Developing AI that Pays Attention to Who You Want to Listen to: Deep-learning-based Selective Hearing with SpeakerBeam

Marc Delcroix, Tsubasa Ochiai, Hiroshi Sato, Yasunori Ohishi, Keisuke Kinoshita, Tomohiro Nakatani, and Shoko Araki

Abstract

In a noisy environment such as a cocktail party, humans can focus on listening to a desired speaker, an ability known as selective hearing. In this article, we discuss approaches to achieve computational selective hearing. We first introduce SpeakerBeam, which is a neural-network-based method for extracting speech of a desired target speaker in a mixture of speakers, by exploiting a few seconds of pre-recorded audio data of the target speaker. We then present our recent research, which includes (1) the extension to multi-modal processing, in which we exploit video of the lip movements of the target speaker in addition to the audio pre-recording, (2) integration with automatic speech recognition, and (3) generalization to the extraction of arbitrary sounds.

Keywords: speech processing, deep learning, SpeakerBeam, selective hearing

1. Introduction

Humans can listen to the person they want to (i.e., a target speaker) in a noisy environment such as a cocktail party by focusing on clues about that speaker such as her/his voice characteristics and the content of the speech. We call this ability *selective hearing*. It has been the goal of speech-processing researchers to reproduce a human's selective hearing ability. When several people speak together, the speech signals of the speakers tend to overlap, creating a speech mixture. It is difficult to distinguish the speech of the target speaker from that of the other speakers in such a mixture since all speech signals share similar characteristics. One conventional approach to address this issue is to use blind source separation (BSS), which separates a speech mixture into the source speech signals of each speaker. Research on BSS has made tremendous progress. However, BSS algorithms usually (1) require knowing or estimating the number of speakers speaking in the speech mixture and (2) introduce an arbitrary permutation between the separated outputs and speakers, i.e., we do not know which output of BSS corresponds to the target speaker. These limitations of BSS can impede the deployment of BSS technologies in certain practical applications.

Target-speech extraction is an alternative to BSS that has attracted attention. Instead of separating all speech signals, target-speech extraction focuses on extracting only the speech signal of the target speaker from the mixture. It uses clues about the target speaker to identify and extract that speaker in the mixture [1, 2, 3]. Several speaker clues have been proposed such as an embedding vector that is derived from a pre-recorded enrollment utterance and represents the

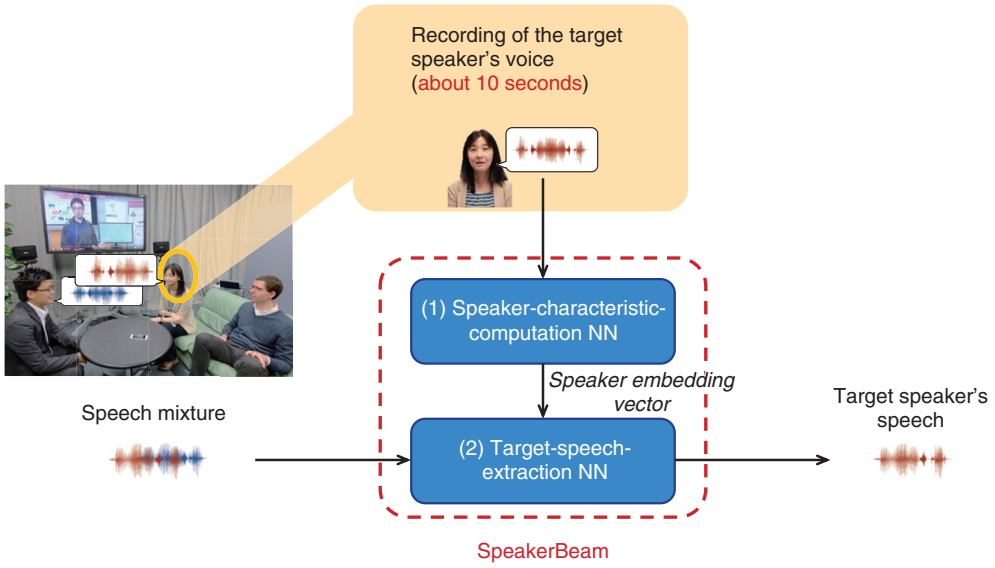


Fig. 1. Principle of SpeakerBeam.

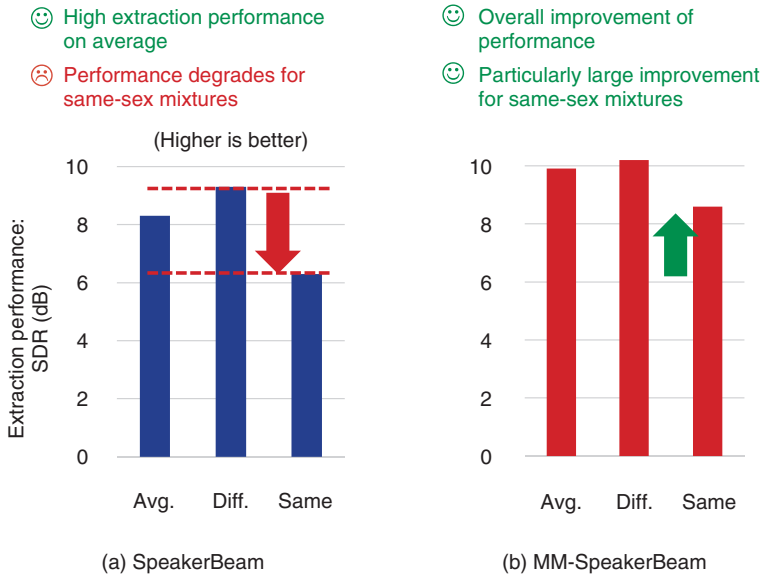
voice characteristics of the target speaker (audio clue) or video data showing the lip movements of the target speaker (video clue). Using such speaker clues, these speech extraction methods focus on only extracting the target speaker without requiring the number of speakers in the mixtures. The output of the methods corresponds to the target speaker, avoiding any permutation ambiguity. Therefore, target-speech extraction naturally avoids the limitations of BSS.

In this article, we briefly review the audio-clue-based target-speech extraction method, SpeakerBeam. We experimentally show one of its limitations, i.e., performance degrades when extracting speech in mixtures of speakers with similar voice characteristics. We then introduce the multimodal (MM) extension of SpeakerBeam, which is less sensitive to the above problem. Finally, we discuss how the principles of target-speech extraction can be applied to other speech-processing problems and expand on future work directions to achieve human's selective hearing ability.

2. SpeakerBeam: Neural-network-based target-speech extraction with audio clues

Figure 1 is a schematic of SpeakerBeam, which is a neural network (NN)-based target-speech extraction method that exploits audio clues of the target speaker. SpeakerBeam consists of two NNs. The *speaker-characteristic-computation NN* accepts an enrollment recording of the voice of the target speaker of about 10 seconds and computes a speaker-embedding vector representing her/his voice characteristics. The *target-speech-extraction NN* accepts the mixture signal and speaker-embedding vector and outputs the speech signal of the target speaker without the voice of the other speakers. The speaker-embedding vector informs the target-speech-extraction NN which of the speakers from the mixture to extract. These two networks are trained jointly to obtain speaker-embedding vectors optimal for target-speech extraction. SpeakerBeam was the first method for target-speech extraction based on audio clues representing the voice characteristics of the target speaker.

We conducted experiments to evaluate SpeakerBeam's performance using two-speaker mixtures generated from a corpus of English read speech utterances. **Figure 2(a)** shows the extraction performance of SpeakerBeam measured with the signal-to-distortion ratio (SDR). The higher the SDR the better the extraction is. SpeakerBeam achieved high extraction performance on average with an SDR of more than 8 dB. However, by breaking down this number in terms of performance for mixtures of speakers of the same or different sexes, we observed a severe degradation in performance by more than 2 dB when extracting speech from same-sex mixtures. This reveals the difficulty of SpeakerBeam to identify and extract the target speech when the speakers in the mixture have



Avg.: average, Diff.: mixtures of different-sex speakers
Same = mixtures of same-sex speakers

Fig. 2. Evaluation of SpeakerBeam performance on two-speakers mixtures.

relatively similar voice characteristics, which occurs more often with same-sex mixtures. One approach to address this issue is to rely on other clues than audio clues to carry out target-speech extraction such as video clues that do not depend on voice characteristics.

3. MM SpeakerBeam

In parallel to audio clues, others have proposed using video clues to carry out target-speech extraction. For example, Ephrat et al. [3] used a video recording of the face and lip movements of the target speaker to extract speech. Their method uses a pre-trained NN, such as FaceNet, to extract features or face-embedding vectors representing the characteristics of the face of the target speaker. These face-embedding vectors form a dynamic representation of the lip movements of the target speaker speaking in the mixture. They are fed to a target-speech-extraction NN, similar to that of SpeakerBeam, to identify and extract the speech signal in the mixture that corresponds to those lip movements. The video clues do not depend on the voice characteristics of the target speaker. Therefore, video-clue-based approaches can be used even when the speakers have similar voice characteristics. For example, in an extreme case, Ephrat et al. [3] showed that video-clue-based approaches could even extract speech in a mixture of two speech utterances of the same speaker as long as the speech content, thus lip movements, were different. However, video clues are sensitive to obstructions, i.e., when the mouth of the target speaker is hidden from the video, which often occurs.

We previously proposed an extension of SpeakerBeam called MM-SpeakerBeam that can exploit multiple clues [4, 5]. For example, by using both audio and video clues, we can combine the benefits of audio- and video-clue-based target-speech extraction, i.e., robustness to obstructions in the video thanks to the audio clue and handling of mixtures of speakers with similar voices thanks to the video clue. **Figure 3** is a schematic of MM-SpeakerBeam. MM-SpeakerBeam exploits both video and audio clues and uses a *face-characteristic-computation NN* to extract a time sequence of face-embedding vectors from the video clue, as in Ephrat et al.'s study [3], and a speaker-characteristic-computation NN to extract speaker-embedding vectors, as in audio-clue-based SpeakerBeam. MM-SpeakerBeam includes a clue-selection mechanism to select the speaker clues based on clue reliability, which dominantly exploits audio clues when the face is obstructed in the video and the video clues when the speakers have similar voice

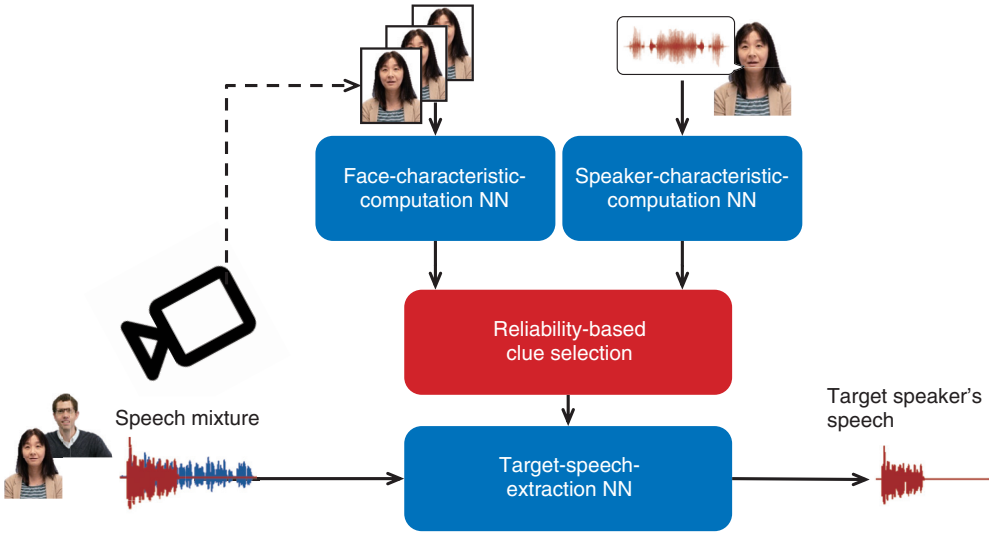


Fig. 3. MM-SpeakerBeam.

characteristics. We implemented the clue-selection mechanism using a similar attention mechanism to that initially proposed for neural machine translation. The target-speech-extraction NN is similar to that of SpeakerBeam. Thanks to the clue-selection mechanism, we can combine clues optimally depending on the situation, making MM-SpeakerBeam more robust than target-speech-extraction methods relying on a single modality.

Figure 2(b) shows the speech-extraction performance of MM-SpeakerBeam. We can see that the overall performance improves and that the largest improvement was achieved for same-sex mixtures. These results reveal that by exploiting multiple modalities (here audio and video), MM-SpeakerBeam can achieve more stable performance. We refer the readers to our demo webpage [6] to listen to various examples of processed signals.

4. Extension to other speech-processing tasks

We can apply the principle of SpeakerBeam to speech-processing tasks other than target-speech extraction. For example, after we proposed SpeakerBeam, others have used a similar method to achieve target-speaker voice-activity detection (TS-VAD) [7], which consists of estimating the start and end timing of speech of the target speaker in a mixture. TS-VAD is an important technology when developing automatic meeting-transcription or minute-generation systems as it enables us to determine who speaks when in a conversation. The use of target-speaker clues is very effective for voice-activity detection under challenging conditions [7]. Another extension of SpeakerBeam consists of target-speech recognition, which outputs the transcription of the words spoken by the target speaker directly, without any explicit speech-extraction step [8].

5. Future perspectives

There are various potential applications for target-speech extraction such as for hearing aids, hearables or voice recorders that can enhance the voice of the speaker of interest, and smart devices that respond only to a designated speaker. Target-speech extraction can also be useful for automatic meeting-transcriptions or minute-generation systems. We plan to extend the capability of SpeakerBeam to get closer to human selective hearing ability, thus open the door for novel applications.

One of our recent research interests is to extend the extraction capabilities of SpeakerBeam to arbitrary sounds. **Figure 4** illustrates the concept of our recently proposed universal sound selector [9]. This system uses clues indicating which sound categories are of interest, instead of audio or video clues. With this system, we can develop hearing devices that can extract different important sounds from the environment (e.g., woman or siren in the figure) while suppressing other disturbing sounds (dog barking, car noise, or man speaking) depending on the user or

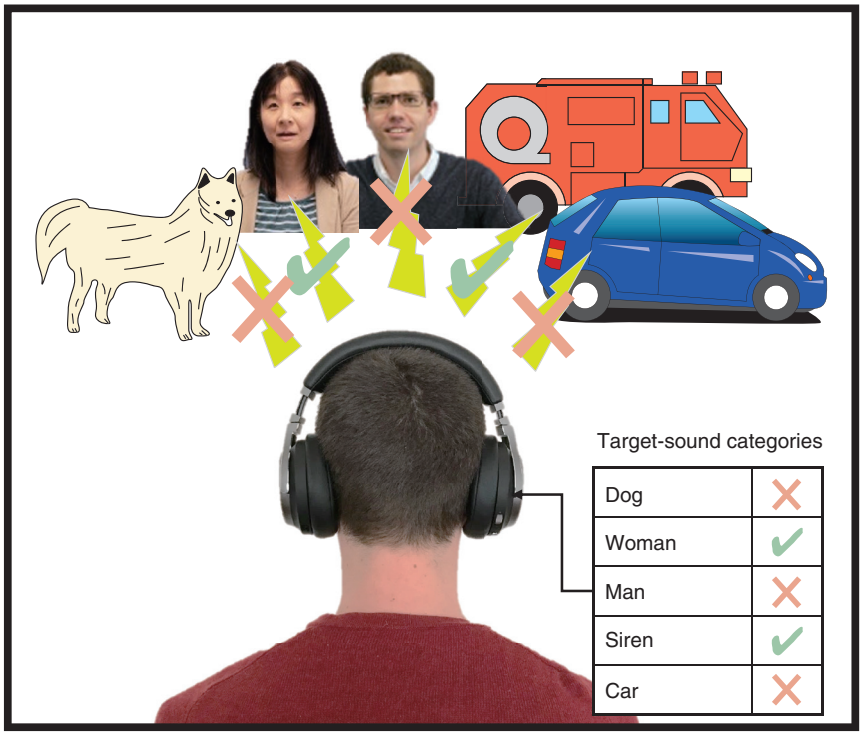


Fig. 4. Universal sound extraction.

situation. Interested readers can find a demo of this system on our webpage [10].

Finally, humans can focus on a conversation depending on its content. A well-known example is that we can easily pick up when someone is saying our name at a cocktail party. Humans can thus exploit more abstract clues, as well as audio and video, to achieve selective listening such as the topic of a conversation or other abstract concepts. To achieve human selective hearing, we should extend SpeakerBeam to speech extraction on the basis of such abstract concepts. This introduces two fundamental research problems. First, how to represent abstract speech concepts. We have made progress in this direction [11]. The second problem consists of how to extract the desired speech signal on the basis of such abstract concept representations. We will tackle these problems in our future research.

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.
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html

[2] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker Aware Neural Network for Target Speaker Extraction in Speech Mixtures," IEEE JSTSP, Vol. 13, No. 4, pp. 800–814, Aug. 2019.

[3] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-independent Audio-visual Model for Speech Separation," ACM Trans. Graph., Vol. 37, No. 4, Article 112, pp. 1–11, Aug. 2018.

[4] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single Channel Target Speech Extraction with Audio-visual Speaker Clues," Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.

[5] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal Attention Fusion for Target Speaker Extraction," Proc. of IEEE Spoken Language Technology Workshop (SLT) 2021, pp. 778–784, Jan. 2021.

[6] Demonstration page of the paper [5], http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html

[7] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, "Target-Speaker Voice Activity Detection: A Novel Approach for Multi-speaker Diarization in a Dinner Party Scenario," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

[8] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, and T. Nakatani, "End-to-end SpeakerBeam for Single Channel Target Speech Recognition," Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.

[9] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to What You Want: Neural Network-based Universal Sound

Selector," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.
[10] Demonstration page of the paper [9], http://www.kecl.ntt.co.jp/icl/signal/member/tochiai/demos/universal_sound_selector/index.html
[11] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Pair Expansion for Learning Multilingual Semantic Embeddings Using Disjoint Visually-grounded Speech Audio Datasets," Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

Marc Delcroix
Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received an M.Eng. from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, in 2007. He was a research associate at NTT Communication Science Laboratories from 2007 to 2008 and 2010 to 2012 then became a permanent research scientist at the same lab in 2012. His research interests include target-speech extraction, robust multi-microphone speech recognition, model adaptation, and speech enhancement. He took an active part in the development of NTT's robust speech recognition systems for the REVERB and CHiME 1 and 3 challenges, which all achieved the best performance results in the tasks. He was one of the organizers of the REVERB challenge 2014 and the 2017 Institute of Electrical and Electronics Engineers (IEEE) Automatic Speech Recognition and Understanding Workshop (ASRU 2017). He is a member of the IEEE Signal Processing Society (SPS) Speech and Language Processing Technical Committee (SLTC). He was a visiting lecturer at the Faculty of Science and Engineering of Waseda University, Tokyo, from 2015 to 2018. He received the 2005 Young Researcher Award from the Kansai section of the Acoustical Society of Japan (ASJ), the 2006 Student Paper Award from the IEEE Kansai section, the 2006 Sato Paper Award from ASJ, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2016 ASJ Awaya Young Researcher Award. He is a senior member of IEEE and a member of ASJ.

Tsubasa Ochiai
Researcher, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a B.E., M.E., and Ph.D. from Doshisha University, Kyoto, in 2013, 2015, and 2018. He was a corporative researcher with National Institute of Information and Communications Technology, Kyoto, from 2013 to 2018 and a research fellow of Japan Society for the Promotion of Science from 2015 to 2018. He has been a researcher at NTT Communication Science Laboratories since 2018. His research interests include speech recognition, speech enhancement, and machine learning. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and IEEE. He was the recipient of the Student Presentation Award from ASJ in 2014, the Awaya Prize Young Researcher Award from ASJ in 2020, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2021.

Hiroshi Sato
Researcher, Voice and Dialog Recognition Technology Group, Cognitive Information Processing Laboratory, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from the University of Tokyo in 2016 and 2018. He joined NTT in 2018 and has since been engaged in research, development, and practical application of speech-processing technologies. His research interests include speech enhancement, robust speech recognition, and speech dialog systems. He is a member of ASJ.

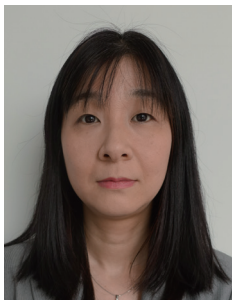
Yasunori Ohishi
Senior Research Scientist, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a Ph.D. from Nagoya University, Aichi, in 2009. Since joining NTT in 2009, he has been researching speech and audio signal processing. His research interests generally concern audio event detection, music information retrieval, and crossmodal learning with audio applications. He received the Awaya Prize Young Researcher Award from ASJ in 2014. He is a member of IEEE, ASJ, the Information Processing Society of Japan (IPSJ), and IEICE.

Keisuke Kinoshita
Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received an M.E. and Ph.D. from Sophia University, Tokyo, in 2003 and 2010. After joining NTT Communication Science Labs in 2003, he has been engaged in fundamental research on various types of speech, audio, and music signal processing, including 1ch/multi-channel speech enhancement (blind dereverberation, source separation, noise reduction), speaker diarization, robust speech recognition, and distributed microphone array processing, and developed several innovative commercial software. He is an author or a co-author of more than 20 journal papers, 5 book chapters, more than 100 papers presented at peer-reviewed international conferences, and an inventor or a co-inventor of more than 20 Japanese patents and 5 international patents. He began serving as an associate editor of IEEE/ACM Transactions on Audio, Speech and Language Processing in 2021 and has been a member of IEEE SPS Audio and Acoustic Signal Processing Technical Committee (AASP-TC) since 2019. He served as the chief coordinator of the REVERB challenge (2014), editor of IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (from 2013 to 2017), and guest editor of EURASIP journal on advances in signal processing (2015). He was honored to receive the 2006 IEICE Paper Award, the 2010 ASJ Outstanding Technical Development Prize, the 2011 ASJ Awaya Prize, the 2012 Japan Audio Society Award, 2015 IEEE ASRU Best Paper Award Honorable Mention, and 2017 Maejima Hisoka Award. He is a member of IEEE, ASJ, and IEICE.

Tomohiro Nakatani

Senior Distinguished Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. from Kyoto University in 1989, 1991, and 2002. Since joining NTT as a researcher in 1991, he has been investigating speech-enhancement technologies for developing intelligent human-machine interfaces. He was a visiting scholar at Georgia Institute of Technology, USA, in 2005. He was a visiting assistant professor in the Department of Media Science, Nagoya University, from 2008 to 2018. He received the 2005 IEICE Best Paper Award, the 2009 ASJ Technical Development Award, the 2012 Japan Audio Society Award, the 2015 IEEE ASRU Best Paper Award Honorable Mention, and the 2017 Maejima Hisoka Award. He was a member of the IEEE SPS AASP-TC from 2009 to 2014 and has been a member of the IEEE SPS SL-TC since 2016. He served as an associate editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing from 2008 to 2010, chair of the IEEE Kansai Section Technical Program Committee from 2011 to 2012, chair of the IEEE SPS Kansai Chapter from 2019 to 2020, a workshop co-chair of the 2014 REVERB Challenge Workshop, and general co-chair of the IEEE ASRU. He is a fellow of IEEE, and member of IEICE and ASJ.



Shoko Araki

Group Leader and Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

She received a B.E. and M.E. from the University of Tokyo in 1998 and 2000, and Ph.D. from Hokkaido University in 2007. Since she joined NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis. She was a member of the IEEE SPS AASP-TC from 2014 to 2019. She has been a board member of ASJ since 2017. She also served as a member of the organizing committee of the International Symposium on Independent Component Analysis and Blind Signal Separation (ICA) 2003, the International Workshop on Acoustic Signal Enhancement (IWAENC) 2003, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2007, the Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA) 2017, IEEE WASPAA 2017, IWAENC 2018, and the evaluation co-chair of the Signal Separation Evaluation Campaign (SiSEC) 2008, 2010, and 2011.

She received the 19th Awaya Prize from ASJ in 2001, the Best Paper Award of the IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, the Academic Encouraging Prize from IEICE in 2006, the Itakura Prize Innovative Young Researcher Award from ASJ in 2008, The Young Scientists' Award of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2014, IEEE SPS Best Paper Award in 2014, and IEEE ASRU 2015 Best Paper Award Honorable Mention in 2015. She is a member of IEEE, IEICE, and ASJ.