Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Performance estimation of speech recognition based on acoustic parameters under reverberation environments with CENSREC-4

Takahiro Fukumori (1), Masanori Morise (2) and Takanobu Nishiura (2)

(1) Graduate School of Science and Engineering, Ritsumeikan University, Japan
(2) College of Information and Science, Ritsumeikan University, Japan

PACS: 43.72.Ne

ABSTRACT

The Corpus and Environment for Noisy Speech RECognition 4 (CENSREC-4) evaluation framework has been distributed for evaluating distant-talking speech under various reverberation environments. The CENSREC-4 includes both real and simulated reverberant speech with convoluting impulse responses in the same environment. In addition, it consists of many room impulse responses to simulate various environments by convolving with clean speech signals and these impulse responses in real environments. How many variable reverberant impulse responses it contains has not, however, been evaluated. We thus try to evaluate CENSREC-4 with our proposed reverberation criterion on the basis of *C* value of ISO3382 Annex A acoustic parameters. We specifically focus on criteria to represent the difficulty of reverberant speech recognition, and also confirm why it is difficult to easily evaluate the recognition performance in a part of CENSREC-4 data sets with our proposed reverberation criterion. We have already proposed the reverberation criterion with *C* value of ISO3382 Annex A acoustic parameters to represent the difficulty of reverberant speech recognition, and we have tried to estimate the performance of distant-talking speech recognition on the basis of the impulse response between the speaker and microphone. First we investigated the relationship between the *C* value and the performance of reverberant speech recognition on the basis of measured impulse responses. We then calculated a regression curve approximated by exponential regression analysis in each reverberant environment. We finally tried to estimate the recognition performance in various reverberant environments with CENSREC-4. We carried out evaluation experiments to confirm the difficulty of easily evaluating the recognition performance in parts of CENSREC-4 data sets. As a result of the evaluation experiments, we confirmed that recognition performance could be estimated with 0.5 % errors in a 250 ms (T_{60}) environment, 2.9 % errors in a 450 ms (T_{60}) environment, 4.6 % errors in a 600 ms (T_{60}) environment, and 20.2 % errors in an 850 ms (T_{60}) environment on reverberant. We accurately estimated the recognition performance of reverberant speech in a light reverberation environment when the relationship between *C* value and the recognition performance is approximated by exponential function. Consequently, we also confirmed that it was difficult to estimate the performance of reverberant speech recognition in a heavy reverberation environment with CENSREC-4. We therefore confirmed that CENSREC-4 contained very challenging and variable reverberant data.

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.

The working group of the Information Processing Society in Japan (IPSJ) has worked on methodologies and frameworks for evaluating Japanese noisy speech recognition. It first released the Corpus and Environment for Noisy Speech RECognition 1 [1] (CENSREC-1) for evaluating speech recognition performance in noisy environments. After that, they released CENSREC-2 [2] (in-car recognition of connected digits), CENSREC-3 [3] (in-car isolated word recognition), and CENSREC-1-C [4] (voice-activity detection under noisy conditions). Thus far, they 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) had previously been actively pursued. However, many researchers have recently returned thoroughly analyzing distorted data to investigate the mechanisms responsible for individual distortions and have tried to address them. Thus, they distributed an evaluation framework, including database and evaluation tools, called CENSREC-4 [5], which is focused on evaluating distant-talking speech under reverberant environments. CENSREC-4 includes both the real and the simulated reverberant speech with convoluting impulse responses in the same environment. In addition, it consists of many room impulse responses to simulate various environments by convolving with clean speech signals and these impulse responses in real environments. How many variable reverberant impulse responses it contains has not, however, been evaluated. Thus, we try to evaluate CENSREC-4 with our proposed reverberation criterion RSR-

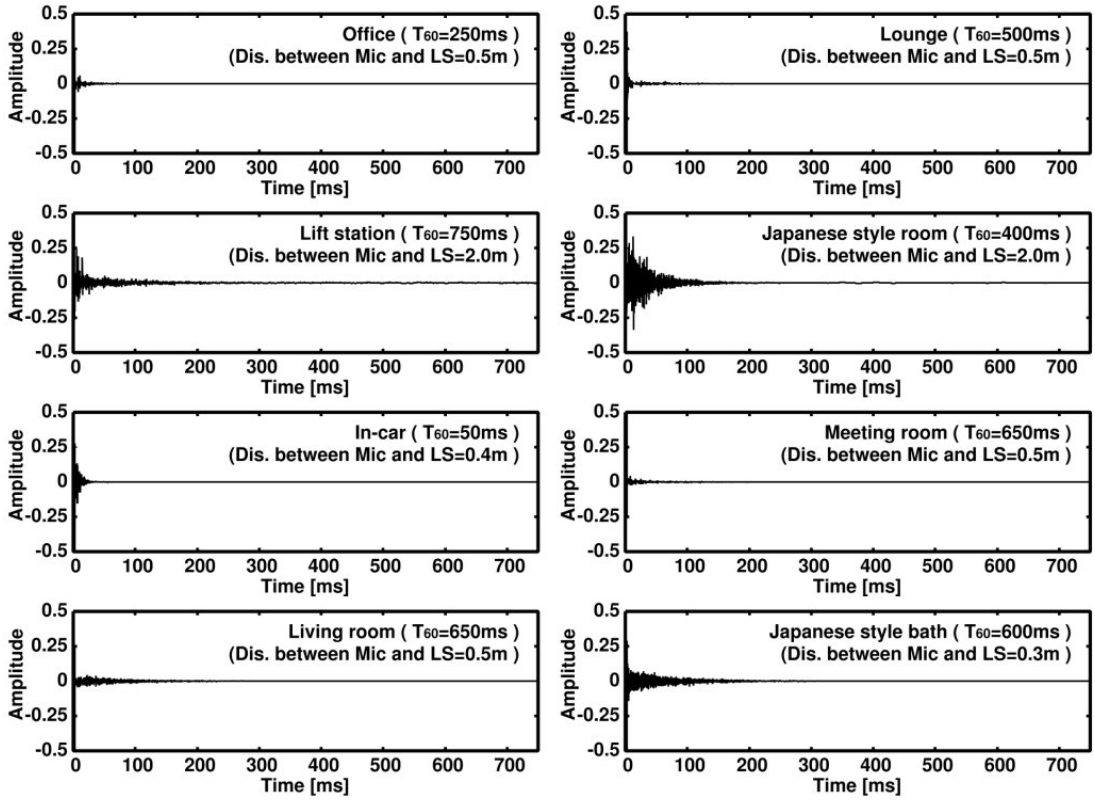

Fig. 1. Impulse responses in eight environments with CENSREC-4

C_n (Reverberant Speech Recognition criteria with C_n) on the basis of ISO3382 Annex A acoustic parameters. We specifically focus on criteria to represent the difficulty of reverberant speech recognition, and also confirm why it is difficult to easily evaluate the recognition performance in parts of CENSREC-4 data sets with our proposed reverberation criterion.

CENSREC-4

CENSREC-4 is a framework for evaluating distant-talking speech under various reverberant environments. The data it contains are connected digit utterances. Two subjects 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 noise in addition to reverberations. An evaluation framework has only been provided for the basic data sets as HTK based HMM training and recognition scripts. The basic data sets are used for the evaluation environment for the room impulse response-convolved speech data. This evaluation framework includes both real reverberant speech and simulated reverberant speech (with convoluting impulse responses) in the same environment.

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 bath (a prefabricated bath). The impulse responses were normalized at 0.5 with an absolute value for the maximum amplitude. Figure 1 gives the impulse responses recorded in these eight kinds of environments. As shown in Fig. 1, this evaluation framework includes impulse responses in many reverberant environments. "LS" in Fig. 1 means LoudSpeaker. However, how variable reverberant impulse responses it contains has not been evaluated. We thus try to evaluate CENSREC-4 with our proposed reverberation criterion on the basis of *C* value ISO3382 Annex A acoustic parameters. In the next section, we focus on criteria to represent the difficulty of reverberant speech recognition, and also explain why it is difficult to easily evaluate the recognition performance in parts of CENSREC-4 data sets with conventional methods such as reverberation time.

PERFORMANCE ESTIMATION BASED ON CONVENTIONAL REVERBERATION TIME

Reverberation time is used to estimate reverberant speech recognition performance. However, it is insufficient to represent the difficulty of reverberant speech recognition. We explain the difficulty in evaluating the speech recognition performance with reverberation time in this section.

Theory of reverberation time

Reverberation time, a parameter that expresses the duration of sound, is the most fundamental concept for evaluating indoor acoustical fields. It is the time required for a sound in a room to decay by 60 dB (conventionally notated as "T_{60}").

Measuring of reverberation time

Schroeder [6] developed a basic method of measuring reverberation by integrating the square of the reverberation's impulse responses. The reverberation time is easily measured with this method. The reverberation curves are derived from Eq. (1) with impulse response $h(t)$.

$$\left\langle y_d^2(t)\right\rangle = N\int_t^\infty h^2(\lambda)d\lambda, \qquad (1)$$

where < > is the ensemble average and N is the power of the unit frequency of random noise. The reverberation time in this reverberation curve is the time it takes to drop 60 dB below the original level.

Tab. 1. Environments to calculate speech recognition performance

Environment	T_{60}	RIRs
Laboratory	450 ms	72
Conference room	600 ms	120
Lift station	850 ms	120

RIRs:Room Impulse Responses

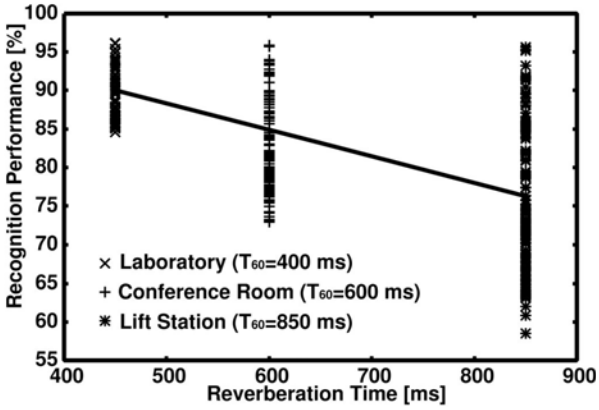

Fig. 2. Speech recognition performance in three reverberant environments

Performance estimation of reverberant speech recognition based on reverberation time

Reverberation time is usually used to estimate reverberant speech recognition performance. However, other reverberant features are altered by the difference between assuming a diffusible sound field in a room and an actual sound field. Thus, it is difficult to estimate speech recognition performance with only reverberation time. In this section, we conducted an evaluation experiment in three reverberant environments as shown in Tab. 1 to examine the relationship between reverberation time and speech recognition performance. We first measured several impulse responses in each environment. After that, we acquired speech recognition performance with a speech recognition engine by using the training data convolved speech sample and each measured impulse response. Figure 2 shows the obtained results. The line in Fig. 2 represents the average speech recognition performance in each reverberant environment. We confirmed the speech recognition performance degraded and the variance increased in the heavy reverberant environment. As a result, we could confirm that it is significantly more difficult to estimate speech recognition performance in a heavy reverberation environment than in light one.

PERFORMANCE ESTIMATION BASED ON NEW REVERBERATION CRITERIA RSR-D_N

In this section, we explain the reverberation criteria developed to solve the problem of estimating reverberant recognition performance with conventional criterion reverberation time.

Early reflections in reverberant speech recognition

In previous research [7], we confirmed two facts about reverberant speech recognition. One is that early reflections within about 12.5 ms after direct sound contributed slightly to recognizing reverberant speech in quiet environments, although early reflections within about 50 ms from the duration of direct sound contributed greatly to human hearing ability. The other is that late reflections over about 12.5 ms after direct sound decreased the recognition of reverberant speech. On the basis of these results, we confirmed that it is difficult to estimate the reverberant speech recognition performance using only reverberation time, since it does not take these factors into consideration. Therefore, we concluded that we would need to use the experimental results we had previously obtained to determine suitable reverberation criteria for recognizing reverberant speech.

ISO3382 acoustic parameters

In 1997, ISO3382 [8] proposed parameters for measuring room acoustics. The ISO3382 standards used of previously defined acoustical parameters to define how reverberation time should be measured in rooms. The ISO3382 standards focus particularly on the clarity (C value) in the category of the balance between early and late arriving energies based on previous research [7].

Clarity (*C* value)

The *C* value expresses the clarity of acoustics and is derived from Eq. (2).

$$C_n = 10\log_{10}\left(\int_0^n h^2(t)dt \Big/ \int_n^\infty h^2(t)dt\right), \qquad (2)$$

where $h(t)$ is impulse response and n is the border time between early and late arriving energies. The C value improves under the conditions of higher direct and early reflections and degrades under the conditions of higher late reverberations.

New reverberation criteria with RSR-*Cn*

We attempted to design the new reverberation criteria RSR-C_n (Reverberant Speech Recognition criteria with C_n) to estimate reverberant speech recognition performance as shown at the top of Fig. 3. First, we examined the relationship between the C value and reverberant speech recognition performance. We then used regression analysis on the basis of correlation coefficients for them to design the RSR-C_n to cover each reverberation time. We used four steps in our approach, explained in detail as follows.

Step 1: We measured many impulse responses in a number of environments to obtain training data. Using the measured impulse responses as a basis, we derive reverberation times from Eq. (1).

Step 2: We next derive the C value from Eq. (2) after performing **Step 1**. In Eq. (2), the border time n is essential for determining the maximum value of the relationship between C value and speech recognition performance. Thus, we determined the suitable border time n as described later and then used the value to calculate C_n.

Step 3: We then acquired speech recognition performance with a speech recognition engine [9] by using the training data obtained by using dry data and measured impulse responses as described in **Step 1**.

Step 4: Finally, we conducted regression analysis on the basis of the *C* value calculated from **Steps 1** and **2** and the speech recognition performance calculated in **Step 3**. We used exponential functions as regression curves calculated with regression analysis.

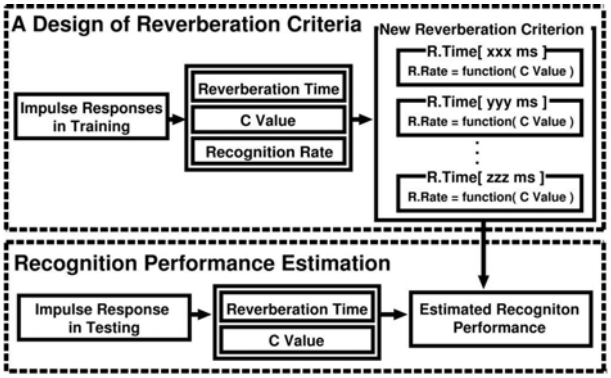

Fig. 3. Proposed method overview

Tab. 2. Environments to design reverberation criterion RSR-*Cn*

Environment	T_{60}	RIRs
Env.A	400 ms	72
Env.B	600 ms	120
Env.C	850 ms	120

Tab. 3. Environments to calculate suitable *n*

Environment	T_{60}	RIRs
Japanese-style room	400 ms	72
Conference room	600 ms	120
Standard stairs	750 ms	56

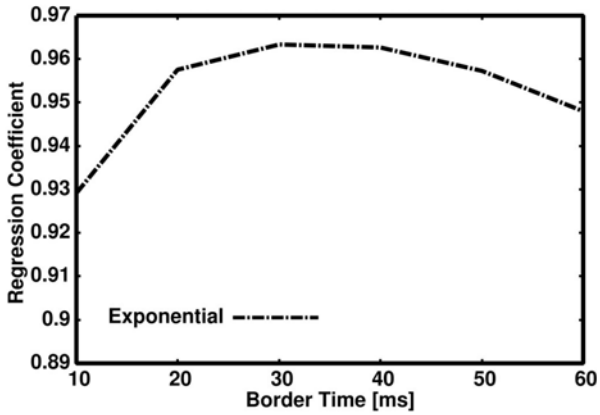

Fig. 4. Relationship between correlation coefficient in each regression curve and border time *n*

Performance estimation with RSR-C_n

As shown at the bottom of Fig. 3, we will try to estimate the speech recognition performance with the RSR-C_n. We first calculate the reverberation time and the *C* value on the basis of impulse responses in test environments. Then on the basis of them, we will try to estimate the speech recognition performance with the RSR-C_n to cover each same reverberation time.

EVALUATION EXPERIMENTS

We used the proposed criteria to estimate the reverberant speech recognition performance. Initially, we measured 312 impulse responses to design the reverberant criteria RSR-C_n in the three training environments shown in Tab. 2. A time-stretched pulse [10] was used to measure the impulse responses. The recordings were conducted with 16 *kHz* sampling and 16 *bit* quantization. All impulse responses were measured for distances ranging between 100~5,000 mm. To estimate speech recognition performance, we used connected digit utterance set in CENSREC-4 as the speech samples that were made up of eleven Japanese numbers ("1:*ichi*", "2:*ni*", "3:*san*", "4:*yon*", "5:*go*", "6:*roku*", "7:*nana*", "8:*hachi*", "9:*kyu*", "0:*zero* or *maru*") that were uttered by 104 speakers (52 females and 52 males).

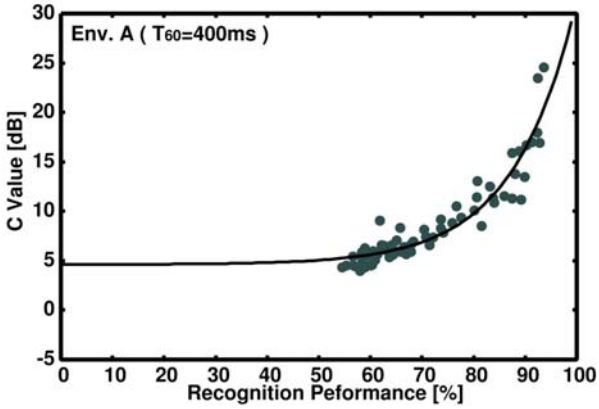

Fig. 5. Relationship between RSR-D20 and speech recognition performance in Env.A

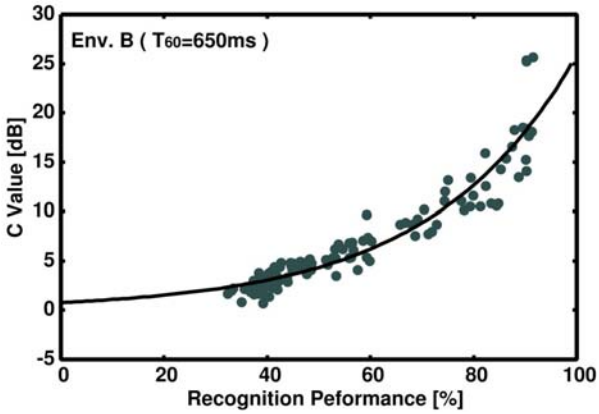

Fig. 6. Relationship between RSR-D20 and speech recognition performance in Env.B

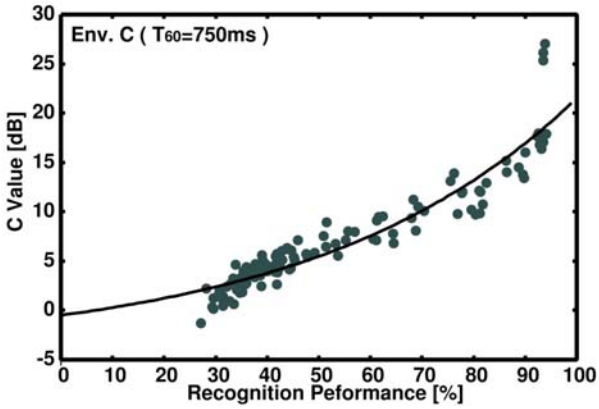

Fig. 7. Relationship between RSR-D20 and speech recognition performance in Env.C

Tab. 4. Correlation coefficients

Environment	Coefficient
Env.A	0.94
Env.B	0.93
Env.C	0.93

Tab. 5. Actual and estimated recognition performance in five test environments with CENSREC-4

Environment	T_{60}	C value	Actual Recognition	Est. Rec. with T_{60}	Est. Rec. with C (Env.)	Error with T_{60}	Error with C
Office	250 ms	19.1 dB	93.4 %	78.2 %	92.6 % (A)	14.9 %	0.5 %
Lift station	750 ms	5.7 dB	30.7 %	52.3 %	50.9 % (C)	21.6 %	20.2 %
Living room	650 ms	6.4 dB	65.3 %	56.0 %	60.7 % (B)	9.3 %	4.6 %
Japanese room	400 ms	5.6 dB	54.3 %	70.5 %	56.9 % (A)	16.2 %	2.6 %
Meeting room	650 ms	15.2 dB	74.1 %	56.0 %	85.2 % (B)	18.1 %	11.1 %

Suitable border time *n* for reverberation criteria RSR-C_n

In Eq. (2), the border time n is essential for determining the maximum value of the relationship between C value and speech recognition performance. Thus, we conducted evaluation experiments in the three environments as shown in Tab. 3, using the C value and exponential function to determine the most suitable border time n. Figure 4 shows the results we obtained. From exponential regression analysis, 30 msec was determined to be the most suitable border time. We therefore used 30 msec as the border time for calculating C_n and designing RSR-C_{30}.

Suitable RSR-C_n design

Figures 5~7 show the relationship between speech recognition performance and C_{30} for the three training environments shown in Tab. 2. Table 4 shows correlation coefficients with exponential function for these three environments. As shown in Table 4, we confirmed that RSR-C_n coefficients are higher than 0.93 in all environments. We thus confirmed that RSR-C_{30} is the suitable criterion for estimating reverberant speech recognition.

Performance estimation with RSR-Cn

Finally, we attempted to estimate the reverberant speech recognition performance for the five test environments in CENSREC-4: office (T_{60}=250 ms), lift station (T_{60}=750 ms), living room (T_{60}=650 ms), japanese style room (T_{60}=400 ms), and meeting room (T_{60}=650 ms).

Table 5 lists the results, where "Est. Rec with T_{60}" means the estimated performance with reverberation time T_{60} criterion. "Est. Rec with C (Env.)" means the estimated performance with RSR-C_n in Envs. A, B and C. In this experiment, RSR-C_n in Envs. A, B and C were selected as the reverberation time environments closest to the test environment. The average performance estimation error of 16.02 % was achieved with reverberation time. On the other hand, the average performance estimation error of 7.8 % was achieved with RSR-C_{30}. As a result, we confirmed that RSR-C_{30} had fewer errors than T_{60} criteria in all reverberation environments. Consequently, we also confirmed that it was difficult to estimate the performance of reverberant speech recognition in heavy environment with CENSREC-4. We therefore confirmed that CENSREC-4 contained very challenging and variable reverberant data.

CONCLUSIONS

To evaluate how many variable reverberant impulse responses CENSREC-4 contains, we tried to estimate recognition performance in CENSREC-4 with our proposed reverberation criterion RSR-C_{30} (Reverberant Speech Recognition criteria with C_{30}), which calculates recognition performance on the basis of C_{30} for ISO3382 acoustic parameters. As a result of experiments, we confirmed that the proposed criterion RSR-C_{30} estimates performance much better than the conventional reverberation criteria, reverberation time. Moreover, CENSREC-4 has impulse responses including various reverberant features. In future work we will attempt to define more suitable reverberation criteria in the frequency domain for reverberant speech recognition.

ACKNOWLEDGMENTS

This work was partly supported by Global COE and Grants-in-Aid for Scientific Research funded by Japan's Ministry of Education, Culture, Sports, Science, and Technology.

REFERENCES

1. 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, Mar. 2005.
2. 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, Sep. 2006.
3. M. Fujimoto, K. Takeda, and S. Nakamura, "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.
4. N. Kitaoka, et. al., "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), pp. 607-612, Dec. 2007.
5. Masato Nakayama, et. al., "CENSREC-4: Development of Evaluation Framework for Distant-talking Speech Recognition under Reverberant Environments", Interspeech, pp. 968-971, Sep. 2008.
6. M. R. Schroeder, "New Method of Measuring Reverberation-Time," JASA, Vol. 37, pp. 409-412, 1965.
7. 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, pp. 1052-1055, Aug. 2007.
8. ISO3382: "Acoustics measurement of the reverberation time of rooms with reference to other acoustical parameters", International Organization for Standardization, 1997.
9. A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," In Proc. European Conf. on Speech Communication and Technology, pp. 1691-1694, 2001.
10. Y. Suzuki, F. Asano, H.-Y. Kim, and Toshi Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am. Vol. 97(2), pp.1119-1123, 1995.