interspeech 2008 incorporating SST 08

CENSREC-4: Development of Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments

Masato Nakayama[1], *Takanobu Nishiura*[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], *Tetsuji Ogawa*[9], *Shigeki Matsuda*[11], *Shingo Kuroiwa*[10], *Kazuya Takeda*[2], *and Satoshi Nakamura*[11]

[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]Waseda University, [10]Chiba University, [11]ATR/NiCT

[1]{mnv28004@fc, nishiura@is, gr021052@se}.ritsumei.ac.jp, [2]{kitaoka, miyajima, 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]ogawa@pcl.cs.waseda.ac.jp, [10]kuroiwa@faculty.chiba-u.jp, [11]{shigeki.matsuda, satoshi.nakamura}@atr.jp

Abstract

In this paper, we newly introduce a collection of databases and evaluation tools called 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 in various 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 the evaluation environment for the room impulse response-convolved speech data. The extra data sets consist of simulated and recorded data. An evaluation framework is only provided for the basic data sets as evaluation tools. The results of evaluation experiments proved that CENSREC-4 is an effective database for evaluating the new dereverberation method because the traditional dereverberation process had difficulty sufficiently improving the recognition performance.

Index Terms: Various environments, Impulse response, Convolution, Real recorded data, Evaluation framework

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 is being focused on. Since October 2001, we from the working group of the Information Processing Society in Japan [1] 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) [2], CENSREC-2 [3], CENSREC-3 [4], and CENSREC-1-C [5].

In this paper, we newly introduce a framework including a database and evaluation tools called 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 record multiplicative noise to investigate reverberant speech recognition. CENSREC-4 also records the ambient noises in each environment.

2. CENSREC-4

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. These 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 the reverberations. An evaluation framework is only provided for the basic data sets as HTK-based HMM training and recognition scripts.

2.1. Basic data sets

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

2.1.1. 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 real environments. Impulse responses were measured using the time stretched pulse (TSP) method [6]. The TSP length was 131,072 points. The number of synchronous additions was 16. 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 prefabricated bath. We measured the room impulse responses based on the conditions shown in Table 1. Figure 1 shows the microphone settings for all environments except the in-car and the Japanese style bath. In all environments except the in-car and the Japanese style bath, we set the microphone near the center of the room.

Table 2 shows the room size, the distance between the mi-

Accepted after peer review of full paper

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

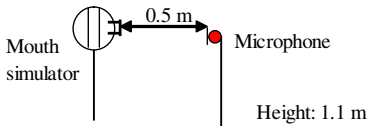

Figure 1: Recording setup for impulse responses

crophone and the loudspeaker (mouth simulator), the reverberation time, temperature, humidity, and the average ambient noise level in each recording room. In Table 2, reverberation time (T_{60}) is displayed with 0.05 sec resolution, and the ambient noise level is displayed with 0.5 dB resolution. The recording subjects in Table 2 were selected based on the variation of reverberation time and the needs of application.

2.1.2. 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 was used; the sampling frequency was 16 kHz for CENSREC-4, whereas it was 8 kHz for CENSREC-1. 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 a little-endian format.

The training and testing data, which were prepared in the same way as in CENSREC-1, were divided into 2 sets: 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). Two sets of training data were prepared: clean and multi-condition. Total utterances were 8,440 by 110 speakers (55 females and 55 males).

2.2. Extra data sets

The extra data sets consist of simulated and recorded data. They are affected by both the additive and multiplicative noise. These data digress from the main topic, as in 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.

2.2.1. 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

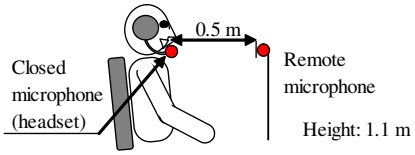

Figure 2: Recording setup for real data

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, four kinds of reverberations (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).

For the training data, total utterances were 6,752 by 88 speakers (44 females and 44 males). 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, which were not utilized as training data.

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

We recorded real data with two microphones (closed and remote) under the conditions shown in Table 1 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 2 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).

2.3. Reference baseline performance

Table 3 shows the CENSREC-4 baseline performance for the basic data sets. In Table 3, its upper half shows the clean training results, its lower half shows the multi-condition training results, its right half shows the digit accuracy, and its left half shows the string correct rate, which is defined as the correct recognition rate for all digits in each connected digit. In Table 3, "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). From Table 3, it can be seen that the longer the reverberation time is, the worse the recognition performance, since a dereverberation process was not used in the CENSREC-4 baseline.

This result is illustrated as a Microsoft Excel spreadsheet to obtain summary tables for evaluating the results. Table 5, which is one example of a summary table with advanced technology, is written with the same format as Table 3. Summary tables of the recognition performance are confirmable as Table 5, because the relative performance with the baseline is calculated automatically by inputting the results into spreadsheets. Published summary tables can be easily compared to other recognition performances.

Table 2: 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}]	Temperature	Humidity	Amb. noise level [dBA]
Office	$A/C/D$	9.0 × 6.0 m	0.5 m	0.25 sec	30℃	40%	36.5
Elevator hall	A	11.5 × 6.5 m	2.0 m	0.75 sec	30℃	50%	39.0
In-car	$A/C/D$	Middle-sized sedan	0.4 m	0.05 sec	29℃	44%	32.0
Living room	A	7.0 × 3.0 m	0.5 m	0.65 sec	30℃	54%	34.0
Lounge	$B/C/D$	11.5 × 27.0 m	0.5 m	0.50 sec	27℃	50%	52.5
Japanese style room	B	3.5 × 2.5 m	2.0 m	0.40 sec	30℃	54%	30.0
Meeting room	$B/C/D$	7.0 × 8.5 m	0.5 m	0.65 sec	27℃	52%	48.5
Japanese style bath	B	1.5 × 1.0 m	0.3 m	0.60 sec	31℃	62%	29.5

Table 3: 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

2.4. Evaluation experiment for extra data sets

We also evaluated the recognition performance for extra data sets. Table 4 shows the recognition performance for the extra data sets, which is written with the same format as Table 3. In Tables 4, "clean" shows the recognition result for the clean speech data, "inf" shows the recognition result for the reverberant speech data (with SNR ∞ dB), "SNR20" shows the recognition result for the reverberant and noisy speech with SNR 20 dB, "SNR10" shows the recognition result for the reverberant and noisy speech with SNR 10 dB, and "SNR5" shows the recognition result for the reverberant and noisy speech with SNR 5 dB. From Table 4, it can be seen that the lower the SNR is, the worse the recognition performance, since the dereverberation and noise reduction processes were not used in this evaluation experiment.

3. Evaluation experiment with advanced technology

Cepstral Mean Normalization (CMN) [7], a traditional dereverberation process with advanced technology, is a simple and effective way of normalizing the feature space and reducing channel distortion. It has, therefore, been adopted in many current systems. To appreciate the difficulties involved for basic data sets, we evaluated the recognition performance improvement with CMN for the basic data sets. Table 5 shows the recognition performance with CMN for basic data sets and is written with the same format as Table 3.

As a result of Table 5, relative performance was improved about 15 to 25% in clean training but was degraded about 7% in multi-condition training. Thus, CMN had difficulty sufficiently improving the recognition performance because it is ineffective under longer reverberant conditions.

Table 5: 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%

In addition, Spectral Subtraction (SS) [8] is one traditional additive noise reduction process with advanced technology. It is a simple and effective way of estimating average noise spectrum and subtracting additive noise. It has, therefore, been adopted in many current systems. To appreciate the difficulties involved for extra data sets, we evaluated the improvement of recognition performance with CMN and SS for the extra data sets. As a result of Table 6, the relative performance of Testset C improved about 15 to 30% for "clean" and "inf" in clean training and about 65 to 85% for "clean" and "inf" in multi-condition training. This result is not sufficient recognition performance improvement in clean training. The relative performance of Testset$C$ with "SNR20" improved about 60 to 75%. The relative performance of Testset$C$ with "SNR10" and "SNR5" improved an average of 5% in clean training and an average of 30% in multi-condition training. This result is not sufficient recognition performance improvement with "SNR10" and "SNR5."

The relative performance of Testset D with close microphones degraded about 0 to 20%, because recorded speech with close microphones is a high SNR condition that doesn't need a noise reduction process. The relative performance of Test-

Table 4: Evaluated recognition performance for extra data sets

Clean training (%STRING)					
C					
Rev/ Noise	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
clean	98.5	98.5	98.5	98.5	98.5
inf	93.1	86.1	43.9	74.1	74.3
SNR20	16.0	4.2	0.2	1.2	5.4
SNR10	0.1	0.2	0.4	0.1	0.2
SNR5	0.1	1.0	1.1	0.6	0.7
D					
Mic.	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
close	92.1	89.5	86.8	89.3	89.4
remote	66.9	62.7	20.9	67.8	54.6

Clean training (%Acc)					
C					
Rev/ Noise	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
clean	99.5	99.5	99.5	99.5	99.5
inf	97.5	95.6	74.0	89.8	89.2
SNR20	45.9	37.5	-5.8	23.5	25.3
SNR10	3.8	4.6	2.0	4.1	3.6
SNR5	3.8	6.9	5.7	4.9	5.3
D					
Mic.	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
close	98.1	97.3	96.5	97.3	97.3
remote	85.1	76.4	43.8	89.2	73.6

Multicondition training (%STRING)					
C					
Rev/ Noise	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
clean	50.1	52.3	50.1	52.3	51.2
inf	73.4	66.7	68.4	73.3	70.5
SNR20	88.2	90.0	24.9	67.1	67.6
SNR10	69.8	76.9	8.3	4.3	39.8
SNR5	44.1	56.9	4.6	0.3	26.5
D					
Mic.	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
close	61.9	60.5	68.4	68.2	64.7
remote	68.4	45.2	40.6	66.9	55.3

Multicondition training (%Acc)					
C					
Rev/ Noise	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
clean	81.5	83.2	81.5	83.2	82.3
inf	90.7	88.7	87.5	90.3	89.3
SNR20	95.1	96.4	53.8	86.2	82.9
SNR10	85.7	90.3	6.8	28.6	52.9
SNR5	67.8	77.3	-4.9	-1.3	34.7
D					
Mic.	Office 0.25 sec.	In-car 0.05 sec.	Lounge 0.50 sec.	Meeting room 0.65 sec.	Average
close	82.4	81.4	86.2	87.9	84.5
remote	88.4	59.4	70.1	86.3	76.0

Table 6: Relative recognition performance with CMN and SS for extra data sets

Relative performance (%STRING)			
C			
Clean training		Multicondition training	
Rev/Noise	Average	Rev/Noise	Average
clean	16.7%	clean	86.0%
inf	26.1%	inf	66.4%
SNR20	62.1%	SNR20	30.7%
SNR10	8.5%	SNR10	31.9%
SNR5	0.1%	SNR5	20.9%
D			
Clean training		Multicondition training	
Mic.	Average	Mic.	Average
close	-20.8%	close	1.8%
remote	42.2%	remote	6.5%

Relative performance (%Acc)			
C			
Clean training		Multicondition training	
Rev/Noise	Average	Rev/Noise	Average
clean	18.0%	clean	86.0%
inf	30.9%	inf	66.3%
SNR20	77.8%	SNR20	37.4%
SNR10	21.3%	SNR10	41.4%
SNR5	-9.5%	SNR5	32.4%
D			
Clean training		Multicondition training	
Mic.	Average	Mic.	Average
close	-20.5%	close	-15.5%
remote	63.8%	remote	3.5%

set D with remote microphones improved about 40 to 60% in clean training and about 5% in multi-condition training because Testset D was recorded in high SNR environments where a traditional noise reduction process can achieve sufficient performance.

Thus, CMN had difficulty sufficiently improving the recognition performance for basic data sets because it is not effective under longer reverberant conditions. Additionally, CMN and SS had difficulty sufficiently improving the recognition performance for the extra data sets because SS is not effective under time-varying noise conditions.

Therefore, we consider that the other conventional post-processes will also experience difficulty sufficiently improving recognition performance with CENSREC-4. This database includes challenging and variable data sets. We hope to develop new dereverberation technology that exceeds conventional post-processes 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 an effective database for evaluating the new dereverberation method because the traditional dereverberation process had difficulty sufficiently improving recognition performance. The framework was released in March 2008, and many studies are being conducted using it in Japan. We will provide baseline script for extra data sets in the near future. CENSREC-4 is being distributed by Speech Resources Consortium in the National Institute of Informatics (NII-SRC), Japan.[9]

5. Acknowledgements

This work was funded by The Ministry of Education, Culture, Sports, Science and Technology of Japan. The authors wish to thank the members of 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 Evaluation Working Group.

6. References

[1] AURORA-J/CENSREC Web site: http://sp.shinshu-u.ac.jp/CENSREC/index.html.en

[2] S. Nakamura et al. “AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition,” IEICE Transactions on Information and Systems, vol. E88-D, no. 3, pp. 535-544, 2005.

[3] S. Nakamura et al. “CENSREC-2: Corpus and Evaluation Environments for In-Car Continuous Digit Speech Recognition,” Proc. ICSLP’06, pp. 2330–2333, Sept. 2006.

[4] M. Fujimoto et al. “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, Nov. 2006.

[5] N. Kitaoka et al. “CENSREC-1-C: An Evaluation Framework for Voice Activity Detection under Noisy Environment,” ASRU 2007.

[6] Y. Suzuki et al. “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, 1995.

[7] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. Acoust. Speech Signal Process., vol.29, no. 2, pp. 254-272, 1981.

[8] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. ASSP, vol. ASSP-27, no. 2, pp. 133-120, Apr. 1979.

[9] NII-SRC Web site: http://research.nii.ac.jp/src/eng/index.html