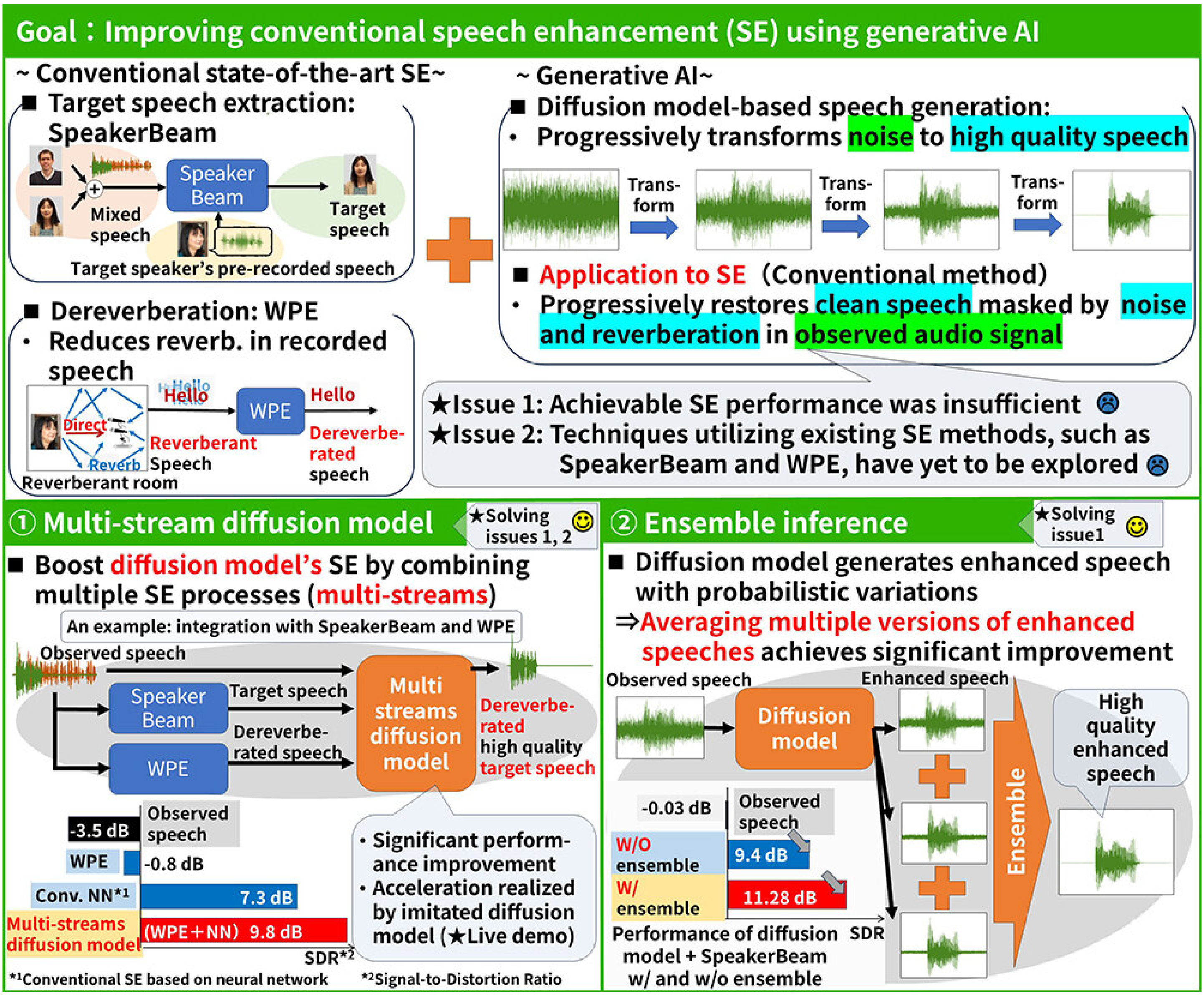
08

Ensemble of multi-stream diffusion model enhances speech

Cleaning-up speech from noisy, reverberant recordings

Abstract

Recent advancements in generative AI have enabled high-quality speech enhancement. This research introduces a speech enhancement method that utilizes the diffusion model, one of the most powerful generative AI models, to effectively remove noise and reverberation from speech recordings. Our approach integrates multiple conventional speech enhancement techniques into a diffusion model-based framework, significantly improving performance. Additionally, we are the first in the world to demonstrate that averaging multiple outputs from the diffusion model, a technique we refer to as “ensemble inference”, greatly enhancing performance. In the future, this technology will enable high-quality speech recording even in noisy environments, making voices sound as if recorded in a studio. This advancement is expected to greatly enhance various speech applications, such as collecting high-quality audio data in everyday environments and enabling more comfortable remote meetings.



Goal：Improving conventional speech enhancement (SE) using generative AI

~ Conventional state-of-the-art SE~

■ Target speech extraction: SpeakerBeam

■ Dereverberation: WPE

- Reduces reverb. in recorded speech

~ Generative AI~

■ Diffusion model-based speech generation:

- Progressively transforms noise to high quality speech

■ Application to SE (Conventional method)

- Progressively restores clean speech masked by noise and reverberation in observed audio signal

★Issue 1: Achievable SE performance was insufficient

★Issue 2: Techniques utilizing existing SE methods, such as SpeakerBeam and WPE, have yet to be explored

① Multi-stream diffusion model

★Solving issues 1, 2

■ Boost diffusion model’s SE by combining multiple SE processes (multi-streams)

An example: integration with SpeakerBeam and WPE

- Significant performance improvement
- Acceleration realized by imitated diffusion model (★Live demo)

*1Conventional SE based on neural network

*2Signal-to-Distortion Ratio

② Ensemble inference

★Solving issue1

■ Diffusion model generates enhanced speech with probabilistic variations

⇒Averaging multiple versions of enhanced speeches achieves significant improvement

Performance of diffusion model + SpeakerBeam w/ and w/o ensemble

References

[1] N. Kamo, M. Delcroix, T. Nakatani, “Target speech extraction with conditional diffusion model,” in *Proc. INTERSPEECH,* pp. 176-180, 2023.

[2] T. Nakatani, N. Kamo, M. Delcroix, S. Araki, “Multi-stream diffusion model for probabilistic integration of model-based and data-driven speech enhancement,” in *Proc. IWAENC,* pp. 65-69, 2024.

Contact

Naoyuki Kamo, Signal Processing Research Group, Media Information Laboratory