SpeakerBeam: A New Deep Learning Technology for Extracting Speech of a Target Speaker Based on the Speaker's Voice Characteristics

Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Shoko Araki, Atsunori Ogawa, and Tomohiro Nakatani

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. Current approaches developed to realize computational selective hearing require knowing the position of the target speaker, which limits their practical usage. This article introduces SpeakerBeam, a deep learning based approach for computational selective hearing based on the characteristics of the target speaker's voice. SpeakerBeam requires only a small amount of speech data from the target speaker to compute his/her voice characteristics. It can then extract the speech of that speaker regardless of his/her position or the number of speakers talking in the background.

Keywords: deep learning, target speaker extraction, SpeakerBeam

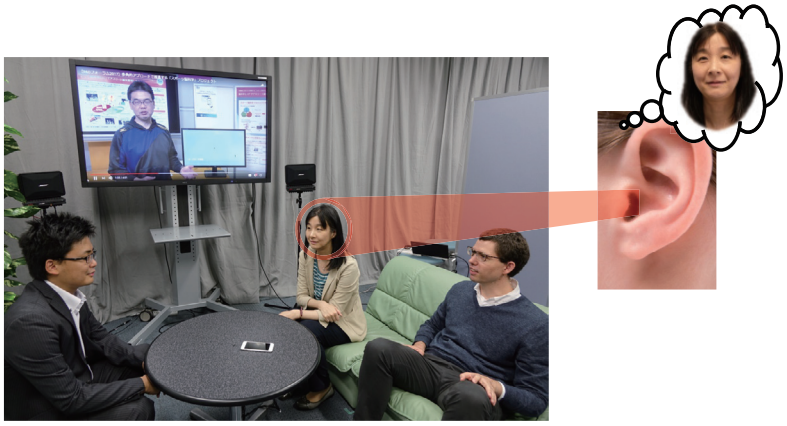
1. Introduction

Automatic speech recognition technology has progressed greatly in recent years, thus enabling the rapid adoption of speech interfaces in smartphones or smart speakers. However, the performance of current speech interfaces deteriorates severely when several people speak at the same time, which often happens in everyday life, for example, when we take part in discussions or when we are in a room where a television is on in the background. The main reason for this problem arises from the inability of current speech recognition systems to focus solely on the voice of the target speaker when several people are speaking [1].

In contrast to current speech recognition systems, human beings have a selective hearing ability (see **Fig. 1**), meaning that they can focus on speech spoken by a target speaker even in the presence of noise or other people talking in the background by exploiting information about the characteristics of the voice and the position of the target speaker.

Previous attempts to replicate computationally the human selective hearing ability used information about the target speaker position [1]. With these approaches, it is hard to focus on a target speaker when the speaker's position is unknown or when he/she moves, which limits their practical usage.

We have proposed SpeakerBeam [2], a novel approach to mimic the human selective hearing ability that focuses on the target speaker's voice characteristics (see **Fig. 2**). SpeakerBeam uses a deep neural network to extract speech of a target speaker from a mixture of speech signals. In addition to the speech mixture, SpeakerBeam also inputs the characteristics of the target speaker's voice so that it can extract speech that matches these characteristics. These voice characteristics are computed from an adaptation utterance, that is, another recording (about 10 seconds long) of the target speaker's voice.



Ability to listen only to a target speaker by focusing on the characteristics of his/her voice (pitch, timbre, etc.) and the direction of arrival of the sound

Fig. 1. Human selective hearing ability.

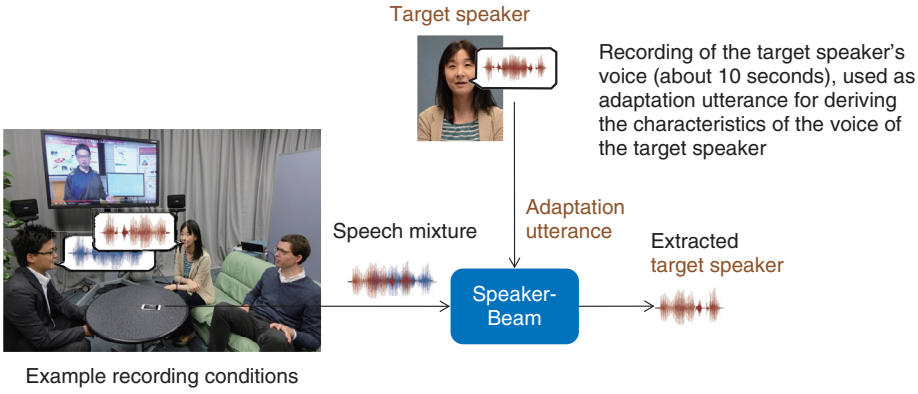


Fig. 2. SpeakerBeam's selective hearing capability.

Consequently, SpeakerBeam enables the extraction of the voice of a target speaker based solely on the target speaker's voice characteristics without knowing his/her position, thus opening new possibilities for the speech recognition of multi-party conversations or speech interfaces for assistant devices.

In the remainder of this article we briefly review conventional approaches for selective hearing. We then detail the principles of the proposed SpeakerBeam approach and present experimental results confirming its potential. We conclude this article with an outlook on possible applications of SpeakerBeam and future research directions.

2. Conventional approaches for computational selective hearing

Much research has been done with the aim of finding a way to mimic the selective hearing ability of human beings using computational models. Most of the previous attempts focused on audio speech separation approaches that separate a mixture of speech signals into each of its original components [1, 3]. Such approaches use characteristics of the sound mixture such as the direction of arrival of the sounds to distinguish and separate the different sounds.

Speech separation can separate all the sounds in a mixture, but for this purpose it must know or be able to estimate the number of speakers included in the mixture, the position of all the speakers, and the

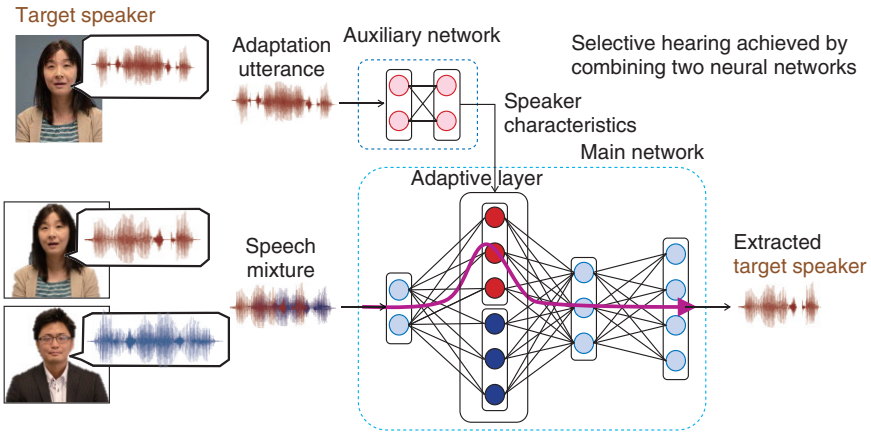


Fig. 3. Novel deep learning architecture developed for SpeakerBeam.

background noise statistics. These conditions often change dynamically, making their estimation difficult and thus limiting the actual usage of the separation methods. Moreover, to achieve selective hearing, we still need to inform the separation system which of the separated signals corresponds to that of the target speaker.

3. Principles of SpeakerBeam

SpeakerBeam focuses on extracting only the target speaker instead of separating all components in the mixture. By focusing on the simpler task of solely extracting speech that matches the voice characteristics of the target speaker, SpeakerBeam avoids the need to estimate the number of speakers, the position, or the noise statistics. Moreover, it can perform target speech extraction using a short adaptation utterance of only about 10 seconds.

SpeakerBeam is implemented by using a deep neural network that consists of a main network and an auxiliary network as described below and shown in **Fig. 3**.

(1) The main network inputs the speech mixture and outputs the speech that corresponds to the target speaker. The main network is a regular multi-layer neural network with one of its hidden layers replaced by an adaptive layer [4, 5]. This adaptive layer can modify its parameters depending on the target speaker to be extracted; namely, it can change its parameters depending on the characteristics of the voice of the target speaker provided by the auxiliary network.

(2) The auxiliary network is a multi-layer neural network that inputs a recording of only the voice of the target speaker (adaptation utterance) that is different from that in the speech mixture. The auxiliary network outputs the characteristics of the voice of the target speaker.

These two networks are connected to each other and trained jointly to optimize the speech extraction performance. Training the auxiliary network jointly with the main network enables the system to learn automatically from data the features that best characterize the target speaker's voice, thus avoiding the complex task of manually engineering features characterizing the target speaker's voice. Moreover, by training the network with a large amount of training data covering various speakers and background noise conditions, SpeakerBeam can learn to achieve selective hearing even for speakers that were not included in the training data. Details of the network architecture and training procedure are explained in our published report [2].

4. Performance of SpeakerBeam

We conducted experiments to evaluate the speech extraction performance of SpeakerBeam and its impact on speech recognition [2]. We used a corpus consisting of sentences read from English newspaper articles and created artificially mixtures of two speakers. Although SpeakerBeam can work with a single microphone, it achieves better performance when using more microphones. In this experiment, we used eight microphones and combined SpeakerBeam with microphone array processing (i.e., beamforming).

An example of processed speech using SpeakerBeam

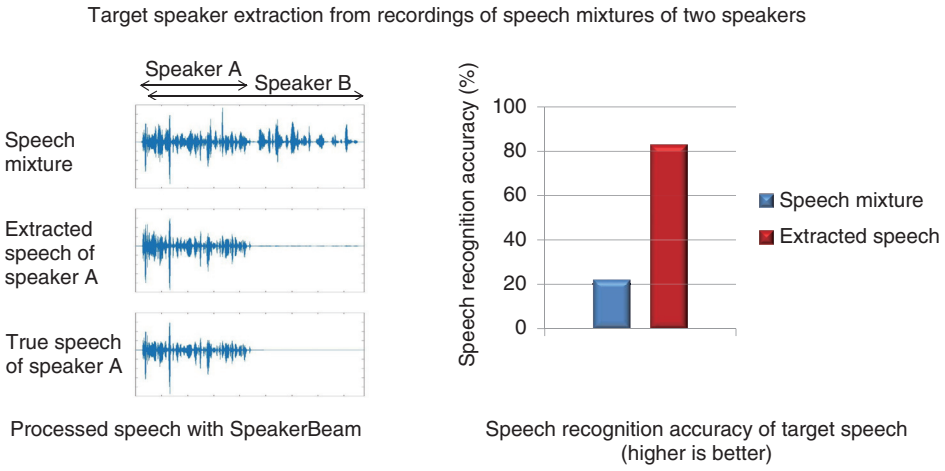


Fig. 4. Evaluation of speech extraction performance and automatic speech recognition with SpeakerBeam.

and the speech recognition accuracy obtained when recognizing mixtures of two speakers with SpeakerBeam (red bar) and without it (blue bar) are shown in **Fig. 4**. We observed a 60% relative improvement in speech recognition performance with SpeakerBeam.

SpeakerBeam can also be employed to improve the audible quality. Interested readers can refer to a video [6] to appreciate the target speaker extraction performance in realistic conditions (real recordings in reverberant conditions with music in the background).

5. Outlook

SpeakerBeam is a novel approach to perform computational selective hearing that offers several advantages compared to previous approaches. For example, it can track a target speaker regardless of the number of speakers or noise sources in the mixture and regardless of the speaker's position. This opens new possibilities for speech recognition of multi-party conversations, speech interfaces for assistant devices such as smart speakers, or for voice recorders and hearing aids that could focus on the speech of a target speaker.

However, there are some issues that need to be addressed before SpeakerBeam can be widely used. For example, speech extraction performance degrades when two speakers with similar voices speak at the same time. To tackle this issue, we plan to investigate improved target speaker characteristics that could better distinguish speakers and to combine target speaker characteristics with location information such as direction-of-arrival features.

Acknowledgment

Part of this development was achieved through a research collaboration with Brno University of Technology in Brno, Czech Republic.

References

[1] T. Hori, S. Araki, T. Nakatani, and A. Nakamura, "Advances in Multi-speaker Conversational Speech Recognition and Understanding," NTT Technical Review, Vol. 11, No. 12, 2013. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa3.html

[2] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning Speaker Representation for Neural Network Based Multichannel Speaker Extraction," Proc. of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017), Okinawa, Japan, Dec. 2017.

[3] S. Makino, H. Sawada, and S. Araki, "Blind Audio Source Separation Based on Independent Component Analysis," In: M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley (eds), Independent Component Analysis and Signal Separation—Proc. of ICA 2007, Lecture Notes in Computer Science, Vol. 4666, Springer, Berlin, Heidelberg, 2007.

[4] M. Delcroix, K. Kinoshita, A. Ogawa, S. Karita, T. Higuchi, and T. Nakatani, "Personalizing Your Speech Interface with Context Adaptive Deep Neural Networks," NTT Technical Review, Vol. 15, No. 11, 2017. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201711fa6.html

[5] M. Delcroix, K. Kinoshita, A. Ogawa, C. Huemmer, and T. Nakatani, "Context Adaptive Neural Network Based Acoustic Models for Rapid Adaptation," IEEE/ACM Trans. Audio, Speech and Lang. Proc., Vol. 26, No. 5, pp. 895–908, 2018.

[6] SpeakerBeam video in English, https://www.youtube.com/watch?v=7FSHgKip6vI
SpeakerBeam video in Japanese, https://youtu.be/BM0DXWgGY5A

Marc Delcroix
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

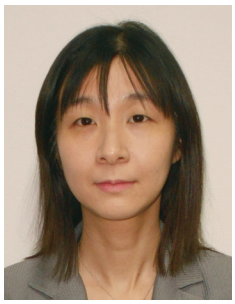
He received an M.Eng. from the Free University of Brussels, Brussels, Belgium, and the Ecole Centrale Paris, Paris, France, in 2003 and a Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, in 2007. He was a research associate at NTT Communication Science Laboratories from 2007–2008 and 2010–2012 and then became a permanent research scientist at the same lab in 2012. His research interests include robust multi-microphone speech recognition, acoustic model adaptation, integration of speech enhancement front-end and recognition back-end, speech enhancement, and speech dereverberation. He took an active part in the development of NTT robust speech recognition systems for the REVERB and CHiME 1 and 3 challenges, which all achieved best performance results in the tasks. He was one of the organizers of the REVERB challenge 2014 and of 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 (SP) Society Speech and Language Processing Technical Committee (SLTC). He is also a visiting lecturer at the Faculty of Science and Engineering of Waseda University, Tokyo. He received the 2005 Young Researcher Award from the Kansai section of the Acoustic 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.

Katerina Zmolikova
Ph.D. student, Brno University of Technology.

She received a Bc. degree in information technology in 2014 and an Ing. degree in mathematical methods in information technology in 2016 from the Faculty of Information Technology, Brno University of Technology (BUT), Czech Republic. Since 2013, she has been part of the Speech@FIT research group at BUT, where she is currently working towards her Ph.D. degree. Her research interests include robust speech recognition, speech separation, and deep learning.

Keisuke Kinoshita
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received an M.Eng. and a Ph.D. from Sophia University, Tokyo, in 2003 and 2010. Since joining NTT in 2003, he has been researching speech and audio signal processing. His research interests include single- and multichannel speech enhancement and robust automatic speech recognition. He received a Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2006, ASJ Technical Development Award in 2009, an ASJ Awaya Young Researcher Award in 2009, a Japan Audio Society Award in 2010, and the Maejima Hisoka Award in 2017. He is a member of IEEE, ASJ, and IEICE.



Shoko Araki
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 a Ph.D. from Hokkaido University in 2007. She joined NTT in 2000 and has since been engaged in researching acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis. She has been on the organizing committees of several conferences, including ICA 2003 (Fourth International Symposium on Independent Component Analysis and Blind Signal Separation), IWAENC 2003 (2003 International Workshop on Acoustic Echo and Noise Control), and WASPAA (IEEE Workshop on Applications of Signal Processing to Audio and Acoustics) 2007 and 2017, and HSCMA2017 (Fifth Joint Workshop on Hands-free Speech Communication and Microphone Arrays). She has also served as the evaluation co-chair of the Signal Separation Evaluation Campaign (SiSEC) 2008, 2010, and 2011. She has been a member of the IEEE Signal Processing Society Audio and Acoustics Technical Committee (AASP-TC) since 2014, and a board member of ASJ since 2017. She received the 19th Awaya Prize from ASJ in 2001, the Best Paper Award of IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, the Academic Encouragement 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 Young Scientists' Prize in 2014, an IEEE Best Paper Award in 2015, and an IEEE ASRU 2015 Best Paper Award Honorable Mention in 2015. She is a member of IEEE, IEICE, and ASJ.

Atsunori Ogawa
Senior Research Scientist, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. and M.E. in information engineering and a Ph.D. in information science from Nagoya University, Aichi, in 1996, 1998, and 2008. He joined NTT in 1998. He is engaged in research on speech recognition and speech enhancement. He is a member of IEEE and the International Speech Communication Association (ISCA), IEICE, the Information Processing Society of Japan (IPSJ), and ASJ. He received ASJ Best Poster Presentation Awards in 2003 and 2006.

Tomohiro Nakatani

Group Leader and 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. Since 2008, he has been a visiting assistant professor in the Department of Media Science, Nagoya University, Aichi. 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 SP Society AASP-TC from 2009 to 2014 and served as the chair of the AASP-TC Review Subcommittee from 2013 to 2014. He has been a member of the IEEE SP Society SL-TC since 2016. 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 WASPAA-2007, Workshop co-Chair of the 2014 REVERB Challenge Workshop, and as a General co-Chair of the IEEE ASRU. He is a member of IEICE and ASJ.