Acoust. Sci. & Tech. **32**, 5 (2011)

TECHNICAL REPORT

CENSREC-4: An evaluation framework for distant-talking speech recognition in reverberant environments

Takahiro Fukumori[1,*], Takanobu Nishiura[1,†], Masato Nakayama[2], Yuki Denda[3], Norihide Kitaoka[4], Takeshi Yamada[5], Kazumasa Yamamoto[6], Satoru Tsuge[7], Masakiyo Fujimoto[8], Tetsuya Takiguchi[9], Chiyomi Miyajima[4], Satoshi Tamura[10], Tetsuji Ogawa[11], Shigeki Matsuda[12], Shingo Kuroiwa[13], Kazuya Takeda[4] and Satoshi Nakamura[12]

[1]*Ritsumeikan University, Kusatsu, 525–8577 Japan*
[2]*Kinki University, Kinokawa, 649–6493 Japan*
[3]*Murata Machinery, Ltd., Kyoto, 612–8686 Japan*
[4]*Nagoya University, Nagoya, 464–8603 Japan*
[5]*University of Tsukuba, Tsukuba, 305–8573 Japan*
[6]*Toyohashi University of Technology, Toyohashi, 441–8580 Japan*
[7]*Daido University, Nagoya, 457–8530 Japan*
[8]*NTT Communication Science Laboratories, NTT Corporation, "Keihanna Science City," Kyoto, 619–0237 Japan*
[9]*Kobe University, Kobe, 657–8501 Japan*
[10]*Gifu University, Gifu, 501–1193 Japan*
[11]*Waseda University, Tokyo, 169–8050 Japan*
[12]*National Institute of Information and Communications Technology, "Keihanna Science City," Kyoto, 619–0288 Japan*
[13]*Chiba University and National Institute of Information and Communications Technology, Chiba, 263–8522 Japan*

(*Received 28 January 2011, Accepted for publication 23 March 2011*)

Abstract: We have been distributing a new collection of databases and evaluation tools called CENSREC-4, which is a framework for evaluating distant-talking speech in reverberant 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 data sets are used for evaluating the room impulse response-convolved speech data to simulate the various reverberations. The extra data sets consist of simulated data and corresponding real recorded data. Evaluation tools are presently only provided for the basic data sets and will be delivered to the extra data sets in the future. The task of CENSREC-4 with a basic data set appears simple; however, the results of experiments prove that CENSREC-4 provides a challenging reverberation speech-recognition task, in the sense that a traditional technique to improve recognition and a widely used criterion to represent the difficulty of recognition deliver poor performance. Within this context, this common framework can be an important step toward the future evolution of reverberant speech-recognition methodologies.

Keywords: Reverberant speech database, Reverberant speech recognition, Various recording environments, Room impulse response, Evaluation framework

PACS number: 43.72.Ne [doi:10.1250/ast.32.201]

1. INTRODUCTION

The performance of speech recognition has been drastically improved by statistical methods and huge speech databases in recent years. Improvements in performance under realistic environments, such as noisy conditions, have become the focus of research, and various projects on evaluating speech recognition in noisy environments have been organized.

*e-mail: cm013061@ed.ritsumei.ac.jp
†e-mail: nishiura@is.ritsumei.ac.jp

The SPeech recognition In Noisy Environment (SPINE) project in the USA established specific tasks including the recognition of a spontaneously spoken English dialog between an operator and a soldier in noisy environments (SPINE1, 2 [1]). The European Telecommunications Standards Institute (ETSI) also developed frameworks for evaluating speech recognition in noisy environments, which were collectively called Aurora. ETSI had distributed Aurora 2 [2], a connected digit-recognition task with various additive noises, Aurora 3, an in-car connected digit-recognition task, Aurora 4 [3], a continuous noisy speech-recognition task, and Aurora 5 [4], a noisy, simulated hands-free and cellular network transmission speech-recognition task.

We, the Working group [5] of the Information Processing Society in Japan (IPSJ), have worked on methodologies and frameworks for evaluating Japanese noisy speech recognition since October 2001. We first conformed to the ETSI Aurora 2 task settings because of their simplicity and generality, and released the Corpus and Environment for Noisy Speech RECognition 1 (CENSREC-1, which was formerly called AURORA-2J) [6], which included a database and evaluation tools. After that, we released CENSREC-2 [7] (in-car recognition of connected digits), CENSREC-3 [8] (in-car isolated word recognition), and CENSREC-1-C [9] (voice-activity detection under noisy conditions), with original evolutions. Thus far, we have developed frameworks for evaluating the performance of additive noisy speech recognition. However, in noisy speech recognition, speech-recognition performance is degraded not only by additive noise but also by multiplicative noise under distant-talking speech conditions. Speech-recognition methods against complex distortion, including additive noise, convolutional distortion, and also individual differences (e.g., [10,11]), have previously been actively pursued. However, many researchers have recently returned to the deep analysis of distorted data to investigate the mechanism responsible for individual distortions, and tried to address these. Thus, we released a new evaluation framework, which includes a database and evaluation tools, called CENSREC-4, which is a framework focusing on the evaluation of distant-talking speech in reverberant environments [12]. This evaluation framework has two main features. First, it includes both real reverberant speech and simulated reverberant speech (with convoluting impulse responses) in the same environment. Second, it includes various reverberant environments. We hope that it will be widely used to enable the development and comparison of new algorithms for the recognition of speech in reverberant environments, and will eventually lead to techniques that effectively deliver good performance under these conditions. Moreover, the database may also be used to investigate techniques of estimating the performance of speech recognition in different reverberant environments.

In this paper, we first introduce a framework including a database and evaluation tools of CENSREC-4, which is an evaluation framework for distant-talking speech under hands-free conditions. We then evaluate improvements in recognition performance with Cepstral Mean Normalization (CMN) [13] for CENSREC-4 data sets, and estimate the reverberant speech recognition performance of CENSREC-4 data sets with reverberant criteria, RSR-D_n [14].

2. DATA SETS OF CENSREC-4

We released a new evaluation framework, including a database and evaluation tools, called CENSREC-4, which is a framework for evaluating distant-talking speech under various reverberant environments. The data it contains are connected digit utterances, the same 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 noise in addition to reverberations. This evaluation framework has two main features.

- It includes both real reverberant speech and simulated reverberant speech (with convoluting impulse responses) in the same environment.
- It includes various reverberant environments.

2.1. Basic Data Sets

The basic data sets were used for the environment to prepare the room impulse response-convolved speech data.

2.1.1. Room impulse response data

Many room impulse responses were measured in real environments so that these could be used to simulate speech recorded in various environments by convolving them with clean speech signals and room impulse responses. The impulse responses were measured by the time stretched pulse (TSP) method [15]. The TSP length was 131,072 points and there were 16 synchronous additions.

The impulse responses were normalized at 0.5 with an absolute value for the maximum amplitude. CENSREC-4 had impulse responses recorded in eight kinds of environments: an office, an elevator hall (the waiting area in front of an elevator), a car, a living room, a lounge, a Japanese-style room (a room with a tatami floor), a meeting room, and a Japanese-style bathroom (a prefabricated bath). Figure 1 gives the impulse responses recorded in these eight kinds of environments. We measured the impulse responses for the environments using the equipment and setup listed in Table 1. Figure 2 presents the microphone settings in all environments except those in the car and the Japanese-style bathroom.

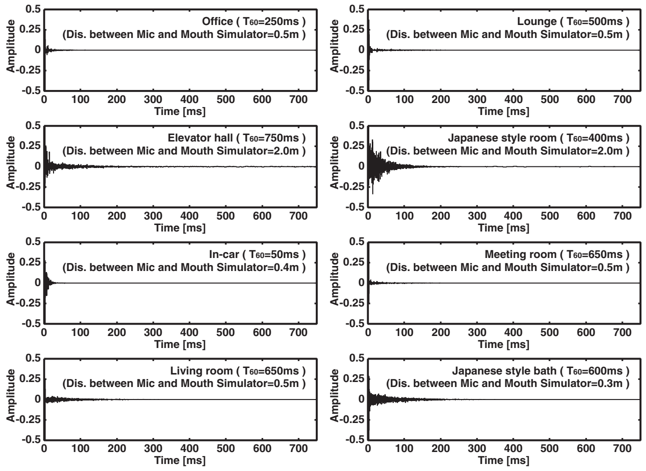

Fig. 1 Impulse responses in eight environments with CENSREC-4.

Table 1 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
Mouth simulator	B&K, Type 4128
Speaker amplifier	YAMAHA, P4050
Sampling frequency	48 kHz (downsampled to 16 kHz before convolving)
Quantization	16 bits

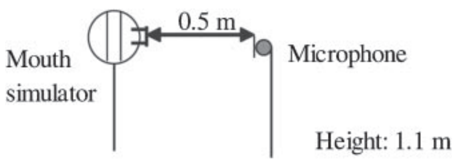

Fig. 2 Recording setup for impulse responses in all environments except those in car and in Japanese-style bathroom.

We positioned the microphone near the centers of the spaces in all environments, except in the car and in the Japanese-style bathroom. For the car environment, we used a medium-sized sedan and positioned the mouth simulator on the drivers seat and the microphone on the sun visor for the environment inside the car. The mouth simulator and microphone were about 0.4 m apart. We positioned the microphone on a coffee table in the lounge environment. For the bathroom environment, we positioned the mouth simulator over the bathtub, which was filled with cold water, and placed the microphone on a sidewall in the Japanese-style bath environment. The mouth simulator and the microphone were about 0.3 m apart. Table 2 lists the recording conditions, including the size of the spaces, the distance between the microphone and the mouth simulator, the reverberation time (T_{60}), the temperature, the humidity, and the average ambient-noise level in each recording environment. In Table 2, the reverberation time (T_{60}) is given with a resolution of 0.05 s, and the ambient-noise level is given with a resolution of 0.5 dB.

2.1.2. Simulated data (Testset A/B)

We simulated reverberant speech by convolving the impulse responses to clean speech. We used the clean speech from CENSREC-1 (the sampling frequency was 16 kHz for CENSREC-4, but 8 kHz for CENSREC-1). The recording conditions are listed in Table 3. The other details on the recording conditions, utterances, and speaking styles were the same as those for CENSREC-1. The vocabulary in the simulated data included in CENSREC-4 consisted of eleven Japanese numbers the same as in CENSREC-1: "ichi (1)," "ni (2)," "san (3)," "yon (4)," "go (5)," "roku (6)," "nana (7)," "hachi (8)," "kyu (9)," "zero (0)," and "maru (0)." The recording was conducted in a soundproof booth.

The training and testing data were prepared in the same way as those for CENSREC-1. The testing data were divided into two sets: Testset A (in an office, in an elevator hall, in a car, and in a living room) and Testset B (in a lounge, in a Japanese-style room, in a meeting room, and in a Japanese-style bathroom). There were a total of 4,004 utterances by 104 speakers (52 females and 52 males). These utterances for Testsets A and B were divided into four groups corresponding to the reverberant conditions.

Table 2 Room size, distance between microphone and mouth simulator (MS), reverberation time, ambient-noise level, humidity, and temperature in recording.

Room	Test set	Room size	Dis. between Mic. and MS	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 s	30°C	40%	36.5 dB
Elevator hall	A	11.5 × 6.5 m	2.0 m	0.75 s	30°C	50%	39.0 dB
In-car	$A/C/D$	Middle-sized sedan	0.4 m	0.05 s	29°C	44%	32.0 dB
Living room	A	7.0 × 3.0 m	0.5 m	0.65 s	30°C	54%	34.0 dB
Lounge	$B/C/D$	11.5 × 27.0 m	0.5 m	0.50 s	27°C	50%	52.5 dB
Japanese style room	B	3.5 × 2.5 m	2.0 m	0.40 s	30°C	54%	30.0 dB
Meeting room	$B/C/D$	7.0 × 8.5 m	0.5 m	0.65 s	27°C	52%	48.5 dB
Japanese style bath	B	1.5 × 1.0 m	0.3 m	0.60 s	31°C	62%	29.5 dB

Table 3 Recording conditions of clean speech for simulated data.

Headset Microphone	SENNHEISER HMD25
Sampling frequency	16 kHz
Quantization	16 bits
Format	Little-endian

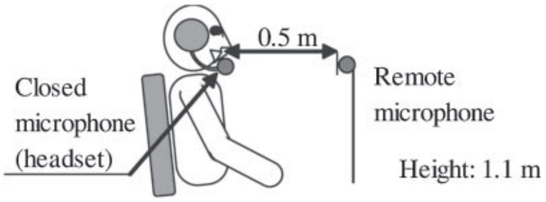

Fig. 3 Recording setup for real speech data in all environments.

Thus, each reverberant condition included 1,001 utterances. The noise in Testset *A* for CENSREC-1 was used for both the test set and the training set (called *known noises*), but that in Testset *B* was only used for the training set (*unknown noises*). Similar to this, the CENSREC-4 basic data sets also had two types of test sets: Testset *A* with *known reverberant environments* and *B* with *unknown reverberant environments*. Two sets of training data were prepared, i.e., clean training and multicondition training. There were a total of 8,440 utterances by 110 speakers (55 females and 55 males). For the multicondition training data, four kinds of reverberation (in an office, in an elevator hall, in a car, and in a living room) were convolved with clean speech. Thus, each reverberant condition included 2,110 utterances.

2.2. Extra Data Sets

Extra data sets consist of simulated and recorded data. They are affected by both additive and multiplicative noise. These data are different from those in the main evaluation environments for reverberant-speech recognition. Thus, we only provide testing/training data as extra data sets and do not provide an evaluation framework with these in the present assessments.

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

We 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 were called Testset C and consisted of four environments: two from Testset A (in an office and in a car) and two from Testset *B* (in a lounge and in a meeting room). In all four environments, we recorded the background noise for about 120 s. The first half of the recorded noise data was used to prepare the testing data, and the second half was used to prepare the training data.

There was a total of 4,004 utterances by 104 speakers (52 females and 52 males) for the testing data, which is identical to those in Testsets *A* and *B*. To prepare Testset *C*, these utterances were divided into four groups, and four kinds of reverberations (in an office, in a car, in a lounge, and in a meeting room) were convolved. Then background noise was added to the reverberant speech at ∞ dB, 20 dB, 10 dB, and 5 dB of the signal-to-noise ratio (SNR). However, when the reverberant and noisy conditions were the same, the utterance content was also the same, regardless of SNR. Thus 1,001 utterances were included for each reverberant condition.

For the training data, there were a total of 6,752 utterances by 88 speakers (44 females and 44 males). To prepare extra training data, these utterances were convolved as four kinds of reverberations (in an office, in an elevator hall, in a car, and in a living room), and background noise was added to the reverberant speech at ∞ dB, 20 dB, 10 dB, and 5 dB of SNR. Thus, the extra training data included 422 utterances for each reverberant condition and SNR. In addition, clean data were prepared as optional data comprising a total of 1,688 utterances by 22 speakers (11 females and 11 males). The data are different from the training data mentioned above and are intended for use in adaptive training.

2.2.2. Real recorded data in real environments (Testset *D*)

We recorded real data with two microphones (close and remote) under the conditions listed in Table 1. We used human speakers instead of a mouth simulator. This data set, called Testset *D*, was recorded in 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 for Testsets *A* and *B*. Figure 3 outlines the recording setup. The recorded speech by each speaker consisted of two major parts: testing data (49 or 50 utterances) and training data for adaptation (11 utterances). Testset *D* had 2,536 utterances (2,536 files).

3. BASELINE SCRIPTS AND EVALUATION OF CENSREC-4

3.1. Reference Baseline Scripts

We produced CENSREC-4 baseline scripts on the basis of CENSREC-1 baseline scripts to carry out HMM training and recognition experiments by using HTK [16] in the same way as had been done in CENSREC-1. They were only provided for the basic data sets, as previously described. As a result of various experiments (with various HMM topologies and various feature vectors) and discussions, we specified six conditions for producing baseline scripts.

- The acoustic model set consisted of 18 phoneme models (/a/, /i/, /u/, /u:/, /e/, /o/, /N/, /ch/, /g/,

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

Clean training (%STRING)					
A					
	Office 0.25 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	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 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	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 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	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 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	Average
w/o	99.5	99.4	99.5	99.4	99.4
w	74.0	89.5	89.8	78.0	82.8

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

Multicondition training (%Acc)					
A					
	Office 0.25 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	Average
w	94.4	90.6	95.0	91.6	92.9
B					
	Lounge 0.50 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	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			
Multicondition training	w			

%Acc				
		A	B	Overall
Clean training	w/o			
	w			
Multicondition training	w			

Relative performance (%STRING)				
		A	B	Overall
Clean training	w/o			
	w			
Multicondition training	w			

Relative performance (%Acc)				
		A	B	Overall
Clean training	w/o			
	w			
Multicondition training	w			

/h/, /k/, /ky/, /m/, /n/, /r/, /s/, /y/, /z/), silence ('sil'), and a short pause ('sp').

- Each phoneme model and 'sil' had five states (three emitting states), and 'sp' had three states (one emitting state). The output distribution of 'sp' was the same as that of the center state of 'sil.'
- Each state of the phoneme models had 20 Gaussian mixture pdfs, and 'sil' or 'sp' had 36 Gaussian mixtures.
- The feature parameter of the baseline system had 39-dimensional feature vectors that consisted of 12 MFCCs, 12 ΔMFCCs, 12 $\Delta\Delta$MFCCs, log power, Δ power, and $\Delta\Delta$ power, which were calculated using the HCopy of HTK. The analysis conditions were pre-emphasis $(1 - 0.97z^{-1})$, a hamming window, a 25-ms frame length, and a 10-ms frame shift.
- Grammar-based connected digit recognition by the HVite of HTK was used for the recognition experiments. Figure 4 shows the recognition grammar, where '|' denotes alternatives, '⟨ ⟩' denotes one or more repetitions, and '[ ]' encloses options. This grammar generates arbitrary repetitions of digits optionally followed by short pauses, and terminal silences are also allowed.

```
$digit = ichi | ni | san | yon |
         go | roku | nana | hachi |
         kyu | zero | maru ;

(
  [sil] < $digit [sp] > [sil]
)
```

Fig. 4 Recognition grammar.

- 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 the recognition experiments.

3.2. Performance of Reference Baseline

Table 4 lists the CENSREC-4 baseline performance for the basic data sets. The upper half has the clean training results and the lower half has the multicondition training results. The right side shows the accuracy of single-digit-level performance, and the left side shows the string-level correct rate, obtained by the connected digit recognition. The "w/o" in Tables 4 and 5 (explained below) indicates the recognition results for the clean speech data (without

Table 6 Recognition performance with CMN for basic data sets.

Clean training (%STRING)					
A					
	Office 0.25 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	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 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	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 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	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 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	Average
w/o	99.42	99.43	99.67	99.63	99.5
w	87.32	92.20	93.25	81.73	88.6

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

Multicondition training (%Acc)					
A					
	Office 0.25 s	Elevator hall 0.75 s, 2m	In-car 0.05 s	Living room 0.65 s	Average
w	92.78	91.90	92.54	90.00	91.8
B					
	Lounge 0.50 s	Japanese room 0.40 s, 2m	Meeting room 0.65 s	Japanese bath 0.60 s	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
Multicondition 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
Multicondition 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%
Multicondition 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%
Multicondition training	w	-20.3%	3.5%	-8.4%

convolving impulse responses), and "w" means the recognition results for the reverberant speech data (with convolving impulse responses).

In Table 4, we can see a tendency that the longer the reverberation time, the worse the recognition performance, since no dereverberating process was used in the CENSREC-4 baseline. However, reverberation time cannot completely explain the degradation in recognition performance. For example, the performance in the living room and meeting room are very different even if the reverberation time is the same. This implies that there are more complex factors involved, which should be considered to address the reverberations. These results were provided on a Microsoft Excel spreadsheet to summarize the tables for evaluating the results. The summarized tables of the recognition performance of basic data sets in CENSREC-4 are listed in Table 5, and were achieved by automatically calculating the relative performance with the baseline by inputting the results into the spreadsheets. Published summary tables can easily be compared with other recognition performance results.

3.3. Evaluation Experiment with Cepstral Mean Normalization

Cepstral Mean Normalization (CMN) [13] is a traditional dereverberating process that uses technology and it 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 understand the difficulties involved with basic data sets, we evaluated improvements in recognition performance with CMN for basic data sets. Table 6 lists the recognition performance with CMN for basic data sets, and Table 7 lists summarized tables of the recognition performance with CMN for basic data sets.

The results in Table 7 indicate relative performance was improved by about 15 to 25% in clean training but was degraded by about 7% in multicondition training. CMN made it difficult to achieve a sufficient improvement in recognition performance because CMN was not effective under longer reverberant conditions. Thus, conventional framewise dereverberation methods for speech recognition could not provide adequate performance. This database contained very challenging data and we hope to develop a new dereverberating technology with this.

3.4. Variability of Performance for Reverberant Speech Recognition with CENSREC-4

In the previous section, we explained that it is difficult to improve the recognition performance of CENSREC-4 data sets with conventional methods such as CMN. In this section, we focus on criteria to estimate the difficulty of reverberant speech recognition, and also explain why it is difficult to estimate the recognition performance of CENSREC-4 data sets. We have already investigated early and late reflections on distant-talking speech recognition with the aim of defining suitable reverberation criteria [17]. We then designed reverberation criteria RSR-D_n (Reverberant Speech Recognition criteria with D_n) [14] to estimate the reverberant speech recognition performance. Thus, we try to estimate the difficulty of reverberant speech recognition of CENSREC-4 and evaluate how many variable reverberant impulse responses CENSREC-4 contains on the basis of RSR-D_n.

3.4.1. Performance estimation based on reverberation criterion RSR-D_n

We designed the reverberation criterion RSR-D_n using the D value based on the ISO3382 acoustic parameter [18] in order to estimate the difficulty of reverberant speech recognition. The D value expresses the clarity of acoustics and is derived from

$$D_n = \int_0^n h^2(t)dt \Bigg/ \int_0^\infty h^2(t)dt, \qquad (1)$$

where $h(t)$ is the impulse response and n is the border time between early-and-late-arriving energies. The D value improves under the condition of higher direct and early reflections and degrades under the condition of higher late reverberations. In previous research, RSR-D_{20}L (RSR-D_{20} with Linear regression function) and RSR-D_{20}Q (RSR-D_{20} with Quadratic regression function) provided much better estimation performance [14]. Thus, we estimate the recognition performance of CENSREC-4 with RSR-D_{20}L in this study.

3.4.2. Estimating performance of reverberant speech recognition

• Experimental conditions

We used RSR-D_{20}L to estimate the reverberant speech recognition performance in five environments of CENSREC-4. We first measured 312 impulse responses to design RSR-D_n in the three training environments (Env. A with 72 RIRs (Room Impulse Responses), Env. B with 120 RIRs, and Env. C with 120 RIRs). On the basis of measured impulse responses, we next derived D_{20} and the performance of speech recognition. Next, we calculated the linear regression curve as the reverberation criterion on the basis of numerous impulse responses and the reverberant speech convolving them. Figure 5 plots the results. We finally attempted to estimate the reverberant speech recognition performance for five test environments in CENSREC-4 (office, elevator hall, living room, Japanese tatami room, and meeting room) on the basis of the designed RSR-D_{20}L in the same or closest reverberation time.

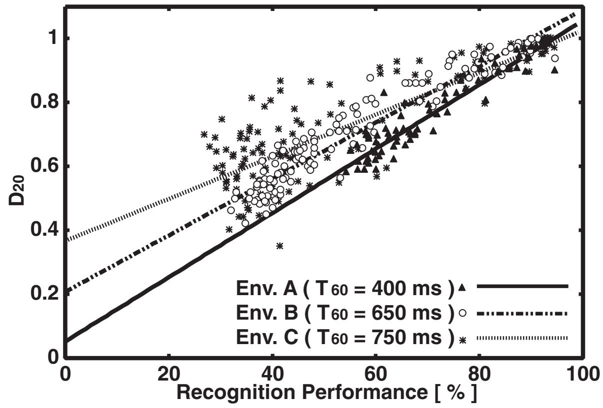

Fig. 5 RSR-D_{20}L (Regression analysis with D_{20} and reverberant speech recognition performance).

Table 8 Actual and estimated recognition performance in five test environments with CENSREC-4.

Test env.	T_{60}	D_{20}	Actual rec.	Est. rec. w RSR-D_{20} (Env.)
Office	0.25 s	0.98	93.4%	92.5%(A)
Elevator hall	0.75 s	0.72	30.7%	53.9%(C)
Living room	0.65 s	0.75	65.3%	69.0%(B)
Japanese room	0.40 s	0.65	54.3%	59.2%(A)
Meeting room	0.65 s	0.96	74.1%	84.7%(B)

Table 9 Errors in estimated recognition performance in five test environments with CENSREC-4.

Test env.	T_{60}	Error w RSR-D_{20}
Office	0.25 s	0.9%
Elevator hall	0.75 s	23.2%
Living room	0.65 s	3.7%
Japanese room	0.40 s	4.9%
Meeting room	0.65 s	10.6%
Average	0.54 s	**8.66%**

• Experimental results

Table 8 lists the results, where "Est. rec. w RSR-D_{20} (Env.)" means the performance estimated with RSR-D_{20} in Env. A, B, and C. In this experiment, RSR-D_{20} in Env. A, B, and C were selected as the reverberation time environments closest to the test environment. Table 9 lists the errors in the recognition performance estimated with RSR-D_{20}. As a result, an average estimation error of less than 10% was achieved in five environments of CENSREC-4 data sets. Estimation error of about 20%, however, was achieved in a longer reverberant environment, Elevator hall. Therefore, we confirmed that CENSREC-4 has very challenging and variable reverberant features which make it difficult to estimate the performance of recognition performance in a particularly heavily reverberant environment.

4. CONCLUSION

We developed a new CENSREC-4, which is an evaluation framework for distant-talking speech under reverberation environments. It is an effective database suitable for evaluating new methods of dereverberation and evaluating performance because the traditional dereverberation process and performance estimation criteria are ineffective in sufficiently improving and estimating recognition performance. The framework was released in March 2008, and not only performance evaluations but also many other studies are being conducted using it throughout Japan. We intend to evaluate extra data sets in the near future. We fervently hope that CENSREC-4 is a first step in supporting the research on effective algorithms for recognition in reverberation.

5. DISTRIBUTION FOR CENSREC-4

The CENSREC-4 is distributed by NII-Speech Resources Consortium (NII-SRC), Japan. The latest information is stored at the following URL.

http://research.nii.ac.jp/src/eng/index.html

ACKNOWLEDGMENTS

We wish to thank the members of the Speech Resources Consortium at the National Institute of Informatics (NII-SRC), Japan, for their generous assistance with our activities. The present study was conducted using the CENSREC-4 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. In addition, this work was also partly supported by Grants-in-Aid for Scientific Research funded by Japan's Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

[1] http://elazar.itd.nrl.navy.mil/spine/
[2] H. G. Hirsh and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000* (2000).
[3] Aurora document no. AU/345/01, "Large vocabulary evaluation of front-ends-baseline recognition system description," Mississippi State University (2001).
[4] H. G. Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a handsfree speech input in noisy environments," Online: http://aurora.hs-niederrhein.de, data available from ELDA: http://www.elda.org, 2007.
[5] AURORA-J/CENSREC Web site: http://sp.shinshu-u.ac.jp/CENSREC/index.html.en
[6] 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, "AURORA-2J, An evaluation framework for japanese noisy speech recognition," *IEICE Trans. Inf. Syst.*, **E88-D**, 535–544 (2005).
[7] S. Nakamura, M. Fujimoto and K. Takeda, "CENSREC2: Corpus and evaluation environments for in car continuous digit speech recognition," *Proc. ICSLP'06*, pp. 2330–2333 (2006).
[8] M. Fujimoto, K. Takeda and S. Nakamura, "CENSREC-3: An evaluation framework for japanese speech recognition in real driving-car environments," *IEICE Trans. Inf. Syst.*, **E89-D**, 2783–2793 (2006).
[9] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda and S. Nakamura, "CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments," *Acoust. Sci. & Tech.*, **30**, 363–371 (2009).
[10] T. Takiguchi, S. Nakamura and K. Shikano, "HMM-separation-based speech recognition for a distant moving speaker," *IEEE Trans. Speech Audio Process.*, **9**, 127–140 (2001).
[11] M. Shozakai, S. Nakamura and K. Shikano, "A speech enhancement approach E-CMN/CSS for speech recognition incar environments," *Proc. ASRU1997*, pp. 450–457 (1997).
[12] T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda and S. Nakamura, "Evaluation framework for distant-talking speech recognition under reverberant environments: Newest part of the CENSREC series," *Proc. LREC2008*, pp. 1828–1834 (2008).
[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, **29**, 254–272 (1981).
[14] T. Fukumori, M. Morise and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria RSR-Dn with acoustic parameters," *Proc. INTERSPEECH 2010*, pp. 562–565 (2010).
[15] Y. Suzuki, F. Asano, H. Y. Kim and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am.*, **97**, 1119–1123 (1995).
[16] http://htk.eng.cam.ac.uk/
[17] T. Nishiura, Y. Hirano, Y. Denda and M. Nakayama, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria," *INTERSPEECH 2007*, pp. 1052–1055 (2007).
[18] ISO3382: "Acoustics measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization (1997).

Takahiro Fukumori received his B.E. degree from Ritsumeikan University in 2010. He is currently a master's student at Ritsumeikan University. His current research interests include criteria design for reverberant speech recognition. He is a member of ASJ, IPSJ and IEICE.

Takanobu Nishiura received his B.E. degree from the Nara National College of Technology in 1997 and M.E. and Ph.D degrees from the Nara Institute of Science and Technology (NAIST) in 1999 and 2001, respectively. From 2001 to 2004, he was a research associate at Wakayama University. He is currently an associate professor at Ritsumeikan University. His current research interests include acoustic sound signal sensor using a microphone array. He received the TELECOM System Technology Award for Students from the Telecommunications Advancement Foundation (TAF) in 2000, and the Best Paper Award from the Virtual Reality Society of Japan (VRSJ) in 2009. He is a member of ISCA, IEICE, ASJ and VRSJ.

Masato Nakayama received his B.E. degree from Kinki University in 2001 and M.E. degree from Wakayama University in 2003. He

completed a graduate course at the Ritsumeikan University in 2008. From 2004 to 2007, he was a research associate at Ritsumeikan University. He was a part-time lecturer at Doshisha Women's College of Liberal Arts in 2007. He is currently a part-time lecturer at Kinki University, and is a postgraduate of Ritsumeikan University. His current research interests include acoustic sound signal sensors using a microphone array. He is a member of the ASJ and ISCA.

Yuki Denda received his B.E. and M.E. degrees in systems engineering from Wakayama University in 2003 and 2005, respectively, and Ph.D degree from Ritsumeikan University in 2008. From 2005 to 2006, he was a research associate at Ritsumeikan University. He is currently a researcher at Murata Machinery, Ltd. He received the TELECOM System Technology Award for Students from the Telecommunications Advancement Foundation in 2007. His current research interests include acoustic signal processing, microphone array, and robust speech recognition. He is a member of IEICE and ASJ.

Norihide Kitaoka received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Engineering degree from Toyohashi University of Technology in 2000. He joined the DENSO CORPORATION, Japan in 1994. He joined the Department of Information and Computer Sciences at Toyohashi University of Technology as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. Since 2006 he has been an associate professor in the Department of Media Science, Graduate School of Information Science, Nagoya University. He was a visiting associate professor of Nanyang Technological University, Singapore in 2009. His research interests include speech processing, speech recognition, and spoken dialog. He is a member of ISCA, ASJ, IEICE, IPSJ, and JSAI.

Takeshi Yamada received his B. Eng. degree from Osaka City University in 1994, and M. Eng. and Dr. Eng. degrees from Nara Institute of Science and Technology in 1996 and 1999, respectively. He is currently with the Graduate School of Systems and Information Engineering, University of Tsukuba, where he is an associate professor. His research interests include speech recognition, sound scene understanding, multichannel signal processing, and media quality assessment. He is a member of IEEE, IEICE, IPSJ, and ASJ.

Kazumasa Yamamoto received his B.E., M.E., and Dr. Eng. degrees in information and computer sciences from Toyohashi University of Technology, Toyohashi, Japan, in 1995, 1997, and 2000, respectively. From 2000 to 2007, he was a research associate in the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan. Since 2007, he has been an assistant professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan. His current research interests include speech recognition and privacy protection of speech signal. He is a member of the IEICE and the IPSJ.

Satoru Tsuge received his B.E., M.E., and Dr. Eng. degrees from the University of Tokushima in 1996, 1998, and 2001, respectively. From 1997 to 1999, he was an intern researcher at ATR Interpreting Telecommunications Research Laboratories, Kyoto. He joined the Faculty of Engineering, the University of Tokushima in 2000 and was Lecturer from 2006 to 2010. Since 2010, he has been an associate professor in the Department of Information Systems, School of Informatics, Daido University. His current research interests include speech recognition, speaker recognition, and information retrieval. He is a member of IPSJ and ASJ.

Masakiyo Fujimoto received his B.E., M.E., and Doctor of Engineering degrees from Ryukoku University in 1997, 2001, and 2005, respectively. From 2004–2006, he worked at ATR Spoken Language Communication Research Laboratories. He is currently a researcher at NTT Communication Science Laboratories. He received the Awaya Award from ASJ in 2003. His current research interests include noise-robust speech recognition. He is a member of IEEE, ISCA, IEICE, IPSJ, and ASJ.

Tetsuya Takiguchi received a B.S. degree in applied mathematics from Okayama University of Science, Okayama in 1994, and M.E. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Nara in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa. He is currently a Lecturer with Kobe University. His research interests include robust speech recognition, auditory scene analysis, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of IEEE, IEICE, IPSJ, and ASJ.

Chiyomi Miyajima received her B.E. degree in computer science and M.E. and Dr. Eng. degrees in electrical and computer engineering from the Nagoya Institute of Technology in 1996, 1998, and 2001, respectively. From 2001 to 2003, she was a research associate in the Department of Computer Science, Nagoya Institute of Technology. Currently she is an assistant professor at the Graduate School of Information Science, Nagoya University. Her research interests include multimodal speech recognition and modeling of driving behavior. She is a member of IEICE, IEEE and ASJ.

Satoshi Tamura received his M.S. and Ph.D. degrees in information science and engineering from the Tokyo Institute of Technology (Tokyo Tech) in 2002 and 2005, respectively. He became a research associate at the Department of Computer Science, Gifu University, in 2005 and has been an assistant professor at Gifu University since 2007. His research interests are speech information processing, such as multimodal (audio-visual) speech recognition, robust speech recognition, and applications of speech recognition to real systems. He is also interested in audiovisual information processing. He is a member of ISCA, IPSJ, ASJ, and JSAI.

Tetsuji Ogawa received his B.S., M.S., and Ph.D. degrees in electric, electronics, and computer engineering from Waseda University, Tokyo, Japan, in 2000, 2002, and 2005, respectively. He has been a research associate (2004–2007) and a visiting lecturer (2007) at Waseda University. He is currently an assistant professor in Waseda Institute for Advanced Study. His research interests include stochastic modeling for pattern recognition, speech enhancement and speech recognition. He is a member of IEICE, IPSJ, and ASJ.

Shigeki Matsuda received his B.S. degree from the Department of Information Science, Teikyo University, in 1997, completed his doctoral program at the Japan Advanced Institute of Science and Technology in 2003, and then joined ATR Spoken Language Communication Research Laboratories as a researcher. He holds a doctoral degree in information science. He is engaged in research on speech recognition, and is a member of ASJ and IPSJ.

Shingo Kuroiwa received his B.E., M.E., and Dr. Eng. degrees in electrocommunications from the University of Electro-Communications, Tokyo in 1986, 1988, and 2000, respectively. From 1988 to 2001 he was a researcher at the KDD R & D Laboratories. From 2001 to 2007 he was an associate professor in the Faculty of Engineering, the University of Tokushima. Currently, he is a professor of the Graduate School of Advanced Integration Science, Chiba University. His current research interests include speech recognition, speaker recognition, natural language processing, and information retrieval. He is a member of IEICE, IPSJ, ASJ, and JSAI.

Kazuya Takeda received his B.S. degree, M.S. degree, and Dr. of Engineering degree from Nagoya University in 1983, 1985, and 1994, respectively. In 1986, he joined ATR (Advanced Telecommunication Research Laboratories), where he was involved in the two major projects of speech database construction and speech synthesis system development. In 1989, he moved to KDD R & D Laboratories and participated in a project for constructing a voice-activated telephone extension system. He joined the Graduate School of Nagoya University in 1995. Since 2003, he has been a professor at the Graduate School of Information Science at Nagoya University. He is a member of IEICE, IEEE, and ASJ.

Satoshi Nakamura received his B.S. in Electronic Engineering from the Kyoto Institute of Technology in 1981 and his Ph.D. in Information Science from Kyoto University in 1992. From 1981 to 1993, he worked for Sharp's Central Research Laboratory in Nara. From 1986 to 1989, he worked for ATR Interpreting Telephony Research Laboratories. From 1994 to 2000, he was an associate professor at the Graduate School of Information Science at the Nara Institute of Science and Technology. In 1996, he was a visiting research professor at the CAIP Center at Rutgers University in New Jersey. From 2000 to 2008, he worked as a director of ATR Spoken Language Communication Research Laboratories. He is currently an executive researcher and a director of the MASTAR project at the National Institute of Information and Communications Technology. He has also served as an honorary professor of University Karlsruhe, Germany since 2004. He received the Awaya Award from the Acoustical Society of Japan in 1992, the Interaction 2001 Best Paper Award in 2001, Yamashita Research Award from the Information Processing Society of Japan in 2001, Telecom System Award, AAMT Nagao Award and Docomo Mobile Science Award in 2007, and IPSJ Kiyasu Achievement Award in 2008. He is a member of IEEE, IPSJ, ASJ and IEICE.