Media Processing Technology for Business Task Support

Takanobu Oba, Kazunori Kobayashi, Hisashi Uematsu, Taichi Asami, Kenta Niwa, Noriyoshi Kamado, Tomoko Kawase, and Takaaki Hori

Abstract

This article introduces two aspects of work toward implementation of services to support business tasks, specifically the use of speech recognition in very noisy environments such as factories and construction sites, and technology for recording minutes of meetings. The most recent audio and speech processing technology that is applied in these services is also described.

Keywords: intelligent microphone, meeting minutes support, speech recognition

1. Introduction

The progress achieved in information and communication technology (ICT) has increased the efficiency of various tasks. For example, conversion of paper documents to electronic form allows information to be accessed and managed over a network and simplifies search and display of statistical information. A disadvantage, however, is that it is sometimes difficult to enter information using a keyboard in environments where there are no desks such as outdoors and in situations where both of a person's hands are occupied by a task such as driving a vehicle or operating machinery. These problems have created a barrier to the use of ICT. Another problem is the recording of meeting minutes and similar situations in which people exchange information verbally. The process involves manually entering information into electronic documents by relying on memory or written notes, which is time-consuming and may produce incomplete results.

2. Speech recognition and noise reduction technology

We describe here two examples of the use of voice recognition in noisy environments such as factories or construction sites and the use of speech recognition in meetings to support the production of minutes.

2.1 Intelligent microphone capable of speech recognition in noisy environments

The intelligent microphone consists of multiple microphones and is based on acoustic signal processing technology that segregates the target speech from other sounds such as other voices or background noise. The intelligent microphone makes it possible to pick up the speaker's voice clearly even in very noisy environments (**Fig. 1**).

This microphone enables high-quality telephone calls (clear hands-free calling) (**Fig. 2**) and highly accurate voice recognition, even in noisy environments such as in factories or construction work sites or in an automobile traveling on a highway.

The acoustic signal processing involves the use of spatial power distribution, frequency characteristics, and temporal fluctuation characteristics to estimate the spectral filter in order to segregate the target speech from other noise (**Fig. 3**). Adapting the signal processing to microphone observations makes it possible to reduce the ambient noise power by a factor of 1/10,000 with less degradation of the target voice signal. This performance enables accurate speech

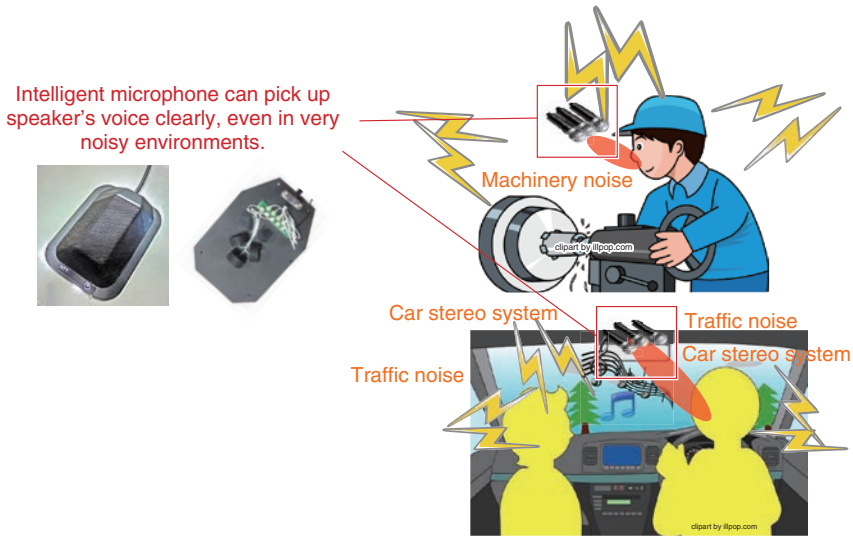


Fig. 1. Concept of the intelligent microphone.

Headset-type intelligent microphone

Fig. 2. Photo and use scenario of intelligent microphone.

recognition and high quality telephone calls even in extreme 100-dB-level noise environments. The steps involved in acoustic signal processing are as follows:

(1) Hybrid of beamforming and spectral filtering

With some microphones, sharp directivity cannot be formed. Thus, enhancement of the target speech cannot be achieved by simply applying beamforming. Adapting a spectral filter to the beamforming output makes it possible to suppress noise efficiently.

(2) Estimation of spatial interference noise power distribution

The desired speech source cannot be enhanced by simply applying conventional beamforming. Therefore, multiple beams are formed to estimate the spatial interference noise power distribution. If the estimated power of the desired sound and interference noise differ, the noise output power will be efficiently reduced.

(3) Reduction of diffuse noise using temporal fluctuation

The observed signals include diffuse noise such as that from an air conditioner. The power spectrum of diffuse noise changes over time. By utilizing these characteristics, we can accurately estimate the frequency spectrum of diffuse noise.

This newly developed intelligent microphone can be implemented for various terminals and contribute to speech services in noisy environments.

2.2. Support for producing meeting minutes

Meetings occur frequently in the business world. What was talked about in detailed and complex discussions is often forgotten a few days later, though, so the content is recorded in the form of minutes. That, however, is not a simple task and requires considerable time. Taking careful notes during the meeting for later preparation of minutes makes it difficult to

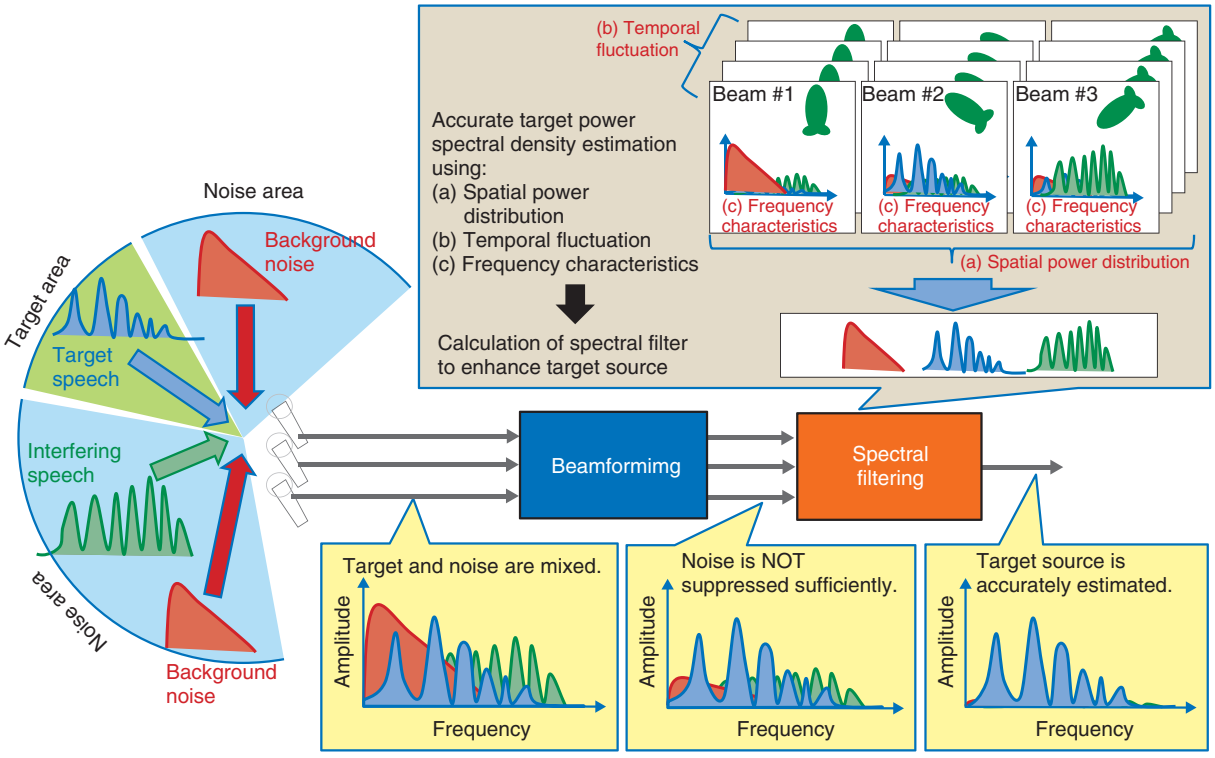


Fig. 3. Acoustic signal processing with the intelligent microphone.

follow the discussion, and priority on note-taking interferes with participation in the meeting.

NTT is developing a service for efficient production of minutes for National Diet sessions and the Assembly meetings of local governments. As a replacement for stenographers, the system supports the creation of minutes by using speech recognition for speech-to-text conversion. The next development target in this project is a system for recording everything that is spoken in business conferences and meetings as text. Having the content in text form would make it possible to rapidly search for particular parts of particular meetings and to retrieve the results. Because the voice recordings also remain, it is also possible to listen to any part of the meeting again.

In this work, real-time speech recognition during meetings is important. This has multiple advantages, one of which is that important comments can be tagged at the meeting (**Fig. 4**). As soon as a speaker utters a comment, the content of the utterance is recognized, transcribed, and displayed on the personal computer. Any participants looking at the display can tag an important utterance simply by clicking it, which is depicted as a green star in Fig. 4. The text consisting of only the tagged comments would appear as simple meeting minutes. Another advantage is that it can be used as a tool for promoting conversation, thus extending use beyond the framework as a simple meeting minutes creation support system. The system in current development has a tag cloud function, which displays words that have been used in the meeting with high frequency (**Fig. 5**). When participants can view the keywords during the discussion, they can keep and organize a view of what has been said as long as the discussion continues, which is likely to stimulate ideas.

Real-time speech recognition also makes it immediately clear which participants speak more and which speak less. Also, speakers may create a negative impression by speaking too rapidly, a manner that we should correct. It is difficult to realize how we are speaking while concentrating on the meeting, but this system may help us by bringing the problem to our attention.

Performing real-time speech recognition in this way can be expected to increase the productivity of meetings. The reason many people think that long

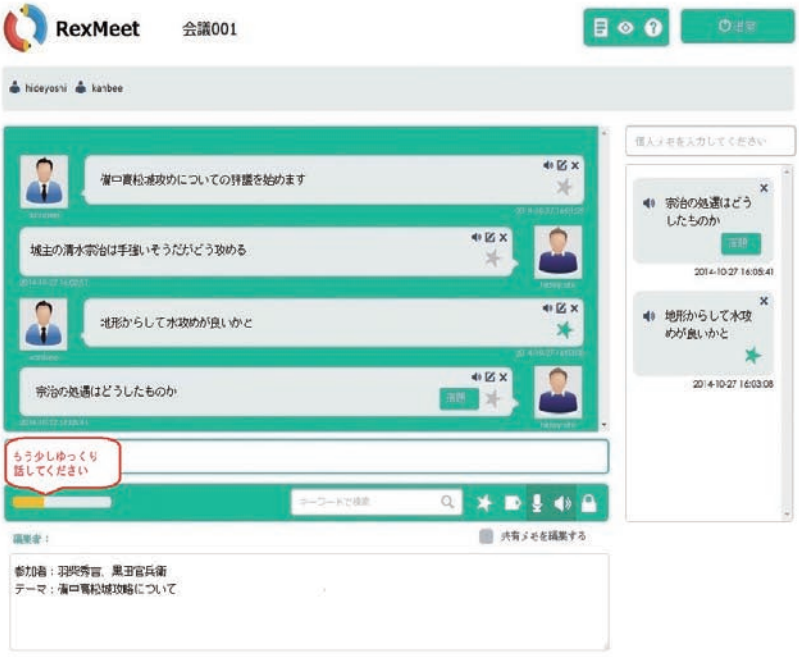


Fig. 4. Screen display of meeting minutes support system.

Fig. 5. Tag cloud.

meetings and conferences are not interesting is probably that daily meetings and other such meetings are low in productivity. We believe that speech recognition can contribute to improving this situation. Although minutes production support is one current target for this technology, we are moving forward with research and development (R&D) aimed at providing support for the meeting process itself.

3. Voice activity detection (VAD)

Voice activity detection (VAD) is a basic and essential

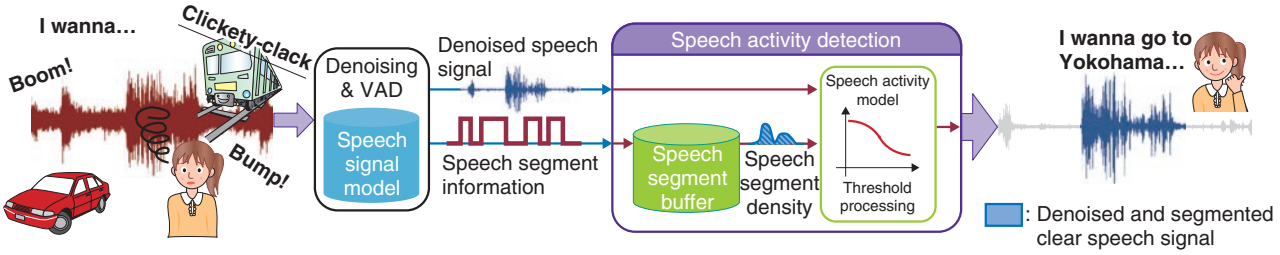


Fig. 6. VAD taking voice activity density into account.

function for many voice applications that involve speech recognition. VAD technology identifies the parts of a microphone signal that actually contain speech by monitoring the microphone signal input and detecting the points where speech begins and ends. Errors in detecting those points are a major problem that greatly reduces the utility of a speech interface for application programs. Detection errors can cause the application to react before the user finishes speaking or result in the application failing to react even after the user finishes speaking, causing a delay.

One difficult problem in VAD is determining whether a short interval is the end of an utterance or part of the utterance. NTT has addressed this problem by developing a technique to identify the temporal density of speech signals that takes the density of speech segments into account (**Fig. 6**). Speech varies in a characteristic way as an utterance progresses. By processing that variation according to our original speech model, we succeeded in extracting the utterance segments of a user with high accuracy. Furthermore, by combining VAD technology previously developed by NTT with technology for simultaneously executing noise suppression, we were able to reduce VAD errors by at least 15% compared to conventional technology, even in a high-noise environment.

4. VoiceRex NX2014 real-time DNN speech recognition engine

The term *deep learning* is currently attracting a lot of interest in media processing fields such as speech recognition and image recognition. Deep learning is mainly pattern processing that uses deep neural networks (DNNs), where *deep* refers to many layers in a neural network. The use of DNNs has remarkably improved the accuracy of speech and image recognition.

This has surprised many researchers because it had been strictly shown mathematically that the representational power of a neural network of three or more layers cannot be increased by adding layers. Nevertheless, when it comes to the question of whether or not the representative capability can be controlled by using an engineering approach, increasing the number of layers while restricting the number of nodes per layer (as in DNNs) is completely different from the conventional approach of increasing the number of nodes while keeping the number of layers at three.

DNNs are used in speech recognition as an acoustic model for determining what phoneme (“ah,” “ee,” etc.) is represented by a particular part of the speech input (**Fig. 7**). Conventional techniques have not been able to achieve a practical level of recognition accuracy when applied to speech recognition in conversations. With the emergence of DNN, however it has become possible to achieve highly accurate conversational speech recognition.

NTT developed the VoiceRex NX2014 speech recognition engine, introducing DNNs at an early stage. One feature of the VoiceRex series is high-speed recognition processing. However, the high computational cost of DNNs resulted in slower processing. That problem spurred the development of various techniques for increasing speed, and the result was a speech recognition engine that is capable of processing speed at the same high level as the original system. From the standpoint of speech recognition service providers, this technology makes it possible to provide services with a speech recognition function that is more accurate yet just as fast as was possible previously. From the user’s viewpoint, the speech recognition engine returns highly accurate results immediately after the user speaks.

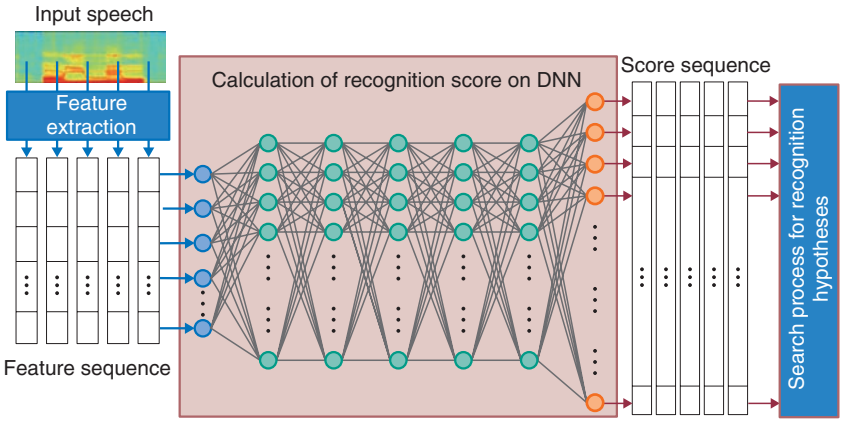
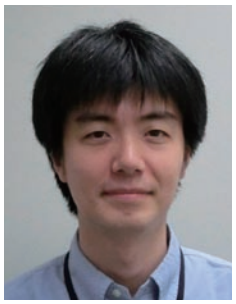


Fig. 7. DNN speech recognition mechanism.

5. Future development

We have described our work on speech processing as a core technology for supporting tasks in business situations. Language processing and image processing also play important roles in supporting business tasks. By gaining a deeper understanding of human language and of the objects that people see and the situations that exist around them, we can achieve more intelligent support in a wider variety of scenarios through the use of ICT equipment. The NTT laboratories are engaged in R&D of various types of media including speech, language, and images. Our objective in the future is to provide more advanced support for work tasks and for people in general through an organic integration of media.



Takanobu Oba
Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E., M.E., and Ph.D. in engineering from Tohoku University, Miyagi, in 2002, 2004, and 2011, respectively. In 2004, he joined NTT, where he engaged in research on spoken language processing at NTT Communication Science Laboratories. He has been with NTT Media Intelligence Laboratories since 2012. He received the 25th Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2008. He is a member of ASJ, the Institute of Electrical and Electronics Engineers (IEEE), and the Institute of Electronics, Information and Communication Engineers (IEICE).

Kazunori Kobayashi
Senior Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E., M.E., and Ph.D. in electrical and electronic system engineering from Nagaoka University of Technology, Niigata, in 1997, 1999, and 2003, respectively. Since joining NTT in 1999, he has been engaged in research on microphone arrays, acoustic echo cancellers, and hands-free systems. He is a member of ASJ and IEICE.

Hisashi Uematsu
Senior Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E., M.E., and Ph.D. in information science from Tohoku University, Miyagi, in 1991, 1993, and 1996. He joined NTT in 1996 and has been engaged in research on psychoacoustics (human auditory mechanisms) and digital signal processing. He is a member of ASJ.

Taichi Asami
Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E. and M.E. in computer science from Tokyo Institute of Technology in 2004 and 2006, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2006 and studied speech recognition, spoken language processing, and speech mining. He received the Awaya Prize Young Researcher Award in 2012 and the Sato Prize Paper Award in 2014 from ASJ. He is a member of ASJ, IEEE, IEICE, and the International Speech Communication Association (ISCA).

Kenta Niwa
Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E., M.E., and Ph.D. in information science from Nagoya University, Aichi, in 2006, 2008, and 2014, respectively. Since joining NTT in 2008, he has been engaged in research on microphone array signal processing. He was awarded the Awaya Prize by ASJ in 2010. He is a member of IEEE, ASJ, and IEICE.

Noriyoshi Kamado
Research Engineer. Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.
He received the B.E. from Nagaoka National College of Technology, Niigata, in 2007, the M.E. in electrical and electronic systems engineering from Nagaoka University of Technology in 2009, and the Ph.D. from Nara Institute of Science and Technology in 2012. He joined NTT Media Intelligence Laboratories in 2012, where he has been working on speech signal processing technologies for a speech recognition system. He is a member of ASJ, IEEE, and the Audio Engineering Society (AES).

Tomoko Kawase
Researcher, NTT Media Intelligence Laboratories.
She received the B.E. and M.E. in system design engineering from Keio University, Kanagawa, in 2011 and 2013, respectively. Since joining NTT in 2013, she has been working on signal processing technologies for a speech recognition system. She is a member of ASJ.

Takaaki Hori
Senior Research Scientist, Signal Processing Group, Media Information Laboratory, NTT Communication Science Laboratories.
He received the B.E. and M.E. in electrical and information engineering from Yamagata University in 1994 and 1996, respectively, and the Ph.D. in system and information engineering from Yamagata University in 1999. Since 1999, he has been engaged in research on spoken language processing at NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories). He was a visiting scientist at the Massachusetts Institute of Technology, USA, from 2006 to 2007. He is currently a senior research scientist at NTT Communication Science Laboratories. He received the 22nd Awaya Prize Young Researcher Award from ASJ in 2005, the 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2009, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan in 2012, and the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2013. He is a senior member of ASJ, IEEE, and IEICE.