ACOUSTICAL LETTER

Evaluation of speech naturalness with steady-state zero padding for improving intelligibility in reverberant environments

Yohei Matsukaze, Takayuki Arai*, Toshimasa Suzuki and Keiichi Yasu

Graduate School of Science and Technology, Sophia University,
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 13 February 2012, Accepted for publication 18 April 2012*)

Keywords: Reverberation, Overlap masking, Zero padding, Speech intelligibility
PACS number: 43.55.Br, 43.71.Gv [doi:10.1250/ast.33.370]

1. Introduction

Reverberation gives depth in music. At the same time, reverberation lowers the intelligibility of speech. For example, speech does not sound clear in a subway station. In addition, it often becomes difficult for students to listen to the voice of their teacher because of reverberation in a classroom. As a result, reverberation may lower the student's concentration. Overlap masking is one of the main reasons why reverberation degrades speech intelligibility [1]. Because of overlap masking, the reverberant tail of one segment masks the following segments.

There are two types of processing to prevent the deterioration of speech intelligibility in a reverberant environment. The first is called "postprocessing," which is done after reverberation is added to a speech signal [2]. The second is "preprocessing," which is done before the reverberation is added to the speech signal [3,4]. Preprocessing can easily be incorporated in a reverberant environment because we do not have to have any device at the listener's side or change the environment.

In this study, we focus on a preprocessing technique to improve speech intelligibility in reverberation. Arai *et al.* proposed steady-state suppression as a preprocessing technique [3,4]. This processing does not suppress important parts of speech, such as transients, but suppresses the amplitude of steady-state parts that are less important for speech comprehension. As a result, steady-state suppression was able to reduce the influence of overlap masking [5]. However, this technique might remove some speech information by suppressing the amplitudes of the original speech signal. To overcome such an effect, Arai *et al.*, aiming at further improvement, proposed steady-state zero padding, which inserts zero sequences, or silence, into the steady-state portions, so that this zero padding reduced the effects of overlap-masking [6]. Similar to the study reported in [6], in the present study, zeros are padded in the center of each steady-state portion and the intelligibility of processed speech in reverberation is examined. In addition, we further investigate whether the processed speech is still natural, because the zero-padding technique reduces the effects of overlap masking with long zeros but also might reduce the naturalness as zeros become longer. Therefore, in this study, we also ask listeners to evaluate the unnaturalness for each stimulus.

2. Experiment

2.1. Steady-state zero padding

The algorithm for steady-state zero padding was the same as that in [6]. First, the original signal is split into 1/3-octave bands. In each of these bands, the logarithmic envelope is extracted. After downsampling, the regression coefficients are calculated from the five adjacent values of the time trajectory of the logarithmic envelope of a subband. Then the mean square of the regression coefficients, D, is calculated over all subbands. After upsampling, we define the speech portion to be in a steady state when D is less than a certain threshold. Once a speech portion is considered to be in the steady state, we pad zeros into the middle of each steady-state portion. The length of the padded zero sequence, or T_z, is variable. (The sampling frequency was 16 kHz, and after downsampling, it was 100 Hz.)

2.2. Stimuli

The speech samples used in this study were based on the 14 Japanese monosyllables [6]. Each target syllable was inserted in the career sentence. In this study, we adopted three different reverberant conditions, where the reverberation time was 2.5, 3.0, and 3.5 s, and three different T_z conditions, where T_z was 50, 75, and 100 ms. We had total of 168 stimuli (14 syllables $\times$ 4 T_z conditions $\times$ 3 reverberation times).

2.3. Procedure

We conducted the experiment in a sound-treated room at Sophia University. The experimental stimuli were presented through headphones (STAX SR-303). The 14 syllables were visually presented on the computer display after each stimulus was presented, and listeners were to select what they heard. The stimuli were randomly presented to each listener. There were total of 168 stimuli. Twenty young normal-hearing listeners participated in the perceptual experiment. Unnaturalness was also evaluated in all trials at the same time as listeners selected monosyllabic response choices.

The definition of "unnaturalness," as well as some other explanations, was told to the listeners before starting the following experiment.

1) What is reverberation?
2) All stimuli contain reverberation.
3) Processed speech samples are also used in the stimuli.

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

Table 1 Evaluation of naturalness.

Reverberation time (s)	2.5				3.0				3.5			
T_z (ms)	0	50	75	100	0	50	75	100	0	50	75	100
Naturalness (%)	71.4	72.5	66.1	60.4	64.6	66.4	62.4	53.2	56.4	55.4	53.9	53.2

4) "Unnatural" is defined as the state that a stimulus sounds "strange" as a result of the processing.

The four instructions were given to the participants beforehand. Participants were told to draw a check mark on the answer sheet when the stimuli sounded unnatural.

2.4. Results

The percentage of naturalness based on the results of the evaluation of "unnaturalness" is shown in Table 1. We calculated the percentage for the 20 participants for each T_z condition ($T_z = 0$ ms means that the stimuli were unprocessed).

The analysis of variance (ANOVA) was carried out using statistical analysis software, SPSS, for the evaluation data of unnaturalness. The Sidak multiple comparison was carried out for each type of processing, but no significant difference among the different T_z conditions was observed; therefore we can conclude that the processing did not reduce the naturalness.

The rate of correct responses to the target stimuli was also calculated for each condition (Table 2). ANOVA was carried out using SPSS. The Sidak multiple comparison showed that the correct rates for $T_z = 75$ ms and $T_z = 100$ ms were both significantly higher than that for the unprocessed stimuli ($p < 0.01$ and $p < 0.05$, respectively).

Table 2 Correct rate for each condition.

	Reverberation time (s)		
	2.5	3.0	3.5
Unprocessed	22.9	19.6	18.2
$T_z = 50$ (ms)	23.3	23.2	21.4
$T_z = 75$ (ms)	28.6	28.2	27.5
$T_z = 100$ (ms)	29.3	30.0	25.0

3. Discussion and Conclusion

This study focused on the naturalness of reverberant speech signals caused by the zero-padding technique. The result of the naturalness evaluation showed that the processed speech signals with $T_z = 50$, 75, and 100 ms were as natural as the original ones. Therefore, when $T_z = 100$ ms or shorter, padding zeros in the steady-state portions does not make speech unnatural. On the other hand, the correct rate was significantly high with $T_z = 75$ and 100 ms for all reverberation times. However, there was no significant difference between the correct rates for $T_z = 75$ and 100 ms. Thus, we conclude that it was not always true that longer T_z yields better performance as the reverberation time becomes longer.

Acknowledgment

This research was partly supported by Sophia University Open Research Center from MEXT.

References

[1] A. K. Nabelek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, **86**, 1259–1265 (1989).

[2] J. B. Allen, D. A. Berkley and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Am.*, **62**, 912–915 (1977).

[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.*, **23**, 229–232 (2002).

[5] 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).

[6] T. Arai, "Padding zero into steady-state portions of speech as a preprocess for improving intelligibility in reverberant environments," *Acoust. Sci. & Tech.*, **26**, 459–461 (2005).