08 Real-time selective listening of everyday sounds

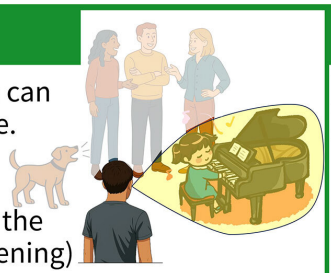
Listening to what you want!

Abstract

Humans can selectively listen to a target sound even when many sounds overlap. This research brings that capability to computers by developing real-time target sound extraction that isolates desired audio from mixed signals on general-purpose PCs while maintaining high accuracy. By incorporating an audio foundation model with general sound representations developed at NTT, the method further improves extraction accuracy and sound quality. We also implement binaural processing to estimate the direction of arrival, making the system closer to human listening. Ultimately, the technology lets users flexibly hear or suppress sounds depending on the context, for example, by reducing household noise in remote-work meetings while preserving meaningful sounds during family calls, enabling more comfortable and effective communication.

Selective Listening

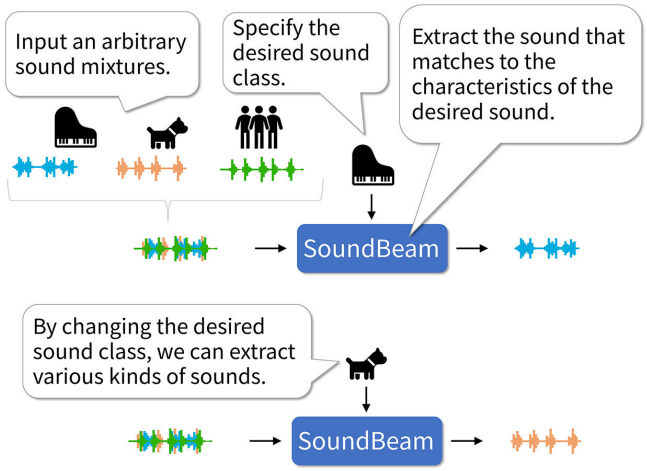
- In daily life, many sounds can be heard at the same time.
- Humans can focus only on the sounds they want to listen to depending on the situation. (= Selective listening)



Goal: Realizing computational selective listening!

SoundBeam Mechanism[1]

SoundBeam realizes selective listening of arbitrary sound using a neural network (NN).

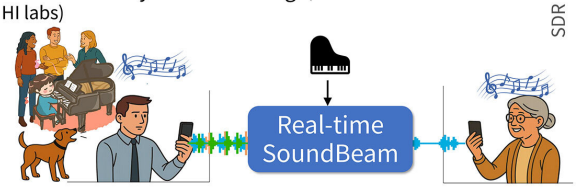


→ We introduce three new functions and performance enhancements of SoundBeam for selective listening toward various applications!

Some parts of the figures were created using generative AI.

① Real-time processing [2]

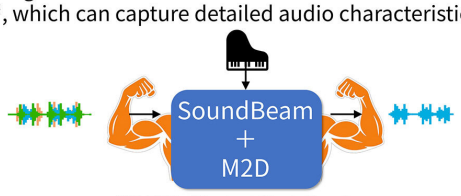
We devised an NN methodology enabling effective selective listening for real-time processing simultaneously with recording. (Collaboration with CD and HI labs)



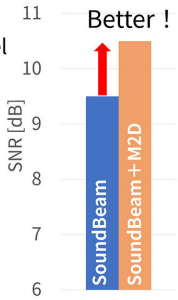
Application: Selectively transmit only the sounds you want others to hear in web conferences.

② High-Quality Sound Extraction [3]

We improved the accuracy of selecting the desired sound and enhanced the quality of the extracted sound, by combining SoundBeam with NTT's audio foundation model M2D※※, which can capture detailed audio characteristics.



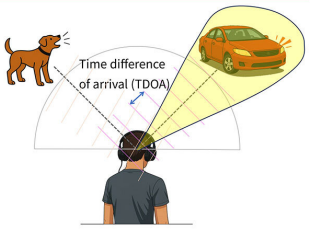
※※ M2D=Masked Modeling Duo
NTT's learning method for audio foundation models

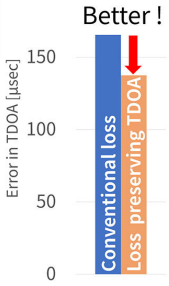


Application: High-quality post-production for recorded audio such as movies, music, and home videos.

③ Reproduction of Sound Direction [4]

By accurately extracting the interaural time difference of sounds reaching the left and right ears, we achieved selective listening that precisely preserves sound source direction information.





Application: Hear important sounds such as warnings, including their directional information while using earphones.

References

[1] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, S. Araki, "SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp.121-136, 2022..

[2] K. Wakayama, T. Ochiai, M. Delcroix, M. Yasuda, S. Saito, S. Araki, A. Nakayama, "Online target sound extraction with knowledge distillation from partially non-causal teacher," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561-565, 2024.

[3] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, D. Niizumi, N. Tawara, T. Nakatani, S. Araki, "SoundBeam meets M2D: Target sound extraction with audio foundation model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[4] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, S. Araki, "Interaural time difference loss for binaural target sound extraction," in *Proc. 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 210-214), 2024. IEEE.

Contact

Marc Delcroix, Media Information Laboratory, Signal Processing Research Group