Listening to topics of interest

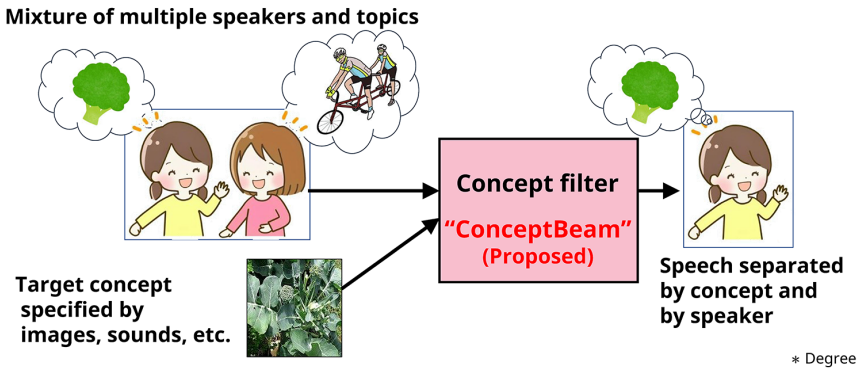
Abstract

We are researching new ways to mathematically express the “meaning” intrinsically communicated by sounds, images, and text. Here, we introduce novel technology for extracting targeted speech from an audio signal containing several speakers talking about several topics all at once by filtering for a specific conceptual “meaning” embodied in images or sounds. This is a new type of sound source separation technology. Historically, research in this field has focused on using the *physical* properties of the signal itself (direction of sound arrival, statistical independence of source signals, speaker characteristics, etc.) to extract the desired signal. In contrast, our method, named ConceptBeam, integrates *semantic* information directly into the signal processing for source separation, enabling content-based semantic extraction of desired signals. In this age of information overload, we aim to realize a society in which information of interest, regardless of format, can be quickly and accurately extracted from vast streams of data by combining semantic processing with traditional signal processing and pattern recognition technologies.

“Concept-based matched filter” extracting speech signals based on their meaning

We have developed a novel signal processing method that separates speech of interest from a mixture of multiple speakers and topics according to target concepts specified by images, sounds, etc..

The system, named ConceptBeam, extracts target speech directly using semantic information, without using speech recognition, which tends to decrease recognition accuracy for mixed speech.



Experimental results

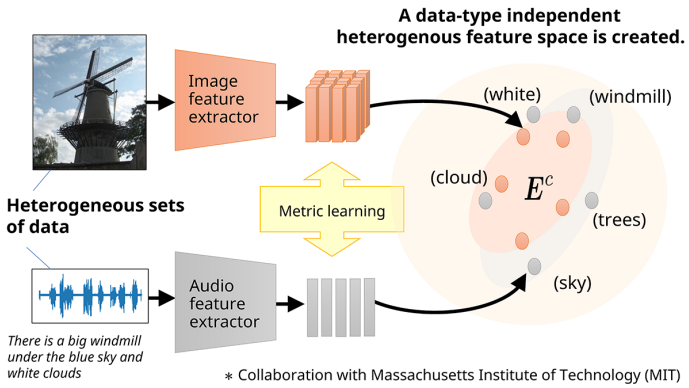
We were able to extract speech with higher accuracy than methods combining existing technologies (Methods 1 and 2).

Method	Concept specifier		Accuracy improvement * (sp.: speakers, con.: concepts)	
	Image	Sound	2 sp. 2 con.	4 sp. 2 con.
Method 1 Separation by word information after speech recognition	✓		1.3 dB	-3.0 dB
		✓	1.1 dB	-1.4 dB
Method 2 Sound source separation and then signal selection	✓		8.6 dB	1.0 dB
		✓	7.9 dB	-0.8 dB
ConceptBeam	✓		10.3 dB	4.0 dB
		✓	11.4 dB	3.8 dB

* Degree of spectral distortion improvement when input signal overlap ratio is 50%.

Point 1: How to represent concepts?

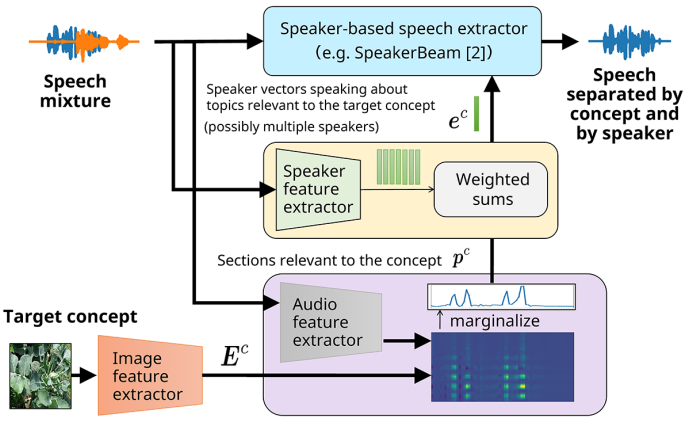
Concepts are represented as vectors in a mathematical feature space independent of specific data modalities such as images or sounds that can be constructed using pre-existing knowledge about the presence or absence of associations (for example, temporal synchronicity) among heterogenous data sets.*



* Collaboration with Massachusetts Institute of Technology (MIT)

Point 2: How to separate targeted signals?

The speaker vectors in the sections related to the target concept are estimated and used for speech extraction.



References

[1] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki, D. Takeuchi, D. Niizumi, A. Kimura, N. Harada, K. Kashino, “ConceptBeam: Concept Driven Target Speech Extraction,” in *Proc. ACM Multimedia*, pp. 4252–4260, 2022.

[2] M. Delcroix, K. Zmolikova, K. Kinoshita, S. Araki, A. Ogawa, 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. 20–24, 2018.

Contact

Kunio Kashino

Biomedical Informatics Research Group, Media Information Laboratory