上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

面向通用场景的语音增强研究

——URGENT 系列挑战赛回顾

张王优

上海交通大学 人工智能学院

2025.10.18

听觉认知与计算声学实验室
Auditory Cognition and Computational Acoustics Laboratory @ SJTU

个人介绍

张王优，上海交通大学助理教授、博导

教育和工作经历

- 2014 年~2018 年　华中科技大学　学士学位
- 2018 年~2024 年　上海交通大学　博士学位
- 2025 年~至今　上海交通大学　助理教授

研究方向

- 语音信号处理：语音增强与修复、语音质量评估
- 语音和音频理解：音频表征学习、鲁棒语音理解

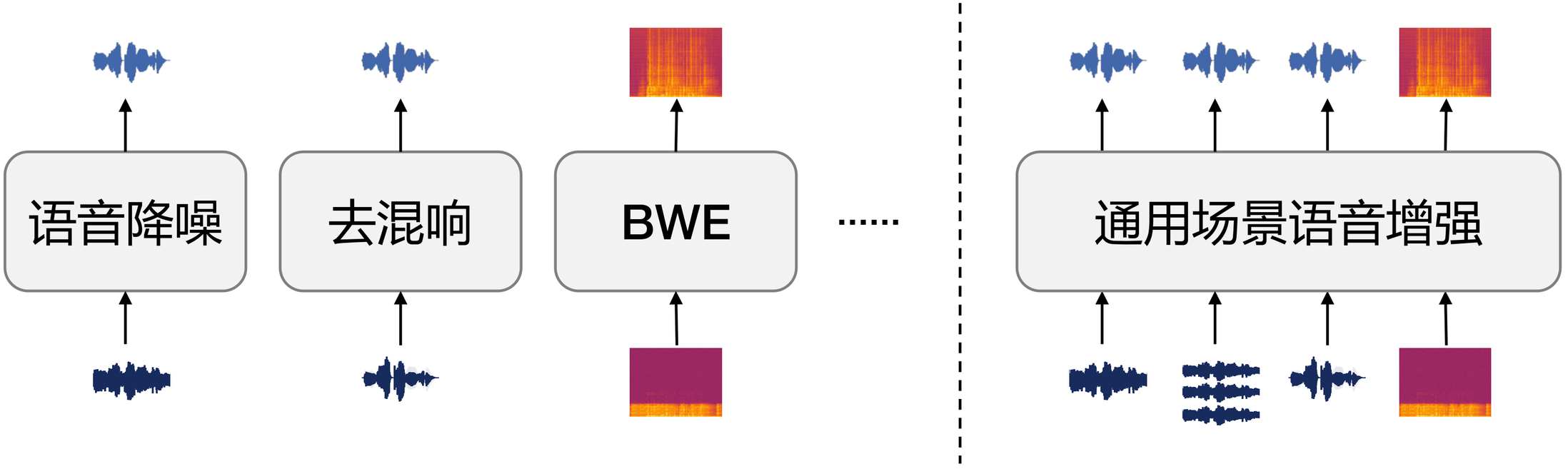
目录

CONTENTS

(1) 面向通用场景的语音增强

任务定义

- 传统的语音增强：用专门的模型处理专门的子任务/场景
- 面向通用场景：　用一个模型统一处理多种子任务/场景



(1) 面向通用场景的语音增强

动机

1. 语音信号的复杂性

环境多样性

- 室内、室外
- 背景噪声、混响
- 自噪声
- 干扰人声

人的多样性

- 性别、年龄
- 口音、语种
- 说话风格
- 人数

信道多样性

- 设备质量
- 麦克风数量、布局
- 采样率
- 静止、移动

(1) 面向通用场景的语音增强

动机

2. 通用模型的研究趋势

- 多项研究发现：在特定方面，通用模型 > 专用模型
  - LID + ASR[1]，SLU[2]，ASR + TTS[3]，ASR + TTS + LID + GID[4]
  - ASR + VSR + AVSR[5]，Diarization + Separation + ASR[6]

[1] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in Proc. IEEE ASRU, 2017, pp. 265–271.
[2] S. Arora, H. Futami, J.-w. Jung, Y. Peng, R. Sharma, Y. Kashiwagi, E. Tsunoo, K. Livescu, and S. Watanabe, “UniverSLU: Universal spoken language understanding for diverse tasks with natural language instructions,” in Proc. ACL, 2024, pp. 2754–2774.
[3] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, “VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks,” in Proc. IEEE ICASSP, 2024, pp. 13326–13330.
[4] R. Yang, H. Yang, X. Zhang, T. Ye, Y. Liu, Y. Gao, S. Zhang, C. Deng, and J. Feng, “PolySpeech: Exploring unified multitask speech models for competitiveness with single-task models,” arXiv preprint arXiv:2406.07801, 2024.
[5] A. Haliassos, R. Mira, H. Chen, Z. Landgraf, S. Petridis, and M. Pantic, “Unified speech recognition: A single model for auditory, visual, and audiovisual inputs,” Advances in Neural Information Processing Systems, vol. 37, pp. 139673–139699, 2024.
[6] M. Shakeel, Y. Sudo, Y. Peng, C.-J. Lin, and S. Watanabe, “Unifying diarization, separation, and ASR with multi-speaker encoder,” Accepted by ASRU 2025.

(1) 面向通用场景的语音增强

动机

3. 实际应用的需求

- 维护成本：单个通用模型 < 多个专用模型
- 可靠性/稳定性
- 尺度定律 ⇦ 通用模型更容易扩大规模
- 可扩展性：通用模型 > 专用模型

(1) 面向通用场景的语音增强

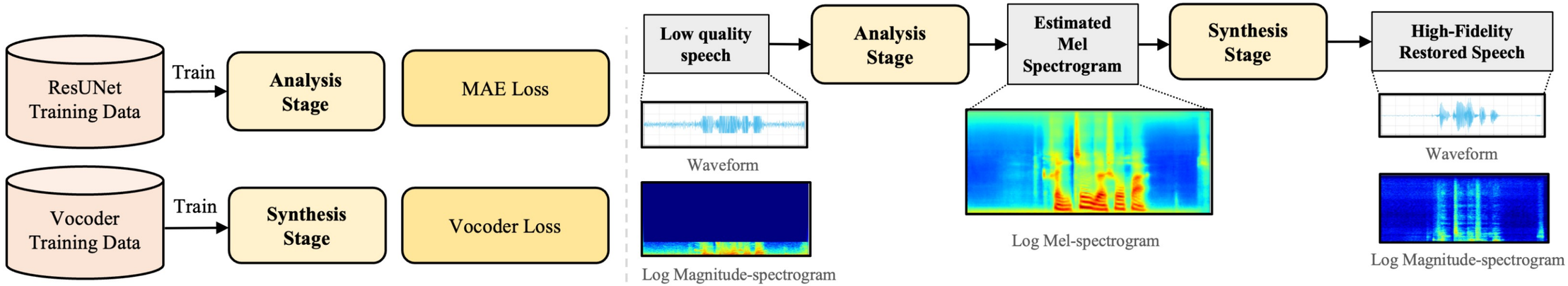
核心问题

1. 如何统一不同子任务？ 数据准备 建模方法
2. 如何应对不同场景数据的差异？ 模型架构 数据调度
3. 如何解决不同子任务的优化冲突？ 优化准则 训练策略
4. 如何准确评估通用场景的增强性能？ 评估指标 benchmark

(1) 面向通用场景的语音增强

如何统一不同子任务?

- VoiceFixer[7]
  - 基于特征增强 + Vocoder 的生成式方法
  - 降噪 + 去混响 + Declipping + BWE

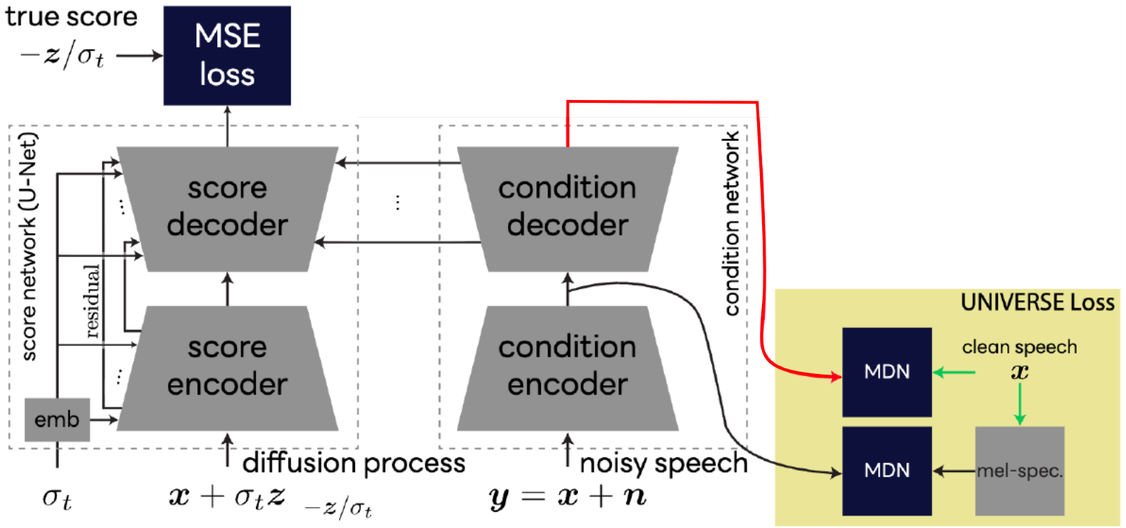


[7] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "VoiceFixer: A unified framework for high-fidelity speech restoration," in Proc. ISCA Interspeech, 2022, pp. 4232–4236.

(1) 面向通用场景的语音增强

如何统一不同子任务?

- UNIVERSE[8]
  - 基于 score-based diffusion 的生成式方法
  - 降噪 + 去混响 + 语音修复（50 余种失真）

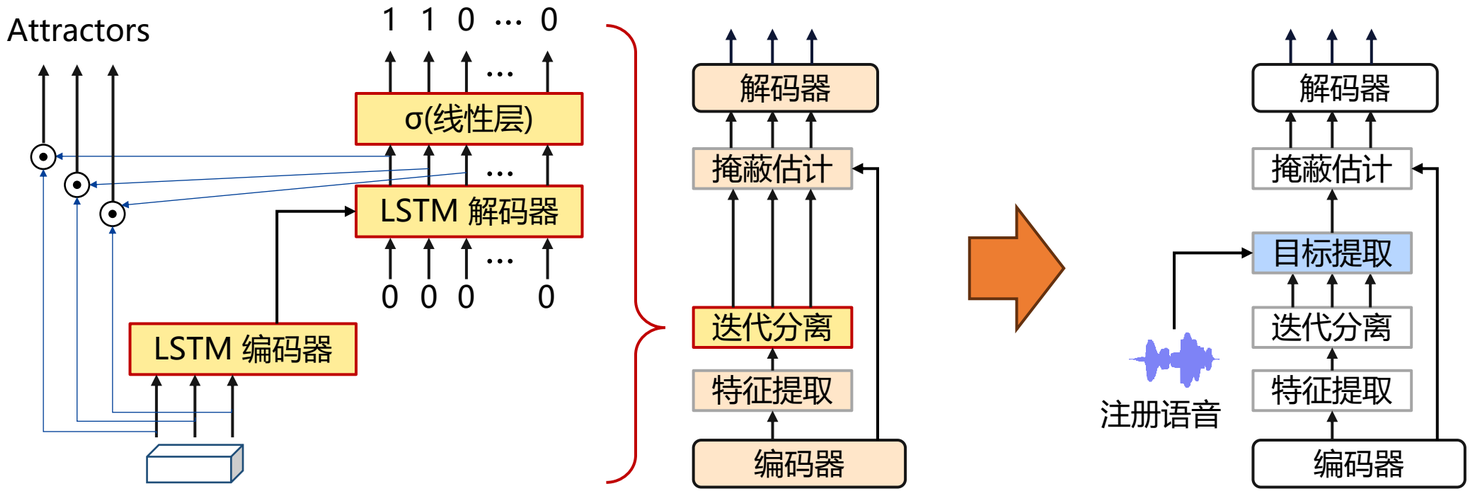


[8] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, “Universal speech enhancement with score-based diffusion,” arXiv preprint arXiv:2206.03065, 2022.

(1) 面向通用场景的语音增强

如何统一不同子任务?

- MUSE[9]
  - 基于 encoder-decoder-attractor (EDA) 的鉴别式方法
  - 语音降噪 + 去混响 + 说话人计数 + 语音分离 + 目标人分离

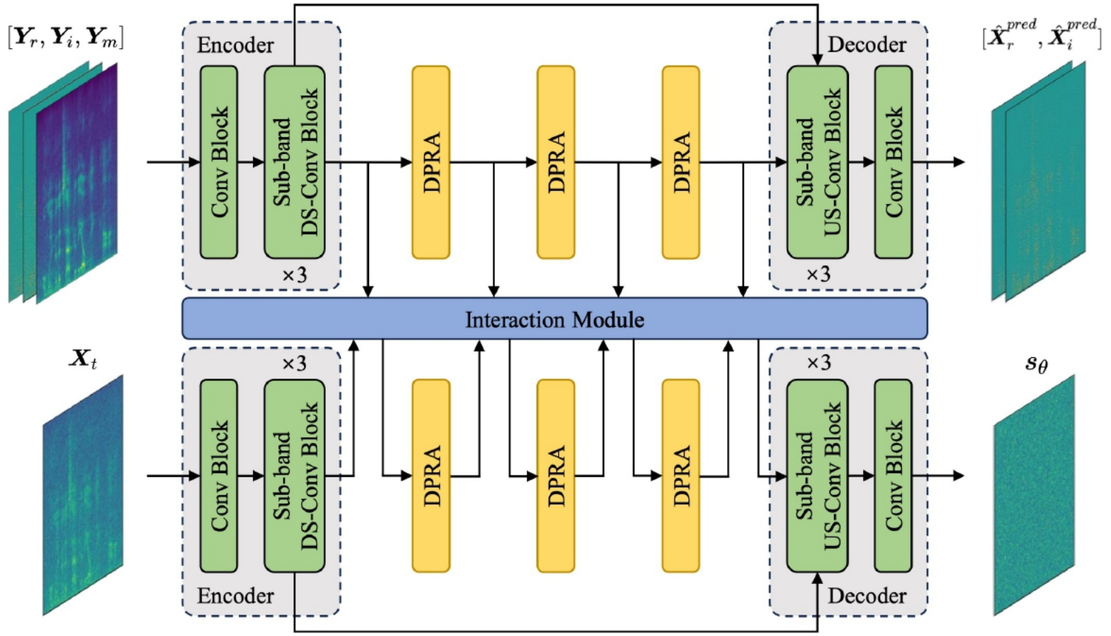


[9] K. Saijo, W. Zhang, Z.-Q. Wang, S. Watanabe, T. Kobayashi, and T. Ogawa, “A single speech enhancement model unifying dereverberation, denoising, speaker counting, separation, and extraction,” in Proc. IEEE ASRU, 2023.

(1) 面向通用场景的语音增强

如何统一不同子任务?

- PGUSE[10]
  - 基于 score-based diffusion + predictive 的鉴别+生成式方法
  - 语音降噪 + 去混响 + 语音修复

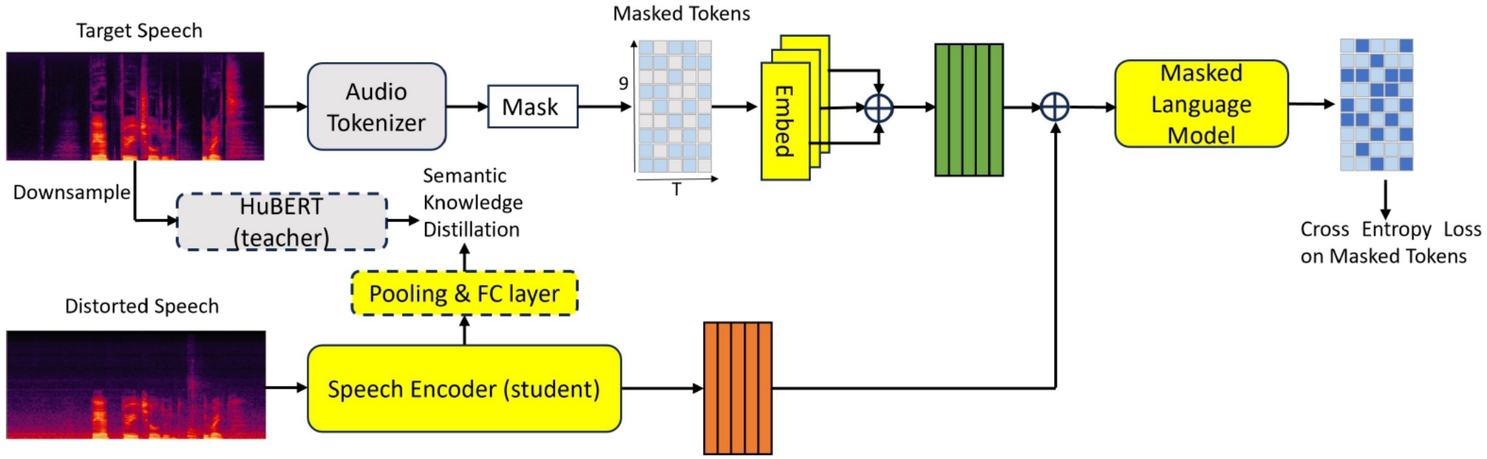


[10] J. Zhang, H. Yan, and X. Li, "A composite predictive-generative approach to monaural universal speech enhancement," IEEE Trans. Audio, Speech, Language Process., vol. 33, pp. 2312–2325, 2025.

(1) 面向通用场景的语音增强

如何统一不同子任务?

- MaskSR[11], MaskSR2[12], AnyEnhance[13]
  - ❑ 基于 masked LM + Decoder 的生成式方法
  - ❑ 语音降噪 + 去混响 + 语音修复 + 目标人分离



[11] X. Li, Q. Wang, and X. Liu, “MaskSR: Masked language model for full-band speech restoration,” in Proc. ISCA Interspeech, 2024, pp. 2275–2279.

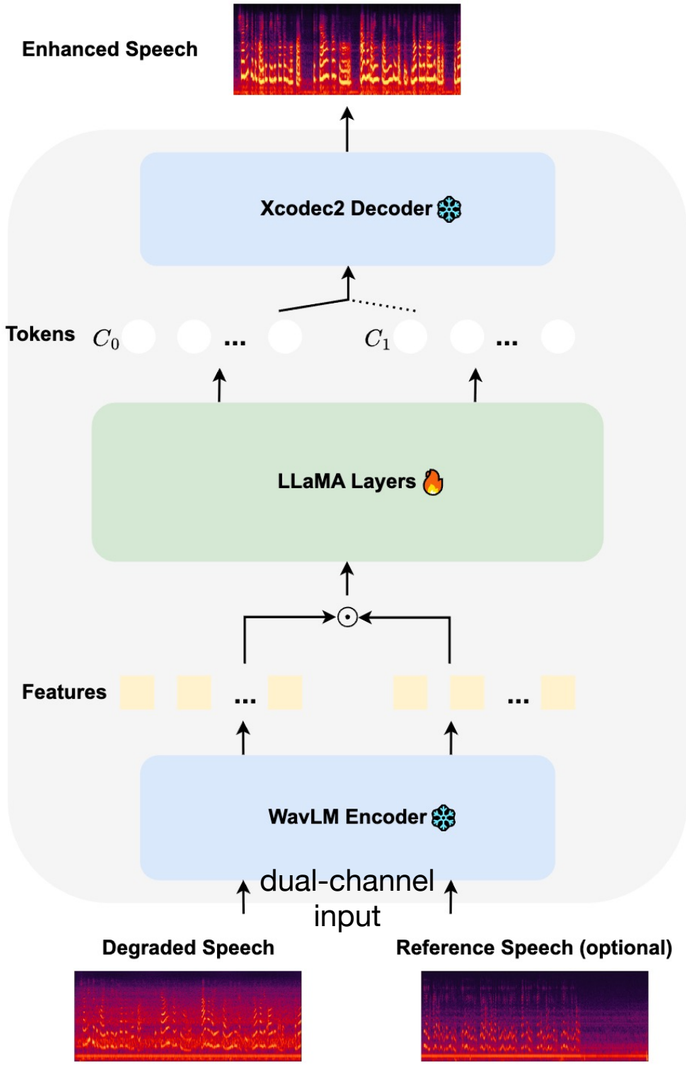
[12] X. Liu, X. Li, J. Serrà, and S. Pascual, “Joint semantic knowledge distillation and masked acoustic modeling for full-band speech restoration with improved intelligibility,” in Proc. IEEE ICASSP, 2025.

[13] J. Zhang, J. Yang, Z. Fang, Y. Wang, Z. Zhang, Z. Wang, F. Fan, and Z. Wu, “AnyEnhance: A unified generative model with prompt-guidance and self-critic for voice enhancement,” IEEE Trans. Audio, Speech, Language Process., vol. 33, pp. 3085–3098, 2025.

(1) 面向通用场景的语音增强

如何统一不同子任务?

- LLaSE-G1[14], SenSE[15]
  - 基于自回归 LM + Decoder 的生成式方法
  - 语音降噪 + 去混响 + AEC + PLC + 语音分离 + 目标人分离



[14] B. Kang, X. Zhu, Z. Zhang, Z. Ye, M. Liu, Z. Wang, Y. Zhu, G. Ma, J. Chen, L. Xiao, C. Weng, W. Xue, and L. Xie, "LLaSE-G1: Incentivizing generalization capability for LLaMA-based speech enhancement," in Proc. ACL, 2025, pp. 13292–13305.

[15] X. Li, H. Xie, Z. Wang, Z. Zhang, L. Xiao, and L. Xie, "SenSE: Semantic-aware high-fidelity universal speech enhancement," arXiv preprint arXiv:2509.24708, 2025.

(1) 面向通用场景的语音增强

如何应对不同场景数据的差异?

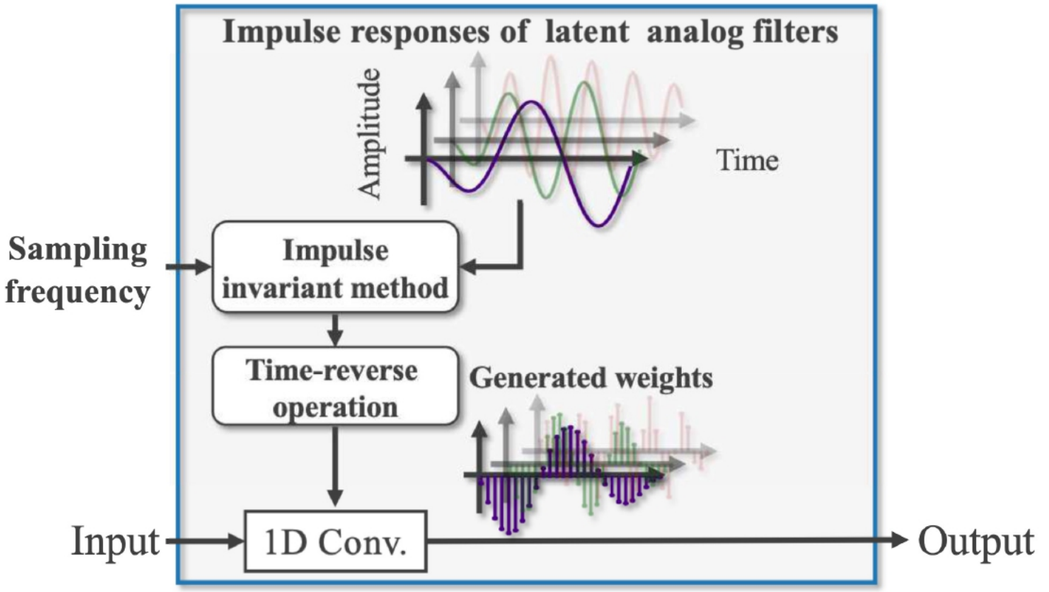
- 说话人、语种、文本内容的差异[16]
- 采样率、编码/压缩/传输方式的差异
  - SFI Conv[17]，SFI STFT[18][19]，USR-BSRNN[20]
- 麦克风数量、阵列结构的差异
  - TAC[21]，DSS[22]，通道间注意力，$TA_{tt}C$[23]

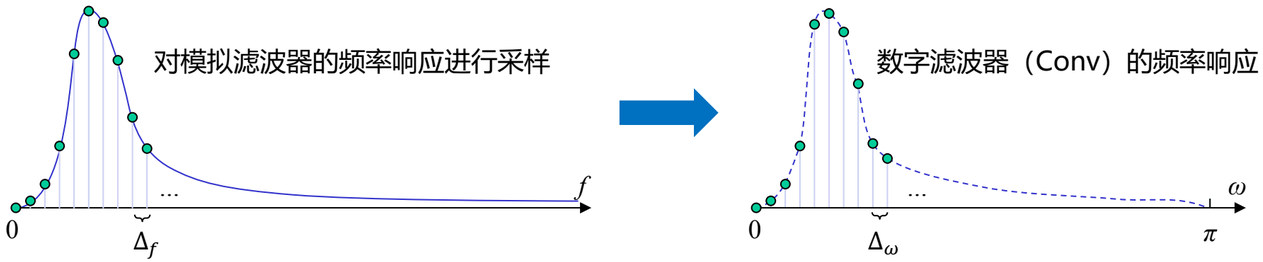
[16] L. Zhang, W. Zhang, C. Li, and Y. Qian, “Scale this, not that: Investigating key dataset attributes for efficient speech enhancement scaling,” arXiv preprint arXiv:2412.14890, 2024.
[17] K. Saito, T. Nakamura, K. Yatabe, Y. Koizumi, and H. Saruwatari, “Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method,” in 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 321–325.
[18] J. Paulus, and M. Torcoli, “Sampling frequency independent dialogue separation,” in 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 160–164.
[19] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in Proc. IEEE ASRU, 2023.
[20] J. Yu, and Y. Luo, “Efficient monaural speech enhancement with universal sample rate band-split RNN,” in Proc. IEEE ICASSP, 2023.
[21] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,” in Proc. IEEE ICASSP, 2020, pp. 6394–6398.
[22] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, “Scene-agnostic multi-microphone speech dereverberation,” in Proc. ISCA Interspeech, 2021, pp. 1129–1133.
[23] W. Zhang, J.-w. Jung, and Y. Qian, “Improving design of input condition invariant speech enhancement,” in Proc. IEEE ICASSP, 2024, pp. 10696–10700.

(1) 面向通用场景的语音增强

如何应对不同场景数据的差异？——采样率

- 采样率无关（SFI）的模型架构
  1) SFI Conv[17]：时域模型
     - 从模拟滤波器采样得到卷积核
  2) SFI STFT[18][19]：频域模型
  3) USR-BSRNN[20]：频域模型



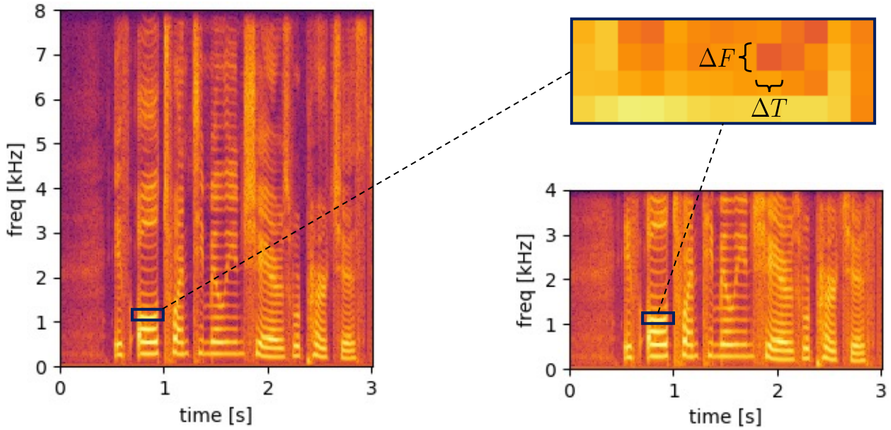


[17] K. Saito, T. Nakamura, K. Yatabe, Y. Koizumi, and H. Saruwatari, “Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method,” in 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 321–325.
[18] J. Paulus, and M. Torcoli, “Sampling frequency independent dialogue separation,” in 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 160–164.
[19] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in Proc. IEEE ASRU, 2023.
[20] J. Yu, and Y. Luo, “Efficient monaural speech enhancement with universal sample rate band-split RNN,” in Proc. IEEE ICASSP, 2023.

(1) 面向通用场景的语音增强

如何应对不同场景数据的差异？——采样率

- 采样率无关（SFI）的模型架构
  1) SFI Conv[17]：时域模型
  2) SFI STFT[18][19]：频域模型
     - 采用时长固定的窗长和窗移 ⇨ 恒定的时间/频率分辨率
  3) USR-BSRNN[20]：频域模型



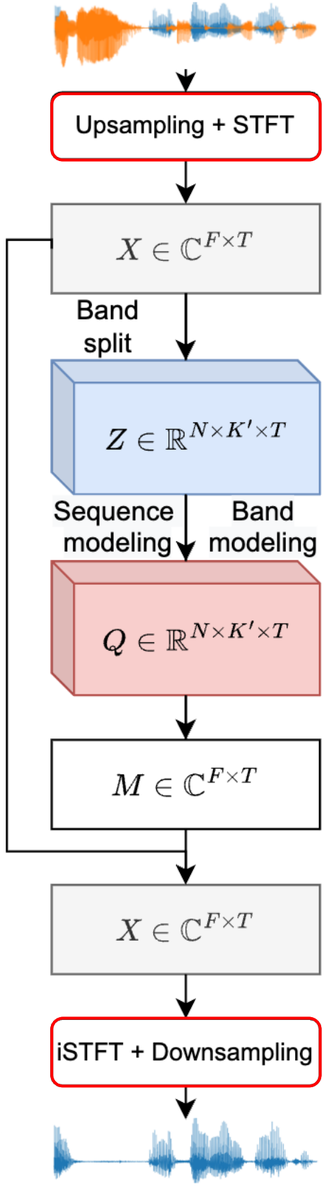
(a) 16kHz (b) 8kHz

[17] K. Saito, T. Nakamura, K. Yatabe, Y. Koizumi, and H. Saruwatari, “Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method,” in 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 321–325.
[18] J. Paulus, and M. Torcoli, “Sampling frequency independent dialogue separation,” in 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 160–164.
[19] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in Proc. IEEE ASRU, 2023.
[20] J. Yu, and Y. Luo, “Efficient monaural speech enhancement with universal sample rate band-split RNN,” in Proc. IEEE ICASSP, 2023.

(1) 面向通用场景的语音增强

如何应对不同场景数据的差异？——采样率

- 采样率无关（SFI）的模型架构
  1) SFI Conv[17]：时域模型
  2) SFI STFT[18][19]：频域模型
  3) USR-BSRNN[20]：频域模型
     - 始终将输入信号上采样到最高采样率
     - 将处理结果重采样回原始采样率



[17] K. Saito, T. Nakamura, K. Yatabe, Y. Koizumi, and H. Saruwatari, “Sampling-frequency-independent audio source separation using convolution invariant method,” in 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 321–325.
[18] J. Paulus, and M. Torcoli, “Sampling frequency independent dialogue separation,” in 30th European Signal Processing Conference (EUSIPCO),
[19] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in Proc. IEEE
[20] J. Yu, and Y. Luo, “Efficient monaural speech enhancement with universal sample rate band-split RNN,” in Proc. IEEE ICASSP, 2023.

(1) 面向通用场景的语音增强

如何应对不同场景数据的差异？——麦克风配置

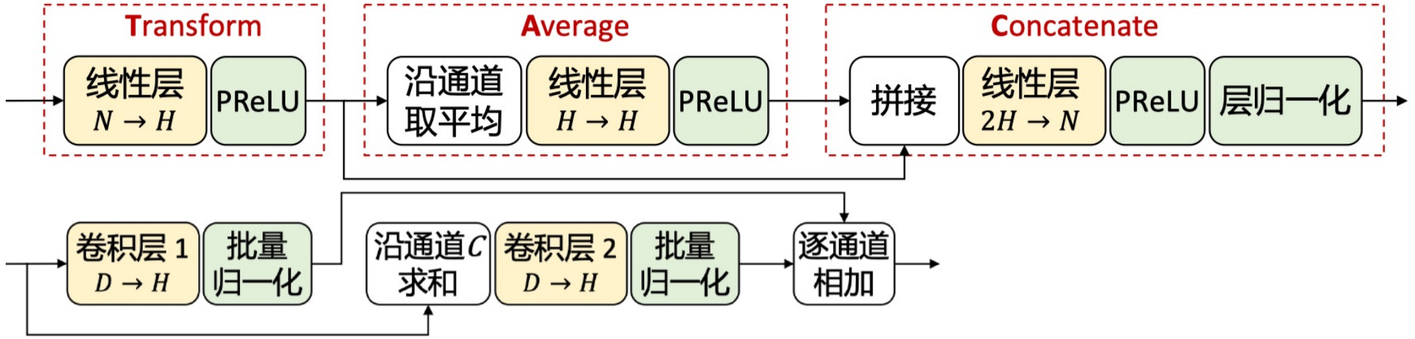
- 麦克风配置无关的模型架构

1) TAC[21]

2) DSS[22]

3) 通道间注意力

4) $TA_{tt}C$[23]



[21] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,” in Proc. IEEE ICASSP, 2020, pp. 6394–6398.

[22] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, “Scene-agnostic multi-microphone speech dereverberation,” in Proc. ISCA Interspeech, 2021, pp. 1129–1133.

[23] W. Zhang, J.-w. Jung, and Y. Qian, “Improving design of input condition invariant speech enhancement,” in Proc. IEEE ICASSP, 2024, pp. 10696–10700.

(1) 面向通用场景的语音增强

如何准确评估通用场景的增强性能?

- 要求
  - 公平可比的训练数据
  - 多样化、高质量的测试数据
  - 多角度综合评估
- 比赛：URGENT 系列挑战赛[24][25]、CCF AATC 2025[26]

[24] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, “URGENT challenge: Universality, robustness, and generalizability for speech enhancement,” in Proc. ISCA Interspeech, 2024, pp. 4868–4872.

[25] K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, Y. Fu, W. Wang, T. Fingscheidt, and S. Watanabe, “Interspeech 2025 URGENT speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 858–862.

[26] J. Zhang, M. Zhu, X. Xu, H. Bu, Z. Ling, and Z. Wu, “The CCF AATC 2025: Speech restoration challenge,” arXiv preprint arXiv:2509.12974, 2025.

目录

CONTENTS

(2) URGENT 挑战赛介绍

URGENT 挑战赛组织方


张王优


Kohei Saijo


李晨达


王　巍


Marvin Sach


付艺辉


Samuele Cornell


Robin Scheibler


倪兆衡


Anurag Kumar


Tim Fingscheidt


Shinji Watanabe


钱彦旻

① 上海交通大学
② 早稻田大学
③ 布伦瑞克大学
④ 卡内基梅隆大学
⑤ Google DeepMind
⑥ Meta

Universality,
Robustness, and
Generalizability
for speech
Enhanceme**NT**

(2) URGENT 挑战赛介绍

URGENT 挑战赛

https://urgent-challenge.com/

- 主题：通用性、鲁棒性、泛化性
- 特点
  - ❑ 首个聚焦于通用场景的语音增强比赛
  - ❑ 基于大规模开源数据，覆盖多语种、多领域、多风格
  - ❑ 多方位的综合评估方式
- 系列：2024 (NeurIPS)、2025 (Interspeech)、2026 (ICASSP)

🔥 **进行中！**

[24] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, “URGENT challenge: Universality, robustness, and generalizability for speech enhancement,” in Proc. ISCA Interspeech, 2024, pp. 4868–4872.
[25] K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, Y. Fu, W. Wang, T. Fingscheidt, and S. Watanabe, “Interspeech 2025 URGENT speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 858–862.

(2) URGENT 挑战赛介绍

URGENT 挑战赛——路线图

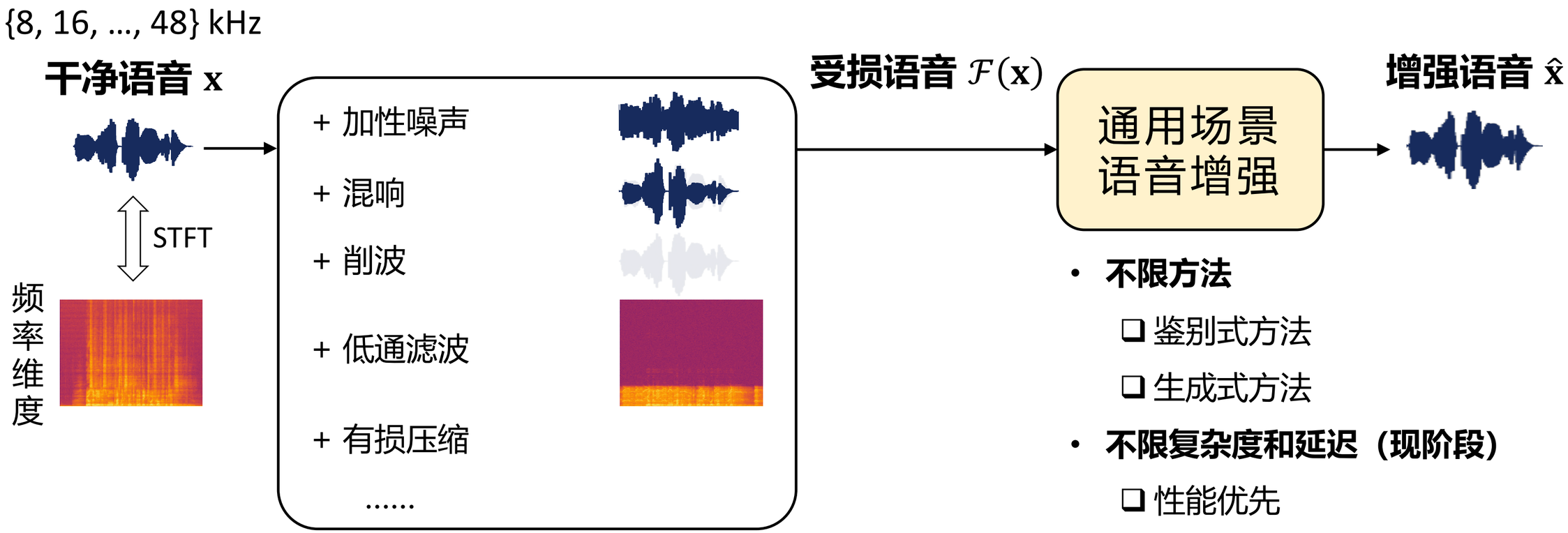
https://urgent-challenge.com/

	传统语音增强研究	URGENT 2024	URGENT 2025	URGENT 2026
规模	中小规模	~1300 h	2500~60k h	精筛 ~800 h
语种	大多英文	英文	5+1 个语种	不变
领域	单一/有限	朗读 对话	不变	✚ 情感 歌声 轻语 长幼
任务	单一/有限	降噪等 4 种	新增 3 种	新增 SQA 赛道
评估	单一/有限	5 大类指标	新增 1 项	新增 3 项

(2) URGENT 挑战赛介绍

URGENT 挑战赛——任务概览

https://urgent-challenge.com/



- **不限方法**
  - ❑ 鉴别式方法
  - ❑ 生成式方法
- **不限复杂度和延迟（现阶段）**
  - ❑ 性能优先
  - ❑ 创新优先

(2) URGENT 挑战赛介绍

URGENT 挑战赛——评估方式

https://urgent-challenge.com/

非侵入式增强指标：DNSMOS、NISQA、UTMOS

侵入式增强指标：POLQA、PESQ、ESTOI、SDR、MCD、LSD

下游任务无关指标：SpeechBERTScore、LPS

下游任务相关指标：SpkSim、CAcc、EmoSim、LidAcc

主观指标：MOS（P.808 ACR → CCR）

逐指标计算排名

↓

计算各大类平均排名

↓

计算总平均排名

目录

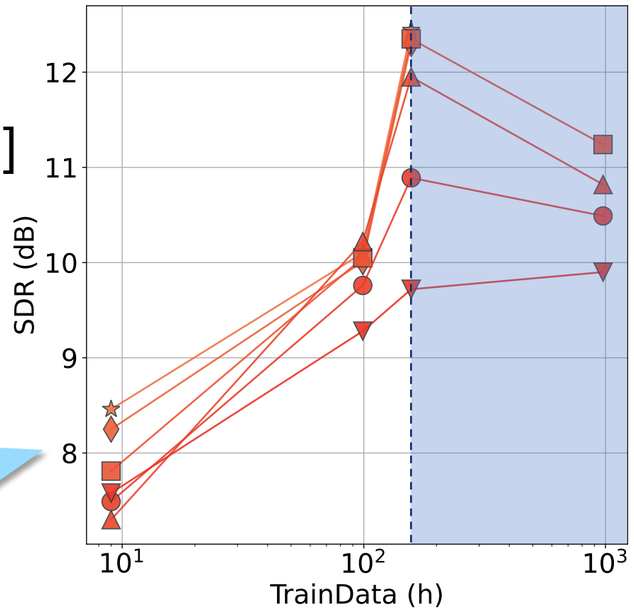
CONTENTS

(3) URGENT 挑战赛经验分享

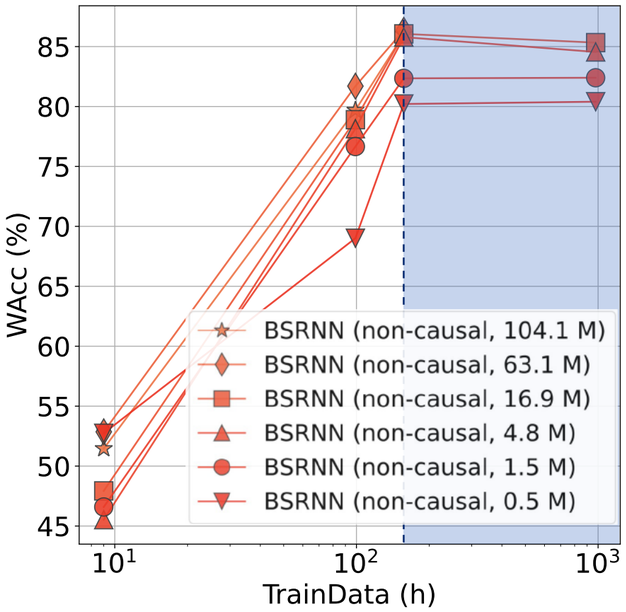
数据准备

- 扩大数据规模是大趋势
  - ❑ 数量 & 多样性同等重要[27]

阴影部分（150h → 1000h）：
仅增加单一领域数据时，
泛化性能不增反降



(a) SDR vs. #Data



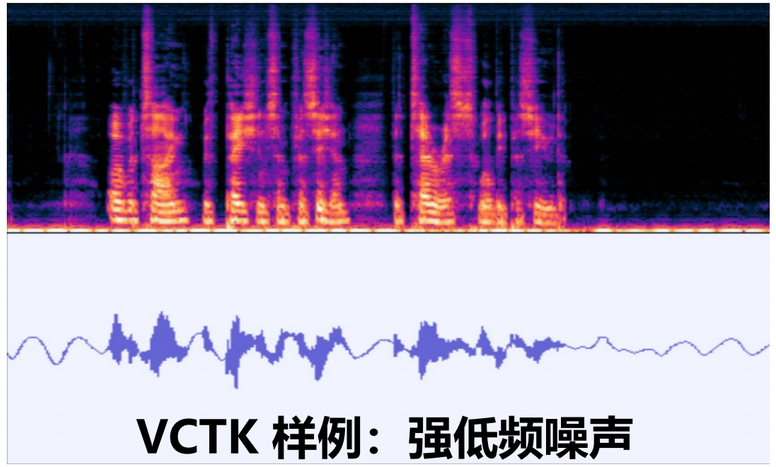
(b) WAcc vs. #Data

[27] W. Zhang, K. Saijo, J.-w. Jung, C. Li, S. Watanabe, and Y. Qian, “Beyond performance plateaus: A comprehensive study on scalability in speech enhancement,” in Proc. ISCA Interspeech, 2024, pp. 1740–1744.

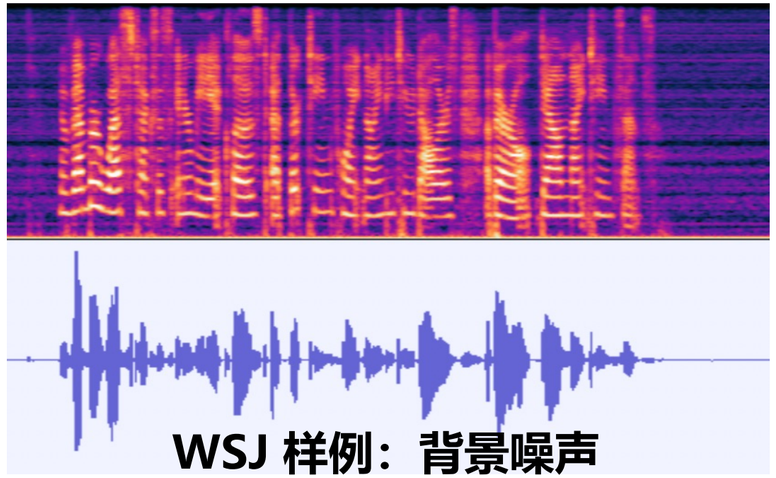
(3) URGENT 挑战赛经验分享

数据准备

- 扩大数据规模是大趋势
- 达到一定数量后，数据质量更重要[28][29]
  - WSJ、VCTK、LibriTTS 等经典数据集都面临标签噪声问题
  - 对生成式方法影响更大



VCTK 样例：强低频噪声



WSJ 样例：背景噪声

[28] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, “Lessons learned from the URGENT 2024 speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 853–857.
[29] C. Li, W. Zhang, W. Wang, R. Scheibler, K. Saijo, S. Cornell, Y. Fu, M. Sach, Z. Ni, A. Kumar et al., “Less is more: Data curation matters in scaling speech enhancement,” Accepted by ASRU 2025.

(3) URGENT 挑战赛经验分享

生成式模型

- 非常擅长提升非侵入式指标，但该类指标无法检测“幻觉”问题
  - “幻觉”：增强语音中出现原本不存在的语音内容、音色等
  - ⇨ 需结合多种指标进行综合评估
- 增强语音通常无法与干净语音对齐 ⇨ 不适宜用 SDR 等指标
- 跨语言泛化能力
  - 在集外语种上，latent diffusion + vocoder 系统容易出现“幻觉”问题（如用英文发音说日语）[25]

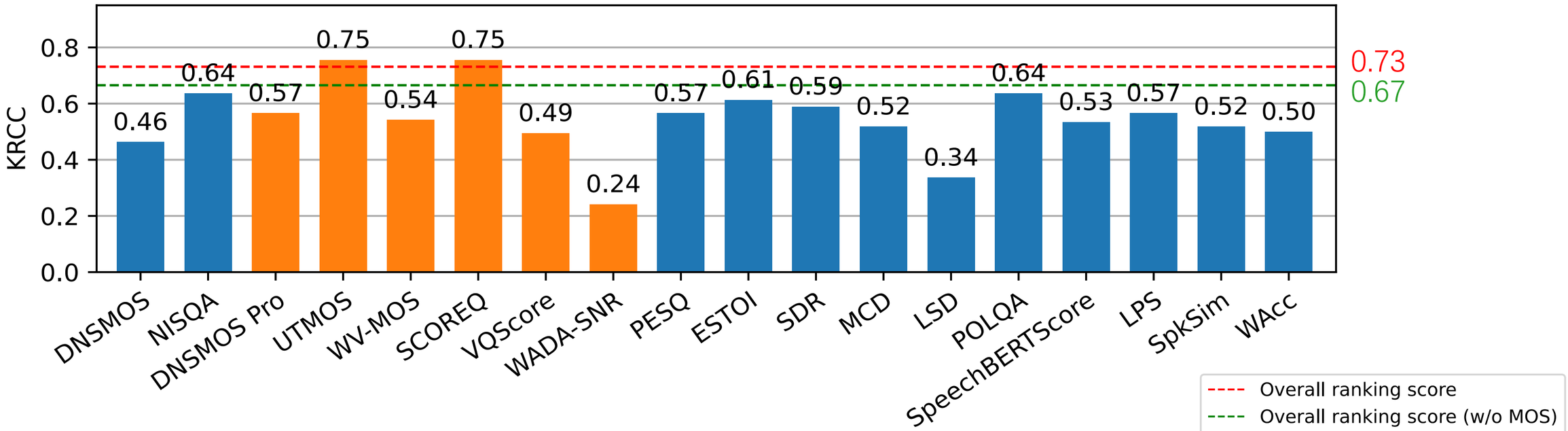
增强前 参考 增强后

[25] K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, Y. Fu, W. Wang, T. Fingscheidt, and S. Watanabe, “Interspeech 2025 URGENT speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 858–862.

(3) URGENT 挑战赛经验分享

评估准则

- 综合利用不同类别的评估指标，结论的可靠性更高
  - URGENT 所采用的多类别排名策略，与 MOS 分数相关性高[28]



[28] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, "Lessons learned from the URGENT 2024 speech enhancement challenge," in Proc. ISCA Interspeech, 2025, pp. 853–857.

(3) URGENT 挑战赛经验分享

评估准则

- 综合利用不同类别的评估指标，结论的可靠性更高
  - URGENT 所采用的多类别排名策略，与 MOS 分数相关性高[28]
- 基于 P.808 ACR 听音测试的 MOS 打分不再是“金标准”[30]
  - 缺少参照，难以检测生成式方法中常见的“幻觉”问题

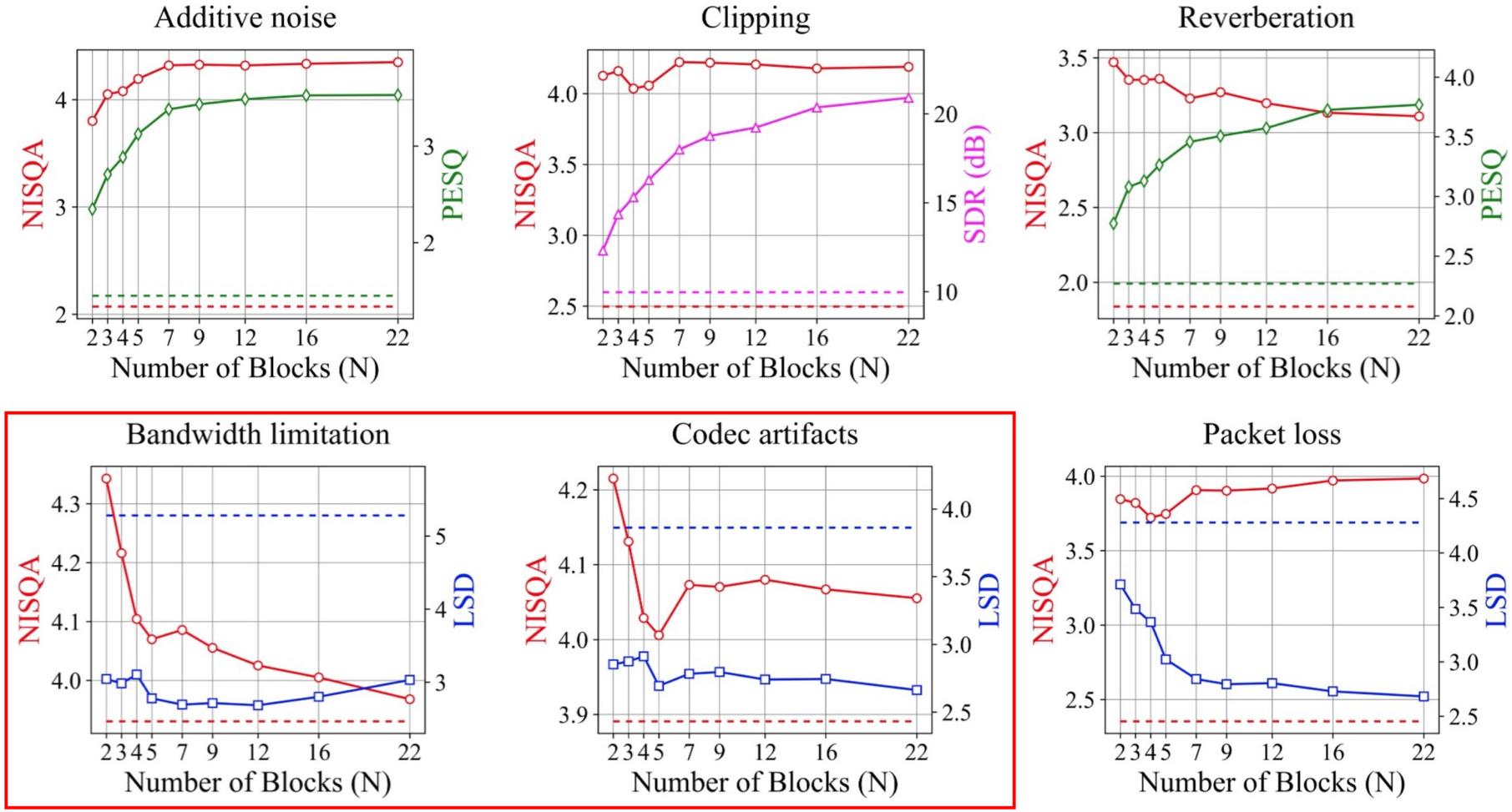
[28] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, “Lessons learned from the URGENT 2024 speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 853–857.
[30] M. Sach, Y. Fu, K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, A. Kumar, W. Wang, Y. Qian, S. Watanabe, and T. Fingscheidt, “P.808 multilingual speech enhancement testing: Approach and results of URGENT 2025 challenge,” arXiv preprint arXiv:2507.11306, 2025.

(3) URGENT 挑战赛经验分享

面向通用场景的语音增强

注：虚线表示未增强的性能

- 是否真正实现了通用语音增强?
- 尺度定律仍有待探索



[31] Z. Sun, A. Li, T. Lei, R. Chen, M. Yu, C. Zheng, Y. Zhou, and D. Yu, “Scaling beyond denoising: Submitted system and findings in URGENT Challenge 2025,” in Proc. ISCA Interspeech, 2025, pp. 873–877.

总结

- 对面向通用场景的语音增强的关注度不断升高
- 最优的模型架构、增强范式仍未有定论
- 扩大规模时，应重视语音数据的多样性和质量
- 不同子任务和不同指标的优化冲突问题有待解决
- 仍需探索更先进、更可靠的评估方法

URGENT 比赛沟通群

谢 谢！

扫码下载报告 PPT

参考文献

[1] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in Proc. IEEE ASRU, 2017, pp. 265–271. [6]

[2] S. Arora, H. Futami, J.-w. Jung, Y. Peng, R. Sharma, Y. Kashiwagi, E. Tsunoo, K. Livescu, and S. Watanabe, “UniverSLU: Universal spoken language understanding for diverse tasks with natural language instructions,” in Proc. ACL, 2024, pp. 2754–2774. [6]

[3] S. Maiti, Y. Peng, S. Choi, J.-w. Jung, X. Chang, and S. Watanabe, “VoxtLM: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks,” in Proc. IEEE ICASSP, 2024, pp. 13326–13330. [6]

[4] R. Yang, H. Yang, X. Zhang, T. Ye, Y. Liu, Y. Gao, S. Zhang, C. Deng, and J. Feng, “PolySpeech: Exploring unified multitask speech models for competitiveness with single-task models,” arXiv preprint arXiv:2406.07801, 2024. [6]

[5] A. Haliassos, R. Mira, H. Chen, Z. Landgraf, S. Petridis, and M. Pantic, “Unified speech recognition: A single model for auditory, visual, and audiovisual inputs,” Advances in Neural Information Processing Systems, vol. 37, pp. 139673–139699, 2024. [6]

[6] M. Shakeel, Y. Sudo, Y. Peng, C.-J. Lin, and S. Watanabe, “Unifying diarization, separation, and ASR with multi-speaker encoder,” Accepted by ASRU 2025. [6]

[7] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, “VoiceFixer: A unified framework for high-fidelity speech restoration,” in Proc. ISCA Interspeech, 2022, pp. 4232–4236. [9]

[8] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, “Universal speech enhancement with score-based diffusion,” arXiv preprint arXiv:2206.03065, 2022. [10]

[9] K. Saijo, W. Zhang, Z.-Q. Wang, S. Watanabe, T. Kobayashi, and T. Ogawa, “A single speech enhancement model unifying dereverberation, denoising, speaker counting, separation, and extraction,” in Proc. IEEE ASRU, 2023. [11]

[10] J. Zhang, H. Yan, and X. Li, “A composite predictive-generative approach to monaural universal speech enhancement,” IEEE Trans. Audio, Speech, Language Process., vol. 33, pp. 2312–2325, 2025. [12]

[11] X. Li, Q. Wang, and X. Liu, “MaskSR: Masked language model for full-band speech restoration,” in Proc. ISCA Interspeech, 2024, pp. 2275–2279. [13]

参考文献

[12] X. Liu, X. Li, J. Serrà, and S. Pascual, “Joint semantic knowledge distillation and masked acoustic modeling for full-band speech restoration with improved intelligibility,” in Proc. IEEE ICASSP, 2025. 13

[13] J. Zhang, J. Yang, Z. Fang, Y. Wang, Z. Zhang, Z. Wang, F. Fan, and Z. Wu, “AnyEnhance: A unified generative model with prompt-guidance and self-critic for voice enhancement,” IEEE Trans. Audio, Speech, Language Process., vol. 33, pp. 3085–3098, 2025. 13

[14] B. Kang, X. Zhu, Z. Zhang, Z. Ye, M. Liu, Z. Wang, Y. Zhu, G. Ma, J. Chen, L. Xiao, C. Weng, W. Xue, and L. Xie, “LLaSE-G1: Incentivizing generalization capability for LLaMA-based speech enhancement,” in Proc. ACL, 2025, pp. 13292–13305. 14

[15] X. Li, H. Xie, Z. Wang, Z. Zhang, L. Xiao, and L. Xie, “SenSE: Semantic-aware high-fidelity universal speech enhancement,” arXiv preprint arXiv:2509.24708, 2025. 14

[16] L. Zhang, W. Zhang, C. Li, and Y. Qian, “Scale this, not that: Investigating key dataset attributes for efficient speech enhancement scaling,” arXiv preprint arXiv:2412.14890, 2024. 15

[17] K. Saito, T. Nakamura, K. Yatabe, Y. Koizumi, and H. Saruwatari, “Sampling-frequency-independent audio source separation using convolution layer based on impulse invariant method,” in 29th European Signal Processing Conference (EUSIPCO), 2021, pp. 321–325. 15, 16, 17, 18

[18] J. Paulus, and M. Torcoli, “Sampling frequency independent dialogue separation,” in 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 160–164. 15, 16, 17, 18

[19] W. Zhang, K. Saijo, Z.-Q. Wang, S. Watanabe, and Y. Qian, “Toward universal speech enhancement for diverse input conditions,” in Proc. IEEE ASRU, 2023. 15, 16, 17, 18

[20] J. Yu, and Y. Luo, “Efficient monaural speech enhancement with universal sample rate band-split RNN,” in Proc. IEEE ICASSP, 2023. 15, 16, 17, 18

[21] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, “End-to-end microphone permutation and number invariant multi-channel speech separation,” in Proc. IEEE ICASSP, 2020, pp. 6394–6398. 15, 19

[22] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, “Scene-agnostic multi-microphone speech dereverberation,” in Proc. ISCA Interspeech, 2021, pp. 1129–1133. 15, 19

参考文献

[23] W. Zhang, J.-w. Jung, and Y. Qian, “Improving design of input condition invariant speech enhancement,” in Proc. IEEE ICASSP, 2024, pp. 10696–10700. 15, 19

[24] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, and Y. Qian, “URGENT challenge: Universality, robustness, and generalizability for speech enhancement,” in Proc. ISCA Interspeech, 2024, pp. 4868–4872. 20, 23

[25] K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, Y. Fu, W. Wang, T. Fingscheidt, and S. Watanabe, “Interspeech 2025 URGENT speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 858–862. 20, 23, 30

[26] J. Zhang, M. Zhu, X. Xu, H. Bu, Z. Ling, and Z. Wu, “The CCF AATC 2025: Speech restoration challenge,” arXiv preprint arXiv:2509.12974, 2025. 20

[27] W. Zhang, K. Saijo, J.-w. Jung, C. Li, S. Watanabe, and Y. Qian, “Beyond performance plateaus: A comprehensive study on scalability in speech enhancement,” in Proc. ISCA Interspeech, 2024, pp. 1740–1744. 28

[28] W. Zhang, K. Saijo, S. Cornell, R. Scheibler, C. Li, Z. Ni, A. Kumar, M. Sach, W. Wang, Y. Fu, S. Watanabe, T. Fingscheidt, and Y. Qian, “Lessons learned from the URGENT 2024 speech enhancement challenge,” in Proc. ISCA Interspeech, 2025, pp. 853–857. 29, 31, 32

[29] C. Li, W. Zhang, W. Wang, R. Scheibler, K. Saijo, S. Cornell, Y. Fu, M. Sach, Z. Ni, A. Kumar et al., “Less is more: Data curation matters in scaling speech enhancement,” Accepted by ASRU 2025. 29

[30] M. Sach, Y. Fu, K. Saijo, W. Zhang, S. Cornell, R. Scheibler, C. Li, A. Kumar, W. Wang, Y. Qian, S. Watanabe, and T. Fingscheidt, “P.808 multilingual speech enhancement testing: Approach and results of URGENT 2025 challenge,” arXiv preprint arXiv:2507.11306, 2025. 32

[31] Z. Sun, A. Li, T. Lei, R. Chen, M. Yu, C. Zheng, Y. Zhou, and D. Yu, “Scaling beyond denoising: Submitted system and findings in URGENT Challenge 2025,” in Proc. ISCA Interspeech, 2025, pp. 873–877. 33