# Research is Like Child Rearing—Many Failures Hold the "Seeds of Research"

*Tomohiro Nakatani*
*Senior Distinguished Researcher, NTT Communication Science Laboratories*

**Overview**

Interfaces using speech recognition have become common practice these days. However, commonly used technologies for this purpose can suffer a drop in recognition performance in noisy environments or when the microphone is too far from the speaker. There is therefore a growing need for technologies that can provide more robust and accurate speech recognition. We asked Dr. Tomohiro Nakatani, Senior Distinguished Researcher at NTT Communication Science Laboratories, whose technology last year achieved the world's highest performance for speech recognition in noisy environments, to tell us about his recent research results and his approach to research.

*Keywords: speech recognition, speech recognition interface, dereverberation*

## Achieving a speech recognition interface that can understand human conversation in diverse environments

*—Dr. Nakatani, please tell us about your current area of research.*

I am researching natural speech recognition interfaces (**Fig. 1**). For example, some of you may have experienced operating a smartphone by speech recognition. Speech recognition is a simpler and more convenient way of inputting information than using a keyboard. In recent years, we have seen wider use of voice-operated smartphones and tablets, so the usefulness of speech recognition interfaces has become well known. Current smart devices, however, require the user to bring the microphone up close and to speak distinctly. But in the real world, people communicate freely with each other without worrying about the existence of a microphone.

I wanted to create a mechanism that enables anyone to access information or to converse with robots simply by speaking without paying attention to electronic devices, so I have been doing research in this field ever since I joined NTT.

This research has been advancing, and as a result, the smart home is on its way to becoming a reality. At the present stage, the goal is to enable users in a smart home with a microphone installed in the living room to operate home appliances through vocal commands. The key to this mechanism is technology that can recognize speech even for utterances made at a

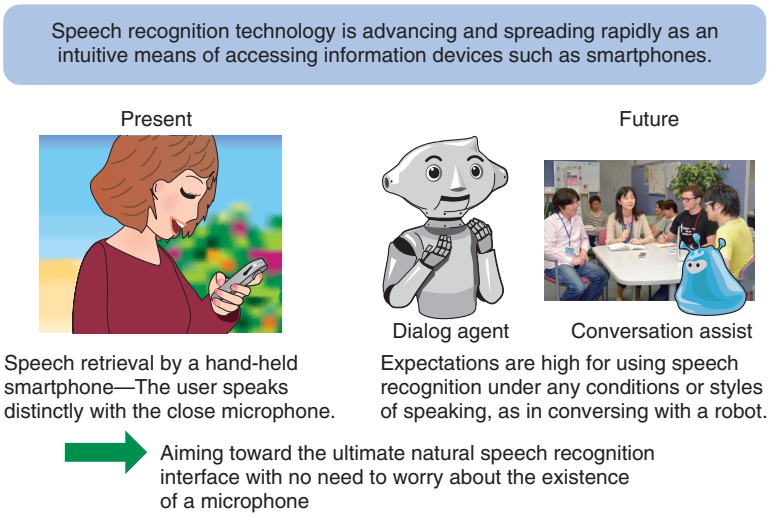


Fig. 1. Natural speech recognition interfaces.

distance from the microphone. An era is approaching when a computer will be able to accurately recognize the speech of each person and extract information in a situation in which a number of people are sitting around a table and conversing.

## New challenges and friendly competition among experts who develop world-class technologies

*—This sounds like technology that can have a profound effect on our lives.*

Actually, NTT has already developed advanced technologies that can accurately recognize and process speech even in noisy environments or when there is reverberation. In fact, NTT has been a world leader in these technologies. For example, in last year's international technology evaluation competition of mobile speech recognition in noisy public spaces (CHiME-3 Challenge), NTT's system won first place in speech recognition accuracy at a level that was significantly higher than the second-place system. This is a result of developing a number of key technologies. These include technology for reducing noise and reverberation without distorting the user's speech, and deep-learning speech recognition technology for accurate modeling of speech even under noisy conditions (**Fig. 2**). Our word error rate (WER) in this competition was about 5%. In contrast, the WER of speech recognition by a conventional deep neural network (a machine learning technique simulating the human brain) was 33%. In addition, an investigation by a major Chinese search engine company found that the rate of mistakes made by people when listening to speech was about 11%. You can see that the WER that we achieved significantly outperformed other systems. This NTT technology should provide a foundation for speech recognition that exceeds human capabilities even in noisy environments.

Let me talk about these technologies in easy-to-understand terms. Please visualize a scene in which a number of people are talking freely inside a room. If an audio recording was made at this time without placing a microphone near the speakers, the accuracy of speech recognition would dramatically drop. There are two main factors behind this deterioration. One is that the quality of the speech recorded via the microphone is poor. This is because of background noise such as from air conditioning and the effects of reverberation caused by speech waves arriving slightly late at the microphone after reflecting from walls or other objects. In addition, overlapping speech from more than one person may be recorded. The other factor is that pronunciations are often ambiguous, and words are often omitted since speakers are talking freely without paying attention to the microphone.

To deal with such factors, we need speech enhancement technologies such as denoising, dereverberation, and speech separation to reduce the effects of noise, reverberation, and other people's voices, respectively. We also need natural-language speech recognition technologies for accurately recognizing unconstrained speech. I myself am mainly in charge

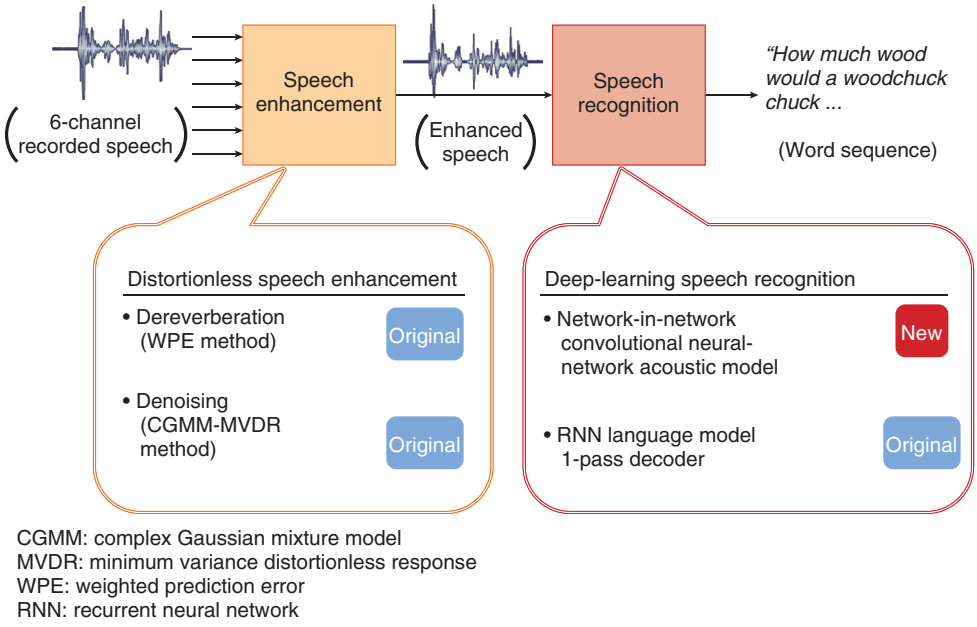


Fig. 2. NTT system features.

of speech enhancement technologies for distinguishing a person's voice.

Since we came out first in the CHiME-3 Challenge, we can now look back with a sense of joy, but at the time, I can honestly say that we were quite apprehensive about what kind of evaluation our technology would receive. However, I felt that we had to live up to the NTT name, so I wanted to do whatever it took to take first place. We established a WER target within 8% and took up the challenge of a new and difficult task. Somehow, we achieved good results thanks to our efforts.

*—There must have been some drama behind the birth of this technology.*

Yes, there were all kinds of dramatic stories, but let me tell you about our Signal Processing Research Group, where I am currently group leader. In 2000, the year following the reorganization of NTT, this group was set up in NTT Communication Science Laboratories (NTT CS Labs) thanks to the efforts of the late Dr. Yoichi Tokura, the director of NTT CS Labs at the time, Dr. Shigeru Katagiri, the first group leader (currently Doshisha University professor), and others. The mission established for this group was to remake speech processing technology starting from the basics and discover new possibilities that exceed conventional limits. The group brought together an elite team of basic researchers in two different technologies—audio processing and speech recognition. I joined this group as a researcher in 2001.

From the start, the aim of this group was to develop natural speech recognition interfaces. To develop this "dream technology," we would have to solve a number of challenging problems that would take a long time to overcome. These include development of speech separation technology to distinguish speaker voices when more than one person is speaking at the same time, dereverberation technology to remove reverberation from speech recorded in a reverberant room and make it easier to hear that speech, speech activity detection technology for detecting speech intervals from recorded sound mixed with noise, speaker diarization technology for estimating who among multiple speakers is speaking and when, acoustic modeling technology for accurately modeling and classifying speech even under noisy conditions, and computationally highly efficient decoding technology for real-time continuous speech recognition that can handle an ultra-large vocabulary of more than 10 million words. These technologies reflect the many problems that had to be overcome.

It's been about 15 years since then, and I can say that many of these problems that we took to be very challenging at the time have been solved. We have

Fig. 3. Demonstration of on-site recognition of recorded conversation.

come a long way, and at the NTT Communication Science Laboratories Open House 2016 held earlier in the year, we introduced a demonstration system that performs on-site computer recognition of the speech of multiple speakers sitting and conversing at a table with a microphone placed at the center (**Fig. 3**). Please take a listen!

*—This speech is really clear after removing noise and reverberation!*

A driving force behind this development and one of its features is an assembly of researchers from the two different fields of audio processing and speech recognition who work side by side in our group. These fields may appear similar at first glance, but their basic technologies and application targets are quite different. They have progressed along different paths. Actually, our group that is carrying out these two lines of research together is unique even on a worldwide basis. Exchanges between these fields in terms of products or academic activities are not that frequent. Through pioneering research that places importance on this boundary region, I believe that we have given birth to a string of new ideas that have helped propel speech research at NTT.

To give an example of just how important the intersection between these two fields has been, let me tell you about dereverberation technology, which I have been deeply involved in. Research into removing reverberation is an area of audio processing that has been progressing for quite some time. The idea here is to mathematically model the propagation of sound in a room and to use that model to remove unnecessary sound such as reverberation from recorded sound. However, this method has not been effective for adequately removing reverberation if the conditions of the room in which sound is to be recorded are unknown.

In speech recognition, in contrast, there is a commonly used technique for learning patterns in speech signals and for processing signals to identify those patterns. (In modern parlance, this process would probably be called a machine learning framework.) With this in mind, we proposed a framework called "pattern-oriented audio processing" that incorporates the idea of pattern processing developed in speech recognition into audio processing. We then proceeded to develop a variety of new algorithms based on this framework. In this way, we developed a technology that automatically recognizes what is happening within recorded sound under unknown conditions and that uses a mathematical model from audio processing to accurately break down the features of the

target sound. Amid this flow of developments, a new dereverberation algorithm was intensively researched around 2006 here at NTT CS Labs in Keihanna (Kyoto-Osaka-Nara) Science City, resulting in a world-first technology.

Another reason as to why the Signal Processing Research Group has been able to produce such breakthrough results is that our research has continued along the path of basic research pioneered by many great senior researchers at NTT. Since the invention of audio coding for mobile phones in the Nippon Telegraph and Telephone Public Corporation era, NTT has a history of being a world leader in speech research. There has also been a strong, trustworthy relationship between NTT researchers and worldwide researchers for many years. This research environment that enables us to share a view of the world with these great senior researchers (at least in part) has been a driving force behind our basic research endeavors.

## Finding the "seeds of research" through many failed attempts

*—What kind of mindset is important when beginning a research project?*

I would like researchers to take three points to heart. First, when trying something for the first time, be prepared for things to go against your plan. Second, even if you face up to the fact that the result was not what you expected, keep in mind that the result may be useful in later research in some form. Third, after long and careful attempts to achieve results, the result that you are seeking may suddenly appear.

For example, we initially experienced a number of failures in our research on dereverberation technology that we eventually achieved as a world-first. Dereverberation had been a major problem for some time in the field of audio signal processing, but a definitive solution had eluded researchers. In the long history of this research, it was believed that the biggest problem in achieving effective dereverberation was determining how to prevent the occurrence of speech whitening, a phenomenon in which the results of processing are a flat speech spectrum and a loss of natural speech characteristics. We ourselves tried all sorts of ideas to solve this problem. However, while applying countermeasures to speech whitening had a somewhat positive effect on improving dereverberation, we were still short of a fundamental solution. It was really one step at a time in the dark.

Amid these trials and tribulations, the pattern-oriented audio processing that I mentioned earlier became our research compass. In particular, we introduced a technique commonly used in speech recognition that learns and distinguishes temporal changes in sound as patterns, and we used it to approach the dereverberation problem. After beginning with some primitive trials, we failed repeatedly in our experiments. But as partial successes began to appear, we started to make discoveries a little at a time. One day, however, after a long period of trial and error, we suddenly arrived at an important discovery. Keeping the temporal structure of the speech from collapsing is more important than preventing spectrum flattening. I really felt as if the veil had fallen from my eyes, even with all the experience I had had as a researcher. With this discovery, we shifted our research to creating a mechanism for removing reverberation without destroying the temporal structure of the speech. Our research into dereverberation began to accelerate from that point on.

*—As an active researcher, what does it take to produce results? What would you say to young researchers?*

Thoroughly testing results that go against your expectations will reveal the next "research seed." Experiencing many failures is an interesting thing. It gives you an opportunity to think about what went wrong. Indeed, from the viewpoint of always failing, I think research and child rearing can be very similar! I myself have two sons in primary school in the third and fifth grades. For example, thinking that I want my kids to be more careful with their time, I promise to give them a reward if they follow a rule that I set down. However, while they may observe the rule perfectly the first time, it is not uncommon for them to completely ignore it the second time. I then wonder why it is that what I expected to happen turned out to be so different from what actually occurred. This sensation is similar to searching for the cause of a failure in research. If I think carefully about it, an idea as to what to do next will eventually come to me. If things go well using this idea, that's great, but if things go bad, I can put another idea to the test. Progress can still be made with failures, though it may happen a little at a time. Conversely, if you think that to fail is common and treat research as something like a game after failing, it might prove to be enjoyable, just like raising kids.

A researcher with little experience will not have

much background in performing tests and may find it difficult to make good decisions. I would therefore like researchers to get in the practice of thinking more about their results. Having one failure after another is not good, but a success should also be questioned as to why it occurred.

Researchers should also make an effort to find topics that they find interesting. Although it is often said, it is nevertheless true that you cannot continue what doesn't interest you for a long period of time. Being a successful researcher requires that one has a clear interest in something and knows one's own strengths.

*—Dr. Nakatani, what is your outlook for the future?*

I envision that the technologies introduced here will be used to develop speech recognition systems that can operate smoothly even in noisy locations where many people are speaking. These might be speech recognition interfaces for public spaces such as cafés and airports and conversation recognition products for offices and home living rooms. We can expect these developments to contribute greatly to the expanded use of smartphone speech agents and communication robots. For example, I would like to see growth in technologies that can support the hearing-impaired or people speaking different languages for whom communication is difficult, and technologies that can help people connect with each other. I am committed to continuing my research efforts while keeping in mind that these future visions will be achieved not that far off—maybe no more than a few years from now!

■ **Interviewee profile**
**Tomohiro Nakatani**
Senior Distinguished Researcher, Supervisor, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.E., M.E., and Ph.D. from Kyoto University in 1989, 1991, and 2002. He joined NTT Basic Research Laboratories in 1991 and moved to NTT Communication Science Laboratories in 2001. During 2005–2006, he was a visiting scholar at Georgia Institute of Technology, USA. Since 2008, he has been a visiting assistant professor in the Department of Media Science, Nagoya University, Aichi. His research interests include speech enhancement technologies for intelligent human-machine interfaces. He received the 1997 JSAI (Japanese Society for Artificial Intelligence) Conference Best Paper Award, the 2002 ASJ (Acoustical Society of Japan) Poster Award, the 2005 IEICE (Institute of Electronics, Information and Communication Engineers) Best Paper Award, and the 2009 ASJ Technical Development Award. During 2009–2014, he was a member of the IEEE (Institute of Electrical and Electronics Engineers) Signal Processing Society Audio and Acoustics Technical Committee (AASP-TC) and has been an associate member since 2015. He has been a member of the IEEE Signal Processing Society Speech and Language Processing TC since 2016. He served as the Chair of the review subcommittee of AASP-TC during 2013–2014, an associate editor of the IEEE Transactions on Audio, Speech, and Language Processing during 2008–2010, the Chair of the IEEE Kansai Section Technical Program Committee during 2011–2012, a Technical Program co-Chair of the IEEE WASPAA-2007, and as a member of the IEEE Circuits and Systems Society Blind Signal Processing Technical Committee during 2007–2009. He is a member of IEEE, IEICE, and ASJ.