ACOUSTICAL LETTER

Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments

Nao Hodoshima*, Tsuyoshi Inoue, Takayuki Arai, Akiko Kusumoto† and Keisuke Kinoshita‡

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

(*Received 1 April 2003, Accepted for publication 11 July 2003*)

Keywords: Speech enhancement, Reverberation, Speech intelligibility, Overlap-masking, Steady-state suppression
PACS number: 43.72.Ew [DOI: 10.1250/ast.25.58]

1. Introduction

In a large auditorium, perceiving speech is often difficult. This is due to reverberation that is caused by a superposition of reflected sounds with various delays and amplitudes. Because reverberation tails affect subsequent segments, an acoustic signal of one segment is masked by the reverberation components of the previous portion, and this effect of overlap-masking degrades speech intelligibility [1,2].

Syllable identification tests show that the spectral transition is crucial for syllable perception [3]. This is likely due to the fact that the information in steady-state portions of the speech signal is relatively redundant with that in transient segments [4]. Both the "delta" processing of cepstral features [3] and the RelAtive SpecTrAl (RASTA) processing [4] enhance transitions of speech and contribute to increase recognition rate in automatic speech recognition.

There are several general approaches for improving speech intelligibility in reverberant environments: microphone-array, pre-processing and post-processing. Microphone-array takes advantage of spatial information about sound source and assures the direction of the desired signal (e.g., [5,6],). Post-processing is a dereverberation technique applied to a signal having already been released into a room and affected by reverberation (e.g., [7,8],). As an example of a post-processing approach, modulation filtering is used. Modulation filtering alters the modulation spectrum of a signal. Langhans and Strube proposed the theoretical inverse modulation transfer function (IMTF) filter, which artificially increased the modulation depth of a reverberated signal in order to account for the decrease in the modulation index of the signal from reverberation [7]. Avendano *et al.* also artificially increased the modulation depth; they employed an IMTF filter derived from their own training data [8].

In the pre-processing approach, a speech signal is processed between a microphone and loudspeaker (e.g., [7,9–12],). As a pre-processing approach, Arai *et al.* suppressed the steady-state portions, as these portions of speech have more energy but are less crucial for speech perception in order to reduce the effect of overlap-masking caused by reverberation [9,10]. Modulation filtering is also used as a part of a pre-processing method ([7,11,12]). As a pre-processing method, Langhans and Strube applied the same technique as they used in with post-processing, but no clear improvement was found [7]. In light of the discovery that the important modulation frequency of a signal for speech perception is around 4 Hz, Kusumoto *et al.* enhanced this particular frequency region in their application of the modulation filter in their pre-processing approach [11]. The results in [9–12] showed promising results for improving speech intelligibility.

Our ultimate goal is to provide a filter for pre-processing which is suitable for an individual auditorium with a distinct reverberation time. In order to achieve this, we need to better understand the effects of the pre-processing filter and reverberation time on speech intelligibility. Because it is difficult to examine the effects of both parameters simultaneously, in a previous study we varied reverberation time while applying a preprocessing filter and evaluated the effects on speech intelligibility [12]. The results showed that modulation filtering was effected by reverberation time, most notably that modulation filtering prevented the degradation of speech intelligibility under specific conditions.

In this paper, we suppress steady-state portions of speech to explore the relationship between speech intelligibility and the reverberation conditions used in [12], which examined the effect of a single pre-processing filter (modulation filter) under various reverberation times. Using the steady-state suppression technique described in [9,10], we conduct a perceptual test with a set of artificial reverberation conditions, in which reverberation times are 0.9 s, 1.0 s, 1.1 s, 1.2 s and 1.3 s while reverberation times of 1.1 s and 1.8 s were used in [9,10].

2. Perceptual experiment

The artificial impulse responses h_n were created as Eq. (1) to obtain the desired reverberation conditions [12]:

$$h_n(t) = e^{-t/\tau} h_o(t) \tag{1}$$

where τ is a time constant. The original impulse response h_o used for this study was measured in the Hamming Hall, Higashi-Yamato City, Tokyo (A reflection board was not used.). Thus, we can obtain the desired reverberation time as a function of τ. Table 1 shows the set of reverberation

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

[†] Currently, Department of Veterans Affairs, Portland VA Medical Center, 3710 SW U.S. Veterans Hospital Road, Portland, OR 97207 USA

[‡] Currently, NTT Communication Science Laboratories, Speech Open Laboratory, 2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0237 Japan

Table 1 Reverberation conditions used in the experiment.

Impulse response	h1	h2	h3	h4	h5
Rev. time (s)	0.9	1.0	1.1	1.2	1.3

conditions used in our experiment (h_o is identical to h3 in Table 1). We used the reverberation time (RT), defined as the time the decay curve of the impulse response decreased 60 dB from steady state.

We applied the same method as in [9,10] to suppress the steady-state portions of speech. In Arai *et al.* [9,10], "half-proc" and "whole-proc" were used, as processing conditions. Half-proc means steady-state suppression applied for the first half of the sentences, whereas whole-proc means steady-state suppression applied for the whole sentences. In this study, we applied whole-proc as a processing condition. In steady-state suppression, first, an original signal was split into 1/3-octave bands. In each band the envelope was extracted. After down-sampling, the regression coefficients were 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, was calculated. We used the D parameter by Furui to measure the spectral transition [3]. After up-sampling, we defined a speech portion as steady-state when D was less than a certain threshold. Once a portion was considered steady-state, the amplitude of the portion was multiplied by the factor of 0.4 (a suppression rate of 40%).

The original samples consisted of nonsense Consonant-Vowel (CV) syllables embedded in a Japanese carrier phrase. The twenty-four CVs used in the experiment are shown in Table 2. The original speech samples were obtained from the ATR Speech Database of Japanese. The CV syllables were selected from the monosyllable data set. The carrier phrase is a combination of two partial sentences taken from a sentence data set. The beginning position of the target vowel was adjusted to 150 ms from the end of the pre-target carrier phrase. The stimuli consisted of two conditions: the original signals with reverberation (Org_rev) and the processed signals with reverberation (Proc_rev).

Twenty-four normal hearing subjects (14 males and 10 females, ages 18 to 26) participated in the experiment. All were native speakers of Japanese.

The experiment, controlled by a computer, was conducted in a soundproof room. The stimuli were presented with headphones (STAX SR-303), and the sound level was adjusted to each subject's comfort level. In the experiment, a stimulus was presented at each trial. Then 24 CVs in Kana orthography were shown on a PC screen. Subjects were forced to choose one of 24 CVs by clicking a button on the PC screen with a mouse. For each subject, 240 stimuli were presented randomly (5 reverberation conditions × 24 CVs × 2 processing conditions).

Table 2 CVs used in the experiment.

	Voiceless Consonants + Vowels	Voiced Consonants + Vowels
Stops + Vowels	/pa/ /ta/ /ka/ /pi/ /ki/	/ba/ /da/ /ga/ /bi/ /gi/
Fricatives + Vowels	/sa/ /ʃa/ /ha/ /ʃi/ /hi/	
Affricates + Vowels	/tʃa/ /tʃi/	/dza/ /dʒa/ /dʒi/
Nasals + Vowels		/ma/ /na/ /mi/ /ni/

Table 3 Mean percent correct in each condition.

	h1	h2	h3	h4	h5
Org_rev (%)	66.5	63.5	61.4	55.1	58.1
Proc_rev (%)	73.1	68.3	67.4	64.2	58.5

3. Results

The mean percent correct for each reverberation and processing condition is shown in Table 3. We analyzed the results for 22 subjects (we excluded two outliers). A 2 × 5 ANOVA for repeated measures was performed, confirming significant main effects of processing ($p < 0.001$) and impulse response ($p < 0.001$). For the comparison of means between processing, a t-test was performed for each impulse response type. A significant difference was obtained for the h1–h4 conditions (h1: $p = 0.049$; h2: $p = 0.026$; h3: $p = 0.003$; and h4: $p < 0.001$).

4. Discussions

We confirmed that the rates for correct responses declined as reverberation time increased, regardless of processing. It was also found that Proc_rev performed better than Org_rev under all reverberant conditions, and a t-test showed that the differences between the correct responses were significant for conditions h1–h4 (RT: 0.9–1.2 s). Our results confirm that the steady-state suppression is useful for improving speech intelligibility as a pre-processing method and that the effect of the steady-state suppression differed with respect to reverberation time.

5. Conclusions

In this paper, we suppressed the steady-state portions of speech based on Arai's technique [9,10] for improving speech intelligibility in reverberant environments. To explore the relationship between the steady-state suppression and several reverberation conditions, we conducted a perceptual test with a set of artificial reverberations. The results showed that the effect of the steady-state suppression depended on reverberation time and clear improvements were obtained with reverberation times of 0.9–1.2 s with the suppression rate of 40%. Thus, we certify that Arai's technique [9,10] is an effective pre-processing method for improving speech intelligibility under reverberant conditions. We predict that the range of reverberation conditions in which clear improvement are observed may be different as we change the suppression rate of steady-state portions. Thus, we would like to use steady-state suppression to investigate what are the upper and lower limits of reverberation time which prevent degradation of speech intelligibility. In other words, the upper and lower limits will be obtained when we 1) increasingly suppress steady-state portions of speech at longer reverberation conditions than those in this study, and when we 2) decreasingly suppress steady-state portions at shorter reverberation conditions than those in this study.

Acknowledgement

We appreciate Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for offering to use the impulse response data.

References

[1] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).

[2] A. K. Nabelek and L. Robinette, "Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects," *J. Acoust. Soc. Am.*, **63**, 187–194 (1978).

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

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, **2**, 578–589 (1994).

[5] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-34**, 1391–1400 (1986).

[6] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E West and M. M. Sondhi, "Autodirective microphone systems," *Acustica*, **73**, 58–71 (1991).

[7] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *Proc. IEEE ICASSP 82*, pp. 156–159 (1982).

[8] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," *Proc. ICSLP 96*, pp. 889–892 (1996).

[9] 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).

[10] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects on suppressing steady-state portions of speech on ntelligibility in reverberant environments," *Acoust. Sci. & Tech.*, **23**, 229–232 (2002).

[11] A. Kusumoto, T. Arai, T. Kitamura, M. Takahashi and Y. Murahara, "Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired," *Proc. IEEE ICASSP 2000*, pp. 853–856 (2000).

[12] N. Hodoshima, T. Arai and A. Kusumoto, "Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments," *Proc. Forum Acusticum*, Sevilla (2002).