Acoust. Sci. & Tech. **26**, 5 (2005)

ACOUSTICAL LETTER

Padding zero into steady-state portions of speech as a preprocess for improving intelligibility in reverberant environments

Takayuki Arai*

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

(*Received 21 February 2005, Accepted for publication 11 May 2005*)

Keywords: Speech enhancement, Reverberation, Speech intelligibility, Overlap masking, Steady-state portion, Speaking rate
PACS number: 43.72.Ew, 43.71.Es, 43.66.Dc, 43.55.Hy, 43, 38.Tj [DOI: 10.1250/ast.26.459]

1. Introduction

Overlap masking is the main reason why reverberation degrades speech intelligibility [1–3]. Because of overlap masking, the reverberant components of one segment mask the segments that follow. As a result, the segments following the reverberating segment are harder to hear. This is particularly true when the reverberating segment has more power, such as a vowel, and the subsequent segments have less power, such as a consonant [4,5].

To reduce overlap masking, Arai and colleagues [4,5] proposed "steady-state suppression" as a preprocess for speech signals in a reverberant environment. In this technique, the steady-state portions of speech, such as the nuclei of syllables, are estimated and suppressed. From the results of several experiments, we have already confirmed that when we apply this process between a microphone and a loudspeaker, it improves speech intelligibility in a reverberant environment [6–8]. In our previous studies, we did not allow the system to modify the total length of an utterance, because this technique can be applied to situations in which the talker is situated in the same room as where processed speech is broadcast.

However, this type of preprocessing in a reverberant environment can also be applied to situations in which the talker is absent from the room in which processed speech is broadcast. In this case, the time scale of a speech signal can be modified. Moreover, we empirically know that speaking slowly helps to increase speech intelligibility, particularly in a large hall with a long reverberation time. Thus, we can also design a new algorithm that stretches a speech signal using a time-scale modification technique to decrease the speaking rate.

Stretching a speech signal is, however, not the best solution. That is, isolating each syllable is more effective for this purpose, because, in theory, it significantly reduces the amount of overlap masking. In practice, isolating vowel-consonant-vowel segments (VCVs) is more natural and more robust to boundary detection errors than isolating CVCs. Thus, in this study, we first detect the steady-state portions of a speech signal by applying the algorithm proposed for steady-state suppression [4,5]. Then, a certain length of the zero sequence is inserted, or padded, into the middle of each detected steady-state portion. It is confirmed that this newly proposed "steady-state zero-padding" technique reduces the amount of overlap masking as a preprocess to prevent the degradation of speech intelligibility in a reverberant environment.

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

2. Steady-state zero-padding method

In the proposed method, we adopted the algorithm for detecting the steady-state portion of speech used in steady-state suppression reported in [4,5]. In this technique, first, an original signal is split into 1/3-octave bands. In each of these bands, the logarithmic envelope is extracted. After down-sampling, 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. D is similar to that proposed by Furui for measuring the spectral transition [9]. After up-sampling, we define a 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 a 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. This method yields discontinuities in the processed speech signal. Once the processed signal is broadcast into a room, however, we find that the discontinuities of the resultant signal are not noticeable due to the smoothing effect by convolving with the impulse response of the room (see Fig. 1). To reduce such discontinuities, we can apply the technique used in the previous studies [4,5], where each edge of the waveform was tapered off by multiplying by a slope.

In one sense, this steady-state zero-padding method can be viewed as a variation of the previously proposed steady-state suppression. In other words, steady-state zero-padding can be mimicked by applying steady-state suppression to a temporally-slowed speech signal.

3. Experiment

We compare the intelligibilities of speech in a room with and without the proposed zero-padding method. As the length of the padded zero sequence T_z was variable in this method, we tested two lengths: 50 and 100 ms. We conducted a perceptual experiment in XEBEC Hall (Kobe, Japan), which has an electrical reverberation system. Using this system, we were able to change and simulate several different room acoustics for the same set of subjects. In this study, we adopted two different reverberant conditions (see Table 1). There were four experimental setups in total, that is, two

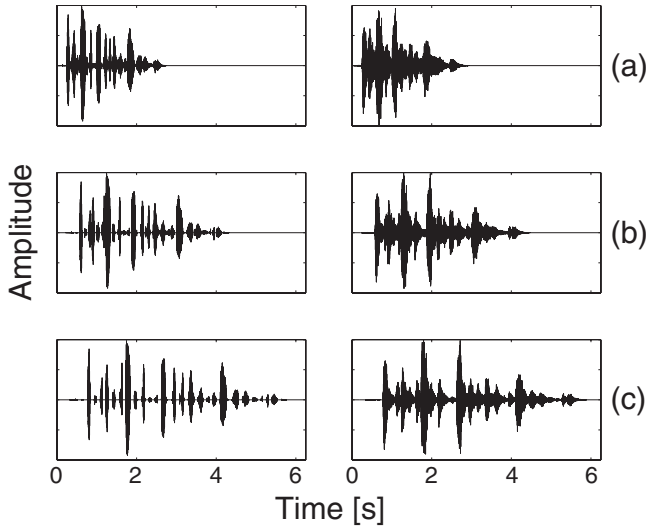

Fig. 1 Original and processed waveforms with (right column) and without (left column) reverberation: (a) original; (b) processed with steady-state zero padding ($T_z = 50$ ms); and (c) processed with steady-state zero padding ($T_z = 100$ ms).

Table 1 Correct rates (%). All differences between pairs were statistically significant (*: $p < 0.05$, **: $p < 0.01$).

	Reverberation time [s]	
	2.9	3.3
Original	44.2]**	37.6]*
Processed ($T_z = 50$ ms)	50.7	42.6
Original	49.3]**	37.6]**
Processed ($T_z = 100$ ms)	57.8	45.2

values of T_z under each of the two reverberant conditions.

The speech samples used in this study were based on the 14 Japanese monosyllables used by Hodoshima and colleagues [6,7], and the speech intelligibilities of two sets of speech samples with and without steady-state zero-padding were compared. The original speech samples consisted of monosyllables embedded in a Japanese carrier phrase.

Thirty-one young normal-hearing subjects participated in four sessions of the perceptual experiment. Each session corresponded to one experimental setup; the four sessions were conducted in order of difficulty. The subjects were asked to identify the target monosyllable in each trial. There were 28 stimulus sentences (14 monosyllables $\times$ 2 for with and without processing) in each session. These were presented only once in random order with a short pause between the stimuli.

Table 1 shows the experimental results. All differences between pairs were statistically significant. From this table, we confirmed that the proposed method prevents the degradation of speech intelligibility even if the reverberation time is relatively long (more than 2.0 s). This table also shows that a longer T_z yields a better improvement, and it is particularly needed for a longer reverberation time.

4. Summary

In this study, we showed that the newly proposed steady-state zero-padding method is effective for preprocessing speech signals in a reverberant environment. We empirically know that reducing the speaking rate improves speech transmission. However, a simple time-scale elongation of speech is not the best method of achieving this. Rather, it is preferable to separate the syllables from each other, so that the amount of overlap masking from the previous syllable can be reduced.

From the results of perceptual experiments using the steady-state zero-padding method, we conclude 1) that the processed signal is smoothed out with reverberation and discontinuities are no longer heard, although the processed signal before broadcasting into a room exhibits discontinuities in the waveform, and 2) the proposed method prevents the degradation of speech intelligibility even if the reverberation time is long. This improvement is explainable in terms of the modulation spectrum of speech, which strongly correlates with the intelligibility of speech [10,11].

We would like to determine how speech intelligibility changes depending on the length of the padded zero sequence. The result of this study supports the notion that the effect of overlap masking decreases as the length increases. The naturalness of speech, however, may decrease concurrently. The determination of the optimum length of the padded zero sequence would be a future work.

The proposed method is more effective than the previously proposed steady-state suppression method [4,5], particularly for a long reverberation time. Furthermore, it is useful when the talker is not situated in the same room as where the processed speech signal is broadcast. A typical situation of such conditions is an emergency broadcast in a tunnel. We would like to test this approach in such reverberant environments.

Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science. I would like to thank all of the people who helped me in various ways, especially Nahoko Hayashi, Nao Hodoshima, Tsuyoshi Inoue, Takahito Goto, Fumihiro Tadokoro and Yusuke Miyauchi of Arai Lab., Sophia University; Kiyohiro Kurisu of TOA Co.; the members of XEBEC Co.; and the subjects who participated in the perceptual experiment performed in this study.

References

[1] V. O. Knudsen, "The hearing of speech in auditoriums," *J. Acoust. Soc. Am.*, **1**, 56–82 (1929).
[2] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).
[3] 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).
[4] 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.*, pp. 449–450 (2001).
[5] 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).

[6] 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*, 1365–1368 (2003).

[7] 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).

[8] N. Hodoshima, T. Goto, N. Ohata, T. Inoue and T. Arai, "The effect of pre-processing approach for improving speech intelligibility in a hall: Comparison between diotic and dichotic listening conditions," *Acoust. Sci. & Tech.*, **26**, 212–214 (2005).

[9] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, **80**, 1016–1025 (1986).

[10] T. Houtgast and H. J. M. Steeneken, "A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, **77**, 1069–1077 (1985).

[11] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, **105**, 2783–2791 (1999).