CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition

Satoshi Nakamura [1,2], *Masakiyo Fujimoto* [3] *and Kazuya Takeda* [4]

[1] National Institute of Information and Communications Technology
[2] ATR Spoken Language Communication Research Laboratories
[3] NTT Communication Science Laboratories
[4] Graduate School of Information Science, Nagoya University
satoshi.nakamura@atr.jp, masakiyo@cslab.kecl.ntt.co.jp, kazuya.takeda@nagoya-u.jp

Abstract

This paper introduces a common database and an evaluation framework for connected digit speech recognition in real driving car environments, CENSREC-2, as an outcome of IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. Speech data of CENSREC-2 was collected using two microphones, a close-talking microphone and a hands-free microphone, under three car speeds and four car conditions. CENSREC-2 provides four evaluation environments which are designed using speech data collected in these car conditions.

Index Terms: noisy speech recognition, common evaluation framework, in-car speech corpus.

1. Introduction

Recently, the progress of speech recognition technology has been brought about by the advent of statistical approaches and large-scale corpora. Furthermore, it is also widely known that progress has been accelerated by the U.S. DARPA projects [1] initiated in the late '80s in terms of project participants competitively developing speech recognition systems for the same task, using the same training and test corpus.

However, current speech recognition performance must still be improved if the system is to be exposed to noisy environments, where speech recognition applications might be used in practice. Therefore, noise robustness is an emerging and crucial factor to be solved for speech recognition techniques.

With regard to the noise robustness problem, there have been two major evaluation projects, SPINE1, 2 [2] and AURORA [3]-[9]. The SPINE (SPeech recognition In Noisy Environments) project was organized by the U.S.'s DARPA, with SPINE1 in 2000 and SPINE2 in 2001. On the other hand, the AURORA was organized by the European Telecommunications Standards Institute (ETSI) [10] AURORA group. To date, AURORA2 [3](a connected digit corpus with additive noise), AURORA3 [4]-[7] (an in-car noisy digit corpus), and AURORA4 [8, 9] (a large-vocabulary continuous-speech recognition corpus with additive noise) have been distributed with HTK (HMM Took Kit) [11] scripts, which can be used to obtain baseline performance[12].

The authors voluntarily organized a special working group in October 2001 under the auspices of the Information Processing Society of Japan in order to assess speech recognition technology in noisy environments. The focus of the working group included the planning of comprehensive fundamental assessments of noisy speech recognition, standardized corpus collection, evaluation strategy developments, and distribution of standardized processing modules. As an outcome of working group, we have already produced the Japanese AURORA-2, AURORA-2J [13], which comprises the English digits translated into Japanese. We have also produced CENSREC-3 (Corpora and Environments for Noisy Speech RECognition) [14], our original evaluation framework CENSREC-3 is designed as the evaluation framework of isolated word recognition in real driving car environments. The main target application of CENSREC-3 is human voice (hands-free) control of car navigation systems. Thus, CENSREC-3 is designed as the evaluation framework that assumes speech-oriented man-machine communication in several car environments.

In this paper, we introduce here, CENSREC-2, a common database and an evaluation framework for connected digit speech recognition in real driving car environments. Speech data of CENSREC-2 was collected using two microphones, a close-talking microphone and a hands-free microphone, under carefully controlled 11 different driving conditions, i.e., combinations of three car speeds and four car conditions. CENSREC-2 provides four evaluation environments which are designed using speech data collected in these car conditions.

2. Data recording

2.1. Vocabulary

The speech recognition task of the CENSREC-2 database constitutes continuous digit recognition in real car driving environments. The vocabulary of CENSREC-2 consists of 11 digit models ("ichi," "ni," "san," "yon," "go," "roku," "nana," "hachi," "kyu," "zero," and "maru," respectively), a silence ("sil"), and a short pause ("sp"). The digit sequence of each utterance and the pronunciation of Japanese digits are the same as the AURORA-2J database [13].

2.2. Speech data recording

In-car speech data was collected in a specially equipped vehicle. Six microphones were mounted on the vehicle as shown in Fig. 1. Microphone no. 1 was a close-talking headset microphone, microphone nos. 3 and 4 were attached to the dashboard, and microphone nos. 5, 6, and 7 were fixed to the ceiling of the vehicle. The speech data recorded with the close-talking (CT) microphone (no. 1: SENNHEISER HMD410 with SONY ECM77B) and the

hands-free (HF) microphone attached to the ceiling of the driver's seat (no. 6: SONY ECM77B) are used for CENSREC-2 [15].

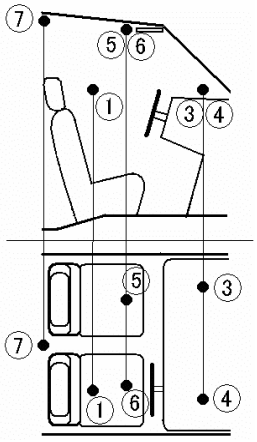

Figure 1: Microphone positions for data collection: Side view (top) and top view (bottom)

The recording conditions for the speech data are shown in Table 1. Speech data were recorded under 11 environmental conditions using combinations of three kinds of vehicle speeds (idling, low-speed driving on a city street, and high-speed driving on an expressway) and four kinds of in-car environments (normal, with air-conditioner (fan) on, with audio CD player on, and with windows open). The speech signals for both training and testing were sampled at 16 kHz, quantized into 16 bit integers, and saved in the little-endian format.

Table 1: Recording environments for in-car speech data

Car speed	In-car conditions
Idling (quiet)	Normal, Fan on, Audio on, Windows open
Low speed	Normal, Fan on, Audio on, Windows open
High speed	Normal, Fan on, Audio on

A total of 17,651 utterances spoken by 104 speakers (52 males and 52 females) were recorded with CT and HF microphones. The training and testing data comprise 14,687 utterances spoken by 73 speakers (33 males and 40 females) with CT (7,492 utterances) and HF (7,195 utterances) microphones, and 2,964 utterances spoken by 31 speakers (19 males and 12 females) with only the HF microphone.

Fig. 2 shows the occurrence frequency of each digit, and Fig. 3 illustrates the occurrence frequency of the number of digits in the training data and test data. There are no six-digit utterances in the database.

Tables 2 and 3 show the average SNR (Signal to Noise Ratio) in each recording condition. In the table, we can see that the average SNR of speech data recorded by the CT microphone is high. The SNR is higher than 15 dB, even when the recording condition is high-speed driving with the "Audio on," which is the worst recording condition with the lowest SNR. On the other hand, the average SNR of speech data recorded by the hands-free microphone is low. The SNR is less than 5 dB in all the recording conditions. In addition to the HF case, the worst recording condition is high-speed driving with the "Fan on," for which the SNR is approximately 0 dB.

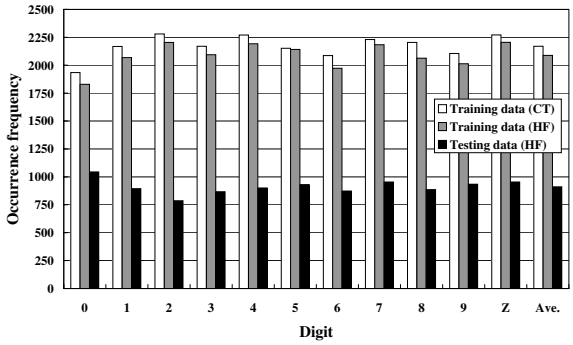

Figure 2: Occurrence frequency of each digit

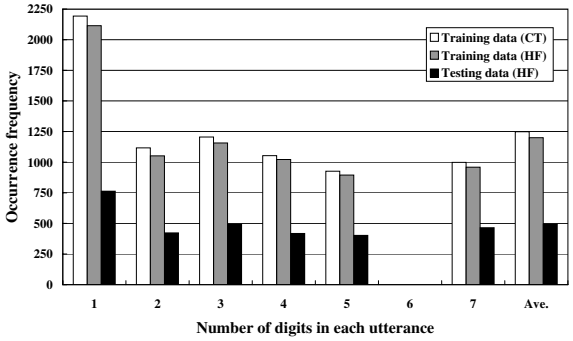

Figure 3: Occurrence frequency of number of digits in each utterance

3. Design of the evaluational framework

CENSREC-2 provides four evaluation environments for speech recognition using the speech data collected in various in-car conditions as described in the previous section. Each evaluation framework consists of the conditions marked by a circle (◯) in Tables 4 and 5. In each condition, the evaluation environments were designed as follows.

Condition 1: The speech data collected by using **same microphones in the same recording environments** were prepared both for training and testing.
Condition 2: The training and testing data were recorded under **different recording environments** by using **the same microphones**.
Condition 3: The training and testing data were recorded under **same recording environments** by using **the different microphones**.
Condition 4: The speech data collected by using **different microphones in the different recording environments** were prepared both for training and testing.

4. Baseline performance

4.1. Baseline scripts for evaluation

The baseline scripts were designed to facilitate HMM training and evaluation by HTK [11]. The evaluation framework was designed

Table 2: Average SNR of training data (dB)

In-car condition	Normal		Fan on		Audio on		Windows open	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF
Idling (quiet)	30.95	1.86	25.68	1.98	23.33	1.46	29.22	3.28
Low speed	23.89	0.29	21.94	0.40	20.53	0.49	18.67	0.29
High speed	18.75	0.19	18.10	0.22	17.71	0.25	—	—

Table 3: Average SNR of testing data (dB)

In-car condition	Normal	Fan on	Audio on	Windows open
Idling (quiet)	2.45	3.12	1.54	3.78
Low speed	0.50	0.33	0.34	0.53
High speed	0.07	0.11	0.28	—

Table 4: Training data for each evaluation condition

Evaluation condition	Condition 1		Condition 2		Condition 3		Condition 4	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF
Idling (quiet)	—	○	—	○	○	—	○	—
Low speed	—	○	—	—	○	—	—	—
High speed	—	○	—	—	○	—	—	—

Table 5: Testing data for each evaluation condition

Evaluation condition	Condition 1	Condition 2	Condition 3	Condition 4
Idling (quiet)	○	—	—	—
Low speed	○	○	○	○
High speed	○	○	○	○

as follows:

- The speech recognition is carried out using whole-word HMMs. In the recognition, a standard pronunciation dictionary and recognition grammar described by the EBNF syntax notation are defined as shown in Fig. 4.
- Each digit HMM had 18 states with 16 output distributions, "sil" had five states with three distributions, and "sp" had three states with one distribution. The output distribution of "sp" was the same as that of the third state of "sil." Each distribution of a digit HMM had 20 Gaussians and that of "sil" or "sp" had 36 Gaussians.
- The feature vector consisted of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients. Analysis conditions were pre-emphasis $1 - 0.97z^{-1}$, hamming window, 20-msec frame length, and 10-msec frame shift. In the baseline performance, cepstral mean subtraction was not applied to the feature vectors.
- In the Mel-filter bank analysis, a cut-off was applied to frequency components lower than 250 Hz.

```
$digit = one | two | three | four |
         five | six | seven | eight |
         nine | zero | oh ;

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

Figure 4: Grammar written in EBNF.

4.2. Baseline recognition results and performance comparison

Table 6 shows the details of baseline recognition results for each car environment for evaluation conditions 1 to 4.

We will also distribute a Microsoft Excel spreadsheet to simplify the comparison of recognition performance. All of the baseline results and the averaged recognition result are shown at the top of Table 7. The data entry for your results (word accuracy) should be made in the middle part of Table 7, after which the relative improvement against the baseline result is automatically given in the bottom part.

5. Conclusion

In this paper, we introduced CENSREC-2, an evaluation framework for Japanese in-car speech recognition.

In the near future, we will develop a series of frameworks for noisy speech recognition, CENSREC-1.5 (AURORA-2.5J): a subset of AURORA-2J with Lombard effect speech, and CENSREC-4: a continuous digit speech database with artificially added non-stationary noise. The CENSREC-4 will be developed as the additional test sets of CESNREC-1 (AURORA-2J).

We also plan to design and distribute the evaluation frameworks of noisy speech recognition gradually made difficult, i.e., reverberant environments, large-vocabulary continuous speech recognition tasks, and so on. Furthermore, we plan to develop and distribute a noise database for noisy speech recognition, evaluation measures instead of word accuracy, and a tool kit of conventionally used noise compensation methods.

We will provide the latest information about CENSREC in the following Web site.

CENSREC Web site:
http://sp.shinshu-u.ac.jp/CENSREC/

6. Acknowledgements

We would like to thank the members of the IPSJ SIG-SLP noisy speech recognition evaluation working group.

Table 6: Details of CENSREC-2 baseline evaluation results (%)

Car speed	In-car condition	Condition 1	Condition 2	Condition 3	Condition 4
Idling (quiet)	Normal	94.06	—	—	—
	Fan on	93.96	—	—	—
	Audio on	68.60	—	—	—
	Windows open	96.46	—	—	—
	Overall	86.38	—	—	—
Low speed	Normal	89.14	79.78	78.98	63.55
	Fan on	88.09	88.60	66.70	56.41
	Audio on	67.04	73.27	60.30	51.26
	Windows open	78.86	77.43	57.10	46.68
	Overall	80.80	79.77	65.84	54.52
High speed	Normal	78.97	57.96	62.45	43.27
	Fan on	79.75	77.20	52.66	39.78
	Audio on	63.76	67.11	51.57	40.71
	Overall	74.14	67.38	55.56	41.25
Overall		**80.58**	**74.49**	**61.46**	**48.87**

Table 7: CENSREC-2 spreadsheet

CENSREC-2 Evaluation Results

CENSREC-2 Baseline Results (%)				
Condition 1	Condition 2	Condition 3	Condition 4	Average
80.58	74.49	61.46	48.87	66.35

CENSREC-2 Word Accuracy (%)				
Condition 1	Condition 2	Condition 3	Condition 4	Average

CENSREC-2 Relative Improvement				
Condition 1	Condition 2	Condition 3	Condition 4	Average

7. References

[1] DARPA project Web site, http://www.nist.gov/speech/publications/

[2] SPINE Web site, http://elazar.itd.nrl.navy.mil/spine/

[3] H.G.Hirsch and D.Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition," *Proc. ISCA ITRW ASR2000*, pp. 18–20, Paris, France, Sept. 2000.

[4] AU/378/01, "Danish SpeechDat-Car Digits Database for ETSI STQ-Aurora Advanced DSR," *Aalborg University*, Jan. 2001.

[5] AU/225/00, "Baseline Results for subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation," *Nokia*, Jan. 2000.

[6] AU/273/00, "Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation," *Texas Instruments*, Dec. 2001.

[7] AU/271/00, "Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation: Description and Baseline Results," *UPC*, Nov. 2000.

[8] AU/337/01, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-Ends on a Large Vocabulary Task: Version 1.0," *Ericsson*, June 2001.

[9] AU/345/01, "Large Vocabulary Evaluation of Front-ends: Baseline Recognition System Description, Final Report," *Mississippi State University*, Jan. 2002.

[10] ETSI Web site, http://www.etsi.org/

[11] HTK Web site, http://htk.eng.cam.ac.uk/

[12] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next," Proc. Eurospeech '01, Aalborg, Denmark, Sept. 2001.

[13] 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 Transactions on Information and Systems*, Vol. E88–D, No. 3, pp. 535–544, Mar. 2005.

[14] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An Evaluation Database for In-Car Speech Recognition and Its Common Evaluation Framework," *Proc. Oriental COCOSDA '05*, pp. 44–49, Jakarta, Indonesia, Dec. 2005.

[15] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara, and F. Itakura, "Construction and Evaluation a Large In-Car Speech Corpus," *IEICE Transactions on Information and Systems*, Vol. E88–D, No. 3, pp. 553–561, Mar. 2005.