

Establishment and Experimental Demonstration of Distant Speech Recognition System for Communication Robot

Shunichi YAMAMOTO*

Naoaki SUMIDA**

Kazuhiro NAKADAI**

ABSTRACT

Distant speech recognition technology for a robot was implemented to communicate with humans at a natural distance. There were two issues to address. First, surrounding reverberation and noise act to reduce recognition performance. In an attempt to resolve this, dereverberation and sound source identification technology were introduced in addition to the existing developed technology. Second, speech frequency response characteristics and wording differ from person to person. To enhance recognition performance, a deep neural network was used in the acoustic model, in addition to class-based N-grams accounting for the properties of speech recognition tasks in the language model. These technologies were then integrated into a robot system to perform a themed experimental demonstration taking reception calls.

The demonstration achieved a reception success rate of 98.3% and, through analysis of the causes of erroneous speech recognition, clarified further issues to address for improving performance.

1. Introduction

Communication is a skill that robots require to coexist with humans and provide them services, and speech is the mode of communication most frequently used in daily interaction, as well as the most natural and direct^{(1), (2)}. As such, verbal communication using speech recognition will be the focus of this paper.

In terms of speech recognition, voice input interfaces based on use of close-talk microphones are now offered in practical applications, such as information retrieval on smartphones, and are gaining public acceptance⁽³⁾. Speech recognition is also a vital function with great application potential for robots required to communicate with humans⁽¹⁾.

However, verbal dialogue between humans and robots at a natural distance will require distant speech recognition, the ability to recognize speech from a distance with the microphone equipped on the robot. One issue to address with distant speech recognition is that noise and reverberations drop the signal-to-noise ratio much more than when using close-talk microphones, greatly reducing speech recognition performance^{(3), (4)}. Noise as addressed

in this paper will refer to all sounds other than the target speaker's voice. As noises by this definition are input when using speech recognition in everyday settings, responding correctly requires the system to distinguish between the target speaker's voice and all other sounds⁽⁵⁾. This issue is particularly important with distant speech recognition, given the high frequency of input from noises and sounds other than the target speaker's voice. In addition, distant speech recognition must address the general issues facing speech recognition: handling individual differences in terms of voice type and wording⁽³⁾.

The individual core technologies of distant speech recognition have been the subject of numerous research studies. There is also the CHiME Challenge, a contest for distant speech recognition using common benchmarks⁽⁶⁾. This contest focuses on handling noise and reverberation, the issues to address for distant speech recognition.

Meanwhile, using ASIMO⁽⁷⁾, our humanoid robot, we at Honda have taken on an application for voice interaction with humans as a theme, aiming to achieve a practical level of distant speech recognition technology. Achieving distant speech recognition with a robot requires not only resolving the various issues facing distant speech recognition, but

* R&D Center X

** Honda Research Institute Japan Co., Ltd.

also integrating the underlying core technologies into robot systems.

Furthermore, extended use of these systems and analysis of the data will be important in evaluating performance of the application and unearthing those issues that only occur when used in the field by actual customers. As such studies are almost nonexistent elsewhere, the data collected and results from the demonstration should contribute greatly to future research and development. This paper will consist of reports on the distant speech recognition technology implemented and the results of the experimental demonstration conducted to verify its effectiveness.

2. Research Goals

The goals of this research are to improve the performance of practical level distant speech recognition and to clarify through actual use the issues to address when using distant speech recognition in a total system.

As detailed in the previous section, the issues to achieving distant speech recognition are numerous. Of these, this paper will be an attempt to improve performance by resolving the following four issues:

- (1) Reduced performance of speech recognition due to reverberation;
- (2) Erroneous speech recognition of non-target sounds;
- (3) Reduced performance of speech recognition due to individual differences in voice type; and
- (4) Reduced performance of speech recognition due to freedom in user wording.

For (1), speech recognition performance deteriorates significantly when reverberation time exceeds the speech recognition processing time unit (25 ms for this system)⁽⁸⁾. In public facilities, one of our targets, reverberation time is expected to exceed this time. Furthermore, compared to close-talk microphones, distant microphones are more affected by reverberation. As observed direct sounds are less powerful, the reverberation component power is almost unchanged. Thus, the dereverberation method proposed by Gomez et al.⁽⁸⁾ was introduced in order to eliminate the reverberation component.

For (2), as erroneous speech recognition processing of noise degrades performance, sound source identification using a Gaussian mixture model (GMM) was introduced as a pre-processing step to speech recognition to differentiate between target speech and non-target sounds⁽⁵⁾.

For (3), the model for speech phonemes (hereinafter, the “acoustic model”) must reflect the characteristics of as many people as possible. Conventionally, the GMM-based Hidden Markov Model (GMM-HMM) has been used for this model⁽⁹⁾, but research⁽¹⁰⁾ has recently increased using deep learning, which offers the potential for more precise modeling. Therefore, in this research, a deep neural network-based Hidden Markov Model (DNN-HMM)⁽¹⁰⁾ was introduced.

For (4), consider the scenario of calling for a worker at

company reception. One person may ask the receptionist, “Mr. Nakadai, please,” while another may say, “Yamamoto here to see Mr. Sumida.” Thus, speech recognition must be able to recognize the varied wording of the user, restricting the applications in which it is used. In light of this, a class-based N-gram language model was introduced⁽¹¹⁾.

In order to make the technology practical, instead of simply evaluating the above technologies individually, we thought it important to integrate them as a robot system and confirm their effectiveness when used by actual customers. Doing so will improve performance by identifying the real issues to address and furthering technology development.

Some examples of robot system demonstrations using distant speech recognition do exist, such as airport guidance using Hitachi, Ltd.’s EMIEW3⁽¹²⁾ and the Haneda Robotics Lab guide robot⁽¹³⁾, but as of April 2017, all such demonstrations have been over shorter terms of approximately one week, and the results, including analysis, remain closed. Therefore, for approximately two months, we demonstrated with the theme of reception calls and analyzed the data in order to clarify the issues to address.

3. Achievement Methods

This section explains the individual technologies used in order to resolve the issues.

3.1. Dereverberation

3.1.1. Dereverberation based on spectral subtraction

This sub-section is a description of the new algorithm used to enable online processing for the spectral subtraction-based dereverberation approach⁽⁸⁾. This approach achieves dereverberation by treating the reverberation component of input speech as noise, and then suppressing this noise with spectral subtraction.

Reverberation components are generally handled separately for early reflections and late reverberations due to the differences in their properties. This approach estimates the late reverberation component and suppresses it through spectral subtraction. The approach only suppresses late reverberation; it does not suppress early reflection. The majority of early reflection, characterized as reflected sounds that arrive relatively quickly after direct sound arrives, falls within the window length that is processing time unit for the latter part of speech recognition. Thus, early reflection can be suppressed with feature normalization techniques, such as cepstral mean normalization (CMN) and spectral mean normalization (SMN).

Late reverberation is estimated from $X(f, t)$, a short-term Fourier transform (STFT) of the input speech, and $H_L(f)$, a transfer function of the late reverberation. Given that the time window of $X(f, t)$ overlaps with the STFT of the past D frame, the amplitude spectrum for late reverberation $\hat{X}_L(f, t)$ at frequency-bin f and time t is approximated by:

$$|\hat{X}_L(f, t)| = \left\{ \frac{1}{D+1} \sum_{d=0}^D |X(f, t-d)H_L(f)|^\gamma \right\}^{\frac{1}{\gamma}} \quad (1)$$

Here, γ is a parameter set to a value of no more than 2, meant to keep the estimate of late reverberation from growing too large. For this research, $\gamma = 1.5$. Disassembling the STFT into frames and using past frames in this fashion allows for online processing with little lag.

The power spectrum $|X_E(f, t)|^2$, which includes the direct sound and early reflection, is calculated by spectral subtraction using the estimated late reverberation as:

$$|\hat{X}_E(f, t)|^2 = \begin{cases} |X(f, t)|^2 - \delta(f) |\hat{X}_L(f, t)|^2 & (> 0) \\ \beta |X(f, t)|^2 & (\text{else}) \end{cases} \quad (2)$$

Here, the β parameter functions to keep the power spectrum from resulting in a negative value. Using the pre-measured impulse response, $\delta(f)$ is calculated by standard optimization of the minimum mean square error (MMSE) from the late reverberation component estimate $\delta(f) \hat{X}_L(f, t)$ and actual value $X_L(f, t)$.

Using phase for input $X(f, t)$, output spectrum $\hat{X}_E(f, t)$ is calculated as:

$$\hat{X}_E(f, t) = \sqrt{|\hat{X}_E(f, t)|^2} e^{j\phi_X(f, t)} \quad (3)$$

Here, $\phi_X(f, t)$ expresses the phase of $X(f, t)$.

3.1.2. Evaluation testing

In order to confirm the effect of dereverberation, word accuracy for the trained DNN-HMM acoustic model was benchmarked, comparing results with and without dereverberation applied. For more details on learning by the DNN-HMM acoustic model and training data, refer to Sub-section 3.3. For evaluation benchmark data, reception call activity was recorded using the robot at reception with a reverberation time of 1.02 seconds with the reverberation time 20 (RT20) method. A sample utterance would be: “Yamamoto of Nakadai Holdings here to see Mr. Sumida.” In total, there were 258 instances of voice activity by 26 people, with 5,629 words recorded. Evaluations are based on word accuracy. The mean for all words and results by class (name of visitor or staff, organization name of visitor, affiliation name of staff; hereinafter collectively the “required vocabulary for reception”) for each voice activity that appeared are given in Fig. 1. Between evaluations with and without dereverberation, the mean word accuracy significantly increased by 2 points, and the level of significance was 5% by binomial test⁽¹⁴⁾. Looking at word accuracy for each class shows that the effect of dereverberation is particularly high for name of visitor or staff and organization name of visitor. As with mean word accuracy, the difference is significant for these

two parameters. Of the classes, name of visitor or staff and organization name of visitor contain many words and are not easy areas for speech recognition to analyze. Presumably, this could explain the greater improvement in recognition rates with dereverberation.

3.2. Sound Source Identification

3.2.1. Sound source identification model

In sound source identification, the input speech for each speech segment is identified as target speech or non-target sound, again referring to noise and sounds other than target speech. Its purpose is to identify target speech with high recall and allow for subsequent noise to be rejected. Target speech and noise were modeled with GMM for use in identification, using 27 dimensions⁽¹⁵⁾ for mel-scale log spectrum (MSLS) features extracted from the input speech. With margins at the start and end of the input speech segment to prevent any speech from being missed, the start and end of segments are unnecessary for sound source identification. Thus, as the power is lower at the start and end of segments than in the target portion, only features for times greater than the average power for the input speech segment were used for sound source identification (Fig. 2).

Next is the GMM model for target speech and non-target sounds. Expressing input features for time t as x_t , the class m (target or non-target) probability density function $p_m(x_t)$ can be expressed as:

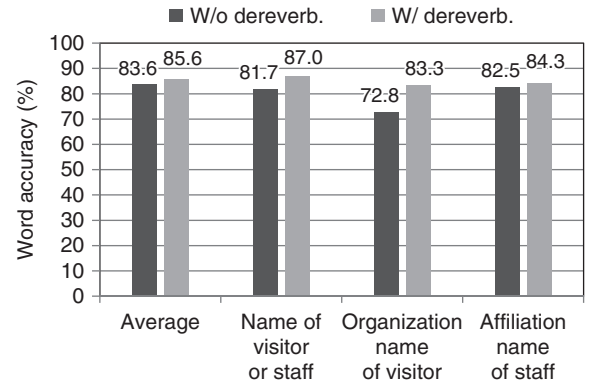


Fig. 1 Advance of word accuracy by using dereverberation

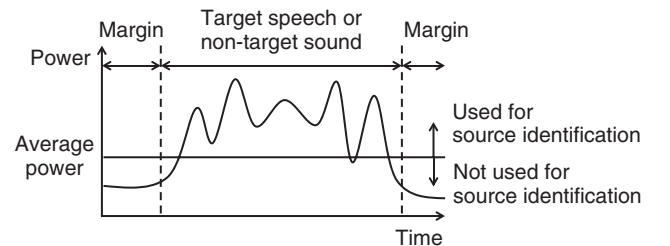


Fig. 2 Frames used for source identification

$$p_m(x_t) = \sum_{k=1}^K \lambda_{mk} N(x_t; \mu_{mk}, \sigma_{mk}^2) \quad (4)$$

In this probability density function, K is the number of mixtures, λ_{mk} is the mixture weight, and $N(x; \mu, \sigma^2)$ is the normal distribution of mean μ and dispersion σ^2 . Parameters λ_{mk} , μ_{mk} , and σ_{mk}^2 are estimated by EM algorithm with the training data.

During sound source identification, a vote is cast for class m with the highest likelihood for each time, using likelihood $p_m(x_t)$ for times of powers greater than average power P_{avg} in the input speech. The overall tally is normalized to arrive at a score, with highest scoring class m identified as the sound source for that voice activity segment. Score S_m is calculated as:

$$S_m = \frac{1}{|T_p|} \sum_{t \in T_p} Presence_m(x_t) \quad (5)$$

$$Presence_m(x_t) = \begin{cases} 1, & m = \operatorname{argmax}_i p_i(x_t) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Where $P(t)$ expresses the power of input speech at time t , it is defined as $T_p = \{t | P(t) > P_{avg}\}$, expressing the collection of times of powers greater than the average power for input speech.

3.2.2. Training data

For the training data for the target speech and non-target sound GMMs, the speech spoken by subjects to the robot and surrounding noise in the reception call usage environment were put through sound source localization, voice activity detection, and sound source separation. Training datasets for each class consisted of target speech from 104 speakers, with 26,164 voice activity segments and 1,577 non-target sounds. Correct labels were assigned to target speech and non-target sounds for each sound separated from the sound source.

3.2.3. Evaluation testing

In evaluation testing for the sound source identification approach of the system, two evaluation sets were used for the recorded speech as spoken by subjects to the robot in the reception call usage environment. The first set was the aforementioned evaluation benchmark data for reception calls with 26 speakers. The second set was the actual data from customers who used the system in the reception call demonstration to be detailed below. Breakdowns of the target speech and non-target sounds for each evaluation set are given in Table 1.

Performance for sound source identification was evaluated as the recall and precision for voice activity units. Here recall refers to the percentage of voice activity where actual target speech (or non-target sound) is predicted as target speech (or non-target sound). Meanwhile, precision

Table 1 Evaluation set for source identification

Evaluation set	Class	Number of sounds
Benchmark	Target speech	1,006
	Non-target sound	239
Reception experiment	Target speech	829
	Non-target sound	7,978

Table 2 Evaluation results of source identification

Evaluation set	Class	Recall (%)	Precision (%)
Benchmark	Target speech	96.5	91.1
	Non-target sound	94.5	80.5
Reception experiment	Target speech	87.1	34.7
	Non-target sound	95.7	98.4

refers to the percentage of voice activity where predicted target speech (or non-target sound) is actual target speech (or non-target sound). The resulting recall and precision for sound source identification in each evaluation set are given in Table 2.

A high recall is important for target speech in reception calls, and high recall was achieved for each evaluation set. However, recall for target speech in reception call usage is lower than the benchmark. Given the many sounds detected as surrounding noise when no speaker was present, due to the demonstration being over a longer term, the recall results are attributed to the high percentage of non-target sounds in the reception call demonstration evaluation set. The high number of noises were successfully identified as non-target sound and rejected at a high recall rate of 95.7%.

3.3. Speech Recognition: Acoustic Model

The approach applied for distant speech recognition in the robot hardware was DNN-HMM, an approach based on DNN reported to dramatically improve speech recognition rates. The following issues arise in DNN-HMM application of distant speech recognition for robot hardware:

- (1) Requires large training dataset; and
- (2) Mismatching between input speech and acoustic model.

Given that deep learning generally requires large training datasets, DNN-HMM applications for distant speech recognition in robots will also presumably require a large training dataset. Conventionally, we use a corpus with 82 hours of read-out speech to train the acoustic model with GMM-HMM. In application of the DNN-HMM, however, we collected a corpus of 904 hours of speech for use in training.

Also, in distant speech recognition using a robot-mounted microphone array, the signal-to-noise ratio of the input speech is decreased by surrounding noise, noise from the robot, and reverberation. Even pre-processing such as noise reduction and dereverberation is unable to

eliminate all noise and reverberation from the input speech. Further, mismatches in the input speech and acoustic model also reduce the rate of speech recognition. Ideally, the acoustic model and input speech will be matched by recording speech using the robot's microphone array in the environment where speech recognition will be used (hereinafter, "actual machine recording"). Preparing the large training dataset in this fashion is no simple task, however. Therefore, in addition to actual machine recording speech, we also used for training data speech simulating input speech during speech recognition usage by using pre-measured impulse response and robot noise, then synthesizing from clean speech and processing with sound source separation and dereverberation (hereinafter, "synthesized speech").

3.3.1. Deep neural network acoustic model

For training and decoding the DNN-HMM acoustic model, the open source software speech recognition toolkit Kaldi⁽¹⁶⁾ was used. The input features were made 297 dimensional, joining the 27-dimension MSLS feature with the 5 frames before and after it. For output, a fully connected DNN with 6 hidden layers of 2,048 units was used to set output as each of the triphone HMM states. State-level minimum Bayes risk was used for the DNN training criteria.

3.3.2. Performance verification through training data volume

In order to confirm performance of the acoustic model created in this research, word accuracy was compared whenever the training data increased. As shown in Table 3, 904 total hours of training data was used. Here "reception speech corpus" and "reception speech corpus recorded with ASIMO" are speech corpora collected for expected reception call tasks. The reception speech corpus comprises task voice activity recorded from expected reception call task scenarios read aloud and using the Wizard of Oz method. In recording task voice activity, speech close to that during system usage can be recorded by giving the speaker a task but leaving wording up to the speaker to recreate conversations between speaker and system. The reception speech corpus recorded with ASIMO is actual machine recordings of expected reception call task scenarios read aloud. For evaluation, the aforementioned evaluation benchmark data was used to evaluate word accuracy of 5,629 words in total. Word accuracy for the training datasets is given in Fig. 3.

A word accuracy of 88.9% was achieved for the 904 hours of training data, representing a vast improvement in performance. For comparison, word accuracy was 29.2% for the evaluation benchmark data when evaluated with the conventional GMM-HMM acoustic model.

3.4. Speech Recognition: Language Model

Language model requirements to be met for the

Table 3 Training data of our acoustic model

Speech corpus	Style	Speech	Hours
Commercially available speech corpus	Reading	Synthesized	329
Reception speech corpus	Reading and spontaneous speech	Synthesized	453
Reception speech corpus recorded with ASIMO	Reading	Recorded with ASIMO	122
Total			904

reception calls were set as follows:

- (1) Allow speakers to speak freely;
- (2) Completely cover all names of staff and affiliation names of staff; and
- (3) Ability to register additional visitor names and organization names of visitors at any time from data collected during the demonstration.

To meet these requirements, a class-based N-gram language model was used. To allow speakers to speak freely in wording, the class-based N-gram language model was trained with a language corpus collected from variously worded reception calls. Next, names of staff, names of visitors, and other headings were grouped into classes, allowing them to be covered completely and simplifying addition of new visitor names or other headings. There are seven classes in total: name of staff, affiliated company name of staff, affiliated office name of staff, affiliated department name of staff, affiliated block name of staff, name of visitor, and organization name of visitor.

The reception call language model as created had a vocabulary of 1,586 words at the start of the demonstration. This expanded to 1,681 words by the end of the demonstration, adding new visitor names and other vocabulary.

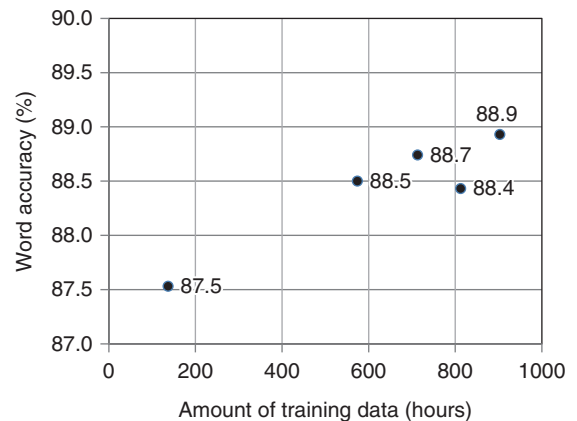


Fig. 3 Advance of word accuracy with increased training data

4. Experimental Demonstration

The technologies from the previous section were integrated with the robot, then used by regular customers and evaluated for performance.

4.1. Target Configuration

In actual performance evaluation, tasks were clearly defined, and then system usage by regular customers was evaluated. Our thinking was that having actual regular people use the system would help in determining current performance for practical use, and that analysis would further clarify the issues to address. Here the target task was set as taking calls at company reception. Customers coming to reception were made to talk to the robot in order to call an associate (person in charge, or PIC), and the robot made the actual call to the PIC. The speech-activated reception call flow is given in Fig. 4. Note here that the demonstration was approved by the Honda R&D Bioethics Committee prior to being conducted.

4.2. Demonstration Environment

For the test, the robot was positioned at reception for the R&D Center X (formerly the Fundamental Technology Research Center). Actual customers visiting the center spoke to the robot, and the robot called the PIC.

4.3. System Configuration

The system configuration used in the demonstration is given in Fig. 5. The areas outlined in dotted lines are the topic technologies treated in this research. In the demonstration, the aforementioned algorithms and data were applied for the dereverberation, sound source identification, acoustic model, and language model.

For the later language understanding step, the vocabulary needed for calling was extracted on a rule basis.

Distant speech recognition systems used in conversations with the robot must be able to respond quickly and process

input speech from multiple sources simultaneously. In order to speed up response times, processing from speech input at the microphone array to speech recognition was moved online. So that the speech recognition system could recognize and process input speech from multiple sources simultaneously, the system was built to track orientation of the multiple sound sources, then assign and manage an ID unique (sound source ID) to each sound source.

Processing from speech input to feature extraction was moved online by using HARK⁽¹⁷⁾, the open source software system for robot audition. For the dereverberation method in particular, as the algorithm required speech input to be analyzed in voice activity units, the algorithm was modified to allow online processing as outlined in sub-section 3.1.1. To create a speech recognition system, the aforementioned

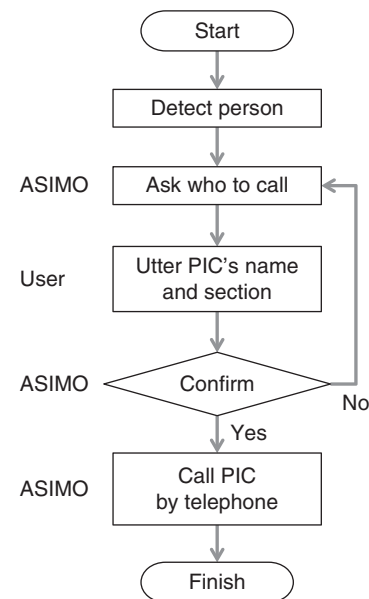


Fig. 4 Reception task flow

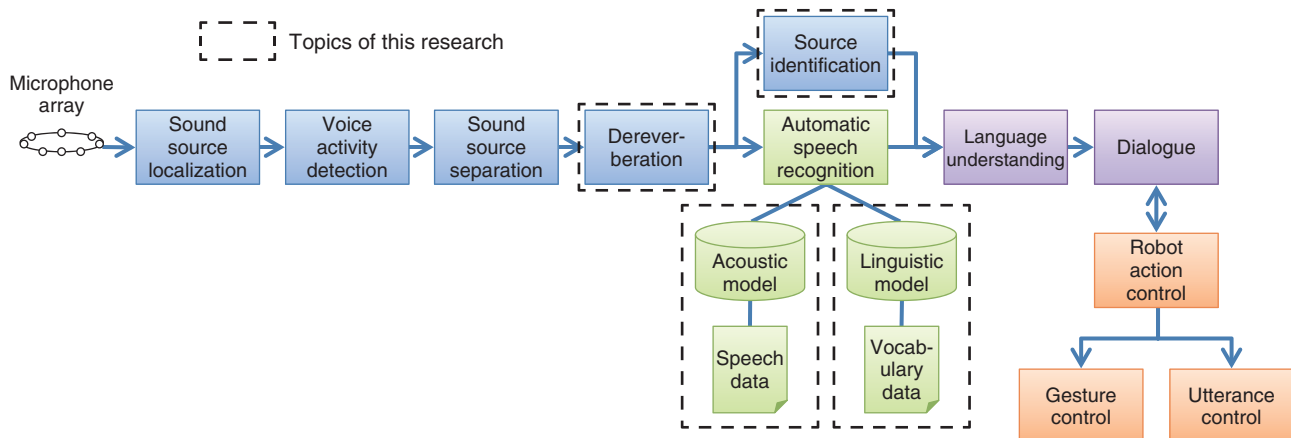


Fig. 5 Receptionist system

Kaldi was used to calculate acoustic likelihoods using a DNN acoustic model and to search word string candidates using a weighted finite state transducer. Newly implemented in this research were sequential processing for input features and simultaneous processing for multiple sound sources with threads. In order to speed up response and simultaneously process multiple sound sources, the package is capable of sequentially processing features input in 10-ms increments only, dividing each pair of input sound source IDs and feature strings into threads, and simultaneously processing the threads.

4.4. Demonstration Scenarios

Demonstration scenarios were built based on the flow given in Fig. 4. For the demonstration, the system was allowed to re-listen three times. Reception was considered successful when the correct PIC was called and unsuccessful in any other case the scenario did not advance, due to erroneous speech recognition or other causes.

4.5. Results

The demonstrations at reception took place over 37 days from January to February 2016. A total of 303 requests by actual speech were received at reception. A total of 4,881 words in actual customer voice activity were used to evaluate speech recognition. Speech recognition was evaluated based on word accuracy and the rate at which the correct PIC was actually called (reception success rate). Overall word accuracy was 66.8%. Of the 303 reception instances, 298 were successful for a reception success rate of 98.3%.

4.6. Discussion

Analyzing the causes for the five unsuccessful reception calls, two were erroneously recognized due to substitution error, and three were due to erroneous sound source identification (target speech discarded as non-target sound). In the substitution errors, the system erroneously recognized a name similar to that of the actual PIC. Regarding the sound source identification errors, an approach with a different model may be needed as the GMM may not fully describe the model to separate target speech with non-target sounds.

In terms of speech recognition, word accuracy dropped from 88.9% in benchmarking to 66.8% in the demonstration. The causes of error in recognition in the demonstration are grouped as given in Table 4. Of these error causes, three represent issues to be addressed in practical application. First, while specific to reception calls, erroneous recognition increases due to first-time visitors' names and company names being unknown words. Second, erroneous recognition increases with regular customers due to increased incidence of phenomena peculiar to spontaneous voice activity, such as filler words and misspeaking. Third, erroneous recognition increases when voice activity is overlapped with robot speech, which

Table 4 Cause of error in reception experiments

Cause of error	Word error rate (%)
Unknown words	10.9
Spontaneous speech	8.8
Non-target sound and VAD	5.7
Lack of training data	4.5
Overlapping speech	3.3
Total	33.2

results from the user not knowing the timing for robot voice activity and other causes. While the robot can recognize overlapping speech, in many cases the person will stop speaking when overlapping occurs, resulting in erroneous recognition. This presumably happens as it can be hard to read when the robot will start speaking.

Most cases of erroneous recognition, however, were not fatal to completing the reception task. The system was able to recognize with high accuracy the most important information, the PIC name, and affiliation name, thus achieving a high reception success rate. Performing tasks more complex than reception calls with voice interaction in the future will require automatic addition of unknown words, support for phenomena peculiar to natural voice activity, and turn-taking that is easier for the user to read.

5. Conclusion

Distant speech recognition technology was created for communicating with humans at a natural distance.

The dereverberation, sound source identification, and acoustic model proposed for the research were evaluated with evaluation benchmark data, and the implemented technology was confirmed to improve performance. In particular, the DNN acoustic model was confirmed to improve performance to 88.9% when trained with 904 hours of data. The proposed approach was integrated into the robot, achieving an overall word accuracy of 66.8% and reception success rate of 98.3% when used by actual customers at reception.

Analysis of the demonstration data revealed the following issues to address in order to further improve performance:

- (1) Support for unknown words;
- (2) Support for speech phenomena peculiar to spontaneous voice activity; and
- (3) Turn-taking and other robot expression.

With continuous daily use by actual customers, this demonstration, which gathered, analyzed, and verified data, should serve as a test case in further experimental demonstrations.

While the research theme of reception took place in a relatively quiet environment, it will be important in the future to also provide support for crowd noise, overhead announcements, and loud noise as found in public facilities and other places where people congregate.

References

- (1) Ministry of Economy: New Robot Strategy - Vision, Strategy, Action Plan -, (2015) (in Japanese)
- (2) Okuno, H. G.: Robot Audition: Present and Future, Journal of the Robotics Society of Japan, Vol. 28, No. 1, p. 2-5, (2010) (in Japanese)
- (3) Kawahara, T.: Recent Progress of Speech Recognition Technology, The journal of the Institute of Electrical Engineers of Japan, Vol. 133, No. 6, p. 364-367, (2013) (in Japanese)
- (4) Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Hori, T., Nakatani, T., Nakamura, A.: Linear Prediction-based Dereverberation with Advanced Speech Enhancement and Recognition Technologies for the Reverb Challenge, REVERB Workshop, (2014)
- (5) Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H., Shikano, K.: Noise Robust Real World Spoken Dialogue System Using GMM Based Rejection of Unintended Inputs, Proceedings of International Conference on Spoken Language Processing, Vol. 1, p. 173-176, (2004)
- (6) Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Yu, M. F. C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S., Nakatani, T.: The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices, IEEE ASRU, (2015)
- (7) Shigemi, S., Kawabe, K., Nakamura, T.: Development of New ASIMO - Realization of Autonomous Machine -, Honda R&D Technical Review, Vol. 24, No. 1, p. 34-41
- (8) Gomez, R., Kawahara, T.: Robust Speech Recognition Based on Dereverberation Parameter Optimization Using Acoustic Model Likelihood, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 7, p. 1708-1716, (2010)
- (9) Shikano, K., Ito, K., Kawahara, T., Takeda, K., Yamamoto, M.: Onseininshiki Sisutemu, Ohmsha, 198p., (2001) (in Japanese)
- (10) Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Processing Magazine, Vol. 29, No. 6, p. 82-97, (2012)
- (11) Brown, P. F., Petra, V. J. D., de Souza, P. V., Lai, J. C., Mercer, R. L.: Class-Based n-gram Models of Natural Language, Computational Linguistics, Vol. 18, No. 4, p. 467-479, (1992)
- (12) Hyumanoid Robotto EMIEW3 wo Katsuyoshita Jissho Jikken, Hitachi Hyoron, Vol. 98, No. 12, p. 742-743, (2016) (in Japanese)
- (13) <https://www.tokyo-airport-bldg.co.jp/hanedaroboticslab/>
- (14) Nakagawa, S., Takagi, H.: Statistical methods for comparing pattern recognition algorithms and comments on evaluating speech recognition performance, The Journal of the Acoustical Society of Japan, Vol. 50, No. 10, p. 849-854, (1994) (in Japanese)
- (15) Nishimura, Y., Shinozaki, T., Iwano, K., Furui, S.: Shuhasutaiikigoto no Omomitsukiyudo wo Mochiita Onseininshiki no Kento, Acoustical Society of Japan 2004 Shunki Kenkyuhappyokai Kouenronbunshu, p. 117-118, (2004) (in Japanese)
- (16) Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J.: The Kaldi Speech Recognition Toolkit, IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, p. 1-4, (2011)
- (17) Okuno, H. G., Nakadai, K.: Robot Audition Open-Sourced Software HARK, Journal of the Robotics Society of Japan, Vol. 28, No. 1, p. 6-9, (2010) (in Japanese)

■ Author ■



Shunichi YAMAMOTO



Naoaki SUMIDA



Kazuhiro NAKADAI