

ACOUSTICAL LETTER

Decreasing speaking-rate with steady-state suppression to improve speech intelligibility in reverberant environments

Takayuki Arai[1,*], Yuki Nakata[1], Nao Hodoshima[1] and Kiyohiro Kurisu[2]

[1]*Department of Electrical and Electronics Engineering, Faculty of Science and Technology, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*TOA Corporation, 2–1 Takamatsu-cho, Takarazuka, 665–0043 Japan*

(*Received 3 December 2006, Accepted for publication 30 January 2007*)

Keywords: speech intelligibility, reverberation, speaking rate slowing, steady-state suppression
PACS number: 43.55.Hy, 43.72.Ew, 43.71.Es, 43.66.Dc, 43.38.Tj [doi:10.1250/ast.28.282]

1. Introduction

It is known that strong reverberation affects speech intelligibility. Although early reflections often help speech intelligibility (the Haas effect, e.g., [1]) late reflections degrade speech intelligibility [2]. Overlap-masking in reverberant environments is the main source of degradation in speech intelligibility [3–5]. Because of overlap-masking, reverberant components of prior speech segments mask successive segments. As a result, speech segments following reverberating segments are more difficult to understand. As the energy of the prior segments increases, the effect of overlap-masking also increases. This fact is particularly important when the reverberating segment is a vowel (which has more power) and the subsequent segments are consonants (which have less power) [6,7].

To reduce overlap-masking, Arai *et al.* [6,7] proposed "steady-state suppression" as a preprocess for speech signals in reverberant environments. Strange *et al.* [8] showed that the information in steady-state portions of a speech signal was relatively insignificant compared with the information in transient portions. Additionally, steady-state portions usually have more energy compared to transients. The "steady-state suppression" technique reduces overlap-masking by estimating and suppressing the more powerful yet less significant steady-state portions of speech, such as the nuclei of syllables.

From the results of several experiments in simulated and actual sound fields, we have already confirmed that when we apply this process between a microphone and loudspeaker, it significantly improves speech intelligibility in reverberant environments (reverberation times of 0.8–1.3 s) [e.g., 6,7,9–12].

On the other hand, we know empirically that speaking slowly helps to increase speech intelligibility, particularly in a large hall with a long reverberation time. In the literature, Bolt and MacDonald [4] reported that speech intelligibility was greatly increased by speaking slowly in a reverberant room. This fact can be explained in terms of overlap-masking. In faster speech, many syllables are preceding a transient part within a short time period, and therefore, the reverberation tails of the preceding syllables mask the transient part. In slow speech, on the other hand, the preceding syllables are far apart, and the number of syllables that affect the following transient part is less, so that the intelligibility of speech increases.

However, Arai [13] pointed out that stretching a speech signal is not the best way to reduce overlap-masking and improve speech intelligibility in reverberant environments. This is because when we speak slowly the steady-state portions get longer and we even tend to elongate such portions in slow speech. Since the steady-state portions, such as syllable nuclei, usually contain more energy than others, the elongated preceding steady-state portions still affect the subsequent transient part in terms of overlap-masking.

Slowing speech by isolating each syllable would be more effective for improving speech intelligibility, because, in theory, it would significantly reduce the amount of overlap-masking [13]. Alternatively, suppressing steady-state portions of speech while decreasing the speaking rate may be effective. Therefore, in the present study, we tried an approach to improve speech intelligibility by applying steady-state suppression in addition to slowing the speaking rate, so that the amount of overlap-masking from the previous syllable can be reduced. We conduct here a perceptual experiment using speech samples processed by speaking-rate slowing with and without steady-state suppression under three reverberant conditions.

2. Steady-state suppression

In this study, we adopted the same algorithm for the steady-state suppression method as used in the previous studies [6,7,9–12]. This technique first splits an original signal into 1/3-octave bands and then extracts the envelope 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 Furui proposed to measure spectral transition [14]. 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 is considered steady state, the amplitude of the portion is suppressed. In this study, the speech portion is suppressed to 40% of the original amplitude, as in previous studies [6,7,9–12].

Figure 1 shows the original speech and the processed speech signals. The original speech sample "He advised

*e-mail: arai@sophia.ac.jp

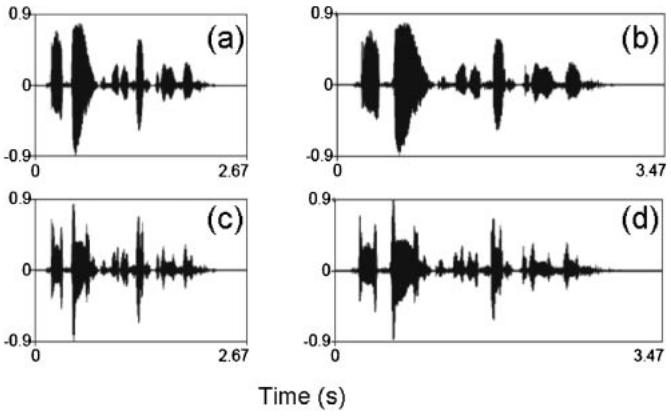

Fig. 1 Original and processed waveforms: (a) original speech sample; (b) slowed speech sample; (c) original speech sample with steady-state suppression; and (d) speech sample processed by steady-state suppression after slowing the speaking rate.

immediate hospitalization." was uttered by a male speaker, taken from the TIMIT corpus [15]. Figure 1(a) is the waveform of the original speech sample, whereas (b) is the waveform of the slowed speech sample by the technique described in Section 3.1 (the expansion rate was 1.3). The waveforms of Fig. 1(c) and (d) are the speech samples processed by steady-state suppression of the speech samples in Fig. 1(a) and (b), respectively.

3. Perceptual experiment

3.1. Speaking-rate slowing

We compared the intelligibilities of with and without steady-state suppression for speech samples with different speaking rates under three reverberant conditions. For speaking-rate slowing, we changed the original speaking rate (SR1: 6 morae/s) to 5 (SR2) and 4 (SR3) morae/s. To decrease the speaking rate, we used Praat [16], which applies the pitch-synchronous overlap and add (PSOLA) method for time-scale modification. The PSOLA method can modify speaking rate without changing the fundamental or formant frequencies of the original speech signal [17].

3.2. Reverberant conditions

We conducted a perceptual experiment under artificial reverberant environments achieved by convolving speech samples with impulse responses. The reverberation times (Ts) of the three impulse responses we used were 1.5 s (Rev1), 2.0 s (Rev2) and 2.5 s (Rev3). These impulse responses were created from a single impulse response (measured at a lecture hall from the database by Sub Working Group on Research in Speech Transmission Quality of the Architectural Institute of Japan) by multiplying an exponential decay as in a previous study [18]. We defined the reverberation time, T, as the time taken for the first 10 dB drop of the decay curve of an impulse response multiplied by six. In our case, an impulse response was first split into octave bands and the mean Ts were calculated for each band having a center frequency of 500, 1,000, 2,000 Hz, respectively.

3.3. Speech samples

The original speech samples consisted of 14 nonsense consonant-vowel (CV) syllables embedded in a Japanese carrier phrase, "Daimoku to shite wa ____ to iimasu" (It is called ____ as a title). 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. Finally, we prepared original speech samples, processed speech samples by steady-state suppression, processed speech samples by speaking-rate slowing, and processed speech samples first by speaking-rate slowing and then by steady-state suppression. All were convolved with each of the three impulse responses used in this study.

3.4. Participants

Twenty-five young participants with normal-hearing (14 males and 11 females, aged 18 to 37 years) participated in the experiment. All were native speakers of Japanese.

3.5. Procedure

The experiment was conducted in a soundproof room. Stimuli were presented diotically through headphones (STAX SR-303) connected to a computer via the digital-to-analog (D/A) converter of a digital audio amplifier (MA-500U, Onkyo) that was connected to the computer via the USB interface. The sound level was adjusted to each listener's comfort level during the training session prior to the experiment. A stimulus was presented in each trial and the listeners were instructed to select one of the 16 options, including 14 CVs, vowel /a/, and 'others,' displayed on the computer screen. The experiment was carried out at each listener's pace. For each listener, 252 stimuli were presented randomly (3 reverberation conditions × 14 CVs × 6 processing conditions).

3.6. Results and discussions

Figure 2 shows the mean percent of correct responses of the perceptual experiment for the three reverberant conditions

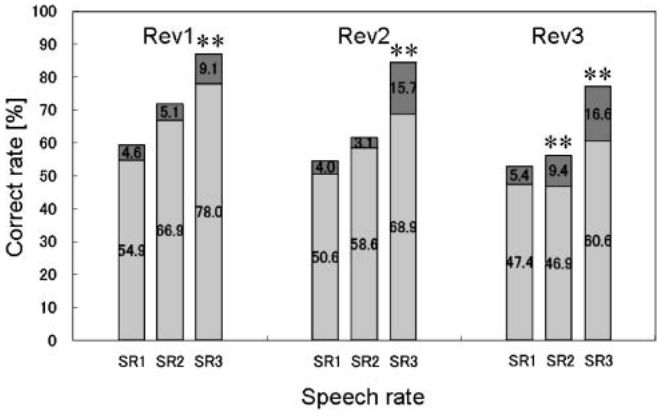

Fig. 2 Mean percent of correct responses of the perceptual experiment for the three reverberant conditions (Rev1: $T = 1.5$ s, Rev2: $T = 2.0$ s, and Rev3: $T = 2.5$ s) and the three speaking rates (SR1: 6 morae/s, SR2: 5 morae/s, and SR3: 4 morae/s). The dark parts of the bars show the improvements in speech intelligibility due to steady-state suppression. The asterisks '**' denote significant improvement ($p < 0.01$) by steady-state suppression.

and the three speaking rates. The dark parts of the bars show the improvements in speech intelligibility due to steady-state suppression. We can see that steady-state suppression improved speech intelligibility under all experimental conditions.

As expected, although the intelligibility of speech decreased as T got longer, it was improved by speaking-rate slowing under each of three reverberant conditions. An ANOVA was carried out with speaking rate (SR1, SR2 and SR3), reverberation (Rev1, Rev2 and Rev3) and processing (unprocessed and processed) as repeated variables, and the mean percent of correct responses as the dependent variable. Results showed that the mean percent of correct responses significantly differed across reverberation [$F(2,48) = 124.51$, $p < 0.01$] and across speaking rate [$F(2,48) = 168.49$, $p < 0.01$]. The mean percent of correct responses was also significantly higher for processed condition than for unprocessed condition [$F(1,24) = 75.57$, $p < 0.01$].

Pairwise comparison showed significant differences between all speaking-rate pairs [$p < 0.01$] both for unprocessed and processed conditions at Rev1. Significant differences were also obtained between SR1 and SR2 [$p < 0.05$], SR2 and SR3 [$p < 0.01$] and SR1 and SR3 [$p < 0.01$] both for unprocessed and processed conditions at Rev2. At Rev3, there were significant differences between SR2 and SR3 [$p < 0.01$] and SR1 and SR3 [$p < 0.01$] both for unprocessed and processed conditions.

The pairwise comparison also showed significant improvements by steady-state suppression for Rev1 with SR3, for Rev2 with SR3, and for Rev3 with SR2 and SR3 [$p < 0.01$]. As the speaking rate decreased, the degree of improvement by steady-state suppression was increased. The improvements by steady-state suppression were statistically significant for all of the slowest conditions (SR3). Moreover, for SR3, we observed higher improvements by steady-state suppression, as the T gets longer. Speech intelligibility was particularly improved by steady-state suppression after speaking-rate slowing under longer reverberant conditions (Ts of 2.0 and 2.5 s).

4. Conclusions

In this study, we investigated the effect of steady-state suppression after slowing the speaking rate of a speech signal. From the results of the perceptual experiment, we confirmed that

(1) slowing the speaking rate improves speech intelligibility in a reverberant environment.

The reason why (1) found is that it separates the adjacent syllables well, and thus, overlap-masking was reduced from the previous syllables. However, it is generally thought that

(2) a simple time-scale elongation of the steady-state portions of speech by slowing the speaking rate cannot effectively reduce overlap-masking.

This might be suggested that the steady-state portions usually contain more energy, so elongating such portions might result in the extra masking effect. In other words, there is a trade-off between (1) decreased overlap-masking due to separating syllables and (2) increased overlap-masking due to elongating steady-state portions. The effect of (1) is larger than that of (2) when the speaking rate is slow enough, so that the simple speaking-rate slowing in the perceptual experiment of this study also showed a certain degree of improvements in terms of speech intelligibility. However, the improvements were not the best, and those with steady-state suppression were even higher. Therefore, we can concluded that

(3) in addition to slowing the speaking rate, it is preferable to suppress the steady-state portions of speech, so that the amount of overlap-masking from the previous syllable can be effectively reduced.

Acknowledgments

We would like to thank the subjects who participated in the perceptual experiment and the 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 Grants-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science.

References

[1] H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc.*, **20**, 145–159 (1972).
[2] A. K. Nábělek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, **17**, 724–739 (1974).
[3] V. O. Knudsen, "The hearing of speech in auditoriums," *J. Acoust. Soc. Am.*, **1**, 56–82 (1929).
[4] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).
[5] A. K. Nábělek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, **86**, 1259–1265 (1989).
[6] 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).
[7] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments," *Acoust. Sci. & Tech.*, **23**, 229–232 (2002).
[8] W. Strange, J. J. Jenkins and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, **74**, 695–705 (1983).
[9] 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. Interspeech*, pp. 1365–1368 (2003).
[10] N. Hodoshima, T. Inoue, T. Arai, A. Kusumoto and K. Kinoshita, "Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments," *Acoust. Sci. & Tech.*, **25**, 58–60 (2004).
[11] N. Hodoshima, T. Arai, A. Kusumoto and K. Kinoshita, "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," *J. Acoust. Soc. Am.*, **119**, 4055–4064 (2006).
[12] N. Hodoshima, T. Goto, N. Ohata, T. Inoue and T. Arai, "The effect of pre-processing for improving speech intelligibility in the Sophia University lecture hall," *Proc. Int. Symp. Room Acoustics: Design and Science*, Hyogo (2004).
[13] T. Arai, "Padding zero into steady-state portions of speech as a preprocess for improving intelligibility in reverberant environ-

ments," *Acoust. Sci. & Tech.*, **26**, 459–461 (2005).
[14] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, **80**, 1016–1025 (1986).
[15] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, **9**, 351–356 (1990).
[16] Praat Homepage (Version 4.3.14): http://www.praat.org/
[17] F. J. Charpentier and M. G. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *Proc. ICASSP*, pp. 2015–2018 (1986).
[18] N. Hodoshima, T. Arai and A. Kusumoto, "Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments," *Proc. Forum Acusticum*, Sevilla (2002).