euronoise

Evaluation of acoustic environments using deteriorated speech sound

Yoshiki Nagatani, Takefumi Sakaguchi and Hiroshi Hosoi

Nara Medical University, 840 Shijo-cho, 634-8522 Kashihara, Japan
naramed-u@nagatani.ne.jp

Aged or hearing-impaired people demand better acoustical environments for higher QOL. Many methods of evaluating acoustical environments with echoes focusing on speech quality have been developed. However, since these methods mainly focus on speech quality in bad acoustic conditions, they are not suitable for evaluations of acoustic environments with high intelligibilities such as normal houses or public facilities for aged people. Therefore, we proposed a new evaluation method using deteriorated speech sounds. In this method, signal-processed speech sounds are presented to subjects in target acoustic environments. In this study, in order to simulate the architectural acoustic environments, Japanese monosyllabic speech sounds were convoluted by the impulse responses of room reverberations, and were presented through a headphone. As a result, it was shown that this new method could detect the small differences of acoustic environments, which the conventional methods could hardly evaluate.

1 Introduction

Aged or hearing-impaired people demand better acoustical environments for higher QOL. Many methods of evaluating acoustical environments focusing on speech quality have been developed. However, since these methods mainly focus on speech quality in bad acoustic conditions (e.g. environments with huge noise or long reverberation), they are not suitable for evaluations in commonplace environments such as normal houses or public facilities for aged people.

For instance, the score of D50 value (deutlichkeit) [1] and speech transmission index (STI) [2, 3, 4] are saturated and not sensitive enough to distinguish such environments. Moreover, perceptual tests on speech intelligibility cannot clarify the differences of room environments because the intelligibility scores always reach almost 100 percent in ordinary room environments.

Morimoto et al. and Sato et al. investigated the evaluating method using the index of "listening difficulty ratings" [5, 6]. The method requires subjects to answer how it is difficult to listen the speech sound in each environment. They claimed that the "listening difficulty" index shows good correlation to the STI. However, this method requires many well-trained subjects to evaluate an acoustic environment.

Therefore, we propose a new evaluation method using deteriorated speech sounds. In our method, signal-processed speech sounds are presented to subjects in target acoustic environments to evaluate. In order to simulate the architectural acoustic environments, Japanese monosyllabic speech sounds are convoluted by the impulse responses of room reverberations, and are presented through a headphone. In this report, we aimed at clarifying the feasibility of our new method focusing on detecting the small difference of acoustic environments, which the conventional methods could hardly evaluate.

2 Materials and Methods

2.1 Subjects

Seven subjects were employed (23-29 years old). They were native-Japanese speakers with normal hearing levels.

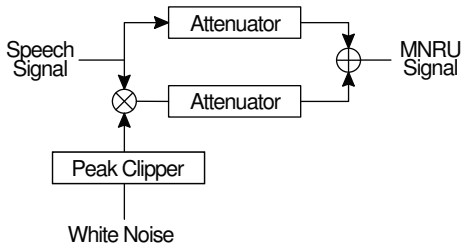

Figure 1: Block diagram of MNRU (modulated noise reference unit) processing. The output is the mixture of the original sound and the white noise whose envelope is equal to that of original sound.

2.2 Speech materials

The 67 Japanese monosyllables pronounced by a professional female speaker included in the speech sound dataset "FW03"[7] were used as speech materials. The monosyllables used in this study consist of 44 resonant sounds, 18 voiced consonants, and 5 semi-voiced consonants. The sound levels of the original speech materials were calibrated based on their perceptual loudness according to the calibration database constructed by Nagatani et al.[8]

2.3 Signal processing of speech materials

For the deterioration of speech sounds, the MNRU (modulated noise reference unit) processing [9] was used. Figure 1 shows the block diagram of the MNRU processing. The output is the mixture of the original sound and the white noise whose envelope is equal to that of original sound. The instantaneous S/N ratio of MNR signal are constant at any slight time, therefore, the qualities of the stimuli are easy to be controlled by choosing the quality factor (S/N ratio) of MNRU-processing. The quality factor used in this study was 0 dB.

2.4 Acoustic environments to reconstruct

In order to simulate the real acoustic environments, the nine types of impulse responses were picked up from the sound source database for environmental/architectural acoustics, "SMILE 2004" [10]. In addition to these, the anechoic condition was also used. The experiments were performed using these ten environments. Table 1 shows the physical properties of these 10 acoustic environments: The speech transmission indexes, D50 values,

Table 1: Physical properties of acoustic environments used in this study. Speech transmission indexes, D50 values, and reverberation times are derived form the impulse responses. The STI insists that the best environment may have a score of 1.0, which is anechoic condition.

No.	Environments	STI	D50 [%]	Rev. [s]
1	Anechoic room	1.00	100.0	0.00
2	Normal house made with wood	0.85	92.7	0.35
3	Movie theater	0.84	92.8	0.29
4	Classroom	0.70	77.9	1.06
5	Multi-purpose hall	0.57	56.5	1.12
6	Hall for lecture meetings 1	0.56	51.7	1.22
7	Hall for classical music 1	0.56	60.0	2.29
8	Hall for lecture meetings 2	0.54	46.5	1.25
9	Hall for classical music 2	0.47	33.1	1.21
10	Event hall	0.43	28.5	2.64

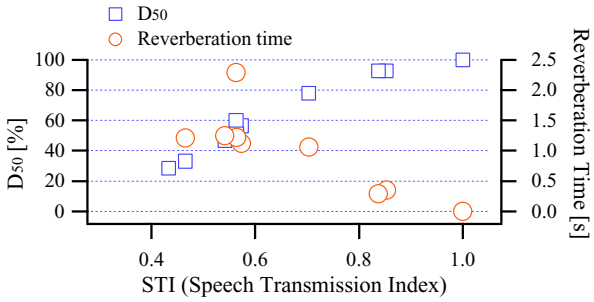

Figure 2: Relationship between the D50 value, the reverberation time and STI of each acoustic environment. The D50 values show good correlation to the STI. The reverberation time doesn't correlate to the STI nor D50 values in some cases.

and reverberation times at 500 Hz derived form the impulse responses are shown. Figure 2 shows the relationship between the D50 value, the reverberation time and STI of each acoustic environment. The STI varies from 0.4 to 1.0. The D50 values show good correlation to the STI.

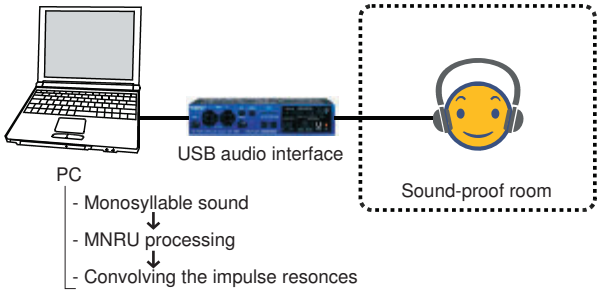

Figure 3: Measurement system. The Japanese 67 monosyllables digitally signal-processed by PC are presented to the subjects in a random order through a headphone in a sound-proof room via an USB audio interface.

2.5 Procedure

In order to reduce the effect of the differences of hearing abilities of each subject, the sound level of stimuli were set to 30 dBSL, where 0 dBSL was the hearing threshold of the intermittent 1 kHz sound. (The intermittent sound were edited using the calibration sound included in FW03 dataset, whose equivalent continuous sound pressure level is equalized to the speech sounds in FW03.) In this study, no ambient noises were presented to the subjects.

The 67 signal-processed speech sounds were presented in a random order to the subjects. The stimuli were presented to both ears from a headphone (SONY, MDR-7506) in a sound-proof room. The subjects were asked to write the monosyllables onto the answering sheet as they heard.

Figure 3 shows the measurement system. The impulse responses were convoluted to the speech sounds by PC. The stimuli were played through an USB audio interface (EDIROL, UA-101).

3 Results and Discussions

3.1 Speech intelligibilities without any echoes

The intelligibilities of each subject without any echoes or reverberations are shown in Table 2. The values indicate the ratio of correct answers among the 67 syllables. When the original speech sounds were presented, the intelligibilities of the all subjects were always almost 100%, which means that the scores are saturated so that they are not sensitive enough in those cases. On the contrary, the scores of the MNRU-processed speech sounds were consistently lower than those of original speech sounds. This implies the possibility that the MNRU-processed speech sounds are suited for the precise evaluation of acoustic environments.

3.2 Speech intelligibilities with echoes

The intelligibilities in each acoustic environment were normalized to those in the anechoic condition, whose

Table 2: Speech intelligibilities of each subjects without any echoes. The original (not deteriorated) speech sounds and MNRU-processed speech sounds are presented in a sound-proof room. The scores of MNRU-processed speech sounds were consistently lower than those of original speech sounds.

Subjects	Score [%]	
	Original	MNRU
1	98.5	79.1
2	95.5	79.1
3	95.5	76.1
4	91.0	77.6
5	95.5	76.1
6	97.0	79.1
7	94.0	79.1
Average	95.3	78.0

STI is 1.0. (The normalized values are called "intelligibility ratio" in this study.) The relationship between the STI and the intelligibility ratio is shown in Fig.4. The all intelligibility ratios of the original (not deteriorated) speech sounds are almost 100%, irrespective of the acoustic environments. This fact indicates that the scores are saturated and can not reflect the difference of environments. On the contrary, the intelligibility ratios of MNRU-processed speech sounds were distinguished from 80 to 100% and showed clear differences between environments compared to that of the original speech. This tells us the perceptual tests using MNRU-processed speech sounds successfully distinguish the acoustic environments.

The intelligibility ratios did not correlate with the STIs in several environments. Onaga et al. also claimed that the early energy of the multiple reflections will cause the discrepancy between the intelligibility and STI [11]. The results suggest that our new method can precisely evaluate the acoustic environments with high intelligibilities, where the STI is not suitable.

In this study, all monosyllables had same weights when calculating the intelligibilities, however, the tendencies of subjects' responses were different on each monosyllable. Therefore, we should check if each monosyllable should be weighted for the scoring. We also have to confirm the reproducibility within each subject for the purpose of practical use of this method. We are also interested in the applicability of our new method to the other languages in addition to Japanese.

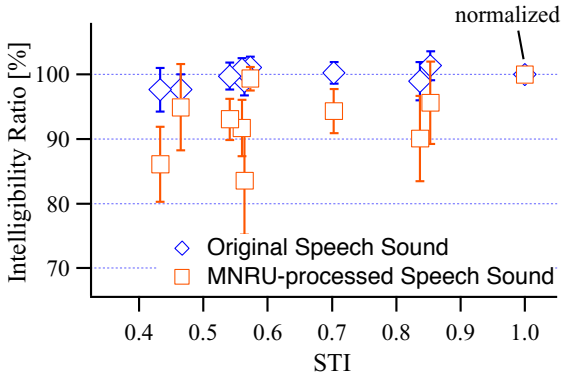

Figure 4: Relationship between the STI and intelligibility ratios in each acoustic environment. The scores in each environment are normalized to that in anechoic condition, whose STI is 1.0. The differences of the scores can be seen only when the MNRU-processed stimuli were presented. The error bars indicate the standard errors of the subjects.

4 Conclusion

We proposed a new method for evaluating the acoustic environments with echoes or reverberation. In this method, the deteriorated speech sounds are presented to subjects in target acoustic environments. As a result, it was shown that our new method could distinguish the small differences of acoustic environments, which the conventional methods could hardly evaluate.

5 Acknowledgments

This work was supported by Grants-in-Aid for Research on Indoor Environmental Medicine of Nara Medical University. This work was also supported by Prof. M. Yanagida and Mr. R. Tachibana of Doshisha University. This work was partly supported by the "Open Competition for the Development of Innovative Technology" of Ministry of Education, Culture, Sports, Science and Technology. The authors would appreciate them.

References

[1] R. Thiele, "Richtungsverteilungs und zeitfolge der schallruckewurfe in raumen," *Acustica, 3*, 291 (1953).

[2] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica, 28*, 66-73 (1973).

[3] M.R. Schroeder, "Modulation transfer functions: definition and measurement," *Acustica, 49*, 179-182 (1981).

[4] IEC 60268-16 Third edition, "Sound system equipment- Part 16: Objective rating of speech intelligibility by speech transmission index" (2003).

[5] M. Morimoto, H. Sato, and M. Kobayashi, "Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces," *J. Acoust. Soc. Am., 116*, 1607-1613 (2004).

[6] H. Sato, M. Morimoto, and M. Wada, "Objective measures for estimating listening difficulty ratings for young and elderly listeners in public spaces," *Proceedings of WESPAC IX*, 1-6 (2006).

[7] S. Amano, K. Kondo, Y. Suzuki, and S. Sakamoto, "Shinmitsudo betsu tango ryoukaidoshiken you onsei dataset (FW03)" (Speech Resources Consortium, National Institute of Informatics, Japan) (2006).

[8] Y. Nagatani, R. Tachibana, T. Sakaguchi, and H. Hosoi, "Loudness calibration of monosyllabic speech sounds in FW03," *J. Acoust. Soc. Jpn.* [submitted, in Japanese].

[9] ITU-T, P810, "Modulated noise reference unit (MNRU)" (1996).

[10] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, and Y. Hidaka, "Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004," *Proc. International Congress on Acoustics*, 1561-1564 (2004).

[11] H. Onaga, Y. Furue, and T. Ikeda, "On the disagreement between speech transmission index (STI) and speech intelligibility," *J. Acoust. Soc. Am. 108*, 2633 (2000).