Performance Estimation of Reverberant Speech Recognition Based on Reverberant Criteria $R_{SR-D_n}$ with Acoustic Parameters

Takahiro Fukumori$^1$, Masanori Morise$^2$, Takanobu Nishiura$^2$

$^1$Graduate School of Science and Engineering, Ritsumeikan University
$^2$College of Information Science and Engineering, Ritsumeikan University

{cm013061@ed, morise@fc, nishiura@is}.ritsumei.ac.jp

Abstract

Reverberation-robust speech recognition has become very important in the field of distant-talking speech recognition. However, as no common reverberation criteria for the recognition of reverberant speech have yet been proposed, it has been difficult to estimate its effectiveness. To address this problem in 2007, we investigated early and late reflections on distant-talking speech recognition to help define suitable reverberation criteria. We here propose a new reverberation criterion $R_{SR-D_n}$ (Reverberant Speech Recognition criteria with $D_n$) based on our previous investigations and ISO3382 acoustic parameters. We then estimated the speech recognition performance obtained with the criteria. Evaluation experiments confirmed that the recognition performance can be accurately and robustly estimated with our proposed reverberation criterion $R_{SR-D_{n20}}Q$ (Reverberant Speech Recognition criterion with $D_{20}$ and Quadratic regression function).

Index Terms: Reverberant speech recognition, Performance estimation, Acoustic parameters, Reverberation

1. Introduction

Rapid improvement in hands-free speech interfaces for capturing distant-talking speech with naturalness and high quality have been made. Reverberation-robust speech capture and recognition are essential for speech interfaces to be useful. Various techniques have been proposed for achieving these, including an inverse filter for speech capture [1], Cepstrum Mean Normalization (CMN) for speech recognition [2], and an acoustic model adaptation with a transfer function for speech recognition [3]. However, it has been difficult to estimate recognition performance for reverberant speech, since no useful criteria for recognizing such speech have yet been proposed. Reverberation time, a parameter that expresses the duration of sound, is usually used to estimate it. Since reverberation time theory assumes a diffusible sound field in a room, the effect of reverberation time does not change even if sound-absorbing material is placed somewhere in the room. By itself, however, reverberation time is insufficient as a criterion for recognizing reverberant speech because recognition performance depends on the distance between a talker and the microphone in the same environment. To overcome this problem, we developed novel reverberation criteria $R_{SR-D_n}$ (Reverberant Speech Recognition criteria with $D_n$) based on ISO3382 acoustic parameters. We previously investigated the effect of early and late reflections on distant-talking speech recognition with the aim of defining suitable reverberation criteria[4] in INTERSPEECH2007. In our recent work, we referred to our previous results to develop the new reverberant-robust criterion $R_{SR-D_{n20}}Q$ (Reverberant Speech Recognition criteria with $D_{20}$ and Quadratic regression function), and used $R_{SR-D_{20}}$ to robustly estimate reverberant speech recognition performance.

2. Performance estimation based on conventional reverberation time

2.1. Theory of reverberation time

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

2.2. Measuring reverberation time

Schroeder [5] developed a basic method that easily measures reverberation time by integrating the square of the reverberation’s impulse responses. The reverberation curves are derived from Eq. (1) with impulse response $h(t)$.

$$< y^2(t) > = N \int_{t}^{\infty} h^2(\lambda)d\lambda,$$

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.

2.3. 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 assumption of a diffusible sound field and an actual sound field. Thus, it is difficult to estimate speech recognition performance using only reverberation time. We conducted an evaluation experiment in the three reverberant environments shown in Table1(b) to investigate the relation 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[6] by using the training data convolved speech sample and each measured impulse response. Figure 1 shows the results. The line in Figure 1 represents the speech recognition performance average in each reverberant environment. We found speech recognition performance degradation and increased variance in heavily reverberant environments. This indicates that
3. Performance estimation based on new reverberation criteria RSR-$D_n$

3.1. Early reflections in reverberant speech recognition

In previous research[4], we showed two facts about reverberant speech recognition. One is that early reflections, within about 12.5 ms after direct sound, contributed only slightly to the recognition of 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. Based on 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.

3.2. ISO3382 acoustic parameters

In 1997, ISO3382[7] proposed parameters for measuring room acoustics. The ISO3382 standards made use of previously defined acoustical parameters to define how reverberation time should be measured in rooms. The ISO3382 standards focus particularly on the definition ($D$ value) in the category of the balance between early and late arriving energies based on previous investigations[4].

3.3. Definition ($D$ value)

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

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

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

3.4. New reverberation criteria with RSR-$D_n$

We attempted to design the new reverberation criteria RSR-$D_n$ to estimate reverberant speech recognition performance as shown at the top of Fig. 2. First, we investigated the relation between the $D$ value and reverberant speech recognition performance. We then used regression analysis based on the correlation coefficients for these to design the RSR-$D_n$ to cover each reverberation time. We used four steps in our approach, explained in detail in Fig. 2.

**Step1:** We measured many impulse responses in a number of environments to obtain training data. Using the measured impulse responses as a basis, we used Eq. (1) to calculate reverberation times.
Step 2: We next calculated the $D$ value with Eq. (2) after performing Step 1. In Eq. (2), the border time $n$ is essential for determining the maximum value of the relation between $D$ value and speech recognition performance. Thus, we determined the suitable border time $n$ as described in Section 4.1 and then used the value to calculate $D_n$.

Step 3: We then acquired speech recognition performance with a speech recognition engine[8] by using the training data obtained using dry data and measured impulse responses as described in Step 1.

Step 4: Finally, we conducted regression analysis using the $D$ value calculated from Steps 1 and 2 and the speech recognition performance calculated in Step 3. We used linear and quadratic functions as regression curves calculated with regression analysis using ordinary least squares.

3.5. Performance estimation with RSR-$D_n$

As shown at the bottom of Fig. 2, we will try to estimate the speech recognition performance with the RSR-$D_n$. We first calculate the reverberation time and the $D$ value using impulse responses in test environments. Using these values, we try to estimate the speech recognition performance with the RSR-$D_n$ in the same reverberation time.

4. Evaluation experiments

We used the proposed criteria to estimate the reverberant speech recognition performance. Initially, we measured 732 impulse responses to design the reverberant criteria RSR-$D_n$ in the nine training environments shown in Table 1(a). A time-stretched pulse[9] 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 from 100 - 5,000 mm. To estimate speech recognition performance, we used an ATR phoneme-balanced set as the speech samples that were made up of 216 isolated Japanese words uttered by 14 speakers (7 females and 7 males). In addition, the recognition performance varies largely depending on the recognition task. Thus, RSR-$D_{20}$ design and performance estimation should be conducted in the same recognition task.

4.1. Suitable border time $n$ for reverberant criteria RSR-$D_n$

In Eq. (2), the border time $n$ is essential for determining the maximum value of the relation between $D$ value and speech recognition performance. Thus, we conducted evaluation experiments in the three environments shown in Table 1(c), using the $D$ value and two regression functions (linear and quadratic) to determine the most suitable border time $n$. Figure 3 shows the results. Linear and quadratic regression analysis showed that 20 msec was the most suitable border time value. We therefore used 20 msec as the border time for calculating $D_n$ and designing RSR-$D_{20}$.

4.2. Suitable RSR-$D_{20}$ design

Figures 5 and 4 show the relation between speech recognition performance and $D_{20}$ for the nine training environments shown in Table 1(a). These figures also show the regression analysis results for the three environments shown in Table 1(d). Figure 6 shows the relation between RSR-$D_{20}$ and speech recognition performance based on the regression analysis results in three environments (Japanese-style room, Conference room and Standard stairs). Table 2 shows correlation coefficients with their respective regression functions for these three environments. We defined that RSR-$D_{20}L$ represents RSR-$D_{20}$ with a linear regression function, and RSR-$D_{20}Q$ represents RSR-$D_{20}$ with a quadratic regression function. The data in Table 2 indicates that both RSR-$D_{20}L$ and RSR-$D_{20}Q$ are the most suitable criteria for estimating reverberant speech recognition.

4.3. Performance estimation with RSR-$D_{20}Q$

Finally, we attempted to estimate the reverberant speech recognition performance for the three test environments shown in Table 1(e). Both closed and open tests were carried out for this purpose. In closed tests, we estimated speech recognition performance on known condition with RSR-$D_n$ designed in the same environment. On the other hand, in open tests, we estimated recognition performance on unknown condition with RSR-$D_n$ designed in the other environments, including the same reverberation time. Figure 7 shows the results. Standard deviations are given in Table 3. The results showed that average estimation error of less than 5% was achieved with RSR-$D_{20}Q$ in all environments. Table 2 shows the correlation coefficients obtained with RSR-$D_{20}L$ and RSR-$D_{20}Q$. As the table shows, both the RSR-$D_{20}L$ and RSR-$D_{20}Q$ coefficients are higher than 0.93 in all environments. Thus, the RSR-$D_{20}$ criteria provided much better estimation performance than conventional reverberant criteria.
5. Conclusions

To facilitate the recognition of reverberant speech, we developed new reverberation criteria RSR-\(D_{20}\) (Reverberant Speech Recognition criteria with \(D_{20}\)), which calculates recognition performance based on \(D_{20}\) for ISO3382 acoustic parameters. Experiments conducted in actual environments confirmed that the proposed criteria (particularly RSR-\(D_{20}\)Q) provide much better estimation performance than conventional reverberation criteria. In future work we will attempt to define more suitable reverberation criteria in the frequency domain for reverberant speech recognition.

6. 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.

7. References