04 Neural beamforming for tracking moving sources

Keeping listening while talker is moving

Abstract

Speech enhancement technology is crucial for machines to correctly recognize human speech in noisy environments because it allows extracting only the voices we want to listen to from the background noise. This research introduces a novel beamforming technique that tracks the speaker's movement and keeps extracting the target speaker's voice, even when the speaker is moving while talking. Beamforming requires information on the direction of arrival of the target source and noise signals (spatial information). In this research, we consider the problem of estimating time-varying spatial information as a problem of segmenting speech according to the speaker's movement and propose a novel framework for solving this problem using deep learning. This framework enables accurate estimation of spatial information even when the target speaker is moving. With this framework, we aim for a future in which people and machines can interact more naturally in any situation, such as when multiple speakers talk freely while walking around.

Speech Enhancement With Beamforming

■ Beamforming
- Utilize multiple microphones (mic array)
- Enhance signal from target source's direction

■ Fundamental technology for speech processing in noisy environments
→ Widely used in commercial devices such as smart speakers
→ Essential for recent speech recognition challenges

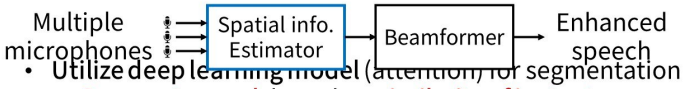
■ Accurate estimation of spatial information is essential for beamforming
(spatial information: information on the direction of arrival of source and noise signals)

Subject ■ Handling moving sources
Spatial information changes when target speaker moves
→ difficult to estimate time-varying spatial information

Beamforming for Moving Sources

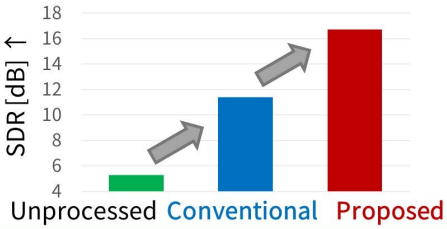
■ Goal
- Accurately enhance target speakers even when they move

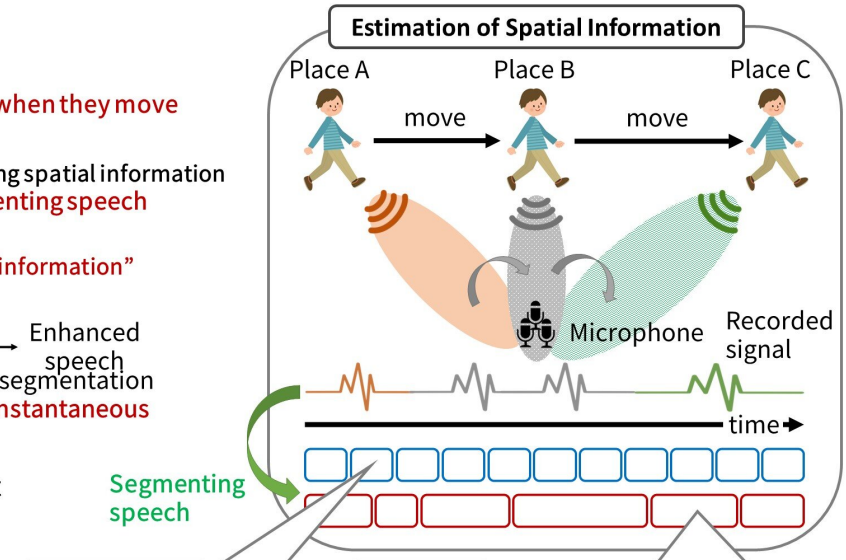
■ Proposed approach
- Optimal estimation framework for time-varying spatial information
→ Estimating spatial information by segmenting speech according to direction and velocity
- "Optimal segmentation" → "Accurate spatial information"



- Utilize deep learning model (attention) for segmentation
→ Segment speech based on similarity of instantaneous spatial information for each time step

■ Experiment with moving source dataset





Conventional
Pre-defined segmentation
☹ Unable to handle variety of moving situations

Proposed
Optimal segmentation with deep learning
☺ Able to handle variety of moving situations
→ Accurate estimation of spatial information & speaker tracking lead to improved enhancement performance

References

[1] T. Ochiai, M. Delcroix, T. Nakatani, S. Araki, "Mask-Based Neural Beamforming for Moving Speakers With Self-Attention-Based Tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 835-848, 2023.

Contact

Tsubasa Ochiai, Signal Processing Research Group, Media Information Laboratory