Deep Learning Based Distant-talking Speech Processing in Real-world Sound Environments

Shoko Araki, Masakiyo Fujimoto, Takuya Yoshioka, Marc Delcroix, Miquel Espi, and Tomohiro Nakatani

Abstract

This article introduces advances in speech recognition and speech enhancement techniques with deep learning. Voice interfaces have recently become widespread. However, their performance degrades when they are used in real-world sound environments, for example, in noisy environments or when the speaker is some distance from the microphone. To achieve robust speech recognition in such situations, we must make progress in further developing various speech processing techniques. Deep learning based speech processing techniques are promising for expanding the usability of a voice interface in real and noisy daily environments.

Keywords: deep learning, automatic speech recognition, speech enhancement

1. Introduction

In recent years, the use of voice-operable smartphones and tablets has become widespread, and their usefulness has been widely recognized. When a user speaks carefully into a terminal, that is, a microphone(s) (**Fig. 1(a)**), his/her voice is usually accurately recognized, and the device works as expected.

On the other hand, there is a growing need for voice interfaces that can work when a user speaks at a certain distance from the microphones. For example, when we record the discussion in a meeting, as shown in **Fig. 1(b)**, we may want to employ a terminal on the table and avoid the use of headset microphones. Furthermore, when users talk to voice-operated robots or digital signage, the users would talk to them from a certain distance.

However, the current speech recognition accuracy of voice-operable devices is generally insufficient when the speaker is far away from the microphone. This is because of the considerable effect of noise and reverberation and because the users speak freely with little awareness of the microphones when the microphones are some distance away. We are therefore studying distant speech recognition and working on speech enhancement and speech recognition techniques to expand the usability of a voice interface in real-world sound environments.

There are two main factors that degrade the automatic speech recognition of distant speech; (1) the quality of speech recorded with a distant microphone is severely degraded by background noise, for example, air conditioners and room reverberation. Moreover, in a multi-person conversation, the speakers' voices sometimes overlap. (2) As the users speak freely without regard to the microphones, their utterances become fully spontaneous and therefore tend to include ambiguous pronunciations and abbreviations. Speech enhancement techniques are essential in order to cope with such complex difficulties, and these include noise reduction, reverberation reduction (dereverberation), speech separation, and spontaneous speech recognition techniques.

Close microphone(s)

(a) Current status

Distant microphone(s)

(b) Our goal

Fig. 1. Current and future status of voice interfaces.

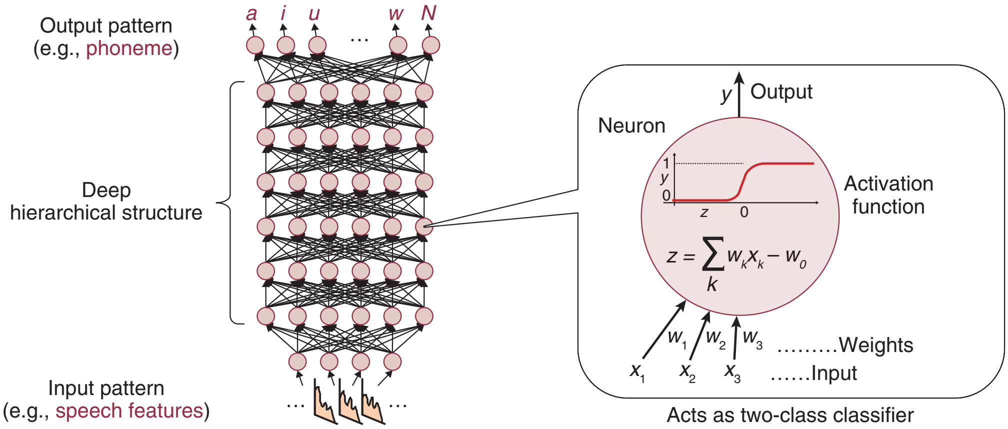


Fig. 2. Acoustic model with deep neural network (DNN).

2. Deep learning in speech processing

We have been studying the aforementioned speech processing techniques in order to achieve distant speech recognition in the real world. In recent years, we have been working on speech processing methods based especially on deep learning. Deep learning is a machine learning method that uses a deep neural network (DNN), as shown in **Fig. 2**. Deep learning has recently come under the spotlight because in 2011 and 2012 it was shown to outperform conventional techniques in many research fields including image recognition and compound activity prediction. High performance has also been achieved with deep learning in speech recognition tasks, and therefore, deep learning based speech processing techniques have been intensively researched in recent years.

In 2011, we began working on deep learning based techniques for automatic recognition of spontaneous speech [1]. It should be noted that a deep learning based real-time speech recognizer developed by NTT Media Intelligence Laboratories has already been released [2]. We have also proven that deep learning improves speech enhancement techniques such as noise reduction when deep learning is effectively leveraged. The remainder of this article describes our speech recognition and speech enhancement techniques that employ deep learning.

3. Speech recognition with deep learning

General automatic speech recognition techniques translate features into phonemes, phonemes into words, and words into sentences, by respectively using an acoustic model, a pronunciation dictionary, and a language model (**Fig. 3**). Originally, deep

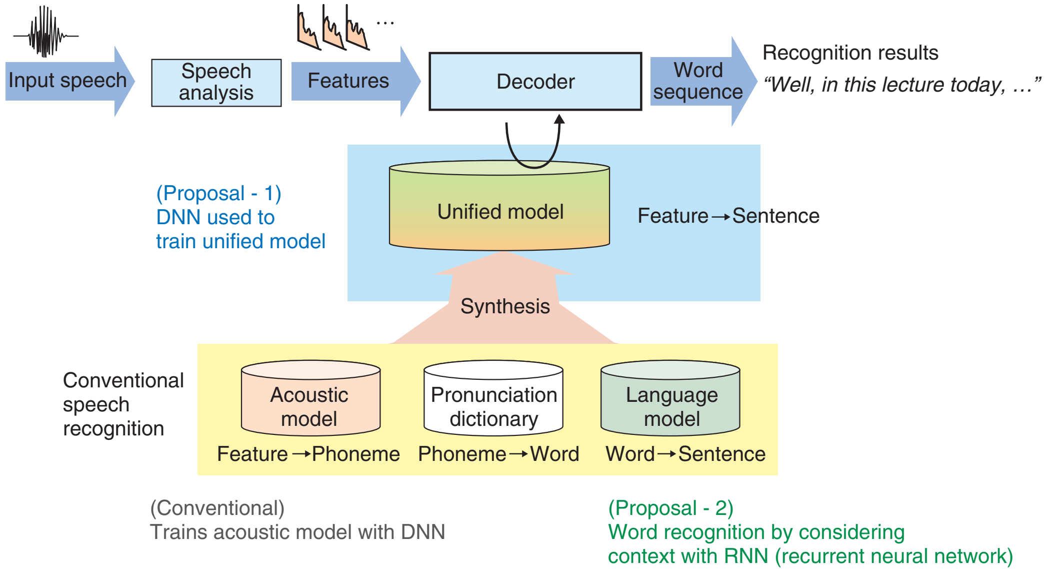


Fig. 3. Speech recognition process.

learning based speech recognition employed a DNN to achieve accurate acoustic modeling, and it outperformed conventional speech recognition techniques that do not use deep learning.

The aforementioned acoustic model, pronunciation dictionary, and language model are usually trained separately, so it has been difficult to consider the interaction between the phonetic and linguistic factors that are present in spontaneous speech. To address these complex factors, we proposed synthesizing the three models into a unified model (Fig. 3) and optimizing it by using a DNN [1]. We demonstrated that this unified model achieves highly accurate spontaneous speech recognition [1].

Moreover, we showed that a recurrent neural network (RNN), which is also a deep learning technique, in a language model provides further improvement in performance. An RNN based language model is effective for spontaneous speech recognition because its ability to hold the history of words enables us to recognize speech by considering a longer context. However, it is generally difficult to achieve fast automatic speech recognition while maintaining the complete contextual history. We therefore proposed an efficient computational algorithm for maintaining contexts and achieved fast and highly accurate automatic speech recognition [3].

The word error rates (WERs) in English lecture speech recognition are shown in **Fig. 4**. *Without DNN* indicates the WER before deep learning was employed. The *DNN acoustic model* shows the large effect of deep learning. We can also see that the *unified DNN*, where the unified model is optimized with a DNN, outperforms the conventional *DNN acoustic model*. Moreover, the *RNN language model* achieves the best performance, which is more than 4 points better than the conventional *DNN acoustic model*. The appropriate use of deep learning techniques significantly improves spontaneous speech recognition performance.

4. Speech enhancement with deep learning

Deep learning also helps to improve speech enhancement performance. This section introduces two noise reduction techniques: a method for use with multiple microphones and a method for use with a single microphone.

The first approach estimates the features of noise-reduced speech by using a DNN (**Fig. 5(a)**). Pairs consisting of clean and noisy speech signals are used to train the DNN to translate noisy speech features into clean speech features. The trained DNN is then used to estimate noise-reduced features when the input consists of noisy features. This method was originally used for noise reduction with a single microphone, however, its extension to multi-microphone use was not obvious. We found that we can improve noise reduction performance by inputting additional features estimated with multi-microphone

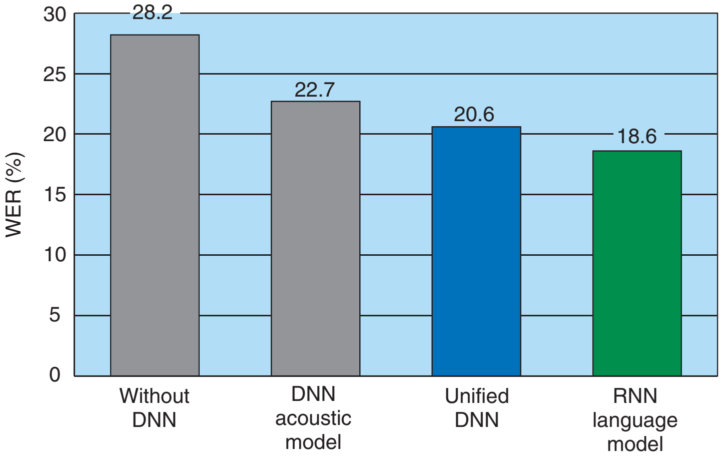
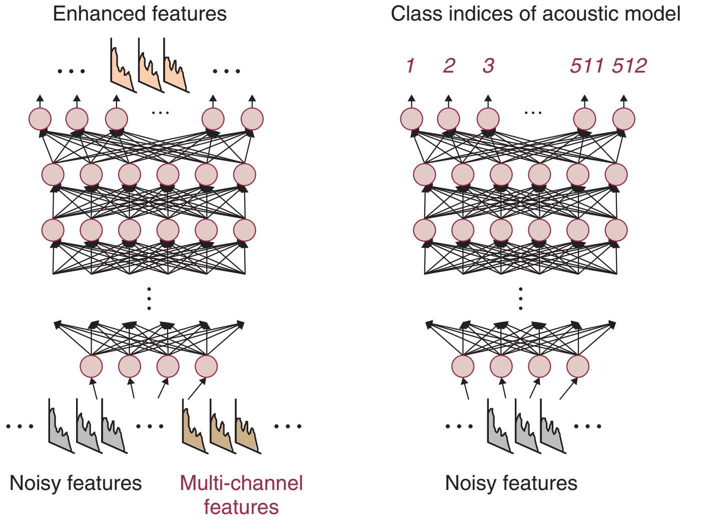


Fig. 4. Word error rates in English lecture speech recognition.



(a) Multi-channel noise reduction

(b) Single channel noise reduction

Fig. 5. Examples of DNN-based noise reduction.

observations into a DNN. We also found that the probability of speech existing at each time-frequency slot, which can be estimated with a microphone array technique, provides us with an effective additional feature [4]. The results of an evaluation conducted under living room noise conditions (PASCAL CHiME challenge task) revealed the superiority of our proposed approach. Specifically, we obtained a reduced WER of 8.8% by using the proposed multi-microphone features compared to a value of 10.7% without them.

The second noise reduction approach is for cases where we can use just a single microphone. This method is applied to calculate noise reduction filter coefficients by using probabilistic models of clean speech and noise without speech. Here, accurate model estimation is important for accurate filter design. We showed that DNN-based clean speech model estimation (**Fig. 5(b)**) achieves high-performance noise reduction [5]. Specifically, we constructed a clean speech model with a set of probabilistic models and utilized a DNN to discriminate the most suitable model for generating the observed noisy speech. With this proposed noise reduction approach, we obtained an improved WER of 19.6% for a noisy speech database (AURORA4), whereas the WER was 23.0% with a conventional method without a DNN.

It is worth mentioning that we do not use a DNN for noise model estimation. This is because it is difficult to obtain a sufficient quantity of noise data for DNN training due to the wide range of variations and momentary fluctuations of noise in the real world. With the proposed method, we estimate the noise model using an unsupervised method, and we simultaneously use a DNN for clean model selection. This approach achieves high-performance noise reduction in real-world sound environments by flexibly considering the variation of noisy signals.

5. Outlook

We believe that distant-talking speech processing is a key technology for expanding the usability of voice interfaces in actual daily life. In particular, conversational speech recognition and communication scene analysis in real-world sound environments are techniques that meet the needs of the times. These techniques should make a significant contribution to artificial intelligence (AI) speech input, which has recently attracted renewed interest for applications such as minute-taking systems in business meetings, intelligent home electronics, and a human-robot dialogue system for use in shopping centers. For these purposes, we need a highly accurate distant speech recognition technique that works in noisy environments. In addition, techniques for identifying the current speakers and for understanding what is going on around the AI device by recognizing, for example, environmental sound events [6], are also becoming more important. We are continuing to work on the development of essential techniques for distant-talking speech processing in order to expand the capabilities of voice interfaces to their fullest extent.

References

[1] Y. Kubo, A. Ogawa, T. Hori, and A. Nakamura, “Speech Recognition Based on Unified Model of Acoustic and Language Aspects of Speech,” NTT Technical Review, Vol. 11, No. 12, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201312fa4.html

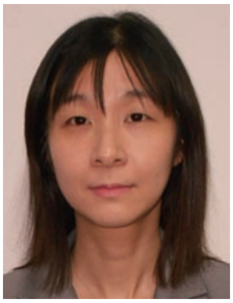
[2] NTT IT CORPORATION News Release, Nov. 11, 2014 (in Japanese).
http://www.ntt-it.co.jp/press/2014/1111/

[3] T. Hori, Y. Kubo, and A. Nakamura, “Real-time One-pass Decoding with Recurrent Neural Network Language Model for Speech Recognition,” Proc. of ICASSP 2014 (2014 IEEE International Conference on Acoustics, Speech, and Signal Processing), pp. 6364–6368, Florence, Italy, May 2014.

[4] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, “Exploring Multi-channel Features for Denoising-autoencoder-based Speech Enhancement,” Proc. of ICASSP 2015, pp. 116–120, Brisbane, Australia, Apr. 2015.

[5] M. Fujimoto and T. Nakatani, “Feature Enhancement Based on Generative-discriminative Hybrid Approach with GMMs and DNNs for Noise Robust Speech Recognition,” Proc. of ICASSP 2015, pp. 5019–5023, Brisbane, Australia, Apr. 2015.

[6] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, “Feature Extraction Strategies in Deep Learning Based Acoustic Event Detection,” Proc. of Interspeech 2015, pp. 2922–2926, Dresden, Germany, Sept. 2015.



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. Since joining NTT in 2000, she has been conducting research on acoustic signal processing, array signal processing, blind source separation, meeting diarization, and auditory scene analysis.

She was a member of the organizing committee of ICA 2003, IWAENC 2003, WASPAA 2007, and the evaluation co-chair of SiSEC 2008, 2010, and 2011. Since 2014, she has been a member of the Institute of Electrical and Electronics Engineers (IEEE) Signal Processing Society Audio and Acoustics Technical Committee. She received the 19th Awaya Prize from the Acoustical Society of Japan (ASJ) in 2001, the IWAENC Best Paper Award in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004 and 2014, the Academic Encouraging Prize from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2006, the Itakura Prize Innovative Young Researcher Award from ASJ in 2008, and the Young Scientists' Prize of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2014. She is a member of IEEE, IEICE, and ASJ.

Masakiyo Fujimoto

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

He received his B.E., M.E., and Dr. Eng. from Ryukoku University, Kyoto, in 1997, 2001, and 2005. From 2004 to 2006, he worked with ATR Spoken Language Communication Research Laboratories. He joined NTT Communication Science Laboratories 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 Special Interest Group Multimedia and Virtual Environments (MVE) in 2008, the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2011, and the ISS Distinguished Reviewer Award from IEICE Information and Systems Society (ISS) in 2011. He is a member of IEEE, IEICE, IPSJ, and ASJ.

Takuya Yoshioka

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

He received his B.Eng., M.Inf., and Ph.D. in informatics from Kyoto University in 2004, 2006, and 2010. In 2005, he interned at NTT, where he conducted research on dereverberation. Since joining NTT in 2006, he has been working on the development of algorithms for noise robust speech recognition, speech enhancement, and microphone arrays. During 2013–2014, he was a Visiting Scholar at the University of Cambridge, Cambridge, UK. He has been a part-time lecturer at Doshisha University, Kyoto, since 2015. He received the Awaya Prize Young Researcher Award and the Itakura Prize Innovative Young Researcher Award from ASJ in 2010 and 2011, respectively, and the Young Researcher's Award in Speech Field from IEICE ISS in 2011.

Marc Delcroix

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

He received an M.E. from the Free University of Brussels, Belgium, and the Ecole Centrale Paris, France, in 2003 and a Ph.D. from the Graduate School of Information Science and Technology, Hokkaido University, Japan, in 2007. He joined NTT Communication Science Laboratories in 2010. He is also a Visiting Lecturer in the Faculty of Science and Engineering of Waseda University, Tokyo. From 2004 to 2008, he was a research associate at NTT Communication Science Laboratories. From 2008 to 2010, he worked at Pixela Corporation developing software for digital television. His research interests include robust speech recognition, speech enhancement, and speech dereverberation. He was one of the organizers of the REVERB challenge 2014. He received the 2005 Young Researcher Award from the Kansai section of ASJ, the 2006 Student Paper Award from the IEEE Kansai section, and the 2006 Sato Paper Award from ASJ. He is a member of IEEE and ASJ.

Miquel Espi

Research Associate, Signal Processing Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E. in computer science from Universidad Politecnica de Valencia, Spain, in 2006, an M.E. in information science from Kagoshima University in 2010, and a Ph.D. in information science and technology from the University of Tokyo in 2013. He joined NTT Communication Science Laboratories in 2013 and has been researching characterization and classification of acoustic events in the context of conversation scene analysis. His current research interests include acoustic scene analysis, acoustic signal processing, and social dynamics. He is a member of IEEE, IEEE Signal Processing Society, and ASJ.

Tomohiro Nakatani

Senior Research Scientist, 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 Poster Award, the 2005 IEICE Best Paper Award, and the 2009 ASJ Technical Development Award. During 2009–2014, he was a member of the IEEE Signal Processing Society Audio and Acoustics Technical Committee (AASP-TC), and he has been an associate member of the AASP-TC since 2015. He has also served as Chair of the review subcommittee of AASP-TC, Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing, Chair of the IEEE Kansai Section Technical Program Committee, Technical Program co-Chair of IEEE WASPAA-2007, and as a member of the IEEE Circuits and Systems Society Blind Signal Processing Technical Committee. He is a member of IEEE, IEICE, and ASJ.