AI Hears Your Voice as if It Were Right Next to You—Audio Processing Framework for Separating Distant Sounds with Close-microphone Quality

Tomohiro Nakatani, Rintaro Ikeshita, Naoyuki Kamo, Keisuke Kinoshita, Shoko Araki, and Hiroshi Sawada

Abstract

When we capture speech using microphones far from a speaker (distant microphones), reverberation, voices from other speakers, and background noise become mixed. Thus, the speech becomes less intelligible, and the performance of automatic speech recognition deteriorates. In this article, we introduce the latest speech-enhancement technology for extracting high-quality speech as if it were recorded using a microphone right next to the speaker (close microphone) from the sound captured using multiple distant microphones. We discuss a unified model that enables dereverberation, source separation, and denoising in an overall optimal form, switch mechanism that enables high-quality processing with a small number of microphones, and their integration with deep learning-based speech enhancement (e.g., SpeakerBeam).

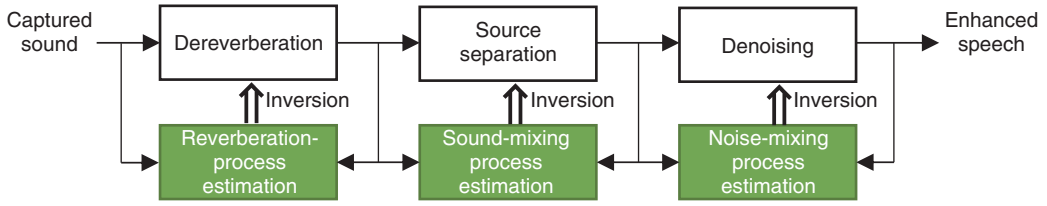
Keywords: speech enhancement, microphone array, far-field speech recording

1. Introduction

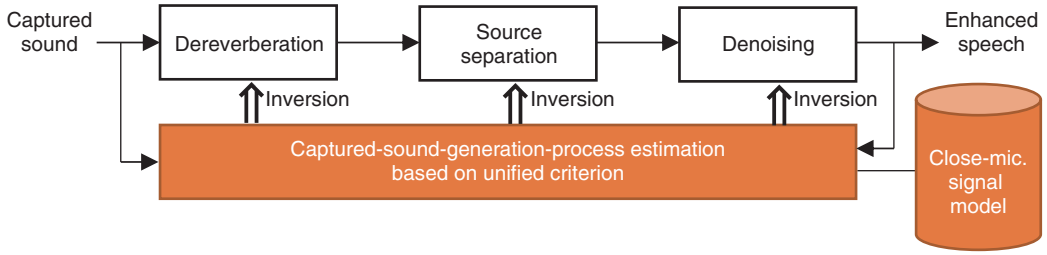
High-quality speech applications using a microphone placed near the speaker's mouth (close microphone) have been widely used, such as automatic speech recognition (ASR) using a smartphone and remote conferencing using a headset. For artificial intelligence (AI) to become a more practical assistant in our daily lives, it is required to handle speech in the same manner even when the speech is captured using microphones far from the speaker (distant microphones). However, with distant microphones, reverberation reflected from the walls or ceilings, voices from other speakers, and background noise become mixed. Therefore, the quality of the captured speech deteriorates significantly, and the performance of speech applications, such as ASR, greatly degrades. To solve this problem, we are developing speech-enhancement technology for extracting a high-quality voice of each speaker as if it were captured with a close microphone from sound captured with distant microphones. This article introduces the latest technology for multi-microphone speech enhancement that uses multiple microphones for achieving higher quality processing than with a single microphone.

2. Challenges for achieving close-microphone quality

To extract a speech with close-microphone quality from sound captured using distant microphones, it is necessary to achieve three types of processing:



(a) Conventional multi-microphone speech enhancement



(b) Unified model-based multi-microphone speech enhancement

Fig. 1. Conventional and unified model-based methods for multi-microphone speech enhancement.

dereverberation, source separation, and denoising. Dereverberation transforms a blurry speech with a distant impression into a clear speech with the impression of being right next to the microphone. When multiple speakers' voices and background noise are mixed, they are separated into individual sounds by source separation and denoising. This makes it possible to extract each speaker's voice with close-microphone quality.

Conventional multi-microphone speech-enhancement methods achieve dereverberation, source separation, and denoising by estimating the generation processes of captured sound, in which sounds propagate from the sources to the microphones and mix, then applying the inverse of the estimated processes to the captured sound (**Fig. 1(a)**). Specifically, the processes of reverberation reflecting from walls or ceilings and reaching the microphones, multiple sounds coming from different directions and mixing, and noise coming from all directions and mixing are estimated, then their inversions are applied.

For example, WPE (weighted prediction error) [1] developed by NTT is the world's first dereverberation method. It can achieve almost perfect dereverberation by estimating the reverberation process of the captured sound without any prior knowledge on what environments in which the sound was captured (i.e., by blind processing), provided the captured sound does not contain noise. Independent component analysis [2, 3], which has been actively studied worldwide by researchers, including NTT, can achieve precise source separation by blind processing, provided the captured sound does not contain reverberation.

However, these conventional multi-microphone speech-enhancement methods cannot be used to solve the three problems (reverberation, multiple sound sources, and noise) at the same time in an overall optimal form. It is impossible to simultaneously estimate all generation processes from the captured sound, which is a mixture of noise, reverberation, and multiple sounds, and execute the inversion of the entire process. Therefore, we have to apply each process in turn. For example, dereverberation is executed first assuming that noise is absent, so precise dereverberation is impossible. We then apply sound-source separation and denoising, assuming that reverberation is wholly suppressed; thus, the best performance cannot be achieved. It is therefore impossible to achieve overall optimal speech enhancement when combining these conventional methods.

The sound captured using distant microphones almost always contains reverberation, multiple sound sources, and noise. For this reason, it has been considered critical to optimally apply the three types of processing, dereverberation, source separation, and

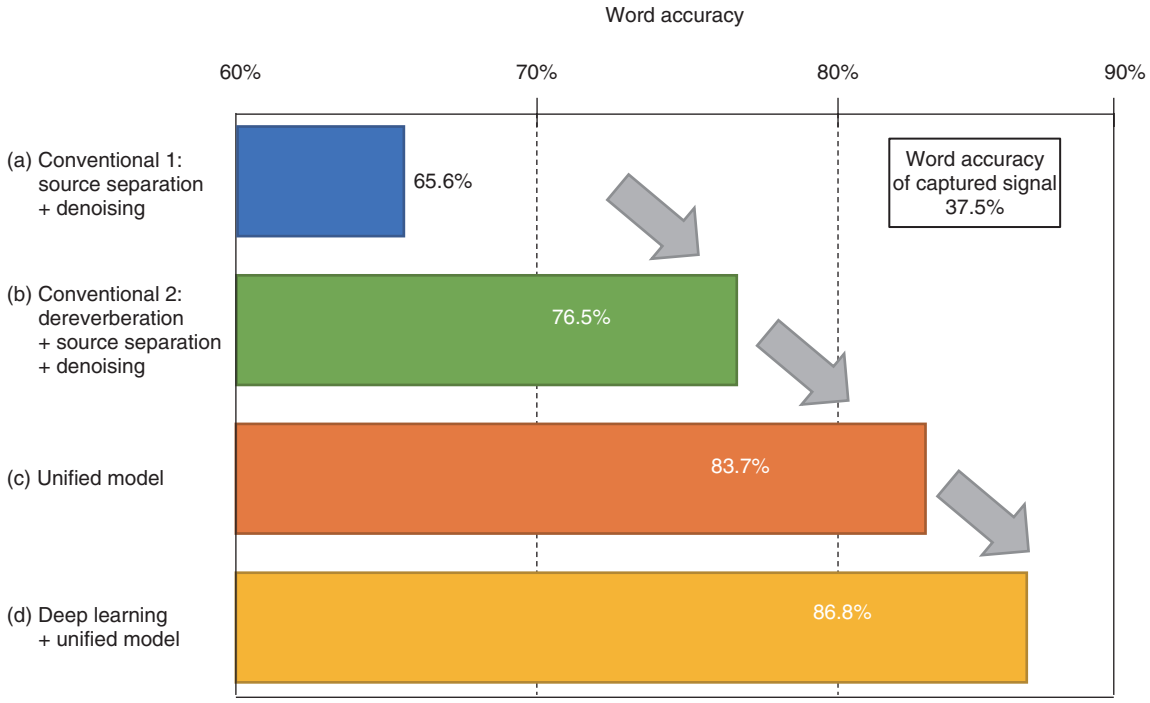


Fig. 2. Improvement in ASR performance using multi-microphone speech enhancement.

denoising, in an overall optimal form.

3. Unified model for dereverberation, source separation, and denoising

In response to this, we devised a unified model that can solve the three problems in an overall optimal form [4, 5]. The unified model first mathematically models the general properties that close-microphone quality speech and noise must satisfy. It can then enable overall optimum processing by optimizing each type of processing on the basis of the *unified criterion* that the sound obtained from combining the three types of processing best satisfies the close-microphone property (**Fig. 1(b)**). For example, we can significantly improve ASR using distant microphones with the unified model (**Figs. 2(a)–(c)**).

Figure 3 shows a spectrogram of two speech signals and noise captured using a close microphone and the mixture of them captured using a distant microphone. The speech signals captured using the close microphone are sparse signals in which the sound concentrates in separate local areas, and are non-stationary signals that change with time. In contrast, noise is a dense and stationary signal in which the sound spreads over a wider area and does not change much with time. However, the mixture captured using a distant microphone has different characteristics. It is denser than the speech signals with close-microphone quality and more non-stationary than noise with close-microphone quality.

The unified model uses the differences in these sound characteristics. It controls dereverberation, source separation, and denoising so that the sound resulting from their application best satisfies the characteristics of speech and noise with close-microphone quality. For example, in dereverberation, we estimate the reverberation-generation process and apply its inversion so that the sound obtained in combination with source separation and denoising best satisfies the close-microphone quality. Similarly, we optimize source separation and denoising by estimating the sound-generation process and applying its inversion to best satisfy the close-microphone quality when combined with dereverberation. With the aim of achieving close-microphone quality, it has become possible to execute overall optimum processing when combining all types of processing.

We have also developed computationally efficient algorithms for unified model-based multiple microphone speech enhancement [6, 7]. For example, the processing using the unified model illustrated in

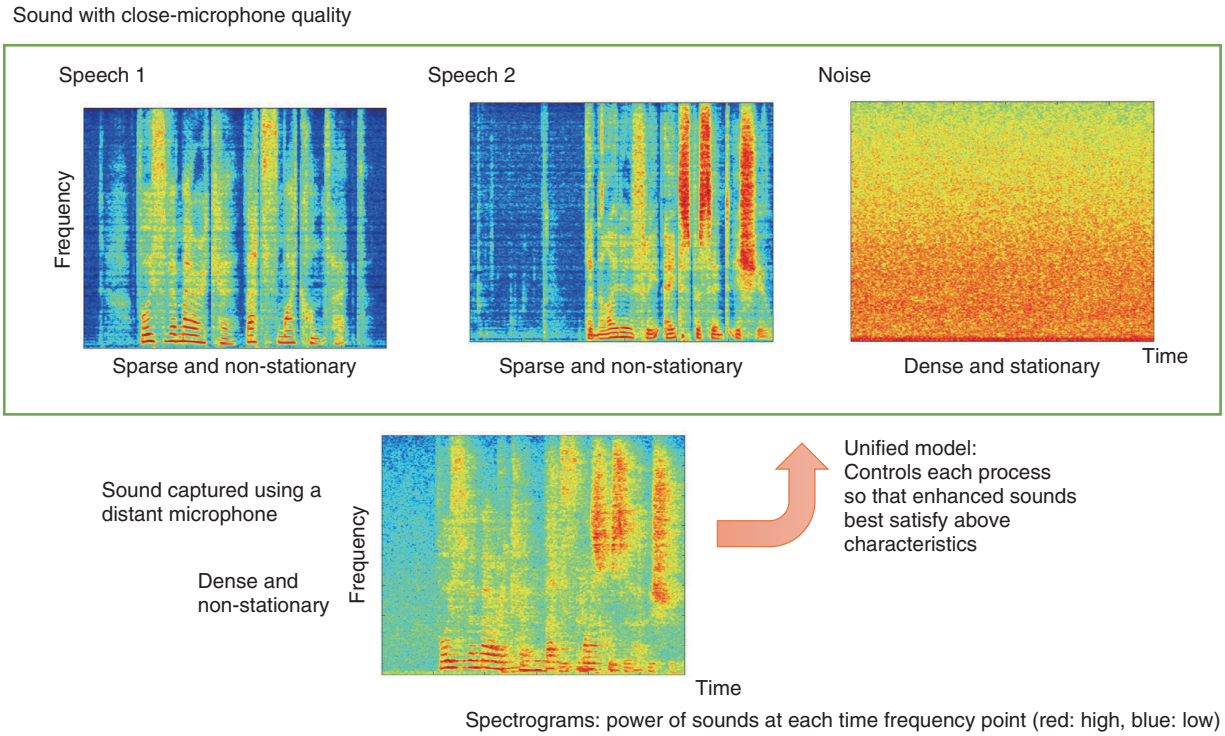


Fig. 3. Spectrograms of sounds captured using close and distant microphones.

Fig. 2 (executing overall optimization of dereverberation, source separation, and denoising using eight microphones) can now be completed in real time using a Linux computer. When we limit the problem to extracting a speaker's voice by blind processing from background noise and little reverberation, we can reduce the computational cost to the extent that real-time processing is possible even with an embedded device.

4. Switch mechanism enabling accurate estimation with a smaller number of microphones

A switch mechanism is an applied technology using the unified model and enables highly accurate estimation even with a relatively small number of microphones [8, 9]. With conventional multi-microphone speech-enhancement methods, it is necessary to use a sufficiently large number of microphones for precise processing compared with the number of sound sources included in the captured sound. This hinders the application of multi-microphone speech enhancement to real-life problems. To solve this problem, we introduce a switch mechanism that can improve estimation accuracy with a small number of microphones.

The idea of this switch mechanism is summarized as follows. Even when the captured sound contains many sound sources, the number of sources appearing simultaneously can be smaller when counting them within each short time interval. Let us explain this using **Fig. 4**. The horizontal axis is time, and a horizontal bar in each color represents when each of the three speakers speaks. When we divide the horizontal axis into short intervals a, b, and c, as shown in the figure, only two speakers are speaking in each time interval even though there are three speakers in total. With this interval division, we can improve multi-microphone speech enhancement by applying it separately to each short interval with the decreased number of speakers. We call this a switching mechanism because we switch speech enhancement for each short interval.

When combined with the unified model, the switch mechanism can perform best. We can use the unified model to optimize the interval-wise application of speech enhancement and the switch mechanism's time interval division. This unified model-based speech enhancement can optimize the all processing

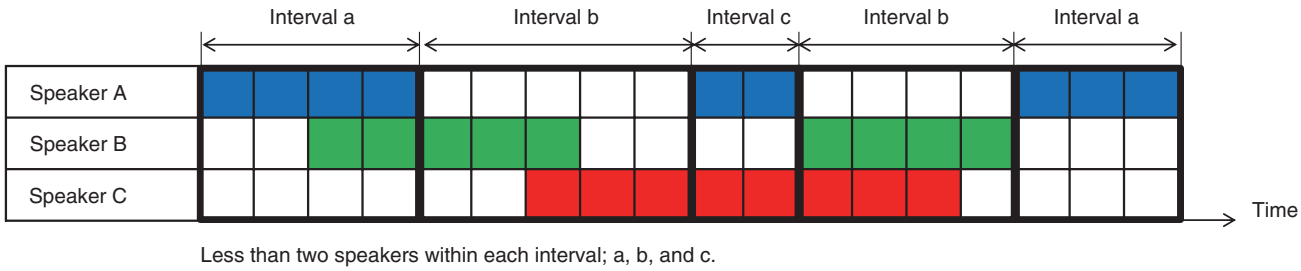


Fig. 4. An example of each speaker's utterance periods in a conversation among three speakers.

types (dereverberation, source separation, and denoising) with the switch mechanism so that the enhanced speech best satisfies the close-microphone quality.

5. Unified model as a versatile technique of audio-signal processing

As described above, our unified model provides theoretically and practically excellent guidelines for integrating the three processing types in speech enhancement that we have conventionally combined in more heuristic ways. The unified model can provide a mechanism to achieve overall optimization even when combining more complicated processing approaches such as the switch mechanism. We can use the unified model as a versatile technique providing a basis for future audio-signal processing-technology development.

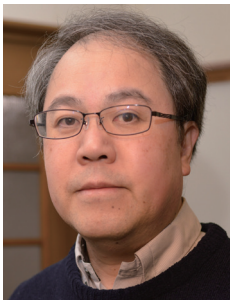
6. Future direction: optimal integration with deep learning

Deep learning is another fundamental approach to speech enhancement, and its integration with multi-microphone speech-enhancement methods is vital for the future development. While deep learning can conduct processing that is difficult with multi-microphone speech enhancement, such as voice characteristics-based selective listening using SpeakerBeam, a deep learning-based approach for computational selective hearing based on the characteristics of the target speaker's voice [10], it also has severe limitations. For example, with deep learning-based speech enhancement, improvement in ASR performance is minimal, and sound quality largely degrades due to reverberation. Therefore, both deep learning and multi-microphone speech-enhancement methods complement each other, thus are indispensable. For example, even when the ASR performance or quality of enhanced speech does not much improve solely by deep learning-based speech enhancement, they can be improved when combined with the multi-microphone speech enhancement. **Figure 2(d)** shows that the combined approach further improves ASR performance compared with solely using the unified model-based multi-microphone speech enhancement. Speech enhancement will have much higher functionality and quality through developing an optimal integration method for both deep learning and multi-microphone speech enhancement.

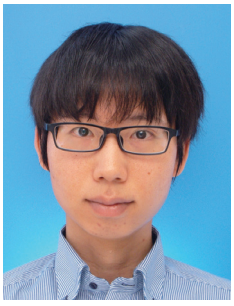
References

[1] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-normalized Delayed Linear Prediction," IEEE Trans. Audio, Speech, and Language Process., Vol. 18, No. 7, pp. 1717–1731, 2010.
[2] N. Ono and S. Miyabe, "Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources," LVA/ICA 2010: Latent Variable Analysis and Signal Separation, pp. 165–172, Springer, 2010.
[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-wise Clustering and Permutation Alignment," IEEE Trans. Audio, Speech, and Language Process., Vol. 19, No. 3, pp. 516–527, 2011.
[4] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly Optimal Denoising, Dereverberation, and Source Separation," IEEE/ACM Trans. Audio, Speech, and Language Process., Vol. 28, pp. 2267–2282, 2020.
[5] R. Ikeshita and T. Nakatani, "Independent Vector Extraction for Fast Joint Blind Source Separation and Dereverberation," IEEE Signal Process. Lett., Vol. 28, pp. 972–976, 2021.
[6] R. Ikeshita, T. Nakatani, and S. Araki, "Block Coordinate Descent Algorithms for Auxiliary-function-based Independent Vector Extraction," IEEE Trans. Signal Process., Vol. 69, pp. 3252–3267, 2021.
[7] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low Latency Online Source Separation and Noise Reduction Based on Joint Optimization with Dereverberation," Proc. of the 29th European Signal Processing Conference (EUSIPCO 2021), pp. 1000–1004, Dublin, Ireland, 2021.
[8] R Ikeshita, N. Kamo, and T. Nakatani, "Blind Signal Dereverberation Based on Mixture of Weighted Prediction Error Models," IEEE Signal Process. Lett., Vol. 28, pp. 399–403, 2021.
[9] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, and S. Araki, "Switching Independent Vector Analysis and Its Extension to

Blind and Spatially Guided Convolutional Beamforming Algorithms," IEEE/ACM Trans. Audio, Speech, and Language Process., Vol. 30, pp. 1032–1047, 2022.
[10] 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, 2018.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html



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 in 2005 and visiting assistant professor in the Department of Media Science, Nagoya University, Aichi from 2008 to 2018. He received the 2005 Institute of Electronics, Information and Communication Engineers (IEICE) Best Paper Award, the 2009 Acoustical Society of Japan (ASJ) Technical Development Award, the 2012 Japan Audio Society Award, the 2015 Institute of Electrical and Electronics Engineers (IEEE) Automatic Speech Recognition and Understanding Workshop (ASRU) Best Paper Award Honorable Mention, the 2017 Maejima Hisoka Award, and the 2018 International Workshop on Acoustic Echo and Noise Control (IWAENC) Best Paper Award. He was a member of the IEEE Signal Processing Society Audio and Acoustic Signal Processing Technical Committee (AASP-TC) from 2009 to 2014 and served as the chair of the AASP-TC Review Subcommittee from 2013 to 2014. He was a member of the IEEE Signal Processing Society Speech and Language Processing Technical Committee (SLTC) from 2016 to 2021. He served as an associate editor of the IEEE Transactions on Audio, Speech and Language Processing from 2008 to 2010, chair of the IEEE Kansai Section Technical Program Committee from 2011 to 2012, technical program co-chair of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2007, workshop co-chair of the 2014 REVERB Challenge Workshop, general co-chair of the 2017 IEEE ASRU Workshop, and chair of IEEE Signal Processing Society Kansai Chapter from 2019 to 2020. He is an IEEE Fellow, and a member of IEICE and ASJ.

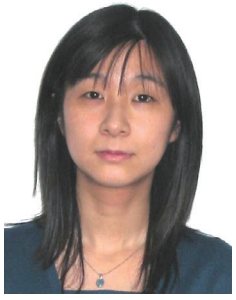


Rintaro Ikeshita
Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a B.E. and M.S. from the University of Tokyo in 2013 and 2015 and joined NTT in 2018. He received the 17th Itakura Prize Innovative Young Researcher Award from ASJ in 2022 and the 49th Awaya Prize Young Researcher Award from ASJ in 2021. He is a member of IEEE and ASJ.

Naoyuki Kamo
Researcher, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received a B.S. and M.S. from Kyoto University in 2012 and 2014. Before he joined NTT in 2020, he developed as a contributor the well-known deep learning-based open software tools for speech-signal processing ESPNET-1 and 2. Since joining NTT, he has developed neural network-based techniques, such as speech enhancement and intelligibility prediction. He is a member of ASJ.

Keisuke Kinoshita
Research Scientist, Google Japan.

He received an M. Eng. and Ph.D. from Sophia University, Tokyo, in 2003 and 2010. After joining NTT Communication Science Labs in 2003, he was 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 types of innovative commercial software. He has been in his current position since October 2022. He has been serving as an associate editor of IEEE/ACM Transactions on Audio, Speech and Language Processing since 2021 and member of IEEE AASP-TC since 2019. He also served as the chief coordinator of the REVERB Challenge (2014), an 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 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 Signal Processing Society AASP-TC (2014–2019) and currently serves as its vice chair. She has been a board member of ASJ since 2017 and currently serves as vice president of ASJ (2021–2022). She also served as a member of the organizing committee of International Symposium on Independent Component Analysis and Blind Signal Separation (ICA) 2003, IWAENC 2003, IEEE WASPAA 2007, Hands-free Speech Communications and Microphone Arrays (HSCMA) 2017, IEEE WASPAA 2017, IWAENC 2018, IEEE WASPAA 2021 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 Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, The Young Scientists' Prize in 2014, IEEE Signal Processing Society Best Paper Award in 2014, and IEEE ASRU 2015 Best Paper Award Honorable Mention in 2015. She is an IEEE Fellow for contributions to blind-source separation of noisy and reverberant speech signals. She is a member of IEEE, IEICE, and ASJ.

Hiroshi Sawada
Executive Manager and Senior Distinguished Researcher, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.E., and Ph.D. in information science from Kyoto University in 1991, 1993, and 2001. His research interests include statistical signal processing, audio-source separation, array-signal processing, machine learning, latent variable modeling, graph-based data structures, and computer architecture. He received the Best Paper Award of the IEEE Circuit and System Society in 2000 and the Best Paper Award of IEEE Signal Processing Society in 2014. He is an IEEE Fellow and member of IEICE and ASJ.