社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2005-166 (2006-03)

The Effects of Speech-Rate Slowing for Improving Speech Intelligibility in Reverberant Environments

Yuki NAKATA†, Yoshiaki MURAKAMI†, Nao HODOSHIMA†, Nahoko HAYASHI†, Yusuke MIYAUCHI†, Takayuki ARAI† and Kiyohiro KURISU‡.

†Department of Electrical and Electronics Engineering, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554 Japan

‡TOA Corporation, 2-1 Takamatsu-cho, Takarazuka, Hyogo 665-0043, Japan
E-mail: † kee_777@hotmail.com

Abstract This study investigated the effects of speech-rate slowing as a pre-processing technique under reverberant conditions. We conducted a perceptual test using speech-rate slowing with and without steady-state suppression (Arai *et al*., Proc. Autumn Meet. Acoust. Soc. Jpn., pp. 449f, 2001, and, Acoust. Sci. Tech., Vol.23, pp. 229f, 2002.) under several reverberant conditions. We hypothesized that speech-rate slowing with steady-state suppression yields greater improvement in speech intelligibility than simple speech-rate slowing. Our results indicated that simple speech-rate slowing improved speech intelligibility significantly at reverberation time of 2.0 s (from 45.2% to 57.7%) and speech-rate slowing after steady-state suppression significantly improved speech intelligibility at reverberation times of 2.0 s (from 45.2% to 70.2%) and 2.8 s (from 43.5% to 56.0%). Furthermore, speech-rate slowing with steady-state suppression was superior to simple speech-rate slowing for improving speech intelligibility at a reverberation time of 2.0 s (from 57.7% to 70.2%).

Keyword Reverberation, Speech intelligibility, Speech-rate slowing, Steady-state suppression

1. Introduction

Speech can be difficult to perceive in a reverberant environment due to overlap-masking, where the reverberant components of prior speech segments mask successive segments [1, 2]. As the speech energy of the prior segments increases, so does the effect of overlap-masking. To reduce overlap-masking, Arai *et al*. [3, 4] proposed a pre-processing approach known as steady-state suppression. This method reduces overlap-masking by suppressing steady-state portions of speech. Steady-state portions have more energy compared to transitions of speech, and they are less crucial for speech perception. Hodoshima *et al*. [5-7] conducted perceptual tests and showed significant improvements in intelligibility using steady-state suppression at reverberation times (RTs) of 0.8-1.3 s.

Following Bolt and MacDonald's [1] report that speech intelligibility is greatly increased by speaking slowly in a reverberant room, Arai [8] suggested an approach that stretches a speech signal using the time-scale modification technique to decrease the speaking rate. However, stretching a speech signal is not the best solution for improving speech intelligibility in reverberant environments in terms of reducing the amount of overlap-masking. Thus, in the present study, we tried an approach based on Arai [8] that intelligibility can be improved by applying steady-state suppression after decreasing the speaking rate. We conduct here a perceptual test using speech samples processed by speech-rate slowing with and without steady-state suppression under three reverberant conditions.

2. Speech-rate slowing

To decrease the speaking rate, we used Praat [9], which applies the PSOLA (Pitch-Synchronous Overlap and Add) method for time-scale modification. The PSOLA method can modify speech rate without changing the fundamental or formant frequencies of the original speech signal [10].

3. Steady-state suppression

In this study, we adopted the same algorithm for the steady-state suppression method as used in previous studies [5-7,11,12]. This technique first splits an original signal into 1/3-octave bands and the envelope is extracted

in each band. After down-sampling, the regression coefficients are calculated from the five adjacent values of the time trajectory of the logarithmic envelope of each band. Then the mean square for the regression coefficients, *D*, is calculated. This parameter *D* is similar to what that which Furui proposed to measure spectral transition [13]. After up-sampling, we define a portion of speech as steady-state when *D* is less than a given threshold (that is the median in this study). Once a speech portion was considered as steady state, the amplitude of the portion was suppressed. In this study, the speech portion was suppressed to 40% of the original amplitude, as in previous studies [5-7, 11, 12].

4. Experiment

Under three reverberant conditions, we compared the intelligibility of 1) original speech samples, 2) speech samples processed by speech-rate slowing, and 3) speech samples processed first by speech-rate slowing and then by steady-state suppression. We conducted a perceptual test under artificial reverberant environments achieved by convoluting speech samples with impulse responses. RTs of the three impulse responses we used were 2.0 s (Rev1), 2.8 s (Rev2; measured at a lecture hall from the database by Sub Working Group on Research in Speech Transmission Quality of the Architectural Institute of Japan) and 3.6 s (Rev3; measured at St. Ignatius Church). Rev1 was created from Rev2 by multiplying an exponential decay as in a previous study [14]. RT is defined as the time the decay curve of an impulse response decreases to 60 dB below its initial level. We used Early Decay Time (EDT), which is the time taken for the first 10 dB drop of the decay curve, and multiplied it by six to estimate the reverberation time. To calculate RT, an impulse response was first split into octave bands, then the mean RTs were calculated for each band having a center frequency of 500, 1000, 2000 Hz, respectively.

The original speech samples consisted of 14 nonsense consonant-vowel (CV) syllables embedded in a Japanese carrier phrase. The vowel was /a/ and the consonants were /p, t, k, b, d, g, s, ʃ, h, dz, dʒ, tʃ, m, n/. The speech samples were obtained from the ATR speech database of Japanese. The ratio of the root-mean square (RMS) in the carrier phrase to that in the CVs was 1:0.7. Stimuli were original speech samples (Original), speech samples processed by speech-rate slowing (TSM) and speech samples processed by steady-state suppression after speech-rate slowing (TSM+SSS). All were convoluted with each of the three impulse responses used in this study. Figure 1 shows the original speech and the processed speech signals.

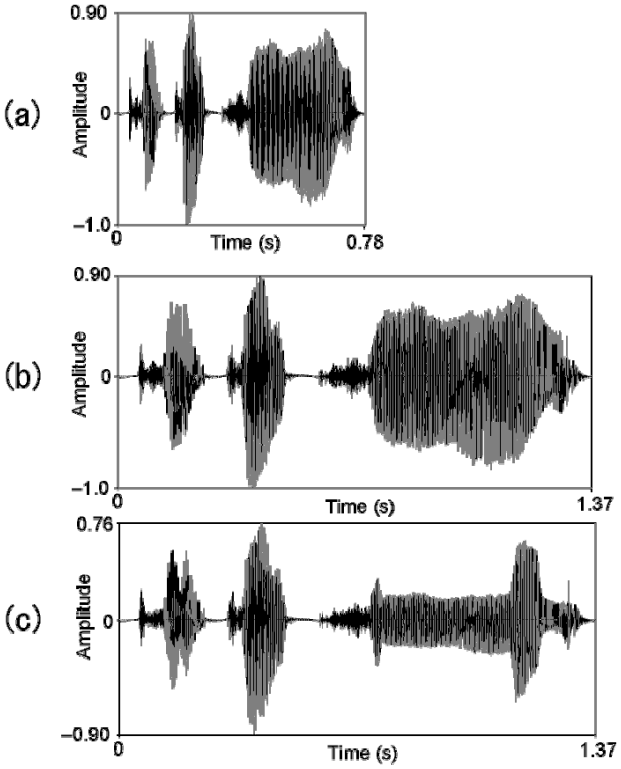

Fig.1: Original and processed waveforms: (a) original speech sample (Original); (b) speech sample processed by speech-rate slowing (TSM); and (c) speech sample processed by steady-state suppression after speech-rate slowing (TSM+SSS).

Twenty-four young normal-hearing subjects (8 males and 16 females, aged 19 to 24 years) participated in the experiment. All were native Japanese speakers.

The experiment was conducted in a soundproof room. Stimuli were presented diotically through headphones (STAX SR-303) connected to a computer. The sound level was adjusted to each subject's comfort level during the training session prior to the experiment. A stimulus was presented in each trial and the subjects were instructed to select one of the 14 CVs displayed on the computer in Kana orthography. The experiment was carried out at each subject's pace. For each subject, 294 stimuli were presented randomly (3 reverberation conditions x 14 CVs x 7 processing conditions). Only three processing conditions are discussed in this paper.

5. Results

Figures 2-4 show the experimental results. Multiple comparison of the three out of seven processing conditions showed significant differences between Original and TSM+SSS ($p < 0.05$) at RT of 2.8 s, Original and TSM ($p < 0.05$) at RT of 2.0 s, and Original and TSM+SSS ($p < 0.01$) at RT of 2.0 s. Also, the difference between TSM and TSM+SSS was significant ($p < 0.05$) at RT of 2.0 s.

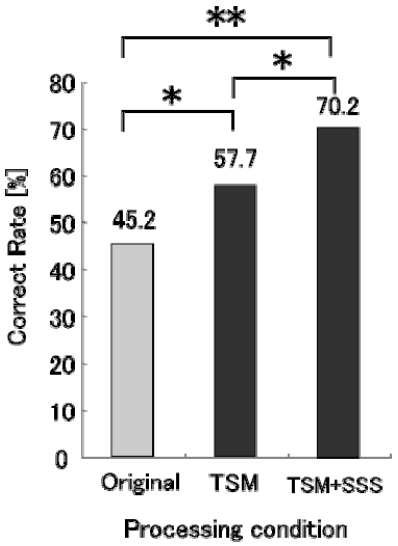

Fig.2: Correct rates (%) at RT of 2.0 s (*: $p < 0.05$, **: $p < 0.01$)

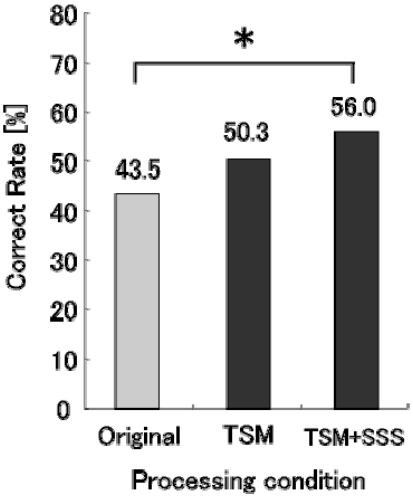

Fig.3: Correct rates (%) at RT of 2.8 s (*: $p < 0.05$)

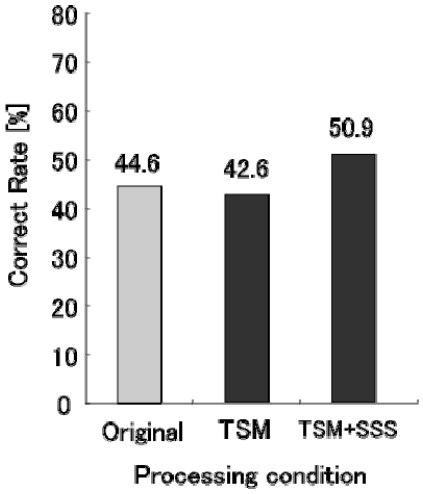

Fig.4: Correct rates (%) at RT of 3.6 s

6. Discussion

The results indicated that speech intelligibility was improved by the speech-rate slowing approach under reverberant conditions. As RT shortened, more positive processing effects for speech intelligibility were obtained under three reverberant conditions. In particular, speech intelligibility was improved by speech-rate slowing after steady-state suppression under moderate reverberant conditions (RTs of 2.0 and 2.8 s). However, speech intelligibility was not improved under the reverberant condition with the longest RT (RT of 3.6 s).

Steady-state suppression after speech-rate slowing was superior to the simple speech-rate slowing approach in terms of speech intelligibility at an RT of 2.0 s (the shortest RT). At an RT of 2.0 s, the correct rate of TSM+SSS was significantly higher than that of TSM. Processing by speech-rate slowing and subsequent steady-state suppression is a superior processing method than by simple speech-rate slowing alone. From this result, it is assumed that overlap-masking was reduced by suppressing the steady-state portions of speech. While previous studies [11,12] found simple steady-state suppression did not yield significant improvements in speech intelligibility at relatively long RTs (more than 2.0 s), the present study showed that a combination of steady-state suppression and speech-rate slowing yielded significantly improved speech intelligibility at RTs of 2.0 and 2.8 s.

7. Conclusions

In this study, we investigated the effects of speech-rate slowing approaches with and without steady-state

suppression on the improvement of speech intelligibility in reverberant environments, and confirmed that they are effective for relatively long RTs (2.0 and 2.8 s). Additionally, the degree of improvement by the combined approach of speech-rate slowing and steady-state suppression was superior to that of simple speech-rate slowing at an RT of 2.0 s. Thus, the effects of speech-rate slowing approaches differ according to the reverberant condition.

As to future work, we plan to determine how speech intelligibility changes depending on the speaking rate, and compare the effects of steady-state suppression between with or without speech-rate slowing.

8. Acknowledgement

We would like to thank the subjects participated in the perceptual experiment, and Prof. Tachibana of the Univ. of Tokyo (at that time), and the members of his lab., especially to Dr. Kanako Ueno and Dr. Sakae Yokoyama, and Sub Working Group on Research in Speech Transmission Quality of the Architectural Institute of Japan for providing the impulse response data. This research was supported by a Grant-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science.

References

[1] R. H. Bolt and A. D. MacDonald, “Theory of speech masking by reverberation,” J. Acoust. Soc. Am., Vol.21, no.6, pp. 577-580, 1949.

[2] A. Nabelek and J. Pickett, “Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners,” J. Speech Hear. Res., 17, pp. 724-739, 1974.

[3] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments,” Proc. Autumn Meet. Acoust. Soc. Jpn., Vol.1, pp. 449-450, 2001 (in Japanese).

[4] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” Acoust. Sci. Tech., Vol.23, pp. 229-232, 2002.

[5] N. Hodoshima, T. Arai, T. Inoue, K. Kinoshita and A. Kusumoto, “Improving speech intelligibility by steady-state suppression as pre-processing in small to medium sized halls,” Proc. Eurospeech, pp. 1365-1368, 2003.

[6] N. Hodoshima, T. Inoue, T. Arai and A. Kusumoto, “Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments,” Proc. China-Japan Joint Conference on Acoustics, pp. 199-202, 2002.

[7] T. Goto, T. Inoue, N. Ohata, N. Hodoshima and T. Arai, “The effect of pre-processing for improving speech intelligibility in the Sophia University lecture hall,” Proc. Autumn Meet. Acoust. Soc. Jpn., Vol.1, pp. 613-614, 2003 (in Japanese).

[8] T. Arai, “Padding zero into steady-state portions of speech as a preprocess for improving intelligibility in reverberant environments,” Acoust. Sci. Tech., Vol.26, no.5, pp. 459-461, 2005.

[9] Praat Homepage (Version 4.3.14): http://www.praat.org

[10] F. J. Charpentier and M. G. Stella, “Diphone synthesis using an overlap-add technique for speech waveforms concatenation,” Proc. ICASSP, pp. 2015-2018, 1986.

[11] N. Hayashi, T. Arai, N. Hodoshima, Y. Miyauchi and K. Kurisu, “Steady-state pre-processing for improving speech intelligibility in reverberant environments: Evaluation in a hall with an electrical reverberator,” Proc. Interspeech, pp.1741-1744, 2005.

[12] Y. Nakata, Y. Murakami, N. Hayashi, Y. Miyauchi, N. Hodoshima, T. Arai and K. Kurisu, “Evaluation of two steady-state processing methods for improving speech intelligibility in reverberant environments,” Proc. Autumn Meet. Acoust. Soc. Jpn., pp. 693-694, 2005 (in Japanese).

[13] S. Furui, “On the role of spectral transition for speech perception,” J. Acoust. Soc. Am., Vol.80, no.4, pp. 1016-1025, 1986.

[14] N. Hodoshima, T. Arai and A. Kusumoto, “Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments,” Proc. Forum Acusticum Sevilla, 2002.