Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments — Newest Part of the CENSREC Series —

Takanobu Nishiura[1], Masato Nakayama[1], Yuki Denda[1], Norihide Kitaoka[2], Kazumasa Yamamoto[3], Takeshi Yamada[4], Satoru Tsuge[5], Chiyomi Miyajima[2], Masakiyo Fujimoto[6], Tetsuya Takiguchi[7], Satoshi Tamura[8], Shingo Kuroiwa[9], Kazuya Takeda[2], Satoshi Nakamura[10]

[1]Ritsumeikan University, [2]Nagoya University, [3]Toyohashi University of Technology, [4]University of Tsukuba, [5]University of Tokushima, [6]NTT Corporation, [7]Kobe University, [8]Gifu University, [9]Chiba University, [10]ATR/NiCT

[1]Kusatsushi, 525-8577 Japan, [2]Nagoya-shi, 464-8603 Japan, [3]Toyohashi-shi, 441-8580 Japan, [4]Tsukubashi, 305-8573 Japan, [5]Tokushima-shi, 770-8506 Japan, [6]"Keihanna Science City", Kyoto-fu, 619-0237 Japan, [7]Kobe-shi, 657-8501 Japan, [8]Gifu-shi, 501-1193 Japan, [9]Chiba-shi, 263-8522 Japan, [10]"Keihanna Science City", Kyoto-fu, 619-0288 Japan

[1]{nishiura@is, gr020040@se, gr021052@se}.ritsumei.ac.jp, [2]{kitaoka@nagoya-u, miyajima@is.nagoya-u.ac, kazuya.takeda@nagoya-u}.jp, [3]kyama@slp.ics.tut.ac.jp, [4]takeshi@cs.tsukuba.ac.jp, [5]tsuge@is.tokushima-u.ac.jp, [6]masakiyo@cslab.kecl.ntt.co.jp, [7]takigu@kobe-u.ac.jp, [8]tamura@info.gifu-u.ac.jp, [9]kuroiwa@faculty.chiba-u.jp, [10]nakamura@slt.atr.co.jp

Abstract

Recently, speech recognition performance has been drastically improved by statistical methods and huge speech databases. Now performance improvement under such realistic environments as noisy conditions is being focused on. Since October 2001, we from the working group of the Information Processing Society in Japan have been working on evaluation methodologies and frameworks for Japanese noisy speech recognition. We have released frameworks including databases and evaluation tools called CENSREC-1 (Corpus and Environment for Noisy Speech RECognition 1; formerly AURORA-2J), CENSREC-2 (in-car connected digits recognition), CENSREC-3 (in-car isolated word recognition), and CENSREC-1-C (voice activity detection under noisy conditions). In this paper, we newly introduce a collection of databases and evaluation tools named CENSREC-4, which is an evaluation framework for distant-talking speech under hands-free conditions. Distant-talking speech recognition is crucial for a hands-free speech interface. Therefore, we measured room impulse responses to investigate reverberant speech recognition. The results of evaluation experiments proved that CENSREC-4 is an effective database suitable for evaluating the new dereverberation method because the traditional dereverberation process had difficulty sufficiently improving the recognition performance. The framework was released in March 2008, and many studies are being conducted with it in Japan.

1. Introduction

Recently, speech recognition performance has been drastically improved by statistical methods and huge speech databases. Now performance improvement under such realistic environments as noisy conditions has become the focus, and some projects for noisy speech recognition evaluation have been organized.

The SPeech recognition In Noisy Environment (SPINE) project in the US established a specific task including the recognition of spontaneously spoken English dialogs between an operator and a soldier in noisy environments (SPINE1, 2).

The European Telecommunications Standards Institute (ETSI) has also developed noisy speech recognition evaluation frameworks called Aurora. ETSI has distributed Aurora 2 (Hirsh and Pearce, Sept 2000), a connected digit recognition task under various additive noises, Aurora 3, an in-car connected digit recognition task, and Aurora 4 (Aurora document-no. AU/345/01, Aug 2001), a continuous noisy speech recognition task.

We, the working group (AURORA-J/CENSREC) in the Information Processing Society in Japan, have worked on evaluation methodologies and evaluation frameworks for Japanese noisy speech recognition since October 2001. We originally followed the ETSI Aurora 2 task setting due to its simplicity and generality, and we have also released CENSREC-1 (Corpus and Environment for Noisy Speech RECognition 1; AURORA-2J) (Nakamura et al., March 2005), which included a database and evaluation tools. After that, we released CENSREC-2 (in-car connected digit recognition) (Nakamura et al., Sept 2006), CENSREC-3 (in-car isolated word recognition) (Fujimoto et al., Nov 2006), and CENSREC-1-C (voice activity detection under noisy conditions) (Kitaoka et al., Dec 2007) with original evolutions.

So far we have developed evaluation frameworks for ad-

Table 1: Noises in CENSREC-1

	Additive noise	Filter
Testset A	subway, babbling, car, exhibition	G.712
Testset B	restaurant, street, airport, train station	G.712
Testset C	subway, street	MIRS

ditive noisy speech recognition performance. But in noisy speech recognition, speech recognition performance is degraded not only by additive noise but also by multiplicative noise under hands-free conditions. In this paper, we newly introduce a framework including a database and evaluation tools named CENSREC-4, which is an evaluation framework for distant-talking speech under hands-free conditions.

2. CENSREC Series

We have developed evaluation frameworks of noisy speech recognition to compare many methods of processing noisy speech. We first review the CENSREC series.

2.1. CENSREC-1/AURORA-2J

CENSREC-1 (AURORA-2J) is a Japan version of AURORA-2, a noisy continuous digit recognition database developed in Europe (ETSI standard document, 2000)(Hirsh and Pearce, Sept 2000). We released it in July 2003, and many researchers have published papers using it. Each utterance ranges in length from 1 to 7 numbers, and the number of speakers (110, 55 females and 55 males) is the same as AURORA-2. The utterance transcriptions are direct translations of AURORA-2. The vocabulary includes eleven Japanese numbers: "ichi," "ni," "san," "yon," "go," "roku," "nana," "hachi," "kyu," "zero," and "maru." There are two training conditions: clean and multi-condition. The test set has three subsets, as shown in Table 1, which is identical to AURORA-2. The noises used in Testset A are also used in multi-condition training, so they are called *known noises*. Only Testset C differs from the others in terms of transmission characteristics.

This database focuses on the effects of additive noises. Training and baseline test scripts based on HTK are also provided.

2.2. CENSREC-2

CENSREC-2 is another database for the evaluation of noisy continuous digit recognition whose data were recorded in actual car driving environments. This database has been distributed since December 2005. All utterances were recorded in a car while driving with close and far (located on the ceiling) microphones. These data are not *simulated* as CENSREC-1; they are *real*. There are 11 recording conditions: combinations of three vehicle speeds (idling, low-speed driving on a city street, and high-speed driving on an expressway) and six in-car environments (normal, with air-conditioner on, with CD player on, and with windows open). A total of 17,651 utterances were spoken by 104 speakers (73 for training data and 31 for test data).

Based on the combination of recording conditions for training and test data, we set the following four evaluation conditions:

Condition 1 microphone: same, environment: same

Condition 2 microphone: same, environment: different

Condition 3 microphone: different, environment: same

Condition 4 microphone: different, environment: different

2.3. CENSREC-3

The CENSREC-3 data, distributed since February 2005, were also recorded in actual car driving environments, but the utterances are isolated words. We selected 50 command words supposedly used for a navigation system. A total of 14,216 utterances were spoken by 18 speakers

Based on the combination of recording environments for training and test data, we set the following six condition categories that correspond to the three conditions, well-matched (WM), moderate-mismatched (MM), and high-mismatched (HM), used in the European AURORA-3 database:

Conditions 1, 2, and 3 microphone: same, environment: same (WM)

Condition 4 microphone: same, environment: different (MM)

Conditions 5 and 6 microphone: different, environment: different (HM)

2.4. CENSREC-1-C

Voice activity detection (VAD) plays an important role in speech processing and includes speech recognition, speech enhancement, and speech coding under noisy environments. We developed an evaluation framework for VAD under noisy environments called CENSREC-1-C. This framework consists of noisy continuous digit utterances and evaluation tools for VAD results.

The simulated speech data of CENSREC-1-C are constructed by concatenating several utterances spoken by one speaker. The number of utterances in the concatenated speech data is either nine or ten. These original utterances are all included in CENSREC-1. A one-second silent signal taken from CENSREC-1 is inserted between the utterances. In CENSREC-1, the number of speakers per noise environment is 104 (52 females and 52 males). Thus, in CENSREC-1-C, the number of speech data per noise environment is 104.

Additionally, we recorded the speech data in two actual noisy environments (a restaurant and near a highway) and in both low and high SNR conditions. We placed a microphone 50 cm from the speaker's mouth. Ten subjects for recording speech were employed. The recorded speech consists of four files for one subject (a total of 38-39 utterances). A single file includes 8-10 utterances in sequence and two-second intervals for each utterance in each noisy environment and each SNR condition. The recorded speech

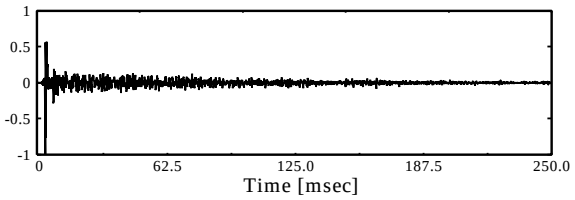

Figure 1: Impulse response data in Japanese style bath

data include 1380 utterances (144 files) for nine subjects in two actual noisy environments and two SNR conditions. One subject tended to put a long time interval between digits in one continuous digit utterance. Therefore the speech data of that subject were not used as evaluation data, but were included as realistic samples in the database.
We defined two evaluation measures: frame-level detection performance and utterance-level detection performance. We also provided evaluation results of a baseline power-based VAD method and an Excel sheet for evaluation.

3. CENSREC-4—Evaluation Framework for Reverberant Speech Recognition

The target evaluation framework of CENSREC-4 is distant-talking speech recognition in various reverberation environments. The data contained in CENSREC-4 are connected digit utterances as in CENSREC-1. Two subsets are included in the data: 'basic data sets' and 'extra data sets.' The basic and extra data sets consist of connected digit utterances in reverberant environments. The utterances in the extra data sets are affected by ambient noises in addition to reverberations. An evaluation framework is only provided for the basic data sets as HTK-based HMM training and recognition scripts.

3.1. Basic data sets

The basic data sets are used as the evaluation environment for the room impulse response-convolved speech data.

3.1.1. Room impulse response data

Many room impulse responses were measured to simulate various environments by convolving with clean speech signals and room impulse responses in actual environments. Impulse responses were measured using the time stretched pulse (TSP) method (Suzuki et al., 1995). The TSP length was 131,072 points, and the number of synchronous additions was 16. Figure 1 shows a sample of impulse responses on the time domain. Impulse responses were normalized at 0.5 with an absolute value of maximum amplitude. CENSREC-4 includes impulse responses recorded in eight kinds of rooms: an office, an elevator hall (a waiting area in front of an elevator), in-car, a living room, a lounge, a Japanese style room (with tatami flooring), a meeting room, and a Japanese style bath (a prefabricated bath). We measured the room impulse responses based on the conditions shown in Table 2. Figure 2 shows the microphone settings for all environments except the in-car and Japanese style bath. Figures 3 and 4 show an example of recording position and landscape in the meeting room environment.

Table 2: Recording equipment and conditions

Microphone	SONY, ECM-88B
Microphone amplifier	PAVEC, Thinknet MA-2016C
A/D board	TOKYO ELECTRON DEVICE, TD-BD-8CSUSB-2.0
Loudspeaker	B&K, Mouth simulator Type 4128
Speaker amplifier	YAMAHA, P4050
Sampling frequency	48 kHz (downsampled to 16 kHz before convolving)
Quantization	16 bits

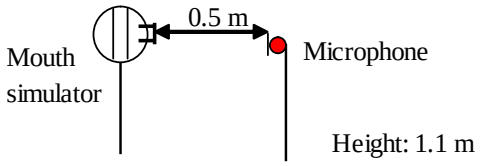

Figure 2: Recording setup for impulse responses

In all environments except in-car and Japanese style bath, we set the microphone near the center of the room, as in Figs. 3 and 4.
For the in-car environment, we used a middle-size sedan and set the mouth simulator on the driver's seat and the microphone on the sunvisor. The distance between the mouth simulator and the microphone was about 0.4 m. In the lounge environment, we set the microphone on a coffee table. In the Japanese style bath environment, we set the mouth simulator over a bathtub filled with cold water and attached the microphone to the side wall. The distance between the mouth simulator and the microphone was about 0.3 m.
Table 3 shows the room size, the distance between the microphone and the loudspeaker (mouth simulator), the reverberation time, temperature, humidity, and the average ambient noise level in each recording room. In Table 3, reverberation time (T_{60}) is displayed with 0.05 sec resolution, and the ambient noise level is displayed with 0.5 dB resolution.

3.1.2. Simulated data (Testset A/B)

We made simulated reverberant speech by convolving the impulse responses to the clean speech. The clean speech of CENSREC-1 (the sampling frequency was 16 kHz for CENSREC-4, whereas it was 8 kHz for CENSREC-1) was used. The details of the recording conditions, utterances, and speaking styles are the same as in CENSREC-1. The vocabulary of the simulated data included in CENSREC-4 consisted of eleven Japanese numbers: "ichi," "ni," "san," "yon," "go," "roku," "nana," "hachi," "kyu," "zero," and "maru. "The recording was conducted in a soundproof booth using a Sennheiser HMD25 headset microphone. The speech data were sampled at 16 kHz, quantized into 16 bit integers, and saved in the little-endian format.
Training and testing data were prepared in the same way as in CENSREC-1. The latter were divided into two sets:

Table 3: Room size, distance between microphone and loudspeaker, reverberation time, ambient noise level, humidity, and temperature in recording

Room	Test set	Room size	Dis. between Mic. and LS	Reverberation time [T_{60}]	Tempe-rature	Humi-dity	Amb. noise level [dBA]
Office	$A/C/D$	9.0 × 6.0 m	0.5 m	0.25 sec	30℃	40%	36.5 dB
Elevator hall	A	11.5 × 6.5 m	2.0 m	0.75 sec	30℃	50%	39.0 dB
In-car	$A/C/D$	Middle-sized sedan	0.4 m	0.05 sec	29℃	44%	32.0 dB
Living room	A	7.0 × 3.0 m	0.5 m	0.65 sec	30℃	54%	34.0 dB
Lounge	$B/C/D$	11.5 × 27.0 m	0.5 m	0.50 sec	27℃	50%	52.5 dB
Japanese style room	B	3.5 × 2.5 m	2.0 m	0.40 sec	30℃	54%	30.0 dB
Meeting room	$B/C/D$	7.0 × 8.5 m	0.5 m	0.65 sec	27℃	52%	48.5 dB
Japanese style bath	B	1.5 × 1.0 m	0.3 m	0.60 sec	31℃	62%	29.5 dB

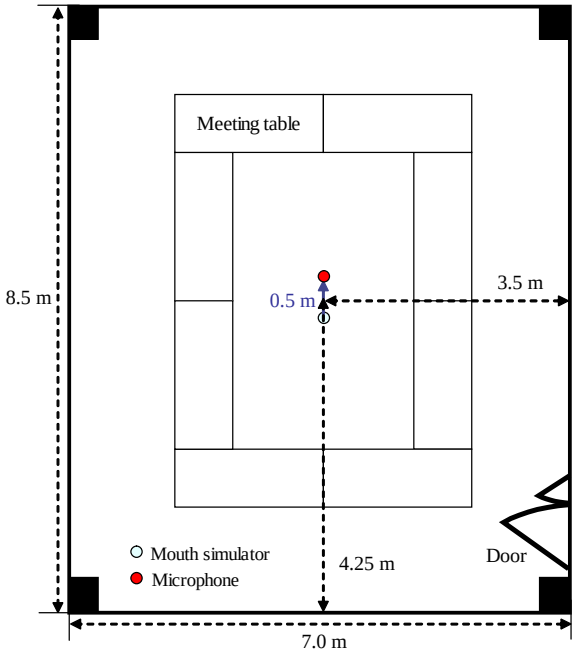

Figure 3: Layout of recording environment in meeting room

Figure 4: Photograph of recording environment in meeting room

Testset A (office, elevator hall, in-car, and living room) and Testset B (lounge, Japanese style room, meeting room, and Japanese style bath). Total utterances were 4,004 by 104 speakers (52 females and 52 males).

For Testset A/B, the utterances were divided into four groups corresponding to the reverberant conditions. Thus each reverberant condition included 1,001 utterances. In CENSREC-1, the noises in Testset A were used for both the testset and the training set (called *known noises*), but those in Testset B were only used for the training set (*unknown noises*). Similar, the CENSREC-4 basic data sets also have two types of testsets: Testset A with *known reverberant environments* and B with *unknown reverberant environments*.

Two sets of training data were prepared, clean and multi-condition. Total utterances were 8,440 by 110 speakers (55 females and 55 males). For the multi-condition training data, four kinds of reverberation (office, elevator hall, in-car, and living room) were convolved to the clean speech. Thus each reverberant condition included 2,110 utterances.

3.2. Extra data sets

The extra data sets consist of simulated and recorded data that are affected by both the additive and multiplicative noise. These data digress from the main topic as the Reverberant Speech Recognition Evaluation Environments. Thus, we only provide the testing/training data as extra data sets and don't provide an evaluation framework with them at the present time.

3.2.1. Simulated data with multiplicative and additive noise (Testset C)

We made simulated reverberant and noisy speech by convolving the room impulse responses and adding noise recorded in real environments to the clean speech. These extra data sets are called Testset C and consist of four environments: two from Testset A (office, in-car) and two from Testset B (lounge, meeting room).

In each environment, we recorded background noise for about 120 sec. The first half of the recorded data was used to make testing data, and the second half was to make training data.

For the testing data, total utterances were 4,004 by 104 speakers (52 females and 52 males), which is completely identical to Testset A/B. To make Testset C, these utterances were quartered, and four kinds of reverbera-

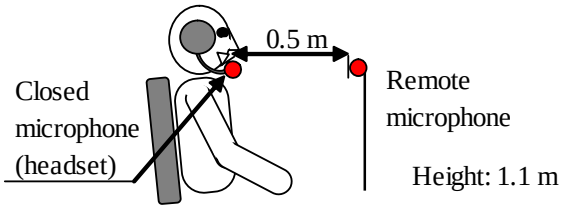

Figure 5: Recording setup for real data

tion (office, in-car, lounge, and meeting room) were convolved, and background noises were added to the reverberant speech at ∞ dB, 20 dB, 10 dB, and 5 dB of the Signal-to-Noise Ratio (SNR). However, if the reverberant and noisy conditions are identical, the utterance contents are also the same, regardless of SNR. Thus 1,001 utterances were included for each reverberant condition.
For the training data, total utterances were 6,752 by 88 speakers (44 females and 44 males). To make extra training data, these utterances were convolved as four kinds of reverberation (office, elevator hall, in-car, and living room), and background noises were added to the reverberant speech at ∞ dB, 20 dB, 10 dB, and 5 dB of SNR. Thus 422 utterances were included for each reverberant condition and SNR. In addition, clean training data were prepared, and the total utterances were 1,688 by 22 speakers (11 females and 11 males) as optional training data that were not utilized as training data.

3.2.2. Real recorded data in real environments (Testset D)

We recorded real data with two microphones (closed and remote) under the conditions shown in Table 2 with human speakers instead of a mouth simulator. This data set, called Testset D, was recorded under the same environments as Testset C by ten human speakers (five females and five males). In each environment, the room size and recording position were the same as Testsets A and B. Figure 5 shows the recording setup. The recorded speech by each speaker consists of two major parts: testing data (49 or 50 utterances) and training data for adaptation (11 utterances). Testset D has 2,536 utterances (2,536 files).

3.3. Reference baseline scripts

We produced CENSREC-4 baseline scripts based on the CENSREC-1 baseline scripts to perform HMM training and recognition experiments by HTK in the same way as CENSREC-1. They were only provided for the basic data sets as described above. As a result of various experiments (with various HMM topology, various feature vectors, and so on) and discussions, we specified the baseline scripts as follows:

- The acoustic model set consists of 18 phoneme models: (/a/, /i/, /u/, /u:/, /e/, /o/, /N/, /ch/, /g/, /h/, /k/, /ky/, /m/, /n/, /r/, /s/, /y/, /z/), silence ('sil'), and short pause ('sp').
- Each phoneme model and 'sil' have 5 states (3 emitting states), and 'sp' has 3 states (1 emitting state). The output distribution of 'sp' is identical as the center state of 'sil'.
- Each state of the phoneme models has 20 Gaussian mixture pdfs, and 'sil' or 'sp' has 36 Gaussian mixtures.
- The feature parameter of the baseline system is 39 dimensional feature vectors that consist of 12 MFCC, 12 ΔMFCC, 12 $\Delta\Delta$MFCC, log power, Δ power, and $\Delta\Delta$ power, calculated by HCopy of HTK. Analysis conditions were pre-emphasis ($1-0.97z^{-1}$), hamming window, 25 ms frame length, and 10 ms frame shift.
- Grammar-based connected digit recognition by HVite of HTK was used for the recognition experiments.
- Almost all the scripts were written as shell scripts and the remainder as Perl scripts. In these scripts, the HMM acoustic models were trained with HTK tools and used for recognition experiments.

3.4. Reference baseline performance

Table 4 shows the CENSREC-4 baseline performance for the basic data sets. In Table 4, its upper half shows the clean training results, its lower half shows the multi-condition training results, its right half shows digit accuracy, and its left half shows the string correct rate, defined as the correct recognition rate for all digits in each connected digit. In Tables 4 and 5, "w/o" shows the recognition result for the clean speech data (without convolving impulse responses), and "w" shows the recognition result for the reverberant speech data (with convolving impulse responses). Table 4 shows that the longer the reverberation time is, the worse the recognition performance, since no dereverberation process was used in the CENSREC-4 baseline.
This result is provided as a Microsoft Excel spreadsheet to get summary tables for evaluating the results. The summary tables of the recognition performance are confirmable as Table 5, because the relative performance with baseline is calculated automatically by inputting the results into spreadsheets. Published summary tables can be easily compared to other recognition performances.

3.5. Evaluation experiment with advanced technology

Cepstral Mean Normalization (CMN) (Furui, 1981), one traditional dereverberation process with advanced technology, is a simple and effective way of normalizing the feature space and thereby reducing channel distortion. It has, therefore, been adopted in many current systems. To appreciate the difficulties involved for basic data sets, we evaluated the improvement of recognition performance with CMN for the basic data sets. Table 6 shows recognition performance with CMN for the basic data sets, and Table 7 shows the summary tables of the recognition performance with CMN for the basic data sets.
As a result of Table 7, relative performance was improved about 15 to 25% in clean training but was degraded about 7% in multi-condition training. Thus, CMN had difficulty achieving sufficient improvement of recognition performance because it is ineffective under longer reverberant conditions. Therefore, we consider that the other traditional

Table 4: CENSREC-4 baseline performance for basic data sets

Clean training (%STRING)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w/o	98.5	98.1	98.5	98.2	98.3
w	93.1	30.7	86.1	65.3	68.8
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w/o	98.5	98.1	98.5	98.2	98.3
w	43.9	74.1	74.1	54.3	61.6

Clean training (%Acc)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w/o	99.5	99.4	99.5	99.4	99.4
w	97.5	57.9	95.6	84.4	83.8
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w/o	99.5	99.4	99.5	99.4	99.4
w	74.0	89.5	89.8	78.0	82.8

Multi-condition training (%STRING)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w	84.0	76.5	85.0	77.4	80.7
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w	52.5	82.3	81.6	62.0	69.6

Multi-condition training (%Acc)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w	94.4	90.6	95.0	91.6	92.9
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w	79.9	93.4	93.6	84.2	87.8

Table 5: Summary tables of recognition performance for basic data sets in CENSREC-4 spread sheet

%STRING				
		A	B	Overall
Clean training	w/o			
	w			
Multi-condition training	w			

%Acc				
		A	B	Overall
Clean training	w/o			
	w			
Multi-condition training	w			

Relative performance (%STRING)				
		A	B	Overall
Clean training	w/o			
	w			
Multi-condition training	w			

Relative performance (%Acc)				
		A	B	Overall
Clean training	w/o			
	w			
Multi-condition training	w			

dereverberation processes will have also difficulty achieving sufficient improvement of recognition performance for the basic data sets. This database includes very challenging and variable data. We hope to develop new dereverberation technology with this database.

4. Conclusion

In this paper, we newly introduced CENSREC-4, an evaluation framework for distant-talking speech under hands-free conditions. CENSREC-4 is a good database suitable for evaluating the new dereverberation method because the traditional dereverberation process had difficulty achieving sufficient improvement of recognition performance. The framework was released in March 2008, and many studies are being conducted with it in Japan. We will evaluate extra data sets in the near future.

5. Acknowledgements

The authors wish to thank the members of the Speech Resources Consortium in the National Institute of Informatics (NII-SRC), Japan, for their generous assistance in these activities. The present study was conducted using the CENSREC-4 database developed by the IPSJ-SIG SLP Noisy Speech Recognition EvaluationWorking Group.

6. References

M. Fujimoto, K. Takeda, and S. Nakamura. Nov. 2006. Censrec-3: An evaluation framework for japanese speech recognition in real driving-car environments. *IEICE Transactions on Information and Systems*, vol. E89-D, no. 11:pp. 2783–2793.

S. Furui. 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 2:pp. 254–272.

H.G. Hirsh and D. Pearce. Sept. 2000. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. *ISCA ITRW ASR2000*.

N. Kitaoka, K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura. Dec. 2007. Censrec-1-c: Development of vad evaluation framework censrec-1-c and investigation of relationship between vad and speech recognition performance. *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU 2007)*, pages pp. 607–612.

S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo. March 2005. Aurora-2j, an evaluation framework for japanese

Table 6: Recognition performance with CMN for basic data sets

Clean training (%STRING)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w/o	98.20	98.40	98.90	98.80	98.6
w	93.40	27.77	96.00	63.24	70.1
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w/o	98.20	98.40	98.90	98.80	98.6
w	66.23	80.32	82.08	60.34	72.2

Clean training (%Acc)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w/o	99.42	99.43	99.67	99.63	99.5
w	97.78	65.96	98.72	83.46	86.5
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w/o	99.42	99.43	99.67	99.63	99.5
w	87.32	92.20	93.25	81.73	88.6

Multi-condition training (%STRING)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w	80.72	77.72	79.02	73.93	77.8
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w	79.62	78.92	80.62	56.04	73.8

Multi-condition training (%Acc)					
	A				
	Office 0.25 sec.	Elevator hall 0.75 sec., 2m	In-car 0.05 sec.	Living room 0.65 sec.	Average
w	92.78	91.90	92.54	90.00	91.8
	B				
	Lounge 0.50 sec.	Japanese room 0.40 sec., 2m	Meeting room 0.65 sec.	Japanese bath 0.60 sec.	Average
w	92.57	91.87	93.14	81.33	89.7

Table 7: Summary table of recognition performance with CMN for basic data sets

%STRING				
		A	B	Overall
Clean training	w/o	98.6	98.6	98.6
	w	70.1	72.2	71.2
Multi-condition training	w	77.8	73.8	75.8

%Acc				
		A	B	Overall
Clean training	w/o	99.5	99.5	99.5
	w	86.5	88.6	87.6
Multi-condition training	w	91.8	89.7	90.8

Relative performance (%STRING)				
		A	B	Overall
Clean training	w/o	13.9%	13.9%	13.9%
	w	16.3%	27.0%	21.7%
Multi-condition training	w	-17.7%	4.2%	-6.8%

Relative performance (%Acc)				
		A	B	Overall
Clean training	w/o	18.1%	18.1%	18.1%
	w	23.9%	31.9%	27.9%
Multi-condition training	w	-20.3%	3.5%	-8.4%

noisy speech recognition. *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3:pp. 535–544.

S. Nakamura, M. Fujimoto, and K. Takeda. Sept. 2006. Censrec2: Corpus and evaluation environments for in car continuous digit speech recognition. *Proc. ICSLP'06*, pages pp. 2330–2333.

Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone. 1995. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *J. Acoust. Soc. Am.*, vol. 97, no. 2:pp. 1119–1123.