



Investigations into Early and Late Reflections on Distant-Talking Speech Recognition Toward Suitable Reverberation Criteria

Takanobu NISHIURA, Yoshiki HIRANO, Yuki DENDA, and Masato NAKAYAMA

College of Information Science and Engineering
Ritsumeikan University, Kusatsu, Japan

{nishiura@is, rs037021@se, gr021052@se, gr020040@se}.ritsumei.ac.jp

ABSTRACT

Reverberation-robust speech recognition has become very important in the recognition of distant-talking speech. However, as no common reverberation criteria for the recognition of reverberant-speech have been proposed, it has been difficult to estimate this. We have thus focused on a reverberation criterion for the recognition of distant-talking speech. The reverberation time is generally currently used as a reverberation criterion for the recognition of distant-talking speech. This is unique and does not depend on the position of the source in a room. However, distant-talking speech recognition greatly depends on the location of the talker relative to that of the microphone and the distance between them. We investigated a suitable reverberation criterion with the ISO3382 acoustic parameters for distant-talking speech recognition to overcome this problem. We first calculated distant-talking speech recognition with early and late reflections based on the impulse response between the talker and microphone. As a result, we found that early reflections within about 12.5 ms from the duration of direct sound contributed slightly to distant-talking speech recognition in non-noisy environments. We then evaluated it based on ISO3382 acoustic parameters. We consequently confirmed that the ISO3382 acoustic parameters are strong candidates for the new reverberation criteria for distant-talking speech recognition.

INDEX TERMS: Distant-talking speech recognition, Reverberation criterion, Acoustic reflection, Room acoustics

1. INTRODUCTION

The recognition of distant-talking speech has rapidly improved in recent years, because many novel speech-recognition techniques have been proposed that are robust against noise and reverberance. The signal to noise ratio (SNR) is generally used as a common criterion in speech-recognition techniques that are robust against noise. SNR is an effective noise criterion for estimating the recognition of speech in noisy environments. As an algorithm based on the perceptual evaluation of speech quality (PESQ) [1] has also been proposed to achieve the same target, we can roughly estimate the recognition of speech in noisy environments. However, no common reverberation criteria have been proposed to attain robust reverberant-speech recognition. It has therefore been difficult to estimate the recognition of reverberant speech. The reverberation time, $T_{[60]}$, [2] is currently generally used to recognize distant-talking speech as a reverberation criterion. It is unique and does not depend on the position of the source in a room. However, distant-talking speech recognition greatly depends on the location of the talker relative to that of the microphone and the distance

between them. Therefore, $T_{[60]}$ is unsuitable for measuring the recognition of distant-talking speech. We propose a new reverberation criterion for measuring the recognition of distant-talking speech to overcome this problem. We investigated a suitable reverberation criterion to enable distant-talking speech to be recognized. We first calculated automatic speech recognition with early and late reflections based on the impulse response between a talker and the microphone. We then evaluated it based on ISO3382 acoustic parameters [3].

2. CONVENTIONAL REVERBERATION CRITERION FOR RECOGNITION OF DISTANT-TALKING SPEECH

2.1. Reverberation time, ($T_{[60]}$)

2.1.1. Reverberation time based on theory of room acoustics

Reverberation time [2] is the most fundamental concept for evaluating indoor acoustical fields and is a parameter that expresses the duration of sound. Reverberation time is the time required for a sound in a room to decay by 60dB (called $T_{[60]}$). As the theory assumes a diffusible sound field in a room, the effect does not change even if sound-absorbing material is placed in any position in the room. The reverberation time is constant for all positions of the sound source and the microphone in the room. However, it alone is insufficient as the criterion for the recognition of distant-talking speech because this depends on the distance between a talker and the microphone in the same environment.

2.1.2. Method of measuring reverberation time

Schroeder developed a basic method [2] of measuring reverberation by integrating the square of the impulse response. The reverberation time is easily measured with his method. The reverberation curves are derived from Eq. (1) with impulse response $h(t)$.

$$\langle y_a^2(t) \rangle = N \int_1^\infty h^2(t) dt \quad (1)$$

where $\langle \rangle$ 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.

2.2. Total Amplitude of Reflection Signals (A value)

The A value [4] is used as a reverberation criterion as often as reverberation time for the recognition of distant-talking speech. It

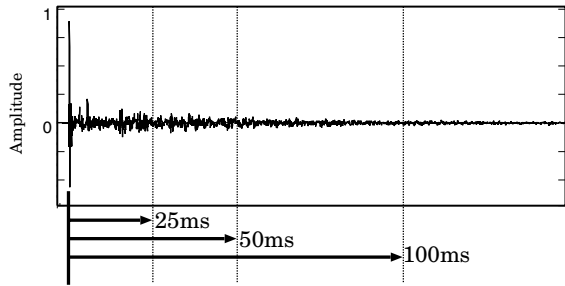


Figure 1: Example of impulse response extraction for three evaluation periods.

is derived from Eq. (2).

$$A = \sqrt{\int_{\epsilon}^t h^2(t)dt / \int_0^{\epsilon} h^2(t)dt}, \quad (2)$$

where ϵ represents the duration of direct sound within approximately 3 – 5 ms. The A value indicates the energy ratio between direction and reflections on the captured signal, and it depends on the distance between the talker and microphone in the same room. However, it does not distinguish early reflections from late reverberations.

3. RELATION BETWEEN EARLY REFLECTIONS AND DISTANT-TALKING SPEECH RECOGNITION

We define early reflections as high-correlation signals with direct sound, especially those that arrive within a few milliseconds of direct sound in this paper. Late reverberations are defined as low or no correlation signals with direct sound, especially those that arrive over a few milliseconds after direct sound.

3.1. Early reflections in distant-talking speech recognition

Early reflections, especially those that arrive within 50 ms of direct sound, are useful to humans when listening to speech [4]. However, the higher the reflection energy becomes, the less effectively speech is recognized, subject to clean acoustic phoneme models. However, it was previously unclear whether early reflections were useful for recognizing speech in the recognition of distant-talking speech because the reverberation time and A values were used as reverberation criteria. We evaluated what relation there was between early reflections and the recognition of distant-talking speech on the basis of impulse responses between a talker and the microphone to develop more suitable reverberation criteria for distant-talking speech recognition,

3.2. Evaluation Experiment

3.2.1. Recording conditions

We measured impulse responses in actual environments. The impulse responses were measured in $T_{[60]} = 0.2$ and 0.7 s environments, subject to distances of 0.1 and 0.5 m between the talker and the microphone. A time stretched pulse [5] was used to measure the impulse responses. The recordings were made with 16 kHz sampling and 16 bit quantization.

Table 1: Experimental conditions for speech recognition.

HMM	IPA monophone model (Gender-dependent)
Feature vectors	12 orders MFCC + 12 orders Δ MFCC + 1 order Δ Power
Frame length	25 ms. (Humming window)
Frame interval	10 ms.

3.2.2. Experimental conditions

An ATR phoneme-balanced set [6] was employed as the speech samples that were made up of 216 isolated Japanese words that were uttered by 14 speakers (7 females and 7 males). We evaluated the relation between early reflections and the recognition of distant-talking speech by convolving speech samples and impulse responses. Impulse responses were extracted during each period of evaluation as shown in Fig. 1 to evaluate the relation between reflections and the recognition of distant-talking speech for all evaluation periods. Table 1 lists the experimental conditions for speech recognition.

3.2.3. Experimental results

Figure 2 plots the experimental results, where $T_{[60]}$ is the reverberation time, $Dis.$ is the distance between the talker and the microphone, and WRR is the word recognition rate. The A value is the energy ratio between the direction and reflections from the duration of direct sound to each evaluation period. We confirmed that early reflections within about 12.5 ms after direct sound only contributed slightly to the recognition of distant-talking speech in quiet environments on the basis of these results, although early reflections within about 50 ms from the duration of direct sound contributed greatly to human hearing ability. We also confirmed that late reflections over about 12.5 ms after direct sound decreased the recognition of distant-talking speech. The higher the A value becomes in Fig. 2, the greater the number of reflections. However, we confirmed that the ability to recognize speech can be improved despite a higher A value. Therefore, we again found that suitable reverberation criteria were necessary for the recognition of distant-talking speech on the basis of our evaluation experiments.

4. TOWARD SUITABLE REVERBERATION CRITERIA

4.1. ISO3382 acoustic parameters

ISO3382 [3] proposed parameters for measuring room acoustics. The ISO3382 standard defines measurements of reverberation times in rooms with reference to other acoustical parameters. Acoustics parameters are classified into four categories on the basis of this standard:

1. Sound level
2. Reverberation time
3. Balance between early and late arriving energies (Clarity, Definition, and Center time)
4. Binaural parameters (IACC, Lateral Fraction)

These parameters are directly calculated based on measured impulse responses. We focused on the third category (balance between early and late arriving energies), because it has a high correlation with clarity and the reverberance of the acoustic sound field.

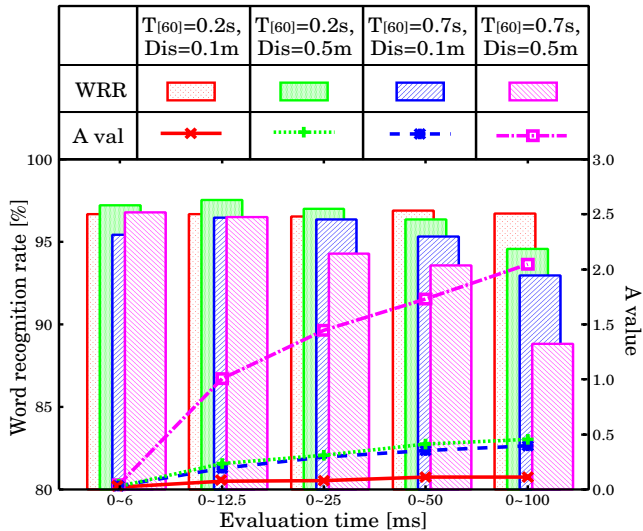


Figure 2: Effects of early reflections on distant-talking speech recognition.

4.2. Balance between early and late arriving energy

“Clarity,” “Definition,” and “Center time” are defined as the acoustic parameters of balance between early and late arriving energies in the ISO3382 standard. The C value expresses the clarity of acoustics and is derived from Eq. (3). The D value expresses the definition of acoustics and is derived from Eq. (4). Center time expresses the center time based on a square impulse response and is derived from Eq. (5).

$$C_{t_1} = 10 \log_{10} \left(\int_0^{t_1} h^2(t) dt / \int_{t_1}^{\infty} h^2(t) dt \right), \quad (3)$$

$$D_{t_1} = \int_0^{t_1} h^2(t) dt / \int_0^{\infty} h^2(t) dt, \quad (4)$$

$$T_s = \int_0^{\infty} t h^2(t) dt / \int_0^{\infty} h^2(t) dt, \quad (5)$$

where, t_1 is the border time between early and late arriving energies. The C value measure and the condition of music are highly correlated with $t_1 = 80$ ms, and the D value measure and the condition of speech are highly correlated with $t_1 = 50$ ms based on the ISO3382 standard. In addition, the larger T_s becomes, the more late reverberations there are.

4.3. Evaluation experiments

We evaluated the relation of the ISO3382 acoustic parameters and the recognition of distant-talking speech to determine suitable reverberation criteria. We also compared all acoustic parameters with regression analysis based on ordinary least squares.

4.3.1. Recording conditions

We measured impulse responses in six environments, i.e., a “Living room” (LV, $T_{[60]} = 250$ ms), a “Conference room” (CR,

Table 2: Regression coefficients for all acoustic parameters.

	LV	CR	CD	PB	EV	SS	AVE.
A	0.81	0.93	0.79	0.89	0.81	0.69	0.82
C_{80}	0.82	0.86	0.85	0.96	0.91	0.89	0.88
D_{50}	0.73	0.91	0.93	0.95	0.92	0.91	0.89
T_s	0.82	0.87	0.95	0.97	0.91	0.77	0.88

$T_{[60]} = 350$ ms), a “Corridor” (CC, $T_{[60]} = 600$ ms), a “Prefabricated bath” (PB, $T_{[60]} = 700$ ms), an “Elevator hall(lobby)” (EV, $T_{[60]} = 700$ ms), and “Standard stairs” (SS, $T_{[60]} = 800$ ms). The distances between the talker and the microphone were between 10 cm and 500 cm in all environments. We measured 307 impulse responses in all. A time-stretched pulse was used to measure the impulse responses as in Section 3.2.1. The recordings were conducted with 16 kHz sampling and 16 bit quantization.

4.3.2. Experimental conditions

The speech recognition experiments were conducted under the same conditions as in Section 3.2.2. An ATR phoneme-balanced set was employed as the speech samples that were made up of 216 isolated Japanese words that were uttered by 14 speakers (7 females and 7 males). Table 1 lists the experimental conditions for speech recognition.

4.3.3. Experimental results

Figures 3-6 plot the experimental results. The horizontal axes represent the word recognition rate, and the vertical axes represent the A value, C_{80} , D_{50} , and T_s . Table 2 lists the results for all acoustic parameters with regression analysis based on ordinary least squares. We found that the ISO3382 acoustic parameters were strong candidates for the reverberation criteria based on these results because the regression coefficients for the C , D , and T_s values were higher than that for the A value.

4.3.4. Discussion

The results from the evaluation experiments proved the ISO3382 acoustic parameters were strong candidates as the reverberation criteria for the recognition of distant-talking speech. We therefore assumed that the early reflection signal, which is the most important factor in the recognition of reverberant speech, does not depend on the total amount of reflection, but on the balance between early and late arriving energies. Our next challenge is to examine the use of suitable reverberation criteria based on the C_{t_1} and D_{t_1} values of ISO3382 acoustic parameters with a suitable border time, t_1 , between early and late arriving energies to prove this hypothesis. If suitable border time t_1 can be estimated, we can easily estimate the recognition of reverberant speech with one impulse response between the talker and microphone.

5. CONCLUSIONS

We evaluated the relation between early reflections and the recognition of distant-talking speech toward suitable reverberation criteria to enable distant-talking speech to be recognized. As a result, we found that early reflections within about 12.5 ms from the duration of direct sound contributed slightly to the recognition of distant-talking speech in non-noisy environments. We also

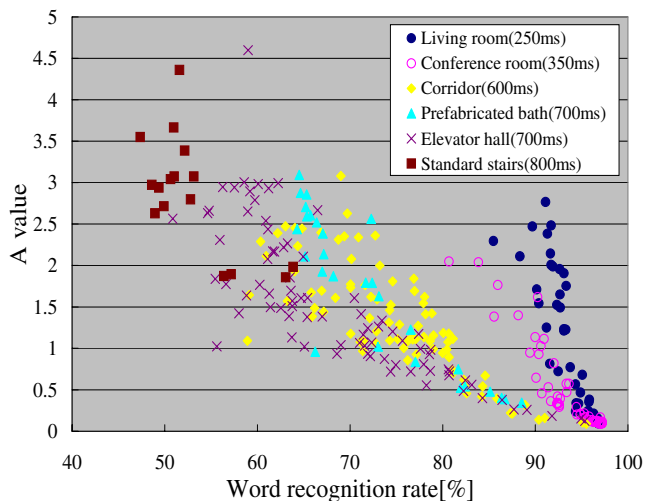


Figure 3: *A value*.

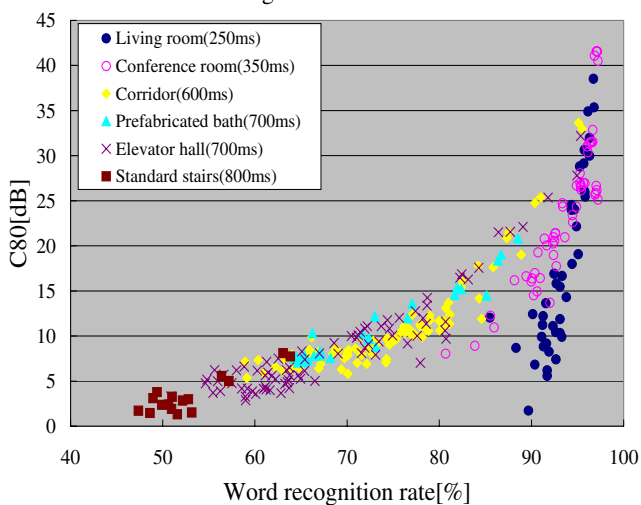


Figure 4: *Clarity (C₈₀)*.

confirmed that the C and D values of ISO3382 were strong candidates for the reverberation criteria of distant-talking speech recognition as a result of evaluation experiments with ISO3382 acoustic parameters. We also intend to investigate suitable reverberation criteria in the frequency domain for distant-talking speech recognition with the Modulation Transfer Function (MTF) [7] in future work.

6. ACKNOWLEDGEMENTS

This work was partly supported by The Leading Project "e-Society" and Grants-in-Aid for Scientific Research No.17700216 and No.17200014, funded by The Ministry of Education, Culture, Sports, Science and Technology of Japan. This work was also partly supported by the "Ono Acoustics Research Fund".

7. REFERENCES

[1] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise condi-

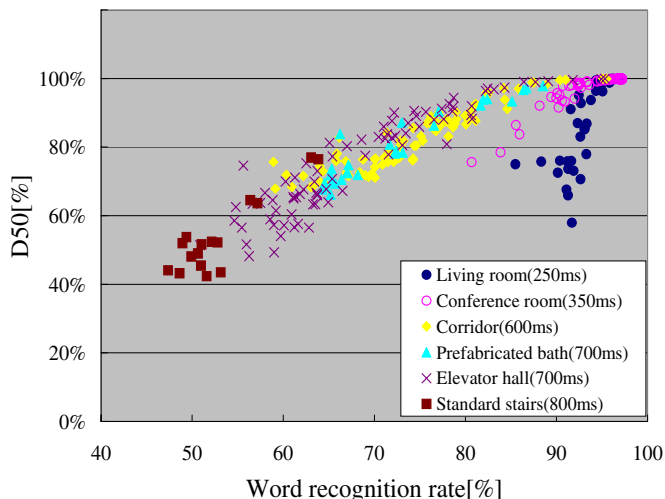


Figure 5: *Definition (D₅₀)*.

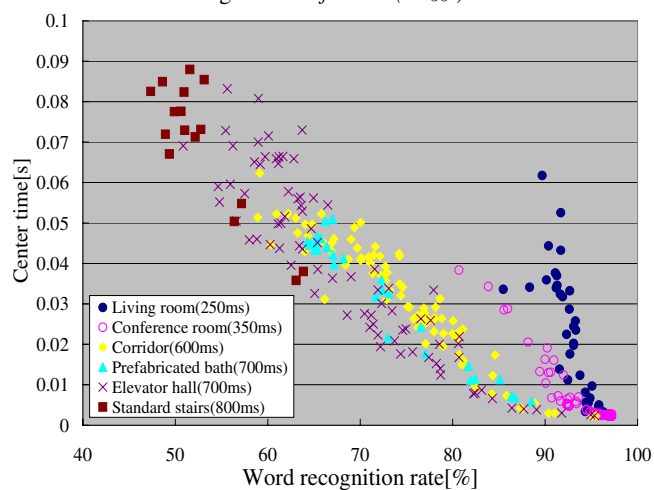


Figure 6: *Center time (T_s)*.

tions using objective quality measures and artificial voice," IEEE Trans. on ASLP, Vol. 14, No. 6, pp. 2006-2013, Nov. 2006.

[2] M. R. Schroeder, "New method of measuring reverberation time," J. Acoust. Soc. Am., Vol. 37, pp. 409, 1965.

[3] ISO3382: Acoustics- Measurement of the reverberation time of rooms with reference to other acoustical parameters. International Organization for Standardization, 1997.

[4] H. Kuttruff, "Room Acoustics," Spon Press, 2000.

[5] Y. Suzuki, F. Asano, H.-Y. Kim, and Toshio 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.

[6] K. Takeda, Y. Sagisaka, and S. Katagiri, "Acoustic-Phonetic Labels in a Japanese Speech Database," Proc. European Conference on Speech Technology, Vol. 2, pp. 13-16, Oct. 1987.

[7] T. Houtgast, H.J.M. Steeneken, and R. Plomp "Predicting speech intelligibility in room acoustics", Acustica, Vol. 46, pp. 60-72, 1980.