Efforts to Enhance Far-field Speech Recognition

Kiyoaki Matsui, Takafumi Moriya, Hiroaki Itou, Takaaki Fukutomi, Shoichiro Saito, Yusuke Shinohara, Satoshi Kobashikawa, Yoshikazu Yamaguchi, and Noboru Harada

Abstract

We collected more than 15,000 hours of pseudo data to enhance training data for robust speech recognition by adding various impulse responses and noises to clean data and re-recording the clean data under actual environments. As a result, we reduced the rate of errors by 43% or more in both close and distant conditions. In this article, we introduce the various approaches that we researched, with the focus being on countering the distant conditions.

Keywords: far-field speech recognition, data augmentation, acoustic modeling

1. Far-field speech recognition

The use of speech recognition systems is increasing rapidly, and in line with this trend, communication agents such as smart speakers and voice dialog robots are being rapidly adopted, as are voice search services on smartphones. With search tasks using smartphones, utterances are often made in the immediate vicinity of the microphone, but when talking to robots or smart speakers, we must assume utterances will be made one to three meters away from the microphone. Furthermore, the speakers will exhibit individual differences in voice volume, and when a speaker speaks softly from some distance in a noisy environment, it is very difficult for current systems to correctly recognize the speech.

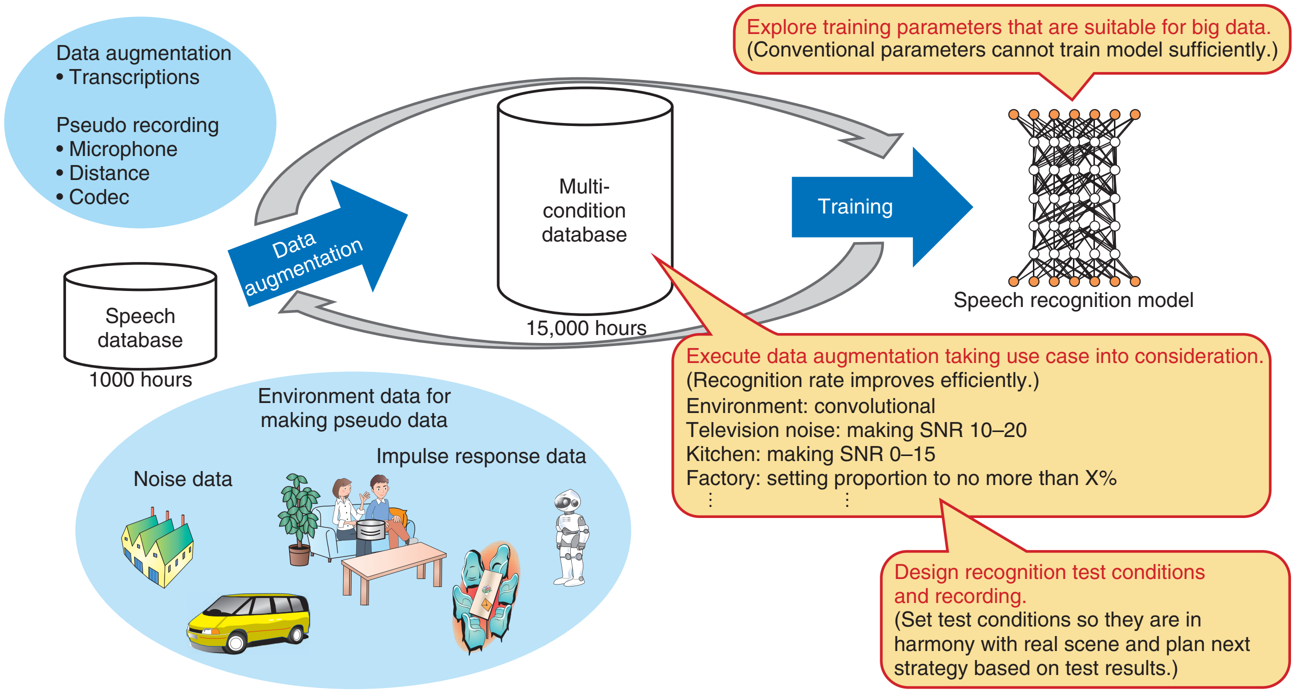
We researched various approaches to counter the noise environment of the home and to achieve robust speech recognition under both close and distant conditions regardless of noise or voice volume. Our aim was to improve system performance by enhancing training data and evaluation data by generating pseudo data and improving playback recording (or, pseudo recording), adjusting training parameters, and improving speech section detection (**Fig. 1**). Tests conducted using these approaches indicated that an error reduction rate of 43% or more was achieved in both close and distant conditions. These approaches are explained in the following sections.

2. Data augmentation by generation of pseudo data and speech samples

Having sufficient training data is necessary to improve system performance, but it is not always easy or practical to obtain the right kind of data. Data augmentation is a common approach to obtain a sufficient amount of data.

2.1 Augmentation of training data

The characteristics of the sound captured by the microphone will vary greatly with the speaker's proximity to the microphone. With close utterances, since the direct sound of the speaker dominates any echoes or noise, recognition can be performed with high accuracy regardless of the ambient noise environment. In contrast, with distant utterances, not only is the target utterance of the speaker affected by factors



Explore training parameters that are suitable for big data.
(Conventional parameters cannot train model sufficiently.)

Execute data augmentation taking use case into consideration.
(Recognition rate improves efficiently.)
Environment: convolutional
Television noise: making SNR 10–20
Kitchen: making SNR 0–15
Factory: setting proportion to no more than X%
⋮ ⋮

Design recognition test conditions and recording.
(Set test conditions so they are in harmony with real scene and plan next strategy based on test results.)

SNR: signal-to-noise ratio

Fig. 1. Overview of our speech recognition enhancement approach.

such as the reverberation characteristics of the room and the attenuation due to distance, but it is also more likely that one or more noise sources that are closer to the microphone than the speaker will drown out the utterances, making recognition very difficult.

The mainstream approach at present is to build an acoustic model using deep neural network (DNN) technology, and it is possible to construct a robust acoustic model if adequate amounts of training data captured in various environments are available. Therefore, to construct acoustic models that are robust against various noises and speakers, it is important to cover as many diverse environments as possible. Normally, data are reinforced by transcribing speech recorded in various environments, but speech transcription takes too long, so gathering sufficient training data is problematic. One promising solution is to add various reverberation effects and noise to existing clean speech data, which yields a sufficient variety of pseudo data. This method of generating pseudo data is important not only for speech recognition but also for DNN training in all fields.

However, increasing the pseudo data will not necessarily yield higher accuracy. For example, with a communication agent that is to be used in the home, we can predict that situations where speaking occurs at distances exceeding 3 m are unlikely to occur. Also, because homes include furniture and sound-absorbing materials such as carpets and curtains, reverberation in the room will not be very strong. The noise environment can be expected to include the noise of television (TV), cooking sounds from the kitchen, and air conditioning. However, modern air conditioners are extremely quiet, and are therefore unlikely to drown out the speech. Thus, we can optimize speech recognition in the target environment by properly reacting to the reverberation strength, type of noise, and volume of noise so as not to generate unrealistic pseudo data.

In addition to generating pseudo data, we also make pseudo recordings by creating corrupted speech as it would be captured by the microphone. This method is very effective when the impulse response cannot be captured due to specifications of the microphone, or when it is necessary to create a more complex reverberation/noise environment. The playback recording configuration is shown in **Fig. 2**. The height, position, and angle of the speaker used for pseudo recording is set considering the practical environment. For example, if the user is assumed to speak while sitting on a

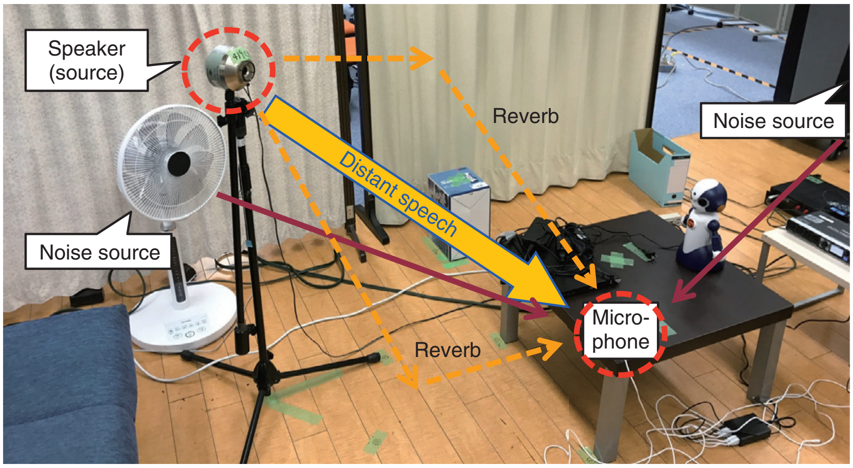


Fig. 2. Pseudo recording setup.

sofa, the speaker's height should match the height of the user's mouth. In addition, the speaker for noise playback is set to replicate the position of the actual noise source. The microphone used to capture the pseudo recording should be set at multiple positions at the same time, as recording in parallel enables more efficient data enhancement. To capture different characteristics, the microphone position can be set at not only different distances to the speaker, but also at different angles with respect to the speaker, position in the room, and other details. Pseudo recording yields training data of higher quality than simulated data because the results are captured after the sounds actually traverse space.

In this way, in addition to the data provided by normal transcription, the training data are reinforced by generating the pseudo data and pseudo recording, and it is possible to increase the training sound data, which originally amounted to only about 1000 hours, to more than 15,000 hours. In addition, at the first stage of speech recognition, utterance sections are accurately extracted by speech segment detection, which greatly improves recognition efficiency. The models used at this time were reinforced by similar data enhancement.

2.2 Creation of evaluation data

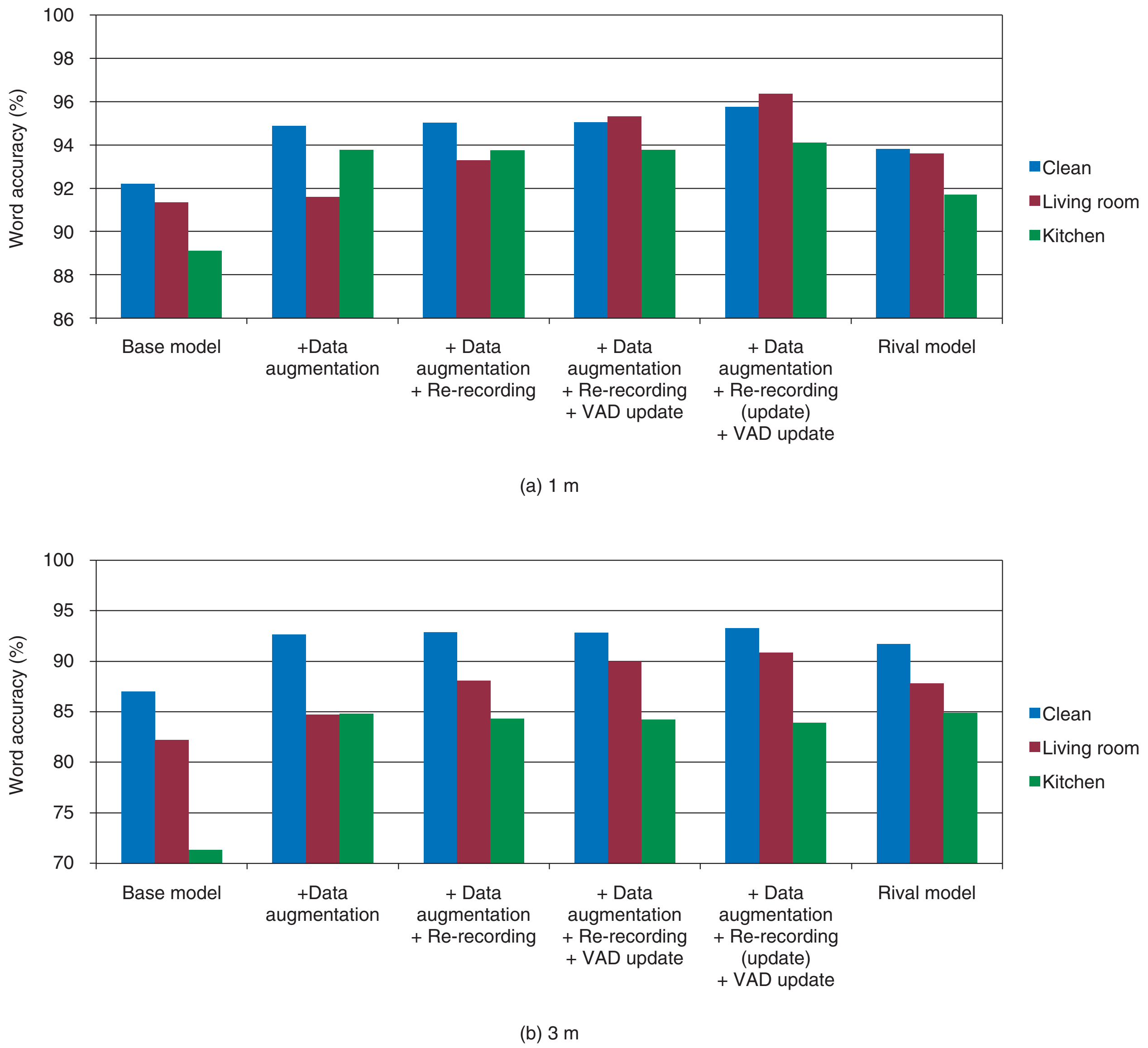
After the model is created from training data, evaluation data are needed to evaluate the model's performance. The ideal evaluation data consist of real data captured in the user's environment, but obtaining such ideal data is impractical, and in most cases evaluation data obtained in a nearly equal environment will be used. Accordingly, pseudo recording is an effective solution to the problem of creating evaluation data.

For this research, we prepared and evaluated a total of 36 patterns created by pseudo recordings using the speech of 20 subjects and setting different conditions of reverberation, distance, voice volume, and noise. The resulting data make it possible to assess system performance in environments for which evaluation data are unavailable. The resulting data revealed the strong and weak points of the trained model, which will make training the next model much more efficient.

3. Determining training parameters

The accuracy of the DNN acoustic model is greatly affected by the training parameters used, and these parameters are very difficult to set to maximize system performance. If the training set is very large scale, say more than 15,000 hours, parameters that are conventionally appropriate will not always work, and it is necessary to review what constitutes the optimal parameter setting.

There are two goals with parameter adjustment: speed enhancement and higher accuracy. The former can be achieved by increasing the number of parallel GPUs (graphics processing units) used to implement training and increasing the number of samples used to calculate each process in training (batch size). However, if the batch size is increased too much, any sample that is used will be close to some entries when the average is determined for each process. This



(a) 1 m

(b) 3 m

VAD: voice activity detection

Fig. 3. Effectiveness of data augmentation.

means that convergence may occur even if training is insufficient.

As for the latter, to incorporate data as diverse as possible into training, it is necessary to increase the amount of data processed in each epoch (one training iteration) or make it easier to select specific condition data from the corpus used for training, which is another of our advances.

By adjusting the training parameters, we were able to develop a more accurate acoustic model while keeping the training time the same as it was when only 1000 hours of speech were used for training.

4. Recognition accuracy under different noise and distance values

In this section, we describe how well our data enhancement and model training techniques increase system performance. The performance results of the base model and the model using the proposed techniques are shown in the graph in **Fig. 3**. The training parameters were optimized for each condition. As can be seen in this graph, pseudo data generation greatly improves the recognition rate in clean and kitchen environments with distances to the microphone of 1 m and 3 m, which confirms the value of data reinforcement. In contrast, in the living room environment, no significant improvement in accuracy was seen. This is most likely because the TV in the living room environment puts out voices, and it is very difficult to judge whether such extraneous speech is noise or not.

Our technique of pseudo recording is very effective in coping with this living room noise, as it enables training to include more practical environmental information and strengthens the detection of speech segments. The result is a strong improvement in recognition rate in the living room environment while maintaining the excellent recognition performance in the clean and kitchen environments. The techniques enable us to build an acoustic model that is significantly better than all conventional alternatives.

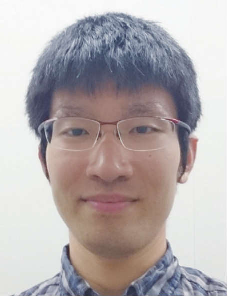
5. Future work

Our research has yielded acoustic models that offer high speech recognition accuracy regardless of the presence or absence of noise in the close and distant speaker conditions. However, in environments where the speaker's voice is mixed with the speech of others, erroneous speech recognition results are generated, so that the system reacts incorrectly or recognizes the user's utterance wrongly. We believe that this problem can be tackled by using direction-based sound collection technology such as the intelligent microphone [1], which will enable robust recognition of the target speaker's voice even in very noisy environments.

This study focused on quite short utterances such as those used for conducting search tasks. Many tasks remain if we are to reliably recognize natural speech as encountered at conferences and call centers. We will continue working to enhance the performance of our techniques to cover these natural utterances.

Reference

[1] T. Oba, K. Kobayashi, H. Uematsu, T. Asami, K. Niwa, N. Kamado, T. Kawase, and T. Hori, "Media Processing Technology for Business Task Support," NTT Technical Review, Vol. 13, No. 4, 2015. https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201504fa6.html



Kiyoaki Matsui

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

He received a B.E. and M.E. from Tohoku University, Miyagi, in 2013 and 2015. Since joining NTT in 2015, he has been researching speech recognition, including voice activity detection and speech enhancement. He is a member of the Acoustical Society of Japan (ASJ).

Takafumi Moriya

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

He received a B.E. from Chiba Institute of Technology in 2014, and an M.E. from Tokyo Institute of Technology in 2016. Since joining NTT in 2016, he has been researching speech recognition, including acoustic modeling. He is a member of ASJ.

Hiroaki Itou

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

He received a B.E. and M.E. from Nagoya University, Aichi, in 2007 and 2009. Since joining NTT in 2009, he has been researching acoustic signal processing for sound reproduction and speech recognition. He was awarded the Awaya Prize by ASJ in 2011. He is a member of ASJ and the Institute of Electronics, Information and Communication Engineers (IEICE).

Takaaki Fukutomi

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

He received a B.E. and M.E. in design from Kyushu University, Fukuoka, in 2006 and 2008. Since joining NTT in 2008, he has been researching speech recognition, including voice activity detection and speech enhancement. He received the Technical Development Award from ASJ in 2014. He is a member of ASJ.

Shoichiro Saito

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

He received a B.E. and M.E. from the University of Tokyo in 2005 and 2007. Since joining NTT in 2007, he has been researching acoustic echo cancellers and hands-free telephone terminals. He is a member of the Institute of Electrical and Electronics Engineers (IEEE), IEICE, and ASJ.

Yusuke Shinohara

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

He received a B.S. and M.S. in mechano-informatics from the University of Tokyo in 2002 and 2004. From 2004 to 2017, he was a research scientist at Toshiba Corporate Research and Development Center, Kanagawa. Since joining NTT in 2017, he has been involved in research and development of speech recognition, in particular, acoustic modeling with various types of deep neural networks. His research interests include speech recognition, computer vision, and machine training. He received the IEICE Pattern Recognition and Media Understanding (PRMU) Young Researcher Award in 2004, and the ASJ Awaya Prize Young Researcher Award in 2013. He has served as a committee member (2013–2015, 2017–present) and a secretary (2015–2017) for the Information Processing Society of Japan Special Interest Group (IPSJ SIG) on Spoken Language Processing. He is a member of IEEE, IEICE, ASJ, and IPSJ.

Satoshi Kobashikawa

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

He received a B.E., M.E., and Ph.D. from the University of Tokyo in 2000, 2002, and 2013. Since joining NTT in 2002, he has been researching speech recognition and spoken language processing. He received the Kiyasu Special Industrial Achievement Award in 2011 from IPSJ, the 58th Maejima Hisoka Award from the Tsushinbunka Association in 2012, and the 54th Sato Paper Award from ASJ in 2013. He is a member of ASJ, IPSJ, and IEICE.

Yoshikazu Yamaguchi

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

He received a B.E. and M.E. in electrical engineering from Osaka Prefecture University in 1993 and 1995. Since joining NTT in 1995, he has been researching automatic speech recognition technologies.

Noboru Harada

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

He received a B.S. and M.S. from the Department of Computer Science and Systems Engineering of Kyushu Institute of Technology, Fukuoka, in 1995 and 1997, and he later received a Ph.D. from University of Tsukuba. Since joining NTT in 1997, he has been researching lossless audio coding, high-efficiency coding of speech and audio, and their applications. He is an editor of the International Organization for Standardization/International Electrotechnical Commission (ISO/ IEC) 23000-6:2009 Professional Archival Application Format, ISO/IEC 14496-5:2001/ Amd.10:2007 reference software MPEG-4 Audio Lossless Coding and ITU-T (International Telecommunication Union - Telecommunication Standardization Sector) G.711.0, and has contributed to 3GPP (3rd Generation Partnership Project) Enhanced Voice Services. He is a member of IEICE, ASJ, IPSJ, the Audio Engineering Society, and IEEE.