09

Deep learning based selective hearing of arbitrary sounds

AI that attends to the sounds you want to listen to

Abstract

Humans can focus on listening to a desired sound in complex sound scenes consisting of many co-occurring sounds to retrieve important information, an ability known as selective hearing. In this work, we introduce SoundBeam, a deep-learning-based approach for computational selective hearing. SoundBeam extends our prior work on selective hearing of speech based on the characteristics of the target speaker's voice to arbitrary sounds. With SoundBeam, users can control which sound to listen to depending on their preference or the environment, e.g., hearing a klaxon when crossing a street but not when working at home, thus creating comfortable and safe sound environments.

Selective hearing

❑ In our daily lives, we are immersed in sound scenes consisting of many co-occurring sounds
- A same sound can carry important information or be a nuisance depending on the situation

A firetruck's siren when driving
→ Important sound

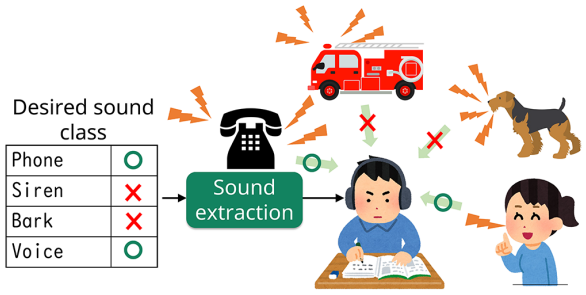
A firetruck's siren when working
→ Nuisance

- Humans can focus on listening to desired sounds = Selective hearing

Research Goal

❑ Develop technology to extract only sounds belonging to a desired sound class specified by the user from a sound mixture.

→ Realize computational selective hearing of arbitrary sounds



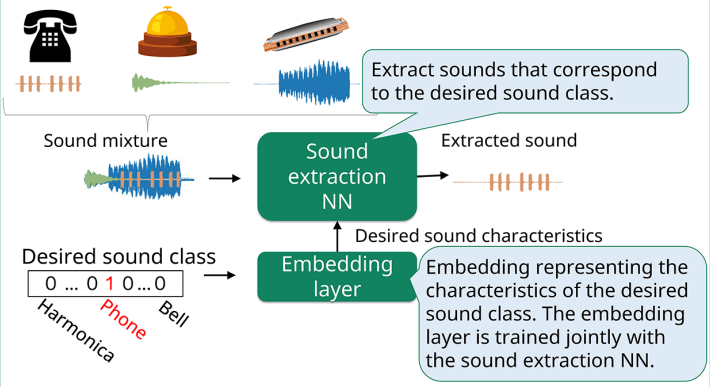
→ Choose the sounds we want to hear depending on the situation or the user's preference.

❑ Applications
- Audio post-editing,
- Listening devices that allow controlling which sounds to hear in an environment

SoundBeam: Selective hearing of arbitrary sounds

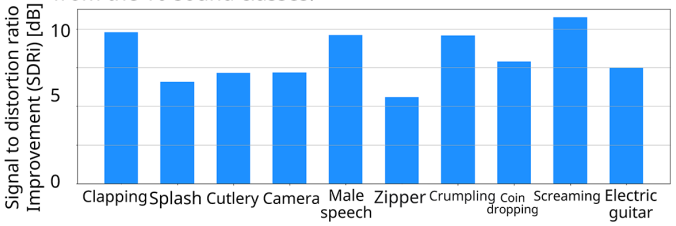
We extend the speech selective hearing technology, SpeakerBeam[1], which we presented at Open-House 2018, to handle arbitrary sounds.

❑ SoundBeam uses a neural network (NN) to realize selective hearing of arbitrary sounds [2,3].
- The NN is trained on data consisting of simulated mixtures of various sounds, the desired sound class labels, and the desired sound signals



The same model can extract various types of sounds simply by changing the desired sound class!

❑ Results of an extraction experiment with 10 sound classes
- Simulated mixtures generated by randomly selecting 3 sounds from the 10 sound classes.



→ Selective hearing of various daily sounds is possible!

SDRi measures how much the desired sound has been emphasized after the extraction process, compared to the mixture. Positive values indicate that the extraction process is successful.

References

[1] M. Delcroix, K. Zmolikova, K. Kinoshita, S. Araki, A. Ogawa, and T. Nakatani, "SpeakerBeam: A New Deep Learning Technology for Extracting Speech of a Target Speaker Based on the Speaker's Voice Characteristics," *NTT Technical Review*, Vol. 16, No. 11, pp. 19–24, Nov. 2018.

[2] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, pp. 2718 - 2722, 2020.

[3] 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 Trans. on Audio, Speech, and Language Processing*, Vol. 31, pp. 121-136, 2023.

Contact

Marc Delcroix
Signal Processing Research Group, Media Information Laboratory