ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095939

一种用于助听器的低延迟混合多通道语音增强系统

Tong Lei[1,2,], Zhongshu Hou[1,2,*], Yuxiang Hu[2], Wanyu Yang[2], Tianchi Sun[1,2], Xiaobin Rong[1,2], Dahan Wang[1,2], Kai Chen[1,2], and Jing Lu[1,2]*

[1] 南京大学声学研究所现代声学教育部重点实验室，中国南京 210093
[2]NJU-地平线智能音频实验室，地平线机器人技术有限公司，中国南京 210038 。

摘要

本文总结了一种用于 2023 年 ICASSP 信号处理大赛“清晰度挑战赛（助听器语音增强）”的混合多通道语音增强系统。该系统由基于规则的去混响模块、多通道增强模块和后处理模块组成。在不使用头部旋转信息和注册语音的情况下，该系统在官方开发集上可达到平均助听器语音感知指数（HASPI）得分为 0.696，助听器语音质量指数（HASQI）得分为 0.320。在用于挑战排名的 Eval1 数据集上，相应的得分分别为 0.729 和 0.316。

索引词 - 2023 年 ICASSP 清晰度挑战赛、语音增强、波束成形、助听器

1. 简介

2023 年 ICASSP 信号处理大赛：清晰度挑战赛（助听器语音增强）旨在促进助听器语音增强系统的竞争性。助听器设备每只耳朵应配备 3 个麦克风（前置、中置和后置），这意味着最多可以利用 6 个输入通道来增强所需的语音。在本文中，我们提出了一种结合基于规则的方法和数据驱动方法的混合多通道语音增强系统。采用在线加权预测误差（online WPE）方法[1]进行去混响，并使用具有短窗的嵌入与波束成形网络（EaBNet）[2]来推断空间滤波器。为了提高恢复语音的质量，我们设计了一种后置滤波器，能够减轻超低信干比（SIR）下的语音损失。

2. 所提议的系统

我们所提出的系统的示意图如图 1 所示。在线语音去混响模块采用了最先进的基于规则的在线 WPE（参见文献[1]）。对于多通道增强模型，我们使用具有短窗的因果 EaBNet（参见文献[2]），以保证低延迟。考虑到期望语音可能会被背景噪声和干扰语音淹没，导致信噪比极低，我们设计了一个后置滤波模块，以补偿在恶劣环境中的语音损失。

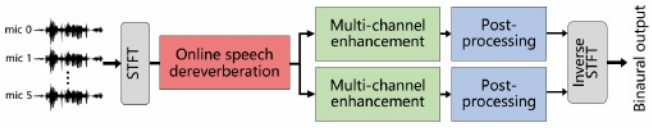

图 1. 我们为 2023 年 ICASSP 清晰度挑战赛所提出的混合系统的框图。

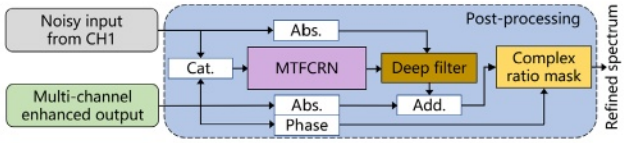

图 2. 基于 MTFCRN 的后处理。

如图 2 所示，将含噪信号的频谱图与左/右耳先前增强的语音进行拼接，然后送入后处理模型以推断出一个深度滤波器[5]和一个复比掩码。推断出的深度滤波器应用于含噪频谱的幅度，其输出与多通道增强输出的幅度相加。然后，将多通道增强输出的幅度补偿值与相位谱一起进一步通过复比掩码进行处理。理论上，任何具有短时傅里叶变换（STFT）输入的语音增强模型都可以用作后处理模型。在本文中，我们选择对多尺度时频卷积轴向自注意力网络（MTFAA）[3]进行修改，将主网络替换为两个双路径循环神经网络（DPRNNs）[4]，并将修改后的模型命名为多尺度时频卷积循环网络（MTFCRN）。

通过不同组合方式构建了四个系统，分别是：EaBNet、EaBNet+DRB、EaBNet+MTFCRN 以及 EaBNet+DRB+MTFCRN，这些组合包含了去混响模块（DRB）、空间滤波模块（EaBNet）以及后处理模块（MTFCRN）。

3. 系统信息及主要特点

3.1. 培训流程

对于随意的 EaBNet 和 MTFCRN，均采用 Adam 优化器进行训练。EaBNet 的初始学习率设为 0.01，若验证损失在 6 个周期内未见改善，则以 0.8 的衰减率逐渐降低。对于 MTFCRN 的训练，则采用了一种预热策略[6]。学习率 αα 随训练步数 ξ 的变化而变化，其变化规律为 αα= √C 1 min（ξ^ψ 1, ξ^3），其中模型大小 CC= 128。

* These authors contributed equally to this work.

预热步骤ψψ为 40000。使用 Adam 优化器，其参数为 ββ$_1$ = 0.9，ββ$_2$ = 0.98， = 10⊠⊠。压缩参数 γγ 为 1。网络使用输入信号段进行训练，要求目标信号和干扰信号至少各持续 1 秒以保证可感知性。

3.2. 损失函数

EaBNet 和 MTFCRN 均是在功率压缩均方误差损失和普通均方误差损失的混合损失上进行训练的。

$$L = p\left|\left|\boldsymbol{S}\right|^{q} - \left|\hat{\boldsymbol{S}}\right|^{q}\right|^{2} + \left|\boldsymbol{S} - \hat{\boldsymbol{S}}\right|^{2}, \quad (1)$$

其中训练目标 $\boldsymbol{S}$ 表示 CH1 去混响信号的频谱，$\boldsymbol{S}$⊠ 表示估计频谱。对于 EaBNet，权重 p 为 2，权重 q 为 0.5；对于 MTFCRN，权重 p 为 7，权重 q 为 0.3。

3.3. 时间延迟

所有模块均在短时傅里叶变换（STFT）域中实现并级联，采用 220 个样本的汉宁窗和 110 个样本的跳距，采样率为 44.1 kHz。总时间延迟为 220/44100 秒，小于本次挑战的要求（5 毫秒）。

3.4. 计算资源

对于每只耳朵，EaBNet 和 MTFCRN 的每秒浮点运算次数（FLOPs）分别为 16.36G 和 8.71G，参数数量分别为 263 万和 200 万。计算统计数据基于 *ptflops v0.6.9 版本*。

3.5. 数据库

ICASSP 2023 官方清晰度挑战数据集包含 6000 个场景用于训练，2500 个场景用于验证，3000 个场景用于评估。每个场景包括 6 通道的耳背式助听器设备录音、头部旋转信号以及一位目标说话人的 4 段短注册语音。我们仅在核心数据库上训练模型，未利用头部旋转数据和注册语音。

Eval1 和 Eval2 数据集各包含 1500 个场景，是主办方用于评估参赛团队所开发模型效果的主要评估集。Eval1 是以与开发集类似的方式生成的评估集，但使用了不同的数据。而具有生态有效性的 Eval2 则包含从真实房间中说话者那里获取的更具挑战性的大量录音。

4. 结果与讨论

表 1 展示了开发集和评估集上的平均 SI-SDR、HASPI 和 HASQI 值。对于双耳客观评估，Clarity 挑战赛采用 HASPI 和 HASQI 作为排名指标，而 SI-SDR 则作为开发集上模型选择的辅助指标。值得注意的是，SI-SDR 是在基于官方 NAL-R 的[7]助听器放大处理之前计算得出的，而 HASPI 和 HASQI 则是在放大处理之后计算得出的。所有指标均使用来自 https://github.com/claritychallenge/clarity 的官方工具进行计算。

表 1.开发和评估测试集的结果。

系统	集合	SI-SDR	HASPI	HASQI	联合的
混合	开发人员	-11.3	0.249	0.138	0.193
	评估1	-	0.266	0.128	0.197
	评估2	-	0.176	0.121	0.149
EaBNet	开发人员	-1.30	0.642	0.291	0.466
+ MTFCRN	开发人员	-1.25	0.661	0.301	0.481
+DRB	开发人员	2.06	0.676	0.308	0.492
+ DRB + MTF CRN	开发人员	2.65	0.696	0.320	0.508
	评估1	-	0.729	0.316	0.522
	评估2	-	0.284	0.133	0.208

我们的消融实验表明，EaBNet+DRB+MTFCRN 系统在开发集的所有指标上都表现出最佳性能，因此我们选择它作为我们最终提交的系统（E030）。尽管我们提交的系统在 Eval1 和 Eval2 中均排名前五，但遗憾的是在 Eval3 中未能进入前五名。

在 Eval1 和 Eval2 数据集上进行评估时，必须指出的是，当在更具挑战性的 Eval2 数据集上进行评估时，其优势会显著降低。

5. 结论

在本文中，我们提出了一种混合多通道助听器语音增强系统，该系统结合了基于规则的去混响、神经空间滤波器以及能够补偿超低信噪比下语音损失的后处理。消融实验的结果表明，去混响模块和后滤波处理模块在评估指标上均有显著提升。我们提交的系统在 Eval1 和 Eval2 数据集上均排名前五。

6. 致谢

本研究得到了国家自然科学基金（项目编号：12274221）的支持。

7. 参考文献

[1] R. Ikeshita, K. Kinoshita, N. Kamo, and T. Nakatani, "Online speech dereverberation using mixture of multichannel linear prediction models," *IEEE Signal Processing Letters*, vol. 28, pp. 1580-1584, 2021.

[2] A. Li, W. Liu, C. Zheng, and X. Li, "Embedding and beamforming: All-neural causal beamformer for multichannel speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6487-6491, 2022.

[3] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 9122-9126, 2022.

[4] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 46-50, 2020.

[5] W. Mack and E. A. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61-65, 2019.

[6] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] D. Byrne and H. Dillon, "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257-265, 1986.