Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters

Takahiro Fukumori*, Masato Nakayama†, Takanobu Nishiura† and Yoichi Yamashita†
*Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan
E-mail: cm013061@ed.ritsumei.ac.jp
†College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan
E-mail: {mnaka@fc, nishiura@is, yama@media}.ritsumei.ac.jp

***Abstract*—The automatic speech recognition (ASR) performance is degraded in noisy and reverberant environments. Although various techniques against degradation of the ASR performance have been proposed, it is difficult to properly apply them in evaluation environments with unknown noisy and reverberant conditions. It is possible to properly apply these techniques for improving the ASR performance if we can estimate the relationship between the ASR performance and degradation factors including both noise and reverberation. In this study, we here propose new noisy and reverberant criteria which are referred as "Noisy and Reverberant Speech Recognition with the PESQ and the D_n (NRSR-PD_n)". We first designed the "NRSR-PD_n" using the relationships among the D value, the PESQ score, and the ASR performance. We then estimated the ASR performance with the designed criteria "NRSR-PD_n" in evaluation experiments. Experimental evaluations demonstrated that our proposed criteria make the well suited for robustly estimating the ASR performance in noisy and reverberant environments.**

I. Introduction

In recent years, robust speech recognition has become very important in the field of distant-talking speech recognition because robust speech capture and recognition are essential for usable speech interfaces. In hands-free speech interfaces, automatic speech recognition (ASR) performance is, however, degraded due to noises and reverberations. Various techniques have been proposed for preventing this degradation, including spectral subtraction method [1] for noisy environments, and an acoustic model adaptation with a transfer function for speech recognition [2]. However, they have a difficulty to completely prevent degradation in environments under unknown noisy and reverberant conditions. It is possible to properly apply these techniques for improving the ASR performance by estimating the relationship between the ASR performance and degradation factors including both noise and reverberation. Estimation methods of the ASR performance in noisy environments [3], [4] have been proposed by using the perceptual evaluation of speech quality (PESQ) score [5]. On the other hand, it can be estimated in reverberant environments by using "Reverberant Speech Recognition criteria with the D_n (RSR-D_n)" [6] which are based on ISO3382 acoustic parameters [7]. However, a method is still needed for robustly estimating ASR performance in an environment with both noise and reverberation. In the study reported here, we developed a method to design criteria for use in accurately estimating the ASR performance which is degraded due to noise and reverberation by using both the D value calculated from the ISO3382 acoustic parameters and the PESQ score.

II. Conventional methods for estimating ASR performance

A. ASR performance estimation in noisy environments

The methods [3], [4] for accurately estimating ASR performance in noisy environments have been proposed by using the PESQ score [5] which is calculated with noisy speech samples. The PESQ score, which takes auditory-psychological effects into account, is used to estimate the subjective quality of speech distorted by ambient noise. To calculate the PESQ score, the original and degraded speech are first transformed into an internal representation by using a perceptual model. A cognitive model then evaluates the difference between the degraded and original speech and estimates the subjective mean opinion score (MOS), which has range from 0.5 to 4.5.

B. ASR performance estimation in reverberant environments

To facilitate a reverberant speech recognition, we previously developed the reverberant criteria RSR-D_n [6], which are used to estimate ASR performance on the basis of the D value calculated from the ISO3382 acoustic parameters [7], 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 of an impulse response. The D value represents the acoustic clarity and is calculated using Eq. (1).

$$D_n = \int_0^n h^2(t)dt / \int_0^\infty h^2(t)dt, \tag{1}$$

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

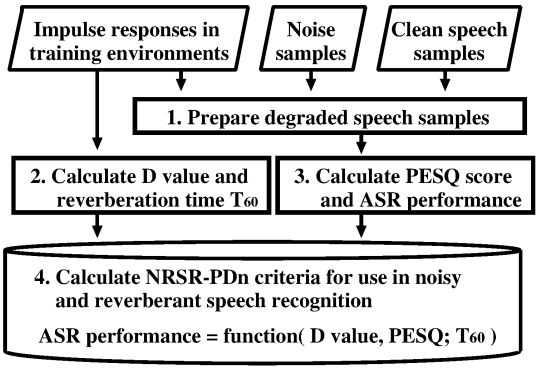


Fig. 1. Design of NRSR-PD_n criteria for use in noisy and reverberant speech recognition.

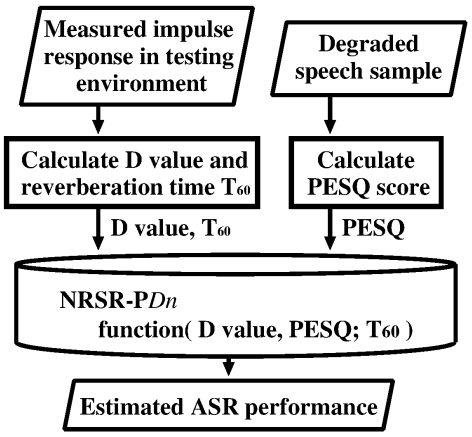


Fig. 2. Estimation of ASR performance using NRSR-PD_n criteria.

III. Design of the Proposed Noisy and Reverberant Criteria NRSR-PD_n

In this section, Noisy and Reverberant Speech Recognition with the PESQ and the D_n (NRSR-PD_n) are proposed as new noisy and reverberant criteria for ASR performance. The NRSR-PD_n are designed in four steps, as illustrated in Fig. 1. They are designed on the basis of the relationships among the D value, the PESQ score, and noisy and reverberant ASR performance. In particular, the multiple-regression analysis is used to design the NRSR-PD_n criteria based on correlation with these calculated value.

Step 1: Prepare degraded speech samples

First, degraded speech samples are prepared as training data to design the NRSR-PD_n criteria used for estimating ASR performance in noisy and reverberant environments in following five steps.

1) Measure many impulse responses in a number of reverberant environments.
2) Create both real and artificial noise samples.
3) Measure speech samples in a clean environment.
4) Create reverberant speech samples which are convolved with 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 with Eq. (1) and the reverberation time T_{60}. Schroeder [8] developed a basic method of measuring reverberation by itegrating the square of the reverberation's impluse responses. The reverberation time is easily measured with his method and is derived on the basis of Eq. (2) with impulse response $h(\lambda)$.

$$< y_d^2(t) > = N \int_t^\infty h^2(\lambda) d\lambda, \tag{2}$$

where $< >$ is the ensemble average and N is the power of the unit frequency of random noise. The reverberation time for a reverberation curve $< y_d^2(t) >$ is the time which 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 with a speech recognition engine, and the PESQ score is calculated using clean and degraded speech samples as described in Sec. II-A.

Step 4: Perform multiple-regression analysis with the D value, the PESQ score, and ASR performance

Finally, multiple-regression analysis is used to design the NRSR-PD_n criteria using the D value and the PESQ score and the ASR performance. The NRSR-PD_n criteria are represented from Eq. (3).

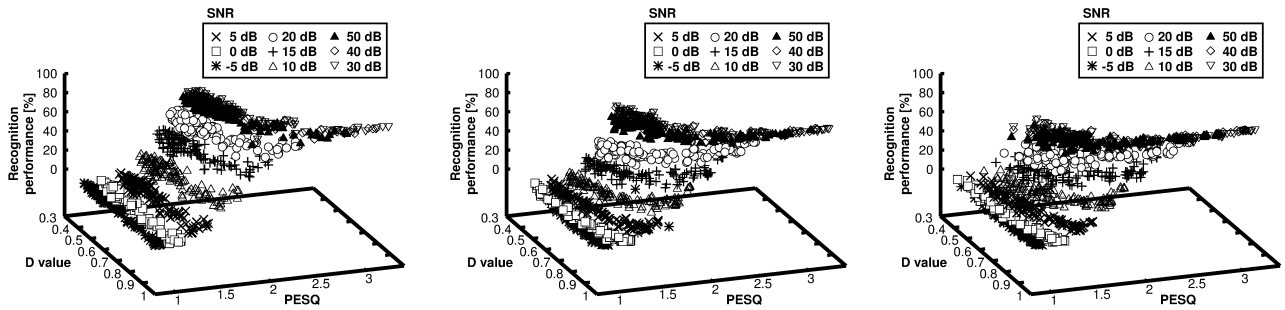
$$y(x_1, x_2; x_3) = a x_1 + b x_2 + c, \tag{3}$$

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 minimum error of the root mean square in the multiple-regression analysis. Moreover, the NRSR-PD_n criteria are designed in each reverberant environment.

IV. Use of NRSR-PD_n to Estimate ASR Performance

The NRSR-PD_n criteria are used to estimate ASR performance in noisy and reverberant environments as illustrated in Fig. 2. As shown in Fig. 2, we can estimate the ASR performance with the NRSR-PD_n in noisy and reverberant environments in just three steps.

1) Measure an impulse response and the degraded speech samples in a test environment.
2) Calculate T_{60} and D value with measured impulse response, and the PESQ score with the degraded speech samples.
3) Estimate the ASR performance using the calculated T_{60}, D value, PESQ score, and the NRSR-PD_n.



(a) Japanese style room (T_{60}=450 ms) (b) Conference room (T_{60}=600 ms) (c) Lift station (T_{60}=850 ms)

Fig. 3. Relationships among D value, PESQ score, and ASR performance in reverberant environments with the factory noise.

TABLE I
EVALUATION CONDITIONS

Speech	ATR phoneme balance 216 words [9]
Speakers	Two females and two males
Decoder	Julius rev. 4.2.1 [10]
HMM	IPA monophone model
Frame length	25 ms (Hamming window)
Shift length	10 ms
Noise	White noise, pink noise, factory noise, human speech like noise (HSLN)

V. EVALUATIONS

We designed the proposed criteria to estimate the noisy and reverberant ASR performance. For estimation of the ASR performance, we used an ATR phoneme-balanced set as the speech samples that consists of 216 isolated Japanese words [9] uttered by four speakers (two females and two males). The recordings were conducted with 16 kHz sampling and 16 bit quantization. All impulse responses were measured for distances rainging between 100 $\sim$ 5,000 mm. A border time (n = 20 ms) of the D in Eq. (1) was used on the basis of a previous study [6]. The conditions of the analysis and recognition processes are also summarized in Table I. Since ASR performance greatly varies with the recognition task, the NRSR-PD_n design and performance estimation were conducted using the same recognition task. In other words, the NRSR-PD_n have to be trained for each different recognition task.

A. Results of NRSR-PD_n design

We designed the NRSR-PD_n in three reverberant environments such as japanese style room (T_{60} = 400 ms, 72 RIRs), conference room (T_{60} = 600 ms, 120 RIRs), and lift station (T_{60} = 850 ms, 120 RIRs). Figure 3 (a)$\sim$(c) shows the relationships among the D value, the PESQ score, and the ASR performance for such three reverberant environments with the factory noise. To design the NRSR-PD_n, we conducted multiple-regression analysis in each reverberant environment as described in Sec. III Step 4. The correlation coefficients obtained by conducting the regression analysis are shown in Table II. As a result of Table II, we confirmed that the NRSR-PD_n are the suitable criteria for estimation of noisy and reverberant ASR performance in comparison with conventional criteria since correlation coefficients with the NRSR-PD_n are higher than that with conventional criteria in all noisy and reverberant environments.

TABLE II
CORRELATION COEFFICIENTS

Reverberant criteria **RSR-D_n**	Reverberation time (T_{60}) 450 ms	600 ms	850 ms
White noise	0.63	0.87	0.89
Pink noise	0.74	0.85	0.87
Factory noise	0.63	0.81	0.82
HSLN	0.72	0.87	0.90

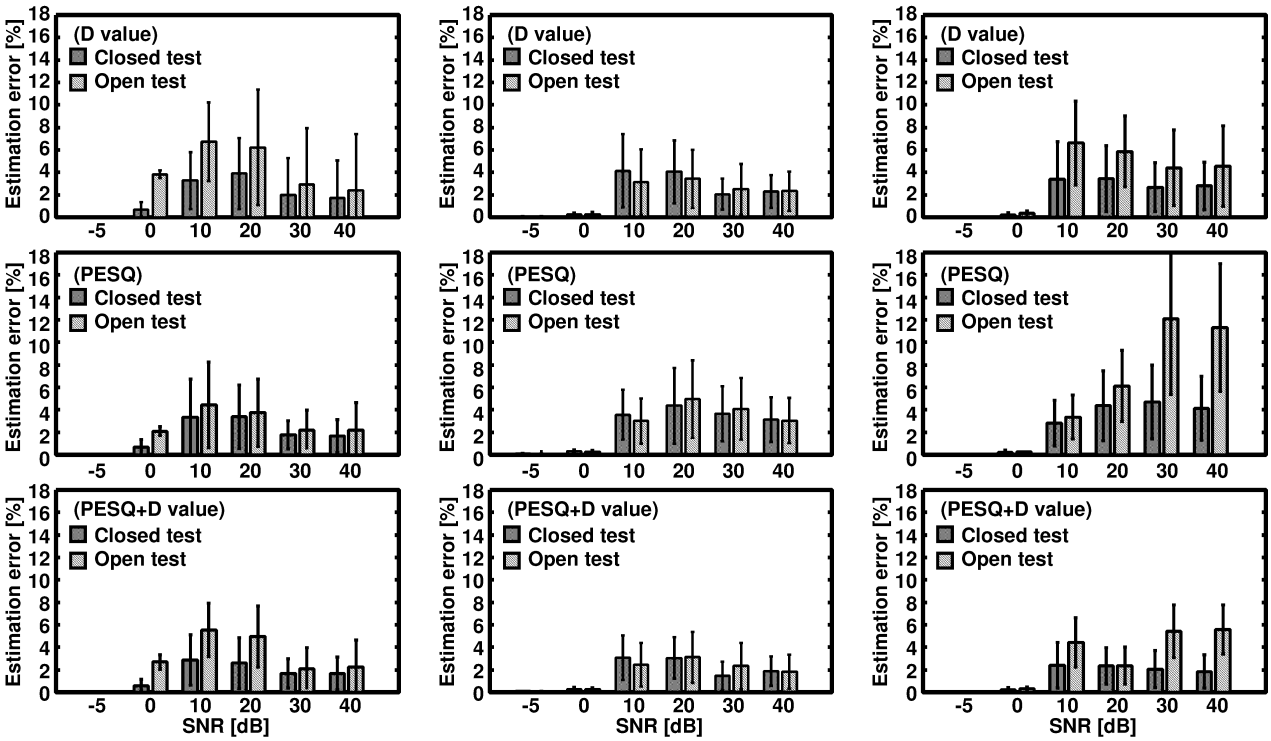
Noisy criteria **PESQ score**	Reverberation time (T_{60}) 450 ms	600 ms	850 ms
White noise	0.69	0.90	0.90
Pink noise	0.66	0.86	0.88
Factory noise	0.64	0.80	0.82
HSLN	0.66	0.89	0.91

Proposed criteria **NRSR-PD_n**	Reverberation time (T_{60}) 450 ms	600 ms	850 ms
White noise	**0.80**	**0.91**	**0.92**
Pink noise	**0.80**	**0.90**	**0.91**
Factory noise	**0.79**	**0.84**	**0.85**
HSLN	**0.77**	**0.92**	**0.93**

B. Results of ASR performance estimation

The performance of noisy and reverberant speech recognition was estimated using closed and open tests in the three test environments such as laboratory (T_{60} = 450 ms, 72 RIRs), corridor (T_{60} = 600 ms, 120 RIRs), and standard stairs (T_{60} = 850 ms, 56 RIRs). In the closed test, we estimated the ASR performance for an known reverberation condition using the NRSR-PD_n designed in the same environment. On the other hand, in the open test, we estimated the recognition performance for an unknown reverberant condition using the NRSR-PD_n designed in the different environment that had the same reverberation time as the test environment. In this experiment, we conducted estimation of the ASR performance with each reverberant speech recognition with RSR-D_n and PESQ score as conventional methods. Figure 4 (a)$\sim$(c) shows the average estimation error for reverberant environments with the factory noise. The average estimation error (ASE) is represented from Eq.(4).

$$ASE = \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y_n}|, \tag{4}$$



(a) Laboratory (T_{60}=450 ms) (b) Corridor (T_{60}=600 ms) (c) Standard stairs (T_{60}=850 ms)

Fig. 4. Average estimation error in reverberant environments with the factory noise.

where y_n and $\hat{y_n}$ represent the actual and estimated ASR performance under the utterance condition n respectively. N is the number of utterance conditions in each noisy and reverberant environment. The ASE degrades under the condition of accurate estimation of the ASR performance. The results showed that ASE was less than 6 % for all environments by using proposed criteria. Moreover, estimation performance with NRSR-PD_n was better than that with either D value or PESQ in most noisy and reverberant conditions. Also, there is same tendency in these reverberant environments with other kinds of noise. This means that the NRSR-PD_n provide accurate estimation results, making them a particular strong candidate for use in recognizing noisy and reverberant speech.

VI. Conclusions

This paper has described a method for estimating ASR performance in noisy and reverberant environments based on NRSR-PD_n. We first designed the NRSR-PD_n using the relationships among the D value, the PESQ, and the ASR performance. We then estimated the ASR performance in noisy and reverberant environments with the NRSR-PD_n. Experiments conducted in actual environments confirmed that our proposed criteria provide accurate estimations, which makes them well suited for use in recognizing noisy and reverberant speech. In future work, we intend to optimize coefficients of the multiple-regression analysis to more accurately estimate noisy and reverberant ASR performance. Moreover, we try to conduct estimation experiments of speech recognition performance in different speech recognition conditions.

VII. Acknowledgements

This work was partly supported by a Grant-in-Aid for Scientific Research funded by MEXT and a Grant-in-Aid for JSPS Fellows funded by JSPS.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, pp. 113–120, 1979.

[2] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Robust speech recognition based on space diversity taking room acoustics into account," *Institute of Electronics, Information, and Communication Engineers*, vol. J83-DII, pp. 2448–2456, 2000.

[3] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Trans. on ASLP*, vol. 14, pp. 2006–2013, 2006.

[4] H. Sun, L. Shue, and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," *IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1, pp. 865–868, 2004.

[5] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codes," ITU-T Rec. P.862, 2001.

[6] T. Fukumori, M. Morise, and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria RSR-Dn with acoustic parameters," *INTERSPEECH 2010*, pp. 562–565, 2010.

[7] "ISO3382: Acoustics measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization, 1997.

[8] M. R. Schroeder, "New method of measuring reverberation time," *JASA*, vol. 37, pp. 409–412, 1965.

[9] K. Takeda, Y. Sagisaka, and S. Katagiri, "Acoustic-phonetic labels in a japanese speech database," *European Conference on Speech Technology*, vol. 2, pp. 13–16, 1987.

[10] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine," *European Conference on Speech Communication and Technology*, pp. 1691–1694, 2001.