注意力机制并不保证最佳的语音增强效果

Zhongshu Hou[1,2,3], Qinwen Hu[1,2,3], Kai Chen[1,2,3], Jing Lu[1,2,3]

[1] 南京大学现代声学重点实验室，中国南京 210093[2] 南大—地平线智能音频实验室，
地平线机器人，中国北京 100094[3] 南京先进人工智能研究院，中国南京 210014

{zhongshu.hou, qinwen.hu}@smail.nju.edu.cn, {chenkai, lujing}@nju.edu.cn

摘要

由于理论上注意力机制能够有效地对时域和频谱域中信号的长期内在联系进行建模，因此它已被广泛应用于语音增强（SE）。然而，通常使用的全局注意力机制可能并非最佳选择，因为在语音增强中，相邻信息的影响自然大于相距较远的信息。本文通过在两个典型的最先进（SOTA）模型中以循环神经网络（RNN）替换注意力机制，验证了这一猜想。这两个模型分别是采用轴向注意力的多尺度时频卷积网络（MTFAA）和基于 Conformer 的度量 GAN 网络（CMGAN）。

索引词—语音增强、注意力机制、循环神经网络

1. 引言

MTFAA 在 DNS4 [1], 中排名第一，而 CMGAN 在公开的 VoiceBank+DEMAND 数据集 [3], 上取得了最佳性能，因此被视为两种最先进的语音增强模型。它们都采用了注意力模块。本文将 MTFAA 在时域和频域中的注意力模块替换为循环神经网络，并将 CMGAN 在频域中的注意力模块替换为循环神经网络，通过比较替换前后系统的性能，验证注意力机制的有效性。

2. 实验

2.1. 数据集

我们在 DNS4 [4] 和 DCASE [5] 数据集上评估系统性能，其中干净片段选自 DNS4 语音数据集，噪声录音来自 DCASE 数据集。将房间脉冲响应与干净语音进行卷积，生成模拟混响语音，并保留早期混响语音作为训练目标。在训练阶段，混响语句与噪声片段以 -15 dB 至 0 dB 的信噪比混合；在测试阶段，信噪比为 -15dB 至 15dB。所有语句在我们的实验中均以 16 kHz 采样。

2.2. 参数设置与训练策略

短时傅里叶变换（STFT）的窗长和跳步长度分别为 32 ms 和 8ms。离散傅里叶变换长度为 512，并使用汉宁窗进行重叠相加波形重建。MTFAA 中的轴向注意力模块被双路径循环神经网络（DPRNN） [6] 替换（记为 MTFAA-DPRNN），CMGAN 在频域中的注意力模块则被传统循环神经网络（RNN）替换（记为 CMGAN-FRNN）。MTFAA 和 CMGAN 的其他参数分别见[1] 和[2], 。

训练时的批量大小为 8，输入音频的时长为 8 s。预热策略[7] 对于训练基于自注意力的模型至关重要，此时学习率α 按以下规则更新： $\alpha = \frac{1}{\sqrt{C}} \times \min\left(\frac{1}{\sqrt{\varphi}}, \frac{\varphi}{\sqrt{\Psi^3}}\right)$，其中$C = 64$、预热步数 $\Psi = 8000$ 和 φ 表示训练步数。我们使用基于预热的 Adam优化器训练模型，其中 $\beta_1 = 0.9$、 $\beta_2 = 0.98$ 和 $\epsilon = 10^{-9}$。压缩参数 γ 为1 3。

2.3. 评估指标

我们选取一组常用的语音指标来评估综合性能，即语音质量感知评估（PESQ） [8]， 短时客观可懂度（STOI） [9], 尺度不变信号失真比（SiSDR），以及基于复合平均意见分数（MOS）的指标[10] ，包括信号失真的 MOS预测（CSIG）、背景噪声侵扰度的MOS预测（ CBAK）和整体效果的MOS预测（COVL）。所有指标的值越高，表示性能越好。

2.4. 实验结果

表1. 实验结果

指标	PESQ				STOI				信号失真比 (dB)			
信噪比(dB)	-15~-5	-5~5	5~15	Avg.	-15~-5	-5~5	5~15	Avg.	-15~-5	-5~5	5~15	Avg.
含噪	1.122	1.052	1.142	1.105	0.432	0.628	0.793	0.617	-12.849	-2.540	5.131	-3.419
MTFAA	1.215	1.657	2.163	1.678	0.591	0.803	0.891	0.762	0.131	7.093	10.263	5.829
MTFAA-DPRNN	1.291	1.831	2.396	1.839	**0.627**	0.831	0.910	0.789	**1.672**	8.380	**11.699**	**7.250**
CMGAN	1.279	1.797	2.375	1.817	0.617	0.828	0.907	0.784	1.340	8.323	10.842	6.835
CMGAN-FRNN	**1.321**	**1.876**	**2.460**	**1.886**	0.625	**0.836**	**0.913**	**0.791**	1.403	**8.394**	11.309	7.035

指标	CSIG				CBAK				COVL			
信噪比(dB)	-15~-5	-5~5	5~15	Avg.	-15~-5	-5~5	5~15	Avg.	-15~-5	-5~5	5~15	Avg.
含噪	1.098	1.333	1.926	1.452	1.036	1.240	1.724	1.333	1.036	1.113	1.447	1.199
MTFAA	2.058	2.821	3.376	2.752	1.801	2.240	2.593	2.211	1.540	2.184	2.738	2.154
MTFAA-DPRNN	**2.252**	**3.075**	3.650	2.992	1.890	2.400	2.775	2.355	**1.680**	2.411	3.001	2.364
CMGAN	2.151	3.029	3.678	2.952	1.890	2.518	2.941	2.450	1.624	2.370	3.009	2.334
CMGAN-FRNN	2.218	3.056	**3.708**	**2.994**	**1.909**	**2.557**	**2.987**	**2.484**	1.679	**2.425**	**3.067**	**2.390**

从表1可以看出，将注意力机制替换为循环神经网络后，MTFAA和CMGAN在所有指标上的表现均有所提升。

4. 结论

本文将两种最先进的语音增强模型中的注意力模块替换为循环神经网络，并验证了注意力机制并不一定能带来最佳性能。

参考文献

[1] 张国昌、余立标、王春亮和魏建强，“采用轴向注意力机制的多尺度时频卷积网络用于语音增强”，发表于国际声学、语音与信号处理会议。电气电子工程师学会，2022年，第9122–9126页。[2] S. Abdulatif、曹瑞哲和杨彬，“CMGAN：用于单声道语音增强的基于Conformer的度量生成对抗网络”，arXiv预印本 arXiv:2209.11112，2022年。[3] C. V. Botinhao、王欣，“探讨用于抗噪文本转语音的基于循环神经网络的语音增强方法”，发表于语音合成研讨会，2016年。[4] Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., ... & Aichner, R.（2022年5月）。ICASSP 2022 深度噪声抑制挑战赛。发表于2022年国际声学、语音与信号处理会议（ICASSP）（第9271–9275页）。电气电子工程师学会。[5] Dohi, K., Imoto, K., Harada, N., Niizumi, D., Koizumi, Y., Nishida, T., ... & Kawaguchi, Y.（2022）。DCASE 2022挑战赛任务2说明与讨论：应用领域泛化技术进行机器状态监测的无监督异常声音检测。arXiv预印本 arXiv:2206.05876。

[6] Luo, Y., Chen, Z., & Yoshioka, T.（2020年5月）。双路径循环神经网络：用于时域单声道语音分离的高效长序列建模。发表于2020年国际声学、语音与信号处理会议（ICASSP）（第46–50页）。电气电子工程师学会。[7] A. Nicolson 和 K. K. Paliwal，“用于因果语音增强的掩码多头自注意力机制”，《Speech Communication》，第125卷，第80–96页，2020年。[8] A. W. Rix、J. G. Beerends 和 M. P. Hollier，“语音质量感知评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法”，发表于国际声学、语音与信号处理会议。电气电子工程师学会，2001年，第2卷，第749–752页。[9] C. H. Taal、R. C. Hendriks 和 R. Heusdens，“时频加权含噪语音的短时客观可懂度测量”，发表于国际声学、语音与信号处理会议。电气电子工程师学会，2010年，第4214–4217页。[10] 胡毅和 P. C. Loizou，“语音增强客观质量度量的评估”，《*IEEE*音频、语音与语言处理汇刊，第16卷，第1期，第229–238页，2007年。