Speech Recognition Technology That Can Adapt to Changes in Service and Environment

Hirokazu Masataki, Taichi Asami, Shoko Yamahata, and Masakiyo Fujimoto

Abstract

Speech recognition is a technology that recognizes spoken words. Speech recognition is a fundamental function for speech and language interfaces, and its quality strongly affects interface usability. Therefore, the recognition accuracy should always be high. However, recognition accuracy can drop significantly if there are changes in the service or the use environment. In this article, we introduce research being carried out to tackle this problem.

1. Introduction

1.1 User interface using speech recognition technology

Speech recognition allows machines to listen to our speech and convert it to text; the technology is thus roughly equivalent to the human ear. As speech recognition is a major component of speech and language interfaces, and its quality strongly influences the usability of the interfaces, the perennial demand is for high recognition accuracy. The NTT laboratories have been researching speech recognition technology for more than 40 years. Around the year 2000, speech recognition became practical in ideal environment owing to the many years of technological innovation and the advances in computer performance [1], [2].

The standard architecture of a speech recognition system is shown in **Fig. 1**. The technology is composed of feature extraction, an acoustic model, and a recognition dictionary. Existing speech recognition systems can recognize our voice only when uttered in a quiet environment using standard voice quality and common words. Consequently, recognition accuracy is significantly degraded in practical use.

1.2 Elements that can cause degradation

(1) Noisy environments

If you speak in a noisy environment such as a busy station or other crowded place, recognition accuracy will be poor because speech features are distorted by noise.

(2) Weak speakers

Even if the acoustic model is trained using the speech from a large number of speakers, recognition performance is poor if the speaker's speech features differ widely from those used to train the acoustic model.

(3) Out-of-vocabulary (OOV) failures

It is impossible to recognize words that are not in the dictionary with current speech recognition technology. This includes newly coined words and infrequently used words.

In this article, we introduce recent research advances that tackle these problems.

2. Voice activity detection and noise suppression

Ambient noise seriously degrades speech recognition accuracy. Thus, effective techniques for detecting voice activity and suppressing noise are critical for improving speech recognition accuracy in noisy environments. Voice activity detection accurately

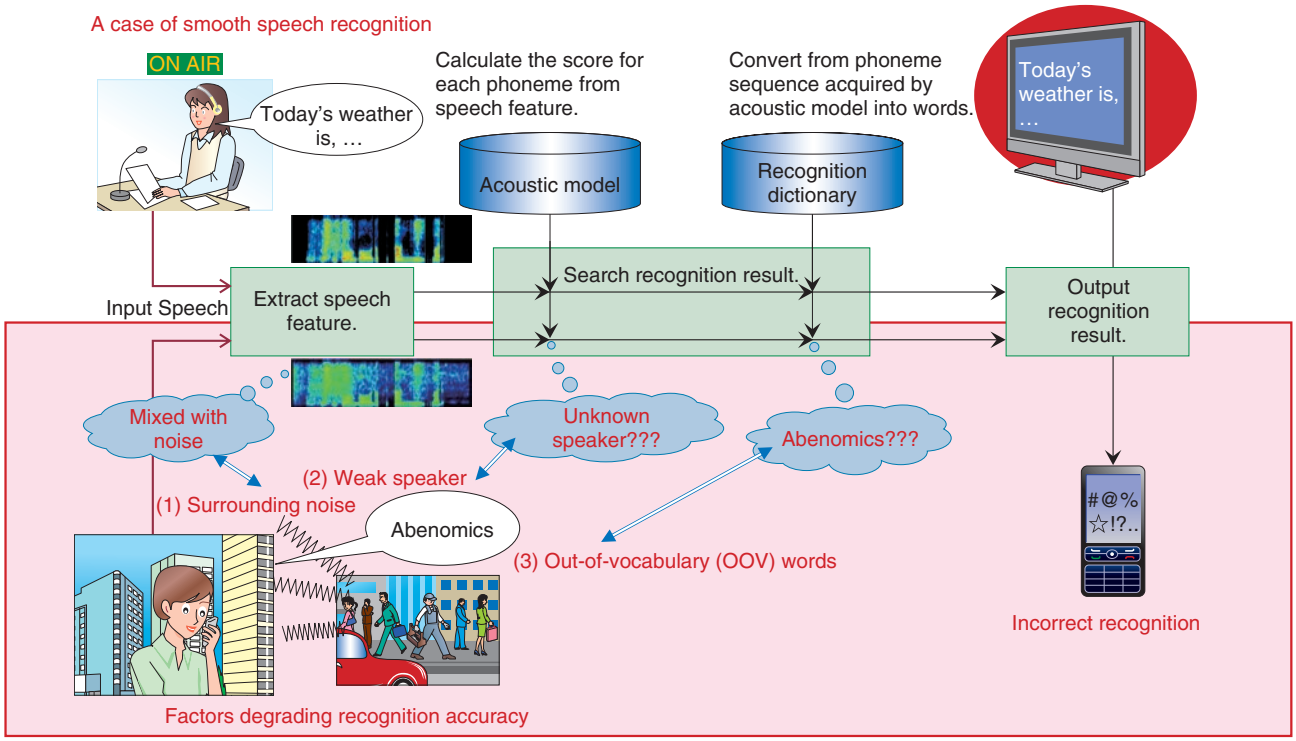

Fig. 1. Principle of speech recognition and problems.

detects periods of human utterances in noise-corrupted speech signals, and noise suppression clearly extracts signals of human utterances from the signals of the detected periods by suppressing the noise components. Typically, these techniques are individually developed and added as discrete front-end processes of a speech recognition system; the output of one is the input of the other in a simple chain. However, since these individual processes cannot share important information, it is difficult to achieve advanced front-end processing. A key problem is the accumulation of errors created by each technique.

To address this problem, we developed an integrated technique for voice activity detection and noise suppression called DIVIDE (Dynamic Integration of Voice Identification and DE-noising). DIVIDE reduces the number of processing errors in both processes and achieves front-end processing with advanced performance. DIVIDE employs statistical models of human speech signals and uses the models in both voice activity detection and noise suppression, as shown in **Fig. 2**. In DIVIDE, the activity probability of a human utterance is calculated from the statistical models in each short time slice.

When the activity probability exceeds a certain threshold, the time slice is tagged as a period of a human utterance. The tagging allows the noise components to be suppressed. Namely, DIVIDE extracts discriminative information that represents the similarity of human speech by utilizing statistical models of human speech signals as a priori knowledge in front-end processing. With this information, voice activity detection and noise suppression are performed simultaneously in DIVIDE.

Speech recognition results in noisy environments are shown in **Fig. 3**. The graph reveals that DIVIDE considerably improves speech recognition accuracy.

3. Automatic adaptation to weak speakers

Acoustic models that map our speech to phonemes* are used in automatic speech recognition. The mapping between speech and phonemes is acquired from manual transcriptions of speech by using machine learning methods ((1) in **Fig. 4**). We preliminarily train the acoustic model by using the speech samples of many people because each person's speech is

* A phoneme is the minimum unit of a speech sound and roughly equivalent to the sound corresponding to a single letter of a Japanese word written in Roman letters.

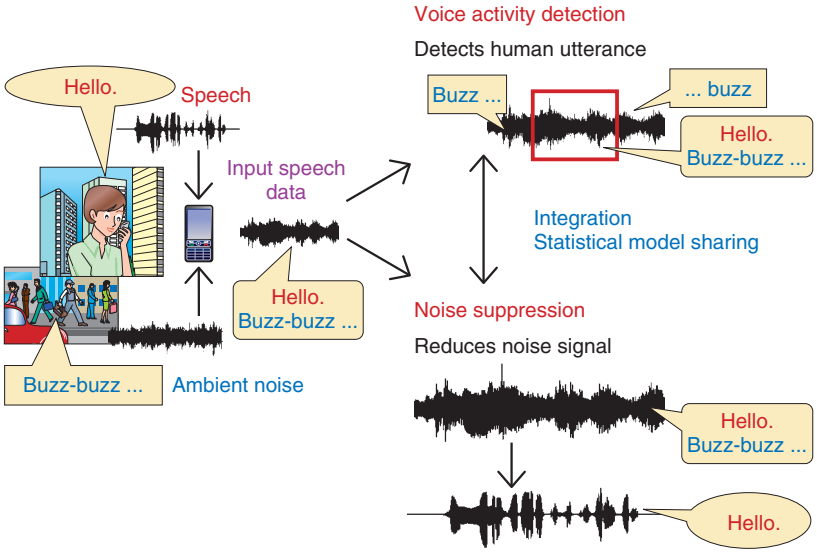

Fig. 2. Overview of DIVIDE.

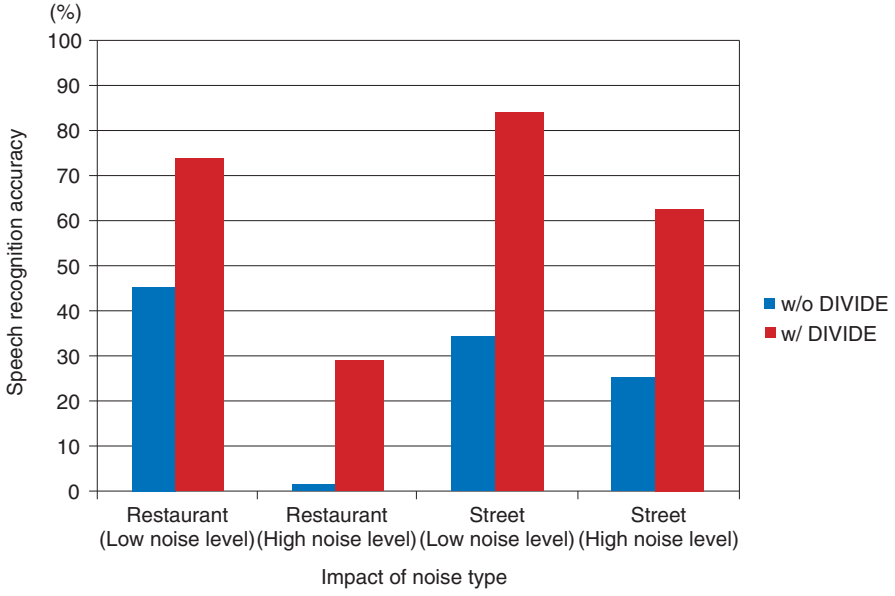

Fig. 3. Results of speech recognition using DIVIDE.

different. Even if the user's speech has not been learned, the acoustic model trained by similar speech can map the user's speech to the correct phoneme. Thus, the speech of many users can be accurately recognized.

However, even if learning involves many speakers, it is impossible to eliminate all blind spots. In actual use, there are always some users who are classified as *weak speakers* in that the acoustic model cannot accurately recognize their speech.

To solve this problem, we have developed a method of automatic adaptation to weak speakers. This method allows in-service speech recognition engines to identify weak speakers, automatically learn their speech, and maintain high recognition accuracy (Fig. 4).

Our speech recognition engine has a function that

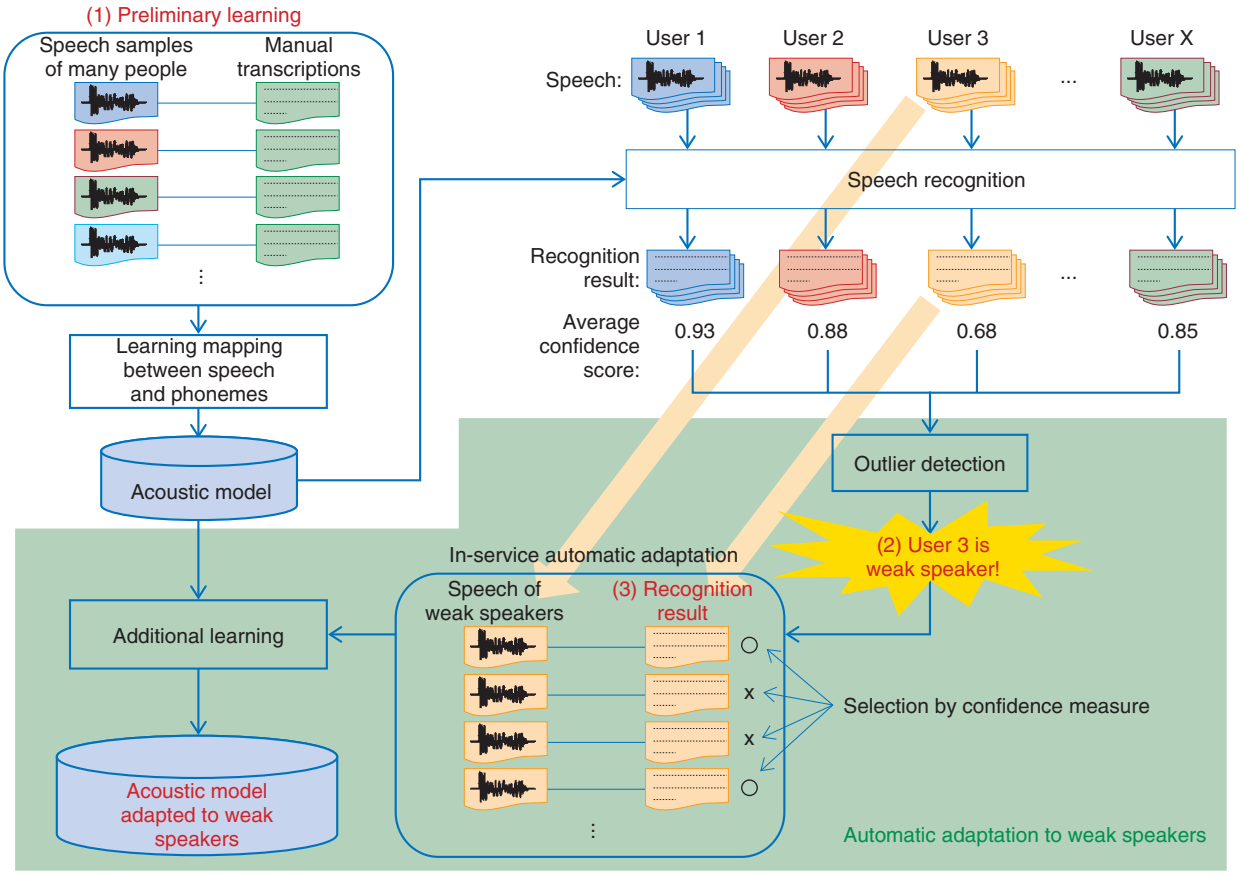

Fig. 4. Overview of automatic adaptation to weak speakers.

grades the correctness of its own recognition results [4]. This confidence score allows the engine to automatically find weak speakers. When the confidence score of a particular user is so low as to be an outlier, the user is taken to be a weak speaker ((2) in Fig. 4).

For the acoustic model to learn the mapping between the speech and phonemes of weak speakers, the direct solution is to obtain manual transcripts of their speech. However, creating manual transcripts is too expensive in terms of cost and time. Thus, our method uses the recognition results instead of manual transcripts to automatically train the acoustic model ((3) in Fig. 4). Unlike manual transcripts, the recognition results include erroneous parts that differ from the actual speech content; this makes learning ineffective. To address this problem, we turn to confidence scores again. Well-recognized results are selected by thresholding the confidence scores and then used for learning. We confirmed through experiments that our method improved the recognition accuracy of 80% of weak speakers to the same level as regular speakers.

4. Automatic vocabulary adaptation

Speech recognition systems include a *recognition dictionary*, which is a list of words to be recognized. All recognition dictionaries are limited in terms of coverage because of cost concerns, so OOV word failures are unavoidable. The current solution is to update the dictionaries manually by adding new words such as the names of new products or the titles of new books.

Several methods have been proposed to avoid having to manually update dictionaries. These methods collect web documents related to user's utterances, extract OOV words from the relevant documents, and add the OOV words to the recognition dictionary (**Fig. 5(a)**). However, these methods register all OOV words in the relevant documents, even words that are not spoken in the target utterances (i.e., redundant word entries). Consequently, useful words may be

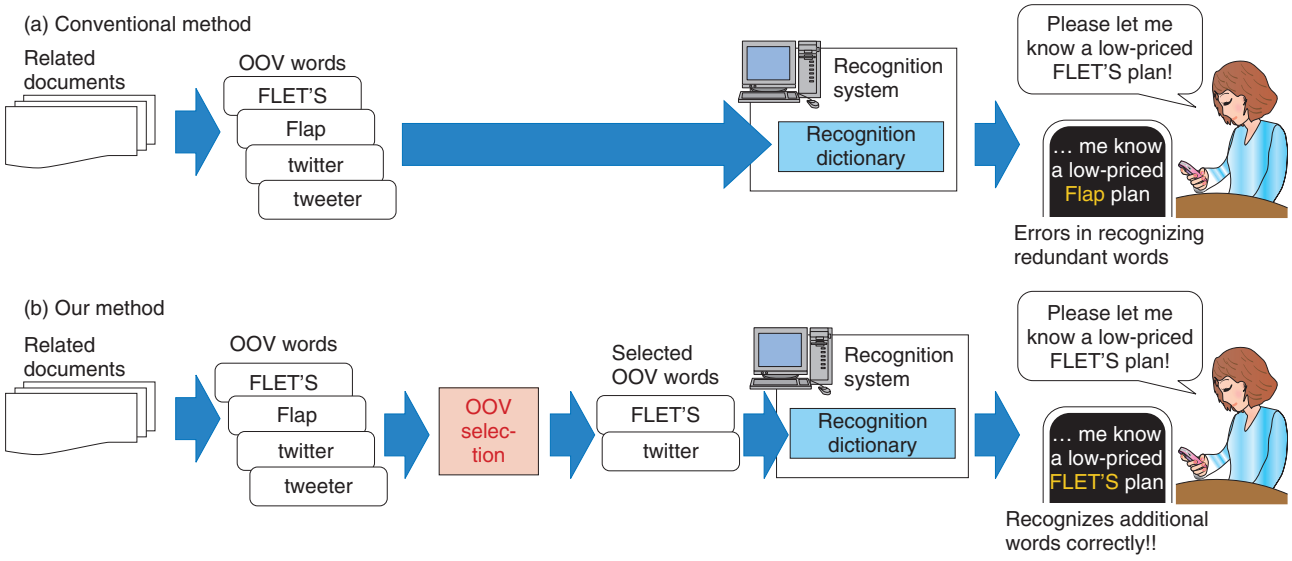

Fig. 5. Comparison of conventional method and our method.

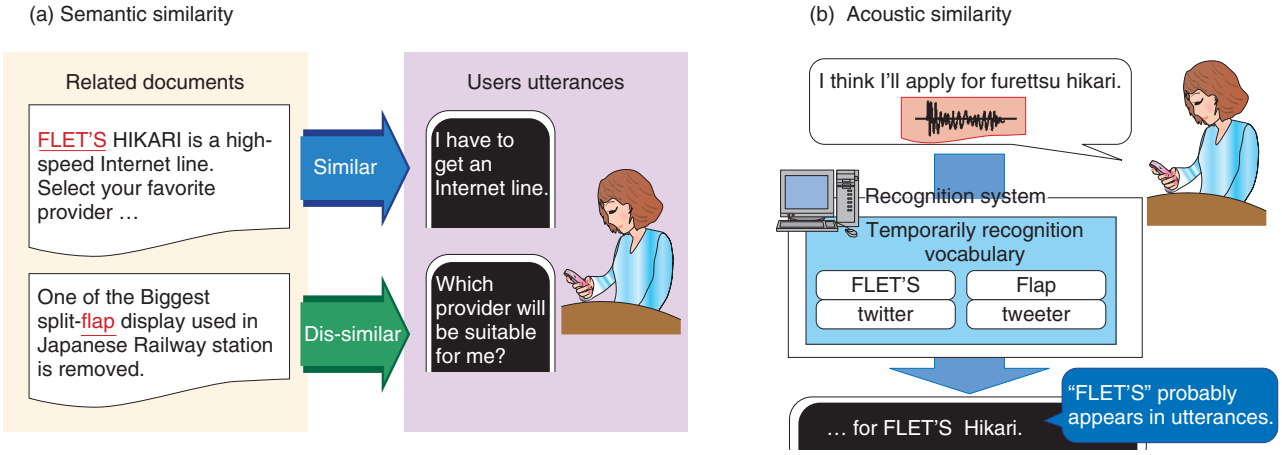

Fig. 6. Mechanism of selecting OOV words using our method.

dropped from the dictionary.

To improve recognition accuracy, we have developed a method that selects only OOV words that will actually be spoken in the user's utterances (**Fig. 5(b)**). Our method yields recognition dictionaries that are suitable for each user or service and that can recognize utterances accurately because redundant words are eliminated.

We use two types of information to select OOV words. First, we use the semantic similarity between each OOV word and the user's utterances (**Fig. 6(a)**). For example, if a user frequently uttered terms relevant to Internet service such as "Internet", "line", and "provider", we select OOV words that co-occur with these terms such as "FLET'S". Second, we use acoustic similarity, which refers to whether or not the pronunciation of OOV words is included in the utterances (**Fig. 6(b)**). For example, if the user uttered "I think I will apply for furettsuhikari", we would assume the word "FLET'S" was probably uttered. To detect the parts of utterances where OOV words appear, we temporarily register all OOV words in the recognition dictionary, and we recognize the utterances using a temporary dictionary to obtain

temporary recognition results. Finally, we select OOV words that appear in the temporary recognition results.

With our method, we were able to reduce the number of recognition errors caused by redundant words by about 10% compared to conventional methods that add all OOV words.

5. Future efforts

We developed the speech recognition engine called VoiceRex to demonstrate the technologies developed at NTT. We plan to implement the technologies introduced in this article in VoiceRex.

We are working on improving the recognition accuracy by enhancing these new technologies to be adapted to each user. We also intend to implement our recognition technologies as cloud services.

References

[1] Y. Noda, Y. Yamaguchi, K. Ohtsuki, and A. Imamura, "Development of the VoiceRex Speech Recognition Engine," NTT Technical Journal, Vol. 11, No. 12, pp. 14–17, 1999 (in Japanese).

[2] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex—Spontaneous Speech Recognition Technology for Contact-center Conversations," NTT Technical Review, Vol. 5, No. 1, pp. 22–27, 2007.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr200701022.pdf

[3] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," INTERSPEECH 2009, pp. 1235–1238, Brighton, UK, Sept. 2009.

[4] T. Asami, N. Nomoto, S. Kobashikawa, Y. Yamaguchi, H. Masataki, and S. Takahashi, "Spoken document confidence estimation using contextual coherence," INTERSPEECH 2011, pp. 1961–1964, Florence, Italy, Aug. 2011.

[5] S. Yamahata, Y. Yamaguchi, A. Ogawa, H. Masataki, O. Yoshioka, and S. Takahashi, "Automatic Vocabulary Adaptation Based on Semantic Similarity and Speech Recognition Confidence Measure," INTERSPEECH 2012, Portland, OR, USA, Sept. 2012.

Hirokazu Masataki

Senior Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E., M.E., and D.Eng. degrees from Kyoto University in 1989, 1991, and 1999, respectively. From 1995 to 1998, he worked with ATR Interpreting Telecommunications Research Laboratories, where he specialized in statistical language modeling for large vocabulary continuous speech recognition. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2004 and engaged in the practical use of speech recognition. He received the Maejima Hisoka Award from the Tsushinbunka Association. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Acoustical Society of Japan (ASJ).

Taichi Asami

Researcher, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. and M.E. degrees 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 from ASJ in 2012. He is a member of the International Speech Communication Association (ISCA), IEICE, and ASJ.

Shoko Yamahata

Research Engineer, Audio Speech and Language Media Project, NTT Media Intelligence Laboratories.

She received the B.E. and M.E. degrees from Waseda University, Tokyo, in 2008 and 2010, respectively. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2010 and studied language models and vocabulary adaptation. She is a member of ISCA and ASJ.

Masakiyo Fujimoto

Researcher, NTT Communication Science Laboratories.

He received the B.E., M.E., and D.Eng. degrees from Ryukoku University, Shiga, in 1997, 2001, and 2005, respectively. From 2004 to 2006, he worked with ATR Spoken Language Communication Research Laboratories. He joined NTT in 2006. His current research interests are noise-robust speech recognition including voice activity detection and speech enhancement. He received the Awaya Prize Young Researcher, Award from ASJ in 2003, the MVE Award from IEICE SIG MVE in 2008, the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2011, and the ISS Distinguished Reviewer Award from IEICE in 2011. He is a member of IEEE, ISCA, IEICE, IPSJ, and ASJ.