一种用于语音增强的卷积门控循环网络

袁文浩,胡少东,时云龙,李 钊,梁春燕
(山东理工大学计算机科学与技术学院,山东淄博 255000)

摘 要: 为了充分利用含噪语音特征来提高语音增强网络的性能,基于含噪语音在时间和频率两个维度上的相关性,本文结合卷积神经网络的局部特征提取能力和门控循环单元的长期依赖建模能力,设计了一种适用于语音增强的卷积门控循环网络.该网络采用卷积网络结构代替全连接网络结构来改进门控循环单元中的特征计算过程,从而能够更好地保留含噪语音特征中的时频结构信息.实验结果表明,与其它语音增强网络相比,本文网络在语音成分的保留和噪声成分的抑制上具有明显优势,增强后语音具有更好的语音质量和可懂度.

关键词: 语音增强;深度神经网络;门控循环单元;卷积神经网络
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2020)07-1276-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.07.005

A Convolutional Gated Recurrent Network for Speech Enhancement

YUAN Wen-hao, HU Shao-dong, SHI Yun-long, LI Zhao, LIANG Chun-yan
(*College of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong* 255000, *China*)

Abstract: In order to improve the performance of speech enhancement networks by making full use of noisy speech features, based on the correlation of noisy speech in time and frequency, by combining the local feature extraction ability of convolutional neural networks and the long-term dependence modeling ability of gated recurrent unit, a convolutional gated recurrent network suitable for speech enhancement is designed in this paper. This network uses a convolutional network structure instead of a fully connected network structure to improve the feature calculation process in the gated recurrent unit, thereby can better retain the time-frequency structure in the noisy speech features. The experimental results show that compared with other speech enhancement networks, the proposed network has obvious advantages in speech component retention and noise component suppression, and the enhanced speech has better speech quality and intelligibility.

Key words: speech enhancement; deep neural network; gated recurrent unit; convolutional neural network

1 引言

语音增强是噪声环境下语音信号处理的关键步骤,传统语音增强方法一般基于统计学原理,其对于平稳噪声具有较好的处理效果,但是在低信噪比和非平稳噪声条件下性能会急剧下降[1,2].针对传统语音增强方法的不足,近年来,研究人员将深度学习技术应用于语音增强,提出了基于深度神经网络的语音增强方法.该类方法利用大量语音和噪声样本数据进行网络训练,建立起含噪语音和增强语音之间的映射关系,相比传统方法显著提高了语音增强性能[3,4].为了提高网络的语音增强性能,研究人员设计了多种不同的网络结构来建立语音增强模型.文献[5~9]均采用全连接神经网络来建立语音增强模型,不同之处在于,文献[5~7]采用了基于时频掩蔽的训练目标,而文献[8,9]则直接将纯净语音的对数功率谱作为训练目标.由于含噪语音在时间维度上具有序列性,文献[10]采用循环神经网络(Recurrent Neural Network, RNN)来建立语音增强模型;而为了更好地利用含噪语音中的长期依赖关系,文献[11,12]利用长短时记忆(Long Short-Term Memory, LSTM)网络来进行语音增强.由于时频域中含噪语音在时间和频率两个维度上具有二维相关性,文献[13~17]采用卷积神经网络(Convolutional Neural Network, CNN)来建立语音增强模型,其中:文献[13]的网络结构为一个由全卷积神经网络构成的编码器-解码器;文献[14]采用了一个由卷积层、池化层和全连接层构成的CNN;文献[15~17]都采用了空洞卷积来提高网络在时频域上的感受野,但是文献[15,16]的网络结

收稿日期:2019-08-02;修回日期:2020-03-30;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61701286, No.11704229);山东省自然科学基金(No.ZR2018LF002);山东省高等学校青年创新团队发展计划(No.2019KJN048)

构中采用了门控机制和残差学习,而文献[17]则采用了密集连接卷积网络.

实际上,无论采用何种网络结构来进行语音增强,提高性能的关键都是对于含噪语音特征的充分利用,而在时频域中,含噪语音最重要的特征就是其在时间和频率两个维度上的相关性[18]. 为了充分利用时频域中含噪语音在两个维度的相关性,提高深度神经网络的语音增强性能,本文通过结合 CNN 对于二维数据的局部特征提取能力和 GRU(Gated Recurrent Unit)对于序列中长期依赖关系的建模能力,设计了一种适用于语音增强的卷积门控循环网络(Convolutional Gated Recurrent Network, CGRN),并通过实验从增强语音的语音质量和可懂度两方面对网络的语音增强性能进行了评估.

2　基于深度神经网络的语音增强

在时频域中,在利用深度神经网络进行语音增强时,为了充分利用含噪语音的上下文信息,网络的输入一般为多帧的含噪语音频域特征,当对第 l 帧含噪语音进行语音增强时,网络的输入可以表示为

$$\boldsymbol{\Psi}_l=\begin{bmatrix} Y_{l-\tau,1} & Y_{l-\tau+1,1} & \cdots & Y_{l+\tau,1} \\ Y_{l-\tau,2} & Y_{l-\tau+1,2} & \cdots & Y_{l+\tau,2} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{l-\tau,K} & Y_{l-\tau+1,K} & \cdots & Y_{l+\tau,K} \end{bmatrix} \tag{1}$$

其中,$Y_{l,k}$表示第 l 帧的第 k 个频带的特征值,$2\tau+1$ 是网络的输入窗长,K 是频域特征的维度. 显然,$\boldsymbol{\Psi}_l$ 是包含时间和频率两个维度的矩阵形式特征.

基于深度神经网络的语音增强就是通过网络训练构造一个函数f_θ来估计纯净语音的频域特征,其中θ是网络的参数集合. 网络的训练过程就是最小化如下的均方误差代价函数的过程:

$$C(\theta)=\frac{1}{M}\sum_{l=1}^{M}\|f_\theta(\boldsymbol{\Psi}_l)-S_l\|_2^2 \tag{2}$$

其中,S_l 是与 $\boldsymbol{\Psi}_l$ 对应的纯净语音的第 l 帧的频域特征,M 是网络训练时采用的 Mini-batch.

3　GRU 网络结构

LSTM 采用门控机制来控制信息在序列中的传递,能够对含噪语音中的长期依赖关系进行建模,表现出了良好的语音增强性能. 而与 LSTM 相比,GRU 具有相似的性能,但结构更加简单. GRU 的基本单元的计算过程包括:

首先,利用输入特征 $\boldsymbol{x}_l$ 和 $l-1$ 帧的隐层特征 $\boldsymbol{h}_{l-1}$ 计算重置门 $\boldsymbol{r}_l$ 和更新门 $\boldsymbol{z}_l$

$$\boldsymbol{r}_l=\sigma(\boldsymbol{W}_r\boldsymbol{h}_{l-1}+\boldsymbol{U}_r\boldsymbol{x}_l+\boldsymbol{b}_r) \tag{3}$$

$$\boldsymbol{z}_l=\sigma(\boldsymbol{W}_z\boldsymbol{h}_{l-1}+\boldsymbol{U}_z\boldsymbol{x}_l+\boldsymbol{b}_z) \tag{4}$$

其中,$\boldsymbol{W}_\#$和 $\boldsymbol{U}_\#$分别是对应隐层特征和输入特征的权重矩阵,$\boldsymbol{b}_\#$是相应的偏置项,σ 代表 Sigmoid 激活函数.

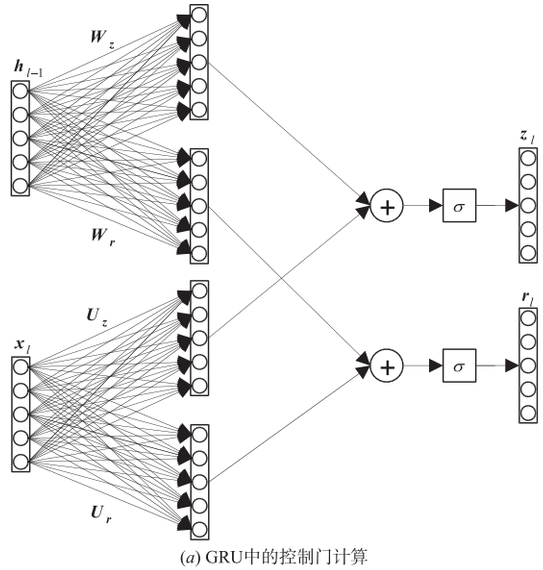
其次,利用计算得到的重置门,结合输入特征和隐层特征更新单元状态

$$\tilde{\boldsymbol{h}}_l=\tanh(\boldsymbol{W}_h(\boldsymbol{r}_l\circ\boldsymbol{h}_{l-1})+\boldsymbol{U}_h\boldsymbol{x}_l+\boldsymbol{b}_h) \tag{5}$$

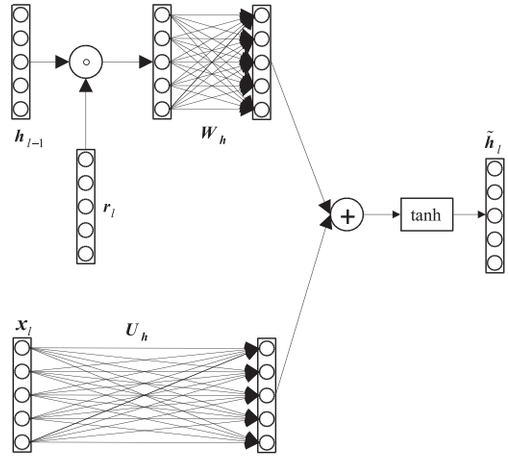
最后,利用更新门,通过一个一阶递归来计算 l 帧的隐层特征

$$\boldsymbol{h}_l=(1-\boldsymbol{z}_l)\circ\boldsymbol{h}_{l-1}+\boldsymbol{z}_l\circ\tilde{\boldsymbol{h}}_l \tag{6}$$

其中,"∘"表示 Hadamard 乘积. 图 1(*a*)(*b*)分别给出了 GRU 中两个控制门和单元状态的计算过程,可见:由于采用了全连接网络进行计算,GRU 中的输入特征、隐层特征、单元状态及两个控制门均为向量形式.



(*a*) GRU中的控制门计算



(*b*) GRU中的单元状态更新

图1　GRU中的控制门计算和单元状态更新

4 CGRN 网络结构

由于含噪语音在相邻帧与相邻频带之间均具有很强的局部相关性,因此式(1)中矩阵形式的输入中的时频结构同样是非常重要的特征信息,而 GRU 中向量形式的特征计算过程显然破坏了含噪语音的时频结构. 为了在特征计算过程中更好地保留含噪语音中的时频结构信息,从而更加充分的利用含噪语音的时频相关性,本文将更适合于矩阵形式特征计算的卷积神经网络与 GRU 进行结合,设计了一种基于卷积特征计算的门控循环神经网络 CGRN. CGRN 与 GRU 采用同样的门控循环机制,不同之处在于,CGRN 利用卷积网络结构代替全连接网络结构来进行特征计算,从而将特征从向量形式转变为矩阵形式. 这种将卷积网络结构深度嵌入循环神经网络结构的网络设计思路,最早出现于文献[19]和文献[20]中,分别用于天气预测和视频特征提取,表现出了良好的网络性能;本文从含噪语音的局部相关特性出发,为了改进语音增强网络中含噪语音特征的计算过程,设计了与文献[20]类似的网络 CGRN.

CGRN 的更新门 $\boldsymbol{Z}_l$ 和重置门 $\boldsymbol{R}_l$ 是通过对输入特征 $\boldsymbol{X}_l$ 和 $l-1$ 帧的隐层特征 $\boldsymbol{H}_{l-1}$ 进行卷积运算得到

$$\boldsymbol{Z}_l = \sigma(\boldsymbol{W}_{\boldsymbol{Z}} * \boldsymbol{H}_{l-1} + \boldsymbol{U}_{\boldsymbol{Z}} * \boldsymbol{X}_l + \boldsymbol{b}_{\boldsymbol{Z}}) \tag{7}$$

$$\boldsymbol{R}_l = \sigma(\boldsymbol{W}_{\boldsymbol{R}} * \boldsymbol{H}_{l-1} + \boldsymbol{U}_{\boldsymbol{R}} * \boldsymbol{X}_l + \boldsymbol{b}_{\boldsymbol{R}}) \tag{8}$$

其中,“ * ”表示卷积运算,$\boldsymbol{b}_{\#}$ 表示相应的偏置项,与 GRU 不同,这里的 $\boldsymbol{W}_{\#}$ 和 $\boldsymbol{U}_{\#}$ 分别表示对应隐层特征和输入特征的卷积滤波器.

同样,单元状态的更新也是通过对输入特征和隐层特征进行卷积运算得到的

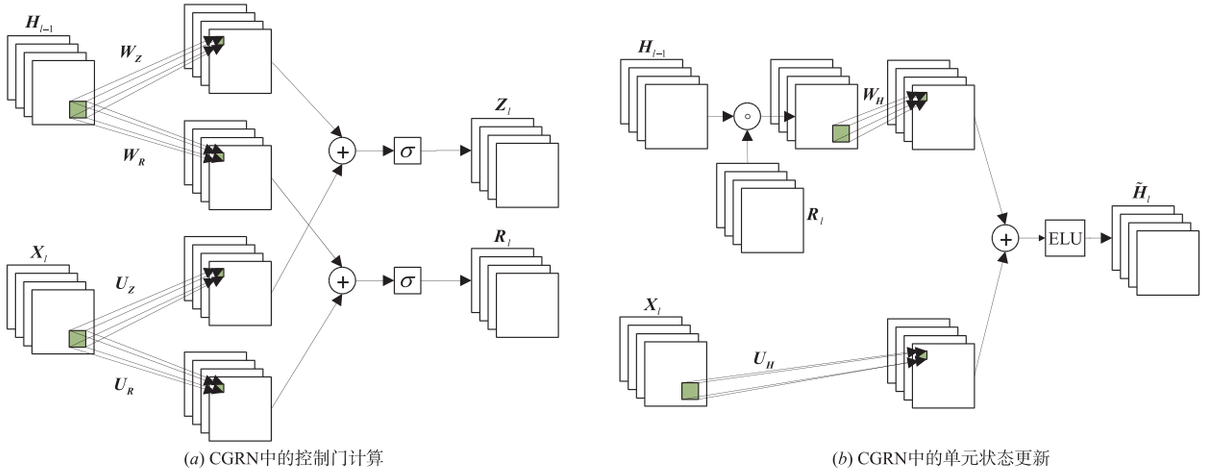
$$\tilde{\boldsymbol{H}}_l = \mathrm{ELU}(\boldsymbol{W}_{\boldsymbol{H}} * (\boldsymbol{R}_l \circ \boldsymbol{H}_{l-1}) + \boldsymbol{U}_{\boldsymbol{H}} * \boldsymbol{X}_l + \boldsymbol{b}_{\boldsymbol{H}}) \tag{9}$$

其中,与 GRU 不同,CGRN 采用收敛速度更快的 ELU (Exponential Linear Unit)激活函数.

最后,CGRN 的基本单元间的时序关系与 GRU 一致,是由更新门控制的递归过程,l 帧的隐层特征为

$$\boldsymbol{H}_l = (1 - \boldsymbol{Z}_l) \circ \boldsymbol{H}_{l-1} + \boldsymbol{Z}_l \circ \tilde{\boldsymbol{H}}_l \tag{10}$$

图 2(a)(b)分别给出了 CGRN 中两个控制门和单元状态的计算过程,可见:由于采用了卷积运算,CGRN 中的输入特征、隐层特征、单元状态及两个控制门均为多个通道的矩阵形式. 显然,这种矩阵形式的特征能够更好的保留含噪语音的时频结构信息.



(a) CGRN中的控制门计算 (b) CGRN中的单元状态更新

图2 CGRN中的控制门计算和单元状态更新

5 实验与结果分析

为了验证 CGRN 相比其它网络结构在语音增强任务上的有效性,将其与 DNN、LSTM 和 GRU 三种网络进行比较. 四种网络均采用 3 层结构,其中,DNN 的结构与文献[5]相同,每个隐层的节点数均为 2048;LSTM 和 GRU 的结构则与文献[10]类似,隐层特征和单元状态的维度均为 512;CGRN 所采用的滤波器个数为 32,滤波器大小为 3 ×9(时间维度 × 频率维度). 四种网络的输入均为窗长为 11 帧的对数功率谱特征,输出为单帧的对数功率谱特征,对数功率谱计算时所采用的短时傅里叶变换的帧长为 32ms (256 点),帧移为 16ms(128 点),相应的对数功率谱特征的有效维度为 129. 为了训练过程稳定,输入输出均进行了均值方差归一化.

为了确保实验结论的充分性,语音增强对比实验分别基于英文的 TIMIT 数据库和中文的 THCHS-30 数据库展开[21,22].

5.1 TIMIT 实验结果

5.1.1 训练集与测试集

训练集包括含噪语音及相应的纯净语音,其中含噪语音是由纯净语音与噪声合成得到的. 纯净语音来自于 TIMIT 数据库的训练集,噪声则采用来自于文献[23]中的 100 类真实噪声. 将纯净语音与噪声的采样

频率均转换为8kHz,然后按照 -10dB、-5dB、0dB、5dB 和10dB 五种不同的信噪比进行合成,从合成得到的含噪语音中随机选取50000段来构成训练集.

测试集所用的纯净语音是来自于TIMIT数据库测试集的192段纯净语音,噪声则采用与训练集完全不同的4类未知噪声,分别为N1(Factory2)、N2(Buccaneer1)、N3(Destroyer engine)和N4(HF channel),这4类噪声来自于Noisex92噪声库[24].将纯净语音与噪声的采样频率均转换为8kHz,然后按照 -7dB、0dB、7dB 三种不同的信噪比进行合成,得到2304(192×3×4)段含噪语音,采用全部的含噪语音段来构成测试集.

5.1.2　语音质量比较

为了比较不同网络的语音增强性能,首先对不同网络增强后语音的语音质量进行客观评估,所采用的指标为语音质量感知评估PESQ(Perceptual Evaluation of Speech Quality)[25]. PESQ的得分范围为 -0.5 到4.5,得分越高表示语音质量越高.

表1　TIMIT数据下不同网络增强语音的平均PESQ

噪声	信噪比(dB)	NOISY	DNN	LSTM	GRU	CGRN
N1	-7	1.62	2.08	2.24	2.15	2.32
	0	2.08	2.62	2.74	2.71	2.85
	7	2.54	3.06	3.10	3.10	3.24
N2	-7	1.29	1.56	1.81	1.72	1.89
	0	1.63	2.07	2.34	2.31	2.44
	7	2.08	2.57	2.80	2.79	2.90
N3	-7	1.49	1.70	1.85	1.91	2.18
	0	1.81	2.21	2.38	2.44	2.70
	7	2.21	2.70	2.83	2.87	3.12
N4	-7	1.30	1.45	1.67	1.69	2.03
	0	1.57	1.82	2.06	2.14	2.55
	7	1.97	2.30	2.49	2.59	2.97
平均		1.80	2.18	2.36	2.37	2.60

表1给出了在TIMIT数据下测试集含噪语音分别采用四种不同的网络进行语音增强后,在4类噪声和3种不同信噪比下的平均PESQ得分,并给出了相应的含噪语音的平均PESQ得分作为对比. 可见:在所有噪声条件下,四种网络增强后语音的平均PESQ得分相比含噪语音均有所提升,表明四种网络均能够有效提高含噪语音的质量;四种网络中,CGRN增强后语音的平均PESQ得分最高,LSTM和GRU次之,DNN最低,表明CGRN增强后的语音具有最高的平均语音质量.

5.1.3　语音可懂度比较

为了进一步比较不同网络的语音增强性能,对不同网络增强后语音的可懂度进行客观评估,所采用的指标为短时客观可懂度STOI(Short Time Objective Intelligibility)[26]. STOI的得分范围为0到1,得分越高表示语音的可懂度越好.

表2给出了TIMIT数据下测试集含噪语音分别采用四种不同的网络进行语音增强后,在4类噪声和3种不同信噪比下的百分比形式的平均STOI得分,并给出了相应的含噪语音的平均STOI得分作为对比. 可见:相比含噪语音,在所有噪声条件下,LSTM、GRU和CGRN增强后语音的平均STOI得分均有明显提升,表明三种网络均能有效提高含噪语音的可懂度;而DNN增强后语音的平均STOI得分相比含噪语音提升非常有限,在低信噪比条件(-7dB N2和-7dB N3)下甚至出现了下降,表明DNN在低信噪比下不能有效提高含噪语音的可懂度. 另外,综合所有噪声条件来看,CGRN增强后语音的平均STOI得分是最高的,LSTM和GRU次之,DNN最低,表明CGRN增强后的语音具有最好的平均可懂度.

表2　TIMIT数据下不同网络增强语音的平均STOI(%)

噪声	信噪比(dB)	NOISY	DNN	LSTM	GRU	CGRN
N1	-7	61.10	66.96	74.19	73.31	73.34
	0	76.16	82.32	85.88	85.44	84.98
	7	87.48	90.12	91.32	90.91	91.05
N2	-7	47.69	46.52	62.37	61.18	61.73
	0	62.50	65.75	77.69	77.39	76.56
	7	79.06	82.09	87.61	87.32	86.51
N3	-7	52.62	52.37	60.49	62.84	68.94
	0	68.91	72.89	78.80	79.55	83.11
	7	84.35	86.51	88.91	88.91	90.63
N4	-7	52.29	52.78	65.61	66.52	67.55
	0	69.04	71.63	79.04	79.92	80.71
	7	84.27	84.78	87.63	87.89	88.36
平均		68.79	71.23	78.29	78.43	79.46

5.1.4　语谱图比较

为了更加直观的比较TIMIT数据下不同网络的语音增强性能,下面以一段含有N4噪声信噪比为0dB的含噪语音为例,对不同网络增强后语音的语谱图进行分析. 图3(*a*)给出了含噪语音的语谱图,图3(*b*)~(*e*)分别给出了采用DNN、LSTM、GRU和CGRN增强后语音的语谱图,图3(*f*)给出了相应的纯净语音的语谱图作为对比. 通过将四种网络的增强语音与纯净语音的语谱图进行比较可见:CGRN具有更好的噪声抑制能力,其它三种网络增强后的语音则存在明显的噪声成分. 上述结论在非正式的试听实验中同样得到验证.

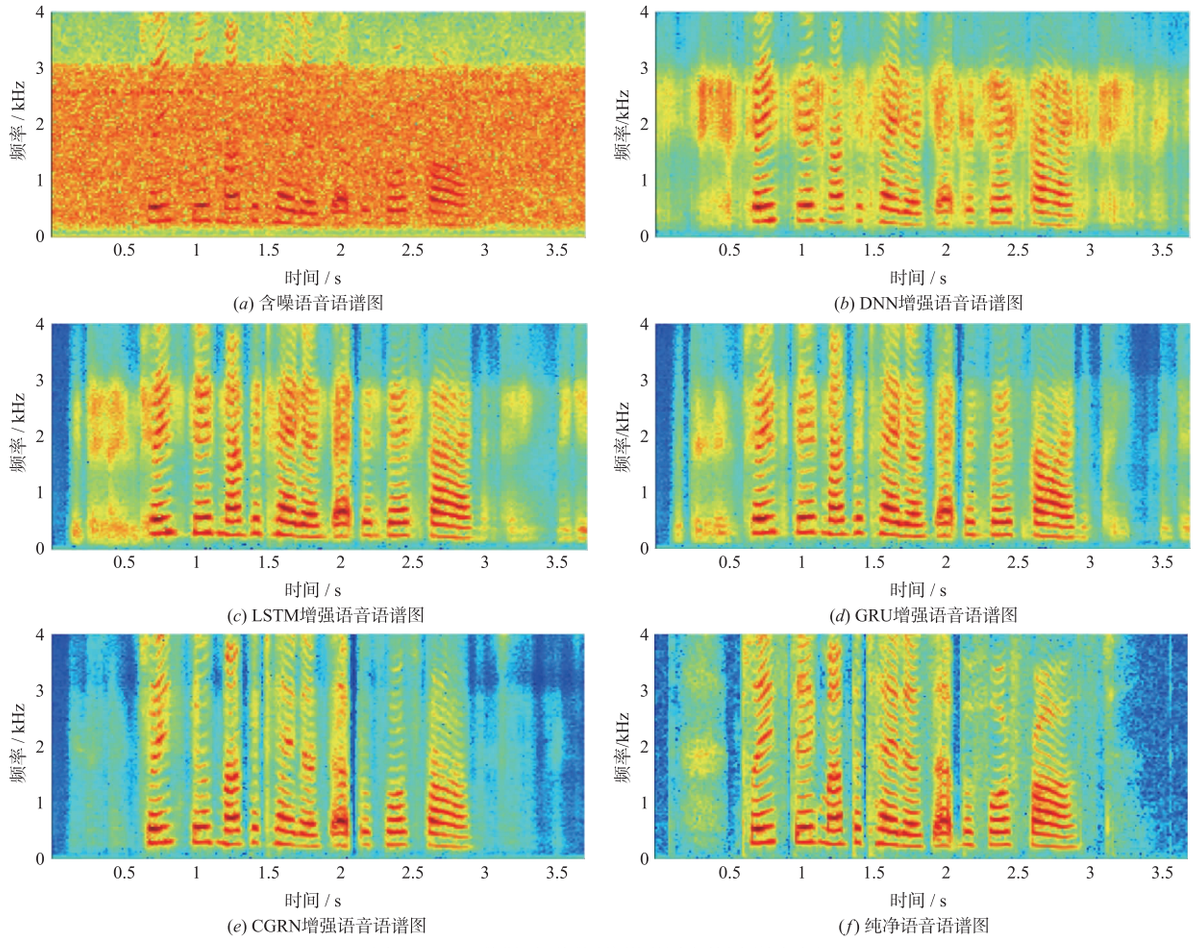


图3 TIMIT数据下不同网络增强语音的语谱图对比

5.2 THCHS-30 实验结果

5.2.1 训练集与测试集

该部分实验的纯净语音来自于清华大学的THCHS-30数据库的训练集,噪声同样采用来自于文献[20]中的100类真实噪声.将纯净语音与噪声的采样频率均转换为8kHz,然后按照-10dB、-5dB、0dB、5dB和10dB五种不同的信噪比进行合成,从合成得到的含噪语音中随机选取25000段来构成训练集.

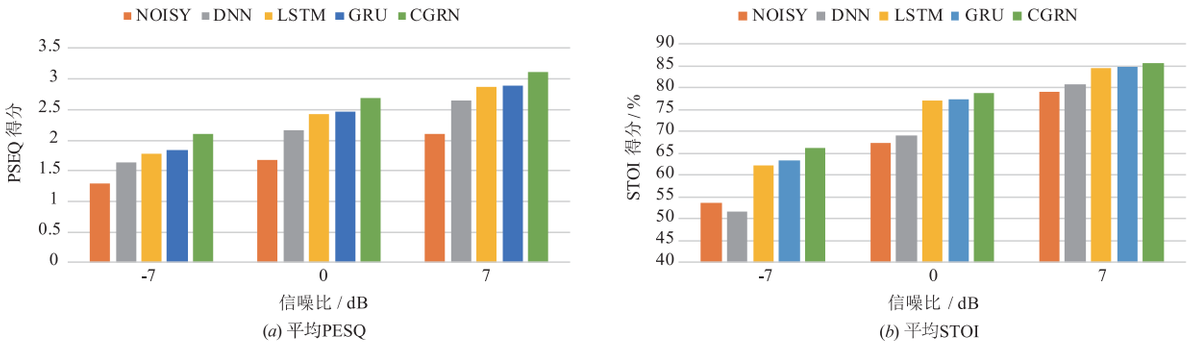
测试集所用的纯净语音是来自于THCHS-30数据库测试集的250段纯净语音,噪声同样采用N1(Factory2)、N2(Buccaneer1)、N3(Destroyer engine)和N4(HF channel).将纯净语音与噪声的采样频率均转换为8kHz,然后按照-7dB、0dB、7dB三种不同的信噪比进行合成,得到3000(250×3×4)段含噪语音,采用全部的含噪语音段来构成测试集.

5.2.2 语音质量和语音可懂度比较

图4(*a*)、(*b*)分别给出了THCHS-30数据下测试集含噪语音分别采用四种不同的网络进行语音增强后,在3种不同信噪比下的平均PESQ得分和百分比形式的平均STOI得分,并给出了含噪语音的相应得分作为对比.可见:与TIMIT数据下的实验结果相同,在两种不同指标下,CGRN在三种不同信噪比下都取得了最好的结果,LSTM和GRU的实验结果相近,DNN的实验结果最差.

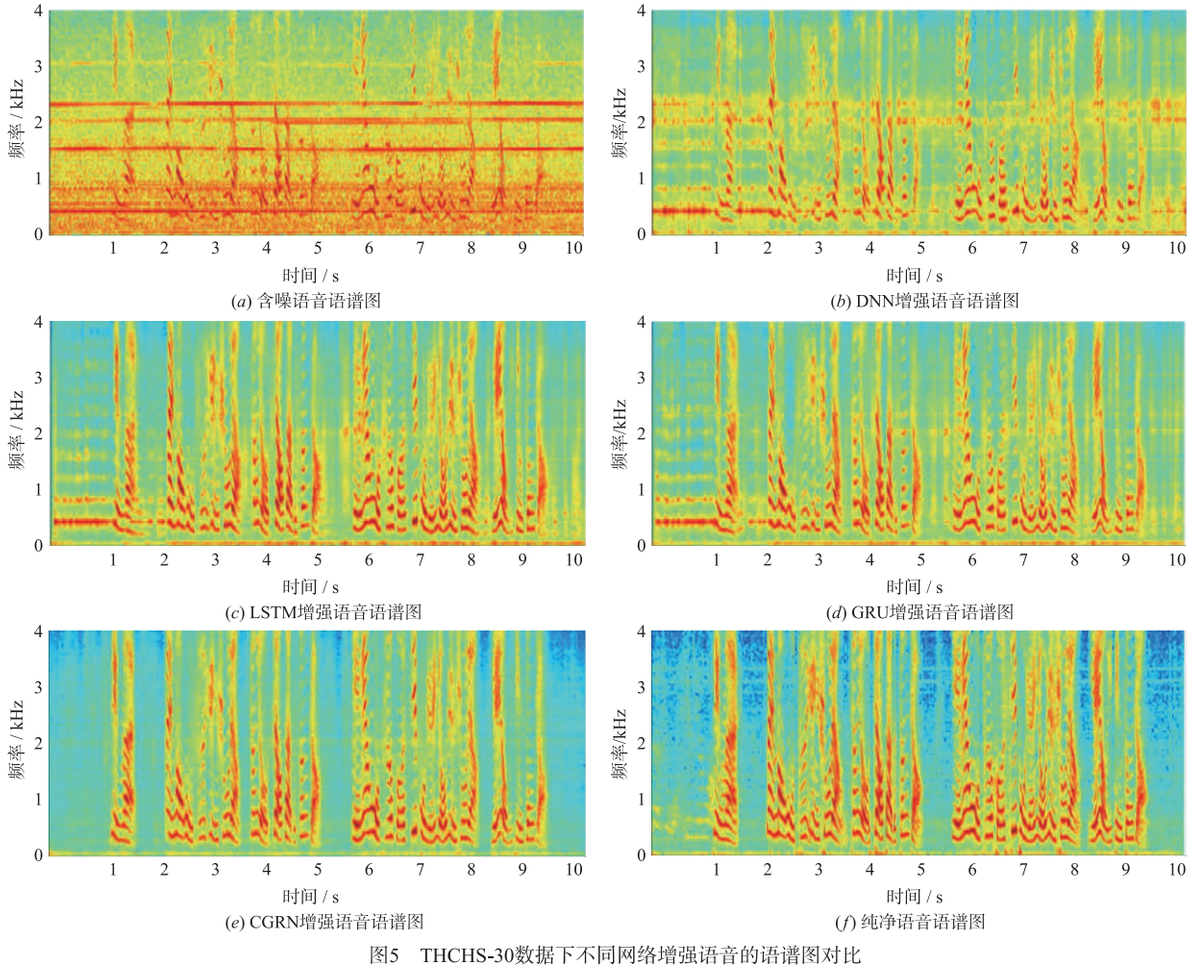
5.2.3 语谱图比较

为了更加直观地比较THCHS-30数据下不同网络的语音增强性能,下面以一段含有N3噪声信噪比为0dB的含噪语音为例,对不同网络增强后语音的语谱图进行分析.图5(*a*)给出了含噪语音的语谱图,图5(*b*)~(*e*)分别给出了采用DNN、LSTM、GRU和CGRN增强后语音的语谱图,图5(*f*)给出了相应的纯净语音的语谱图作为对比.通过将四种网络的增强语音与纯净语音的语谱图进行比较,可以得到与TIMIT数据下相同的结论:相比其它三种网络,CGRN明显具有更好的噪声抑制能力.上述结论在非正式的试听实验中同样得到验证.



(*a*) 平均PESQ　　(*b*) 平均STOI

图4　THCHS-30数据下不同网络增强语音的平均PESQ和平均STOI(%)



(*a*) 含噪语音语谱图　(*b*) DNN增强语音语谱图

(*c*) LSTM增强语音语谱图　(*d*) GRU增强语音语谱图

(*e*) CGRN增强语音语谱图　(*f*) 纯净语音语谱图

图5　THCHS-30数据下不同网络增强语音的语谱图对比

综合分析 TIMIT 数据和 THCHS-30 数据下的实验结果可知:相比 DNN、LSTM 和 GRU 三种网络,CGRN 无论是在增强语音质量和可懂度的客观评价上,还是在语谱图的主观分析上,都取得了更好的结果,表明 CGRN 具有更好的语音增强性能. 另外,表 3 给出了四种不同网络的参数规模,可见:CGRN 与 GRU 的参数量近似,少于 DNN 和 LSTM,表明 CGRN 在没有明显增加网络参数量的前提下提高了语音增强性能.

表 3　不同网络的参数规模

模型	DNN	LSTM	GRU	CGRN
参数量(M)	8.92	8.22	6.18	6.28

6　总结

虽然 GRU 利用门控机制能够有效的对含语音中的长期依赖关系进行建模,但是其中的特征计算均采用

全连接网络结构进行,破坏了含噪语音输入中的时频结构信息.针对这一问题,本文采用卷积网络结构来代替 GRU 中的全连接网络结构,提出了适用于语音增强的 CGRN,并采用不同语音数据集开展实验对 CGRN 的语音增强性能进行了评估.实验结果表明,相比 DNN、LSTM 和 GRU,CGRN 具有更好的语音增强性能,增强后语音的残留噪声更少,语音质量和可懂度更好.

参考文献

[1] 陈楠,鲍长春. 基于双耳线索编码原理的语音增强方法[J]. 电子学报,2019,47(1):227 – 233.
CHEN Nan, BAO Chang-chun. Speech enhancement method based on binaural cues coding principle[J]. Acta Electronica Sinica, 2019, 47(1): 227 – 233. (in Chinese)

[2] OU Shifeng, SONG Peng, GAO Ying. Laplacian speech model and soft decision based MMSE estimator for noise power spectral density in speech enhancement[J]. Chinese Journal of Electronics, 2018, 27(6): 1214 – 1220.

[3] 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.
LIU Wenju, NIE Shuai, LIANG Shan, et al. Deep learning based speech separation technology and its developments [J]. Acta Automatica Sinica, 2016, 42(6): 819 – 833. (in Chinese)

[4] WANG D L, CHEN J. Supervised speech separation based on deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702 – 1726.

[5] WANG Y, WANG D L. Towards scaling up classification-based speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1381 – 1390.

[6] WANG Y, NARAYANAN A, WANG D L. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849 – 1858.

[7] WILLIAMSON D S, WANG D L. Time-frequency masking in the complex domain for speech dereverberation and denoising[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(7): 1492 – 1501.

[8] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 2014, 21(1): 65 – 68.

[9] 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, 2015, 23(1): 7 – 19.

[10] 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 Transactions on Audio, Speech, and Language Processing, 2015, 23(12): 2136 – 2147.

[11] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[A]. Proceedings of International Conference on Latent Variable Analysis and Signal Separation[C]. Liberec: Springer International Publishing, 2015. 91 – 99.

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

[13] PARK S R, LEE J. A fully convolutional neural network for speech enhancement[A]. Proceedings of the Eighteenth Annual Conference of the International Speech Communication Association[C]. Stockholm: ISCA, 2017. 1993 – 1997.

[14] FU S W, TSAO Y, LU X. SNR-aware convolutional neural network modeling for speech enhancement[A]. Proceedings of the Seventeenth Annual Conference of the International Speech Communication Association[C]. California: ISCA, 2016. 3768 – 3772.

[15] TAN K, CHEN J, WANG D. Gated residual networks with dilated convolutions for supervised speech separation [A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. Alberta: IEEE, 2018. 21 – 25.

[16] TAN K, CHEN J, WANG D. Gated residual networks with dilated convolutions for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 189 – 198.

[17] LI Y, LI X, DONG Y, LI M, XU S, XIONG S. Densely connected network with time-frequency dilated convolution for speech enhancement[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. Brighton: IEEE, 2019. 6860 – 6864.

[18] 杨绪魁,屈丹,张文林,闫红刚. 基于长时信息的自适应话音激活检测[J]. 电子学报,2018,46(4):878 – 885.
YANG Xu-kui, QU Dan, ZHANG Wen-lin, YAN Hong-gang. Adaptive voice activity detection based on long-term information[J]. Acta Electronica Sinica, 2018, 46(4): 878 – 885. (in Chinese)

[19] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[A]. Advances in Neural Information Processing Systems[C]. Morgan Kaufmann, 2015. 802 – 810.

[20] BALLAS N, YAO L, PAL C, et al. Delving deeper into convolutional networks for learning video representations [J]. arXiv Preprint, 2015, arXiv:1511.06432.
[21] GAROFOLO J S, LAMEL L F, FISHER W M, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus[M]. Linguistic Data Consortium, Philadelphia, 1993, 33.
[22] WANG D, ZHANG X, ZHANG Z. THCHS-30: A Free Chinese Speech Corpus [OL]. http://arxiv.org/abs/1512.01882, 2015/2019-08-10.
[23] HU G. 100 Nonspeech Environmental Sounds[OL]. http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html, 2004/2019-08-10.
[24] 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.
[25] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. Utah: IEEE, 2001. 749 – 752.
[26] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125 – 2136.

作者简介

袁文浩(通信作者)　男,1985 年出生,山东寿光人. 2013 年毕业于华东理工大学获博士学位,现为山东理工大学计算机科学与技术学院讲师. 主要研究方向为语音信号处理,语音增强.
E-mail: why_sdut@126.com

胡少东　男,1996 年出生,山东泰安人. 2019 年毕业于山东理工大学,现为山东理工大学计算机科学与技术学院硕士研究生. 主要研究方向为语音信号处理,语音增强.
E-mail: 1764513896@qq.com