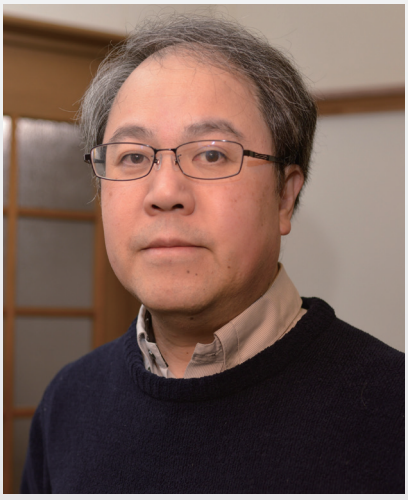
I Want to Create a World Like That of Astro Boy, Where People and Computers Share the Same Sound Space and Can Freely Cooperate with Each Other

Tomohiro Nakatani
Senior Distinguished Researcher, NTT Communication Science Laboratories



Abstract

Automatic-speech-recognition technology has developed rapidly and is now commonly used in voice interfaces such as those of smartphones and smart speakers; however, the technology must be further improved to enable smooth interaction between computers and humans. Tomohiro Nakatani, a senior distinguished researcher at NTT Communication Science Laboratories, has been at the forefront of research regarding speech enhancement, which removes ambient noise and reverberation from various sounds and accurately extracts only the sound that the person wants to hear. We asked him about the progress of his research and attitude as a leading researcher.

Keywords: voice-user interface, speech enhancement, convolutional beamformer, selective listening

Creating a voice-user interface that understands human speech in any environment

—Would you tell us about your current research?

My research goal is to create a technology that enables computers to distinguish the desired sound from a variety of sounds in an environment and recognize conversational speech. I talked about the same goal during my previous interview in 2016. Compared to back then, operating smartphones and other devices using a voice-user interface has become commonplace. However, even with the current speech-recognition technology based on artificial intelligence (AI), when a person speaks to a computer equipped with such technology, he or she needs to follow a special procedure and change the way of speaking so that the AI can understand. In the future, for robots and other devices to be able to integrate more deeply into our lives, it is necessary to create a natural voice-user interface that can properly recognize natural conversations of people to make users

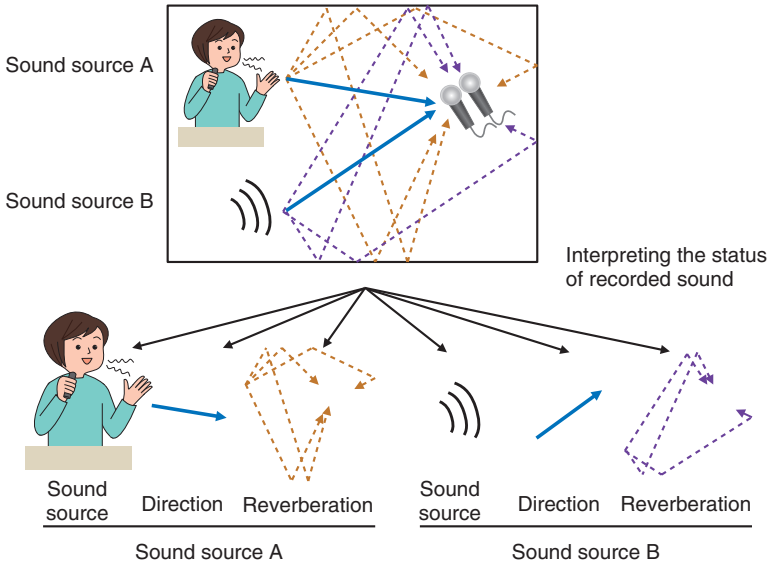


Fig. 1. Elemental decomposition of a recorded sound using multiple microphones.

feel as if they are talking to people in daily conversation.

To make such a natural voice-user interface a reality, I'm currently researching *speech enhancement*, which is a technology to distinguish speech of a target speaker from a recorded sound. In our daily lives, we don't always talk near a microphone. Speech data recorded with a microphone at a distance from the person speaking will include sound reflected off walls (reverberation), the voices of multiple people (speech sources), and background noise. By applying dereverberation, source separation, and denoising, speech enhancement produces speech with quality equivalent to that recorded with a microphone near the mouth of a particular speaker.

We take a two-pronged approach to researching speech enhancement. The first approach is *elemental decomposition* of a recorded sound (**Fig. 1**) with which the sound is decomposed into individual sound elements on the basis of the different physical and statistical properties of the elements. This elemental decomposition is a technological area suited to computers; in other words, it is something that humans cannot do, but computers excel at. One of our recent achievements is the implementation of the world's first convolutional beamformer (BF) that can simultaneously and optimally execute dereverberation, source separation, and denoising. It had only been possible to execute each process individually; thus, it was not possible to get the best performance out of each process when the processes were combined. Our convolutional BF makes it possible to further improve the quality of the enhanced speech from a recorded sound.

The second approach to speech enhancement is *selective listening* (**Fig. 2**) with which only the speech of the person one wants to hear is extracted. This is a technique that people apply in their daily lives; however, until recently, it has been impossible for computers. This approach is represented by our technology called SpeakerBeam, which uses deep learning to extract only the speech that matches the voice characteristics of a particular speaker (as described in articles in the November 2018 [1] and September 2021 [2] issues of this journal). It was the world's first demonstration of human-like selective listening. We have also developed a technology called Recurrent Selective Attention Network, which can accurately estimate who said what and when from ever-changing conversations (conversation analysis).

I believe that a natural voice-user interface can be developed by taking advantage of the technological aspects that computers excel at while simultaneously stepping into the processes that humans excel at.

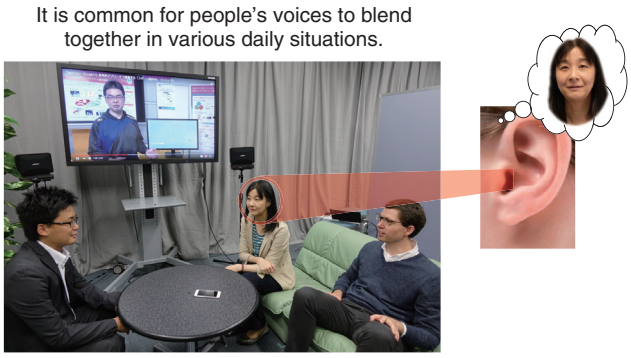


Fig. 2. Selective listening.

It is a good idea to listen to your seniors and others even if you feel that their way differs from your own

—Was there a reason for starting to work on this topic? What do you keep in mind when searching for a research problem and topic?

Since I joined NTT, I have been working on source separation. When I began my career at NTT, research on source separation was still in its infancy, and there was an atmosphere in which people said, "There's no way we can do that; there are other problems that need to be solved first." Nevertheless, I continued to pursue source separation, believing that if humans can do it, computers should be able to do it, too. Currently, it may not be easy for me to continue research that does not always get the approval of others; however, at that time, I believed in my feeling and pursued research on a topic that I wish I could pursue rather than something I could do right away. As a result, I have been able to continue my research in this field for 30 years, so I think it must have been a really good topic.

I think that, in a sense, searching for research topics and problems is like solving a puzzle, that is, exploring different possibilities until the puzzle is solved. For example, a technology for distinguishing between multiple voices did not exist. Accordingly, after wondering how such a task could be achieved, I made a hypothesis about the characteristics of a voice from the frequency distribution of the speech signals, repeated this step for other speech signals, and identified a core element that is common to many hypotheses, setting it as a research topic.

This process is a matter of knowledge and intuition. Regarding knowledge, you have to study hard, which means reading the latest technical papers and studying textbooks for traditional techniques. The more you do that, the more your skills will improve. There is no other way to improve your knowledge. Regarding intuition, I think it depends on whether you have the vision to think "If it were me, I would do this." I think that is an important ability to have both as a researcher and as a person.

—Even if you become a senior distinguished researcher, you are studying every day.

As the world develops technologically, we cannot afford not to study. In the field of neural networks, the speed of computers has increased dramatically, and the number of researchers has increased so much that research is progressing at a rapid pace, so if you don't study hard, you will never keep up. However, just studying haphazardly will not contribute to actual research. You need to use your intuition and knowledge at the same time, and always think about how you can apply them to what you want to do.

It is also important to have opportunities for discussions with fellow researchers at international conferences. You can only collect a limited number of ideas with your eyes and ears; therefore, it is undoubtedly important to broaden your horizons through discussions with other researchers. In the past, I was instructed by my seniors to "meet people" and "visit

research institutions" when I went to international conferences and other events, and I was almost forced to visit researchers halfway around the world. At that time, I had little experience as a researcher and no confidence in communicating in English, so meeting those demands was a very high hurdle for me. Regardless, through such efforts, I was able to come into contact with matters that I had no idea about while getting to know the researchers involved in those matters. Although getting to know someone at a conference is important, visiting their research institutions is the only way to learn many things and discuss them in depth. Even during the current COVID-19 pandemic, I continue to keep up with the human relationships that I nurtured through email and teleconferences.

It may seem like a hurdle to contact a researcher you don't know to ask for a meeting, but the first step is to send an email. In fact, researchers are generally very kind and basically accept meeting with you because they themselves want to get to know other researchers. Sometimes other researchers contact me for the first time, and I never turn them down. By learning the importance of and actually contacting people, I was able to connect with leading researchers in the field of dereverberation and hold workshops with like-minded researchers. Researchers are caught in a world of self-satisfaction if they are not recognized by others; therefore, if you want to be recognized worldwide, you should listen to and refer to the opinions of your seniors and others around you, even if you feel their opinions differ from your own.

Researchers may be like boxes

—Would you tell us what type of attitude should researchers have?

First, I think it is vital to let society know what kind of research you are working on and what problems you want to solve. When you declare your intentions and engage in activities, you can gather a variety of information, and sometimes, you are given solutions that you would not have found on your own.

I believe that researchers are like boxes. A label written on the box states the research agenda. I collect various data myself for the box, and other people send me data and information to add to the box. This action facilitates a project to develop a certain technology, and how much the technology can be developed may depend on the ability of the researcher. Moreover, to collect beneficial information, the way the label is written is also important, so you have to think carefully about how to express it.

The biggest concern of researchers is whether the problem you are working on is something you should really be working on or something you really need to solve, and you must constantly verify it and make the best of it. Fortunately, I found a very good research topic at the beginning of my career and have been working on it for 30 years with almost the same basic *label* and goal. However, as the world has been rapidly developing technologically, I am beginning to think that some parts of my research have become outdated. It may be necessary to make a major upgrade to my research topic and face a new challenge. Assuming that I can continue my research until I am 65 years old, I have about 10 years left as a researcher. I want to think about how I can make those 10 years worthwhile and contribute to society and my research field.

As a researcher, I continue to be interested in enabling a computer to understand sound in the same manner as humans. I don't think that interest will change in the future. A computer should also be able to do what a human can do. In reality; however, it is quite difficult to do that, which makes it all the more interesting, and it has become my life's work. Many problems remain to be solved; even so, in cooperation with many researchers, I want to create a world like that of Astro Boy in which people and computers can share the same *sound space* in any situation, and talk appropriately and cooperate freely while understanding the surrounding circumstances.

—Do you have any words of advice for junior researchers?

My basic advice is to find fellow researchers and create labels for yourself. You can have a cohort of fellow researchers in many situations. For example, in the workplace, there are people you can trust and discuss things with. If you attend a domestic conference, there are fellow researchers who are your rivals. If you participate in an international conference outside Japan, you can link up with fellow researchers who are doing cutting-edge research. As I mentioned before, the technological-development speed of the world and research activities is ever increasing, so we need to work in unison with our fellow researchers to keep up. I have been fortunate to have had many opportunities to collaborate with overseas research institutions and build mutually beneficial relationships.

If you feel stuck in your life in general, you should ask for help and enjoy your life. Because researchers sometimes challenge themselves too seriously when confronting a problem, they tend to shut themselves away if left alone. To prevent that situation from occurring, it may be necessary to ask your seniors to give you support.

Research activities mostly end in failure. At the time of my previous interview, I had a young child, so I compared research to raising a child and said that even if you fail, a seed of research will always be found in that failure. That feeling has not changed. If the result was not what you expected, a different law of nature from what you believed must exist there. By thoroughly examining the situation, you may get the chance to find the next seed of research. When you are inexperienced, you may have a tough time making decisions due to the lack of materials to verify. Nevertheless, you should face the consequences and accumulate experience through repeated failures. If you don't fail, you won't succeed, so don't let failure discourage you from finding the seeds of success.

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, 2018.
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr201811fa2.html

[2] M. Delcroix, T. Ochiai, H. Sato, Y. Ohishi, K. Kinoshita, T. Nakatani, and S. Araki, "Developing AI that Pays Attention to Who You Want to Listen to: Deep-learning-based Selective Hearing with SpeakerBeam," NTT Technical Review, Vol. 19, No. 9, pp. 39–45, 2021.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202109fa4.html

■ Interviewee profile

Tomohiro Nakatani 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 and a visiting assistant professor in the Department of Media Science, Nagoya University, 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, and the 2017 Maejima Hisoka Award. He was a member of the IEEE Signal Processing Society (SPS) Audio and Acoustic Signal Processing Technical Committee (AASP-TC) from 2009 to 2014 and a member of the IEEE SPS Speech and Language Processing Technical Committee (SL-TC) from 2016 to 2021. 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 2017 IEEE ASRU. He is a fellow of IEEE, and member of IEICE and ASJ.