1aSCb21. Performance estimation of speech recognition based on Perceptual Evaluation of Speech Quality (PESQ) and acoustic parameters under noisy and reverberant environments with Corpus and Environment for Noisy Speech RECognition 4 (CENSREC-4)

Takahiro Fukumori*, Masato Nakayama, Takanobu Nishiura and Yoichi Yamashita

*Corresponding author's address: Ritsumeikan University, Kusatsu, 525-8577, Shiga, Japan, cm013061@ed.ritsumei.ac.jp

CENSREC-4 evaluation framework has been distributed for evaluating distant-talking speech under various noisy and reverberant environments. It however has not been evaluated how variable noisy and reverberant features in this contains. We thus try to evaluate CENSREC-4 with our designed noisy and reverberant criteria based on PESQ and acoustic parameters. We specifically focus on criteria to represent the difficulty of noisy and reverberant speech recognition, and also confirm why it is difficult to easily evaluate the recognition accuracy in various noisy and reverberant environments with CENSREC-4. We carried out evaluation experiments to confirm the difficulty to easily evaluate the recognition accuracy in a part of CENSREC-4 corpus with our proposed noisy and reverberant criteria. We first designed the noisy and reverberant criteria using the relationship among the D value, the PESQ, and the ASR performance. We then tried to estimate the recognition accuracy in various noisy and reverberant environments with CENSREC-4. We carried out evaluation experiments to confirm the difficulty to easily evaluate the recognition accuracy in a part of CENSREC-4 corpus. As a result of evaluation experiments, we confirmed that it was difficult to estimate the accuracy of noisy and reverberant speech recognition in heavy noisy and reverberant environment with CENSREC-4. We therefore confirmed that CENSREC-4 contained very challenging and variable noisy and reverberant data.

Published by the Acoustical Society of America through the American Institute of Physics
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 noisy and reverberant features it contains has not, however, been evaluated. Thus, we try to evaluate CENSREC-4 with our proposed noisy and reverberant criteria NRSR-PDn (Noisy and Reverberant Speech Recognition criteria with PESQ and Dn) [6] on the basis of the PESQ score and ISO3382 Annex A acoustic parameters [7]. We specifically focus on criteria to represent the difficulty of noisy and reverberant speech recognition, and also confirm why it is difficult to easily evaluate the recognition performance in parts of CENSREC-4 corpus with our proposed noisy and reverberant criteria.

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, a lift station (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 noisy and reverberant features it contains has not been evaluated. We thus try to evaluate CENSREC-4 with our proposed noisy and reverberant criteria on the basis of the PESQ score and D value ISO3382 Annex A acoustic parameters. In the next section, we focus on criteria to represent the difficulty of noisy and reverberant speech recognition.
CONVENTIONAL METHODS FOR ESTIMATING ASR PERFORMANCE

ASR Performance Estimation in Noisy Environments

Methods for accurately estimating the ASR performance in noisy environments have been proposed that use the PESQ score [8, 9], which is calculated using noisy speech samples. The PESQ score, which takes the auditory-psychological effects into account, is used to estimate the subjective quality of speech distorted by ambient noise. The original and degraded speech samples are first transformed into an internal representation by using a perceptual model to calculate the PESQ score. A cognitive model then evaluates the difference between the degraded and original speech and estimates the subjective mean opinion score (MOS), which ranges from 0.5 to 4.5.

ASR Performance Estimation in Reverberant Environments

We previously developed the reverberant criteria RSR-$D_n$ [6] to facilitate a reverberant speech recognition, which are used to estimate the ASR performance on the basis of the $D$ value calculated from the ISO3382 acoustic parameters, which were formulated for measuring room acoustics. The definition ($D$ value) is particularly important in terms of the balance between early and late arriving energies from an impulse response. The $D$ value is calculated using Eq. (1).

$$D_n = \frac{\int_0^n h^2(t)dt}{\int_0^{\infty} h^2(t)dt},$$

where $h(t)$ is an impulse response and $n$ is the border time between the early and late arriving energies. The $D$ value improves when there are higher direct and early reflections and degrades when there are higher late reverberations. We demonstrated that the average estimation error was less than 5% when these criteria were used in reverberant and noiseless environments.

DESIGN OF PROPOSED NOISY AND REVERBERANT CRITERIA NRSR-$PD_n$

Noisy and Reverberant Speech Recognition with the PESQ and the $D_n$ (NRSR-$PD_n$) are proposed in this section as new noisy and reverberant criteria for the ASR performance. The NRSR-$PD_n$ are designed in four steps, as illustrated in Fig. 2. They are designed on the basis of the relationships among the $D$ value, the PESQ score, and the
noisy and reverberant ASR performance. In particular, a multiple-regression analysis is used to design the NRSR-P\textsubscript{Dn} criteria based on a correlation with these calculated values.

**Step 1:** Prepare degraded speech samples

First, the degraded speech samples are prepared as the training data to design the NRSR-P\textsubscript{Dn} criteria used for estimating the ASR performance in noisy and reverberant environments using the following five steps.

1. Measure many impulse responses in a number of reverberant environments.
2. Create noise samples.
3. Measure speech samples in a clean environment.
4. Create reverberant speech samples that are convolved with the measured impulse responses and clean speech samples.
5. Prepare degraded speech samples by adding noise samples and reverberant speech samples.

**Step 2:** Calculate reverberation time \( T_{60} \) and \( D \) value

Next, the measured impulse responses are used to calculate the \( D \) value using Eq. (1) and the reverberation time \( T_{60} \). The reverberation time is derived on the basis of Eq. (2) using impulse response \( h(t) \).

\[
\langle y_d^2(t) \rangle = N \int_{-\infty}^{\infty} h^2(\lambda) d\lambda, \tag{2}
\]

where \( \langle \cdot \rangle \) is the ensemble average and \( N \) is the power of the unit frequency of random noise. The reverberation time for a reverberation curve \( \langle y_d^2(t) \rangle \) is the time that it takes for the level of a sound to drop 60 dB below its original level (conventionally notated as “\( T_{60} \)”).

**Step 3:** Calculate ASR performance and PESQ score

The degraded speech samples prepared in Step 1 are used to calculate the ASR performance and the PESQ score. The ASR performance is acquired using a speech recognition engine, and the PESQ score is calculated using clean and degraded speech samples.

**Step 4:** Perform multiple-regression analysis using \( D \) value, PESQ score, and ASR performance

Finally, a multiple-regression analysis is used to design the NRSR-P\textsubscript{Dn} criteria using the \( D \) value, the PESQ score, and the ASR performance. The NRSR-P\textsubscript{Dn} criteria are represented as Eq. (3).

\[
y(x_1, x_2, x_3) = ax_1 + bx_2 + c | x_3, \tag{3}
\]

where \( y(x_1, x_2, x_3) \), \( x_1 \), \( x_2 \), and \( x_3 \) represent the estimated ASR performance, the \( D \) value, the PESQ score, and the reverberation time \( T_{60} \), respectively. Coefficients \( a \), \( b \), and \( c \) are calculated by taking the minimum error of the root mean square in the multiple-regression analysis into account. Moreover, the NRSR-P\textsubscript{Dn} criteria are designed for each reverberant environment.

**FIGURE 2.** Design of the NRSR-P\textsubscript{Dn} criteria for use in noisy and reverberant speech recognition.
FIGURE 3. Estimation of ASR performance using NRSR-P\textsubscript{Dn} criteria.

TABLE 1. Evaluation conditions.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ATR phoneme balance 216 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>Two females and two males</td>
</tr>
<tr>
<td>Decoder</td>
<td>Julius rev. 4.2.1 [11]</td>
</tr>
<tr>
<td>HMM</td>
<td>IPA monophone model</td>
</tr>
<tr>
<td>Frame length</td>
<td>25 ms (Hamming window)</td>
</tr>
<tr>
<td>Noise</td>
<td>White noise, Pink noise, Moving in-car noise</td>
</tr>
</tbody>
</table>

USE OF NRSR-P\textsubscript{Dn} TO ESTIMATE ASR PERFORMANCE

The NRSR-P\textsubscript{Dn} criteria are used to estimate the ASR performance in noisy and reverberant environments as illustrated in Fig. 3. As shown in this figure, we can estimate the ASR performance using the NRSR-P\textsubscript{Dn} in noisy and reverberant environments in just three steps.

\textbf{Step 1:} Measure an impulse response and the degraded speech samples in a test environment.

\textbf{Step 2:} Calculate $T_{60}$ and the $D$ value using the measured impulse response, and the PESQ score using the degraded speech samples.

\textbf{Step 3:} Estimate the ASR performance using the calculated $T_{60}$, $D$ value, PESQ score, and the NRSR-P\textsubscript{Dn}.

EVALUATION EXPERIMENTS

We try to evaluate CENSREC-4 with our proposed noisy and reverberant criteria NRSR-P\textsubscript{Dn}. First, we designed the proposed criteria to estimate the noisy and reverberant ASR performance. We used an ATR phoneme-balanced set as the speech samples that consist of 216 isolated Japanese words [10] uttered by four speakers (two females and two males) to estimate the ASR performance. A border time ($n=20$ ms) was used on the basis of a previous study [6]. The conditions for the analysis and recognition processes are also summarized in Table 1. Since the ASR performance greatly varies with the recognition task, the NRSR-P\textsubscript{Dn} design and performance estimation were conducted using the same recognition task.

Suitable NRSR-P\textsubscript{Dn} Design

We measured 312 room impulse responses (RIRs) to design the noisy and reverberant criteria NRSR-P\textsubscript{Dn} in the three training environments, Japanese style room ($T_{60}=400$ ms, 72 RIRs), a conference room ($T_{60}=600$ ms, 120 RIRs), and a lift station ($T_{60}=850$ ms, 120 RIRs). A time-stretched pulse [10] was used to measure the impulse responses. The recordings were conducted with 16 kHz sampling and 16 bits quantization. All impulse responses were measured for distances ranging between 100–5,000 mm. Figure 4 shows the relationships among the $D$ values, the PESQ scores, and ASR performances for these three reverberant environments with white noise. As a result, we
confirmed that the NRSR-\(PDn\) are suitable criteria for the estimation of noisy and reverberant ASR performances since the correlation coefficients obtained by conducting the multiple-regression analysis are higher than 0.9 in all noisy and reverberant environments.

**Performance Estimation of ASR Performance in CENSREC-4 with NRSR-\(PDn\)**

Finally, we attempted to estimate the noisy and reverberant speech recognition performance for the four test reverberant environments in CENSREC-4: Japanese style room (\(T_{60} = 400\) ms), living room (\(T_{60} = 650\) ms), meeting room (\(T_{60} = 750\) ms), and lift station (\(T_{60} = 750\) ms) with three kinds of noise samples. In this experiment, we conducted the estimation of the ASR performance for each noisy and reverberant speech recognition using the \(D\) value and PESQ score as the conventional methods. Figure 5 shows estimation error in reverberant environments with white noise, pink noise, and noise of moving in-car. As a result, we confirmed that NRSR-\(D_{20}\) and PESQ score had fewer errors than \(D\) value in all noisy and reverberant environments. Moreover, there was no difference of the estimation performance between the NRSR-\(PDn\) and the PESQ in all environments. However, we also confirmed that it was difficult to estimate the performance of noisy and reverberant speech recognition in meeting room. We therefore confirmed that CENSREC-4 contained very challenging and variable reverberant data.

![Figure 4](image1.png)

**FIGURE 4.** Relationships among \(D\) value, PESQ score, and ASR performance in reverberant environments with white noise.

![Figure 5](image2.png)

**FIGURE 5.** Estimation error in reverberant environments with white noise, pink noise, and noise of moving car.
CONCLUSIONS

To evaluate how many variable noisy and reverberant features CENSREC-4 contains, we tried to estimate recognition performance in CENSREC-4 with our proposed noisy and reverberant criteria NRSR-P\textsubscript{D20} (Noisy and Reverberant Speech Recognition criteria with PESQ and D\textsubscript{20}), which calculates recognition performance on the basis of the PESQ score and D\textsubscript{20} for ISO3382 acoustic parameters. As a result of experiments, we confirmed that the proposed criteria NRSR-P\textsubscript{D20} estimate performance much better than the conventional reverberant criteria, D value. Moreover, CENSREC-4 has impulse responses including various reverberant features. In future work, we will attempt to define more suitable noisy and reverberant criteria in the frequency domain for noisy and reverberant speech recognition.

ACKNOWLEDGMENTS

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

REFERENCES