times2$, $c=4$、8、16、32、64 |
| Reshape2 | $T\times4\times64$ | $T\times256$ | – |
| 1D-Conv1 | $T\times256$ | $T\times128$ | $k=1$, $s=1$, $c=128$ |
| 门控残差网络 | $T\times128$ | $T\times128$ | $\left.\begin{matrix}k=5,\ s=1,\ c=128\\k=1,\ s=1,\ c=128\end{matrix}\right\}\times15$, $(r=1,2,4,8,16)\times3$ |
| 1D-Conv2 | $T\times128$ | $T\times256$ | $k=1$, $s=1$, $c=256$ |
| Reshape3 | $T\times256$ | $T\times4\times64$ | – |
| 解码层 | $T\times4\times64$ | $T\times128\times1$ | $k=3\times3$, $s=1\times2$, $c=32$、16、8、4、1 |
| Reshape4 | $T\times128\times1$ | $T\times128$ | – |

表 1 中，我们给出了网络模型具体参数设置，整体上，网络输入和输出特征维度为 $T\times128$，$T$ 表示时间帧、128 表示频率维度，而第 3 个维度表示特征图通道数。超参数中，$k$ 表示卷积核的大小，$s$ 表示卷积步长，$c$ 表示输出通道数，$r$ 表示膨胀率。由于门控残差模块的输入输出维度保持不变，具有很好的可移植特性，可以根据模型需求来添加模块数量，通过实验对比，本文门控残差网络由 15 个门控残差模块组成，具体由 3 组膨胀率 $r$ 分别为 1、2、4、8、16 的门控残差模块叠加得到（相同颜色模块膨胀率相同）。最后，选择合适的损失函数训练得到最优网络模型，通过网络映射得到增强语音幅度谱 $\hat{\boldsymbol{S}}_n$，结合人耳对相位信息不敏感的特性，利用含噪语音相位 $\angle\boldsymbol{Y}_n$ 重构得到时域增强语音 $\hat{\boldsymbol{s}}_n$。

### 2.6　损失函数

本文网络采用小批量梯度下降法进行训练。由于频域损失能够更好关注语音的时频相关特征，我们在解码层输出端计算了频域损失，经文献[17]的实验证明，平均绝对值误差函数（MAE，Mean absolute error）在改善语音质量和可懂度方面表现更好，基于 MAE 的频域损失函数表达式为：

$$L_{\mathrm{MAE}}=\frac{1}{M}\sum_{n=1}^{M}\left\|\hat{\boldsymbol{S}}_n-\boldsymbol{S}_n\right\|_1 \tag{12}$$

其中 $\|\cdot\|$ 表示向量的范数，$\boldsymbol{S}_n$ 与 $\hat{\boldsymbol{S}}_n$ 分别表示第 $n$ 帧原始语音与增强语音幅度谱向量，$M$ 表示批处理大小。

由于语音客观评价指标的计算公式与网络训练损失函数具有一定差异，可能会存在损失函数与评价指标失配的问题，即当损失函数下降到一定程度，部分评价指标可能不会继续变化[18]。针对上述问题，提出了以语音评价指标作为网络损失函数，能够在一定程度上进一步提升语音增强效果，根据文献[19]，若采用语音评价指标——比例不变信号失真比（SI-SDR，Scale-invariant signal-to-distortion ratio）作为网络的训练函数，增强语音的客观指标提升明显，SI-SDR 的计算公式为：

$$\text{SI-SDR}=10\log_{10}\left(\frac{\left\|\alpha\boldsymbol{s}_n\right\|^2}{\left\|\alpha\boldsymbol{s}_n-\hat{\boldsymbol{s}}_n\right\|^2}\right) \tag{13}$$

$\boldsymbol{s}_n$ 与 $\hat{\boldsymbol{s}}_n$ 分别表示第 $n$ 帧原始语音与增强语音波形，$\alpha$ 为纯净语音的加权因子，计算公式为：

$$\alpha=\arg\min_{\alpha}\left\|\alpha\boldsymbol{s}_n-\hat{\boldsymbol{s}}_n\right\|^2=\frac{\hat{\boldsymbol{s}}_n^T\boldsymbol{s}_n}{\left\|\boldsymbol{s}_n\right\|_2^2} \tag{14}$$

把式（14）代入式（13）可得 SI-SDR 的优化函数：

$$L_{\text{SI-SDR}}=-\text{SI-SDR}=-\frac{1}{M}\sum_{n=1}^{M}10\log_{10}\left(\frac{\boldsymbol{s}_n^T\hat{\boldsymbol{s}}_n}{\boldsymbol{s}_n^T\boldsymbol{s}_n\hat{\boldsymbol{s}}_n^T\hat{\boldsymbol{s}}_n-\boldsymbol{s}_n^T\hat{\boldsymbol{s}}_n}\right) \tag{15}$$

$L_{\text{SI-SDR}}$ 采用时域信号计算，其同时利用了含噪语音与纯净语音的相位信息，能够进一步优化网络权值，减小幅度谱的偏移。利用频域 MAE 与时域 SI-SDR 的优点，我们对 MAE 与 SI-SDR 进行联合优化，采用平衡因子 $\gamma$ 对两个函数进行平衡，最终网络优化函数为（Joint）：

$$L_{\text{Joint}}=\gamma L_{\text{MAE}}+(1-\gamma)L_{\text{SI-SDR}},0\leqslant\gamma\leqslant1 \tag{16}$$

其中平衡因子 $\gamma$ 通过实验比较取得，通过最小化优化函数更新网络梯度并将误差传递至网络各层，从而更新迭代网络的权值参数与偏置量。

## 3　实验及结果分析

### 3.1　数据集及参数设置

本文训练集的纯净语音选自 TIMIT 语料库中的 800 条不同说话人语音，其中男女声各占一半。训练噪声选自 Noise92 噪声库以及一些常见环境噪声，共 10 种，分别为 Factory1、Babble、White、Pink、Tank、Office、Street、Car、Machinegun、Buccaneer1。含噪语音由等长的纯净语音与噪声按不同信噪比混合得到，将 800 条纯净语音与 10 种噪声按 7 种信噪比（−9 dB、−6 dB、−3 dB、0 dB、3 dB、6 dB、9 dB）混合得 56000 条含噪语音，全部的含噪语音作为训练集，以训练集的 10% 作为验证集，每个轮次后，通过验证集来验证性能。

为了评估模型的增强效果，我们构建了不同的测试集。测试集的纯净语音选自 TIMIT 语料库中的另外 200 条不同说话人语音，男女声各占一半，测试噪声选用 6 种匹配噪声（参与训练的噪声，分别为 Factory1、Babble、White、Pink、Street、Car）和 4 种不匹配噪声（未参与训练的噪声，分别为 F16、Factory2、Volvo、Restaurant）。将 200 条纯净语音与 6 种匹配噪声按 4 种信噪比（−5 dB、0 dB、5 dB、10 dB）混合得到 4800 条含噪语音作为匹配噪声测试集。此外，为了评估模型的泛化能力，将 4 种不匹配噪声与上述 200 条纯净语音按 4 种信噪比混合得到 3200 条含噪语音，令其作为不匹配噪声测试集。

实验中，语音和噪声波形的采样频率为 8 kHz，

短时傅里叶变换的帧长为 31.875 ms,即 255 个采样点,帧移为 8 ms,由于 STFT 后的频谱具有共轭对称性,则采用一半的频谱来减少计算量,这使得每帧语音频率维度为 128,然后通过维度重塑,得到维度为 $T\times128\times1$ 的含噪语音幅度谱作为编码层的输入特征。

本文网络模型通过 Keras 在 Tensorflow 后端进行搭建,对网络进行有监督式训练,联合频域损失函数 MAE 和时域评价指标 SI-SDR 作为网络优化函数,采用小批量梯度下降法来更新网络权值参数,批处理大小为 32,使用 Adam 作为网络优化器,初始学习率为 0.001。

### 3.2 评价指标及对比方法

我们采取 3 种指标对语音进行客观评估。采用国际电联 ITU-T 推荐的语音质量感知评价(PESQ, Perceptual evaluation of speech quality)来衡量语音质量[20],其得分区间为[-0.5,4.5],得分越高语音质量越高,该客观指标能够很好近似主观听觉效果;采用短时客观可懂度(STOI, Short time objective intelligibility)衡量语音可懂度[21],其得分区间为[0,1],得分越高语音被理解的概率越大;采用 SI-SDR 衡量语音的失真程度[22],得分越高语音失真越小。

实验部分采用 2.1 节中的卷积编解码网络 CED 为基线网络,以 CNN[8] 方法和 CRN[13] 方法作为对比方法。CED 对应层级参数设置与表 1 相同,CNN 包含三个卷积层、三个最大池化层、两个全连接层,CRN 即在 CED 的基础上把中间层替换为 3 层单向 LSTM。CNN 与 CRN 的具体参数与训练方式遵循原论文设置,基线 CED 与本文方法(Propose)的训练方式保持一致。所有方法均使用本文的数据集来进行训练和测试。

### 3.3 结果对比及性能分析

图 5 为 10 种测试噪声下本文网络在联合优化函数(Joint)的不同 $\gamma$ 取值下得到的增强语音的 PESQ 均值。对比图 5 可知,不同的 $\gamma$ 取值能够对模型的增强效果产生一定影响,$\gamma=0$ 表示仅采用频域 MAE 对网络训练,$\gamma=1$ 表示仅采用时域 SI-SDR 对网络训练,对比可看出,时域 SI-SDR 在语音质量上的表现要优于频域 MAE,PESQ 约提升了 0.04 ~ 0.05。此外,联合 SI-SDR 与 MAE 后,网络指标进一步提高,这体现了本文联合优化策略的有效性,且当 $\gamma=0.3$ 时,网络取得了相对更高的 PESQ,因此我们把 $\gamma$ 设置为 0.3。

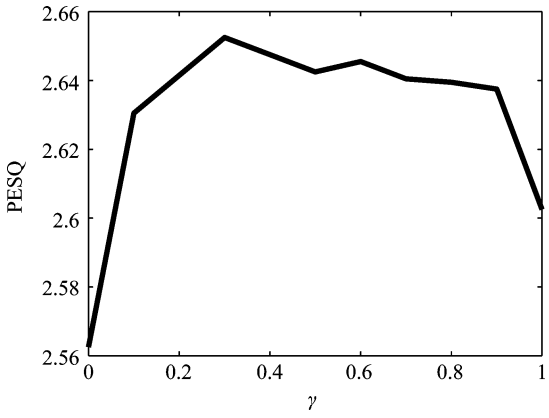


图 5 目标函数的不同 $\gamma$ 取值对增强语音的 PESQ 均值影响

Fig. 5 The effect of different $\gamma$ values of the objective function on the mean PESQ of enhanced speech

图 6 为基线 CED 与本文方法(Propose)在优化函数 MAE、SI-SDR 与 Joint 下得到的增强语音的 PESQ 与 STOI(PESQ、STOI 为 10 种测试噪声与 4 种信噪比下的均值)。由图 6 可看出,无论在何种优化

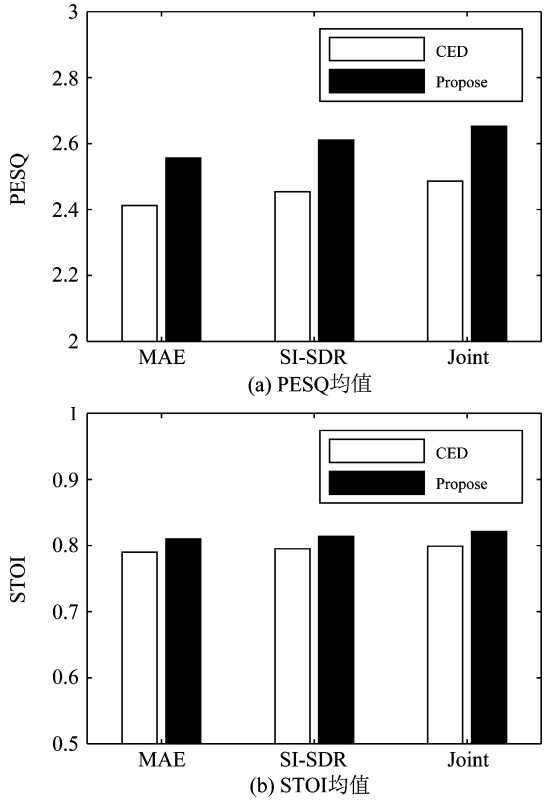


图 6 基线 CED 与本文方法在优化函数 MAE、SI-SDR 与 Joint 下得到的增强语音的 PESQ 与 STOI 均值

Fig. 6 The PESQ and STOI mean values of enhanced speech obtained under the optimized functions MAE, SI-SDR and Joint of the baseline CED and the Propose method

表 2　各方法的平均 PESQ 和 STOI

Tab. 2　The average PESQ and STOI of different methods

| 指标 | 模型 | 匹配噪声 | | | | | 不匹配噪声 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 信噪比/dB | | | | | 信噪比/dB | | | | |
| | | −5 | 0 | 5 | 10 | 均值 | −5 | 0 | 5 | 10 | 均值 |
| PESQ | Noisy | 1.58 | 1.82 | 2.09 | 2.28 | 1.943 | 1.49 | 1.75 | 2.12 | 2.31 | 1.918 |
| | CNN | 1.95 | 2.21 | 2.54 | 2.80 | 2.375 | 1.79 | 2.05 | 2.29 | 2.56 | 2.173 |
| | CED | 2.22 | 2.49 | 2.71 | 2.91 | 2.583 | 2.04 | 2.28 | 2.55 | 2.69 | 2.390 |
| | CRN | **2.29** | 2.54 | 2.77 | 2.95 | 2.638 | 2.10 | 2.31 | 2.52 | 2.72 | 2.413 |
| | Propose | 2.25 | **2.68** | **2.86** | **3.09** | **2.720** | **2.19** | **2.48** | **2.74** | **2.93** | **2.585** |
| STOI | Noisy | 0.54 | 0.65 | 0.76 | 0.82 | 0.693 | 0.55 | 0.63 | 0.75 | 0.83 | 0.690 |
| | CNN | 0.68 | 0.76 | 0.82 | 0.86 | 0.780 | 0.68 | 0.74 | 0.79 | 0.85 | 0.765 |
| | CED | **0.73** | 0.79 | 0.84 | 0.88 | 0.810 | 0.70 | 0.77 | 0.83 | 0.85 | 0.788 |
| | CRN | 0.72 | 0.80 | 0.85 | **0.90** | 0.818 | 0.71 | 0.77 | 0.82 | 0.86 | 0.790 |
| | Propose | 0.72 | **0.83** | **0.87** | **0.90** | **0.830** | **0.73** | **0.79** | **0.85** | **0.88** | **0.813** |

函数下，本文方法都能取得优于基线 CED 的 PESQ 和 STOI，这说明门控残差模块以及门控残差网络的应用促进了网络性能提升，使得网络能够捕获更多整体语音特征，且基线 CED 在各优化函数下的增强效果依然遵循图 5 所得结论，因此基线 CED 同样采用联合优化函数 Joint。

表 2 为匹配噪声和不匹配噪声下各对比方法在各信噪比下的平均 PESQ 和 STOI。对比表 2 可知，在匹配噪声下，CNN 的取得的指标要低于 CED、CRN 与本文方法，其增强效果有限。相比于基线 CED 方法，本文方法的指标得到全面提升，PESQ 均值约提升了 0.137，STOI 均值提升了 0.02，这说明本文网络对噪声的抑制效果较好，增强语音具有较高的质量与可懂度，但在−5 dB 时，CED 取得了更高的 STOI，这说明低信噪比下本文方法对语音可懂度的改善有限。相比于 CRN 方法，本文方法在高信噪比下的 PESQ 提升较多，但在低信噪比下，CRN 取得的 PESQ 略高于本文方法，这说明低信噪比下 CRN 与本文网络的降噪能力接近，同时我们发现，−5 dB与 10 dB 时，CRN 与本文方法取得了相同的 STOI，但在 0 dB 与 5 dB 时，本文方法的 STOI 取得较多领先，这说明在面对中等强度的噪声干扰时，本文方法对语音的增强效果更好，分析原因，可能是在这几种信噪比下，本文网络对语音序列相关信息的捕获能力更强，学习了更加丰富的语音特征。

在不匹配噪声下，各对比方法的指标均有一定下降，但经对比发现，本文方法指标下降较少，PESQ 均值下降了 0.135（4.96%）、STOI 均值下降了 0.017（2.05%），而 CRN 的 PESQ 均值下降了 0.225（8.53%）、STOI 均值下降了 0.028（3.42%），且相较于 CNN 与基线 CED 方法，本文方法的平均 PESQ 和 STOI 都取得明显领先，这说明本文方法的泛化能力较强，具有更好的鲁棒性。

图 7 为十种测试噪声下各方法在四种信噪比下的平均 SI-SDR。对比图 7 可知，在低信噪比下，本文方法的 SI-SDR 得分与其他几种方法接近，但在高信噪比下，SI-SDR 取得了较为明显的提升，这说明本文方法得到的增强语音失真更小，对语音有较好的恢复效果。

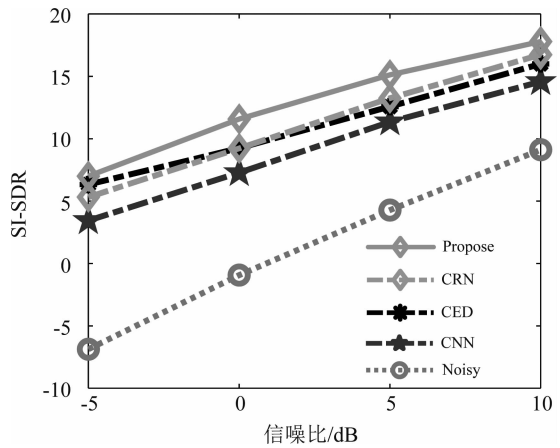


图 7　十种测试噪声下各方法在四种信噪比下的平均 SI-SDR

Fig. 7　The average SI-SDR of different methods in four SNR under ten test noises

为了更直观的比较各方法的增强效果，我们对比了各方法的语谱图（0 dB 的 babble 噪声下得到），如图 8 所示。对比可知，CNN 恢复了部分浊音成分，

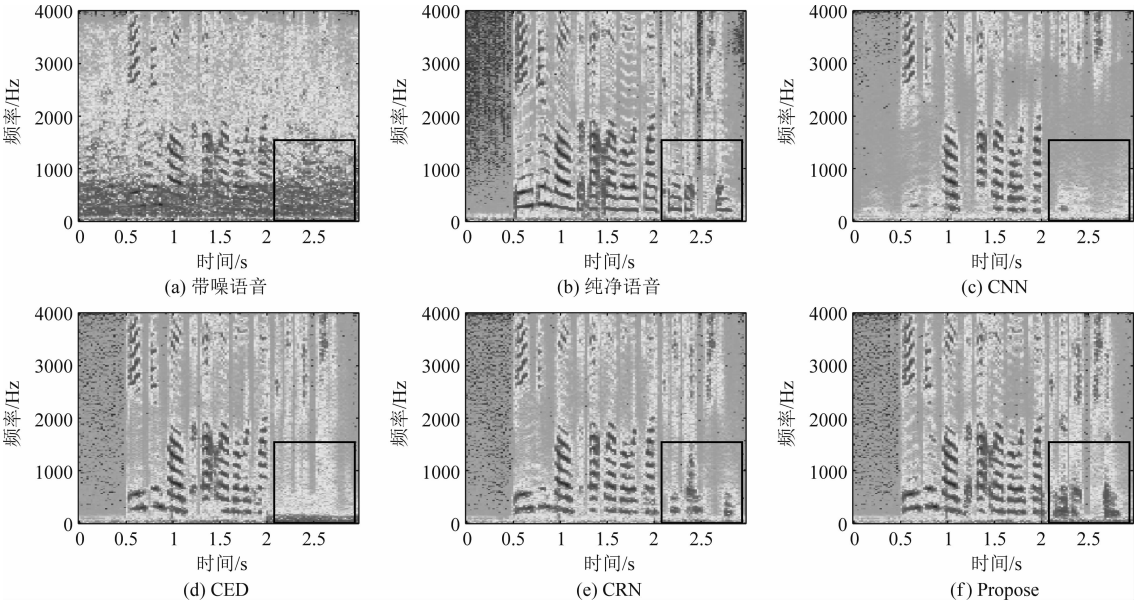


图 8 语谱图

Fig. 8 Spectrogram

但清音部分恢复较差,存在一定背景噪声,相比于 CNN,CED 对语音成分恢复和降噪效果更为明显,但仍有部分语音难以恢复,语音的谐波成分保留较少,而 CRN 在 CED 的基础上进一步提升了增强效果,清浊音部分都能得到有效恢复。相比于以上方法,本文方法明显恢复了更多语音成分,保留了更多谐波特征,同时对噪声的抑制效果更好。

## 4 结论

本文在卷积编解码网络的基础上进行了改进,引入了门控机制、膨胀卷积以及残差连接,搭建了一种新的网络结构,该网络保留原来网络特性的同时能够更好地捕获语音时序相关信息,同时采用损失函数与评价指标联合优化的策略,进一步提升网络增强效果。实验部分,对比了本文方法与基线 CED、CNN 以及 CRN 方法的客观评价指标。结果表明,本文方法能够取得更高的平均 PESQ、STOI 和 SI-SDR,在语音成分的恢复以及噪声抑制上,本文方法取得了明显优势,且具有相对较强的泛化能力,增强语音具有较高的质量与可懂度。

在后续研究中,还需进一步调整中间层门控残差模块的结构以及数量,用以保证在较小参数量的情况下最大化提升网络增强效果,此外,可以进一步优化网络目标函数,提升其在语音客观评价指标上的表现。

**参考文献**

[1] MIYAZAKI R, SARUWATARI H, INOUE T, et al. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(7): 2080-2094.

[2] WIENER N. Extrapolation, interpolation, and smoothing of stationary time series[M]. Cambridge, MA, USA: The MIT Press, 1949.

[3] KIRUBAGARI B, PALANIVEL S, SUBATHRA N. Speech enhancement using minimum mean square error filter and spectral subtraction filter[C]// International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, India, 2014: 1-7.

[4] WANG Yuxuan, WANG Deliang. Boosting classification based speech separation using temporal dynamics[C]// Proc of the 13th Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2012: 1528-1531.

[5] XU Yong, DU Jun, DAI Lirong, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65-68.

[6] WANG Qing, DU Jun, DAI Lirong, et al. Joint noise and mask aware training for DNN-based speech enhancement with SUB-band features[C]// 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA). San Francisco, CA, USA. IEEE, 2017: 101-105.

[7] 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. Speech enhancement combined with sparse non-negative

matrix factorization and neural network[J]. Computer Research and Development, 2018, 55(11): 2430-2438. (in Chinese)
[8] KYONG H L, DO H K. Design of a convolutional neural network for speech emotion recognition[C]//2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020: 1332-1335.
[9] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 234-241.
[10] PARK S R, LEE J W. A fully convolutional neural network for speech enhancement[C]//Interspeech 2017. ISCA: ISCA, 2017: 1993-1997.
[11] STOLLER D, EWERT S, DIXON S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation[J]. arXiv preprint arXiv:1806.03185, 2018. https://arxiv.org/abs/1806.03185.
[12] PANDEY A, WANG Deliang. 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.
[13] TAN Ke, WANG Deliang. A convolutional recurrent neural network for real-time speech enhancement[C]//Interspeech 2018. ISCA: ISCA, 2018: 1405-1408.
[14] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks[C]//International Conference on Machine Learning. PMLR, 2017: 933-941. http://proceedings.mlr.press/v70/dauphin17a.
[15] OORD A V D, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[J]. arXiv preprint arXiv: 1609.03499v2.2016. https://arxiv.org/abs/1609.03499.
[16] VENKATARAMANI S, HIGA R, SMARAGDIS P. Performance based cost functions for end-to-end speech separation[C]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018: 350-355.
[17] PANDEY A, WANG Deliang. On adversarial training and loss functions for speech enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5414-5418.
[18] FU S W, WANG Taowei, TSAO Y, et al. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(9): 1570-1584.
[19] KOLBAEK M, TAN Zhenghua, JENSEN S H, et al. On loss functions for supervised monaural time-domain speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 825-838.
[20] ITU-T, Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs, International Telecommunications Union (ITU-T) Recommendation, 2001: 862.
[21] 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.
[22] ROUX J L, WISDOM S, ERDOGAN H, et al. SDR-half-baked or well done? [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom. IEEE, 2019: 626-630.
[23] GABBASOV R, PARINGER R. Influence of the receptive field size on accuracy and performance of a convolutional neural network[C]//2020 International Conference on Information Technology and Nanotechnology (ITNT), 2020: 1-4.

## 作者简介

**张天骐**　男，1971 年生，四川人。重庆邮电大学通信与信息工程学院教授、博士生导师，主要研究方向为通信信号的调制解调、盲处理、图像语音信号处理、神经网络实现以及 FPGA、VLSL 实现。
E-mail: zhangtg@cqupt.edu.cn

**柏浩钧**　男，1997 年生，四川人。重庆邮电大学通信与信息工程学院硕士研究生，主要研究方向为语音信号处理、语音增强。
E-mail: 1114549265@qq.com

**叶绍鹏**　男，1996 年生，湖北人。重庆邮电大学通信与信息工程学院硕士研究生，主要研究方向数字水印、信息隐藏技术。
E-mail: yspysp77@163.com

**刘鉴兴**　男，1997 年生，重庆人。重庆邮电大学通信与信息工程学院硕士研究生，主要研究方向为信道编码参数盲识别技术研究。
E-mail: 30919117@qq.com