ENHANCING TEMPORAL DYNAMICS OF SPEECH TO IMPROVE INTELLIGIBILITY IN REVERVERANT ENVIRONMENTS

PACS: 43.72.Ew

Hodoshima, Nao; Arai, Takayuki; and Kusumoto, Akiko
Dept. of Electrical and Electronics Engineering, Sophia University
7-1 Kioi-cho, Chiyoda-ku, Tokyo
102-8554
Japan
Tel: +81-3-3238-3417
Fax: +81-3-3238-3321
E-mail: n-hodosh@splab.ee.sophia.ac.jp

ABSTRACT

Due to reverberation, perceiving speech in a large auditorium is usually difficult. We attempted to suppress degradation of speech intelligibility in reverberant environments by enhancing in advance some temporal dynamics of speech (around 4 Hz) crucial for speech perception. Because the effect of our proposed method varies in reverberation conditions, we conducted a nonsense-word perceptual test with artificial reverberation to explore the relation between the degree of enhancement and reverberation time. The results indicate that clear improvement (up to 6.7%) was obtained especially when the targets were fricatives. We concluded that our proposed method is effective under specific conditions.

INTRODUCTION

Due to reverberation, perceiving speech in a large auditorium is often difficult especially for hearing-impaired and elderly people. Reverberation is essential for music, but it degrades speech intelligibility. The speech transmission index (STI) and rapid STI (RASTI) are widely used to objectively measure speech intelligibility in an auditorium [1]. Speech transmission indices are calculated with the modulation transfer function (MTF). MTF is expressed by the modulation depth reduction of the modulation spectrum of an observed signal relative to that of an original signal, as a function of the modulation frequency. The modulation spectrum is derived from a frequency analysis of the temporal envelope of the band-pass signal. The important modulation frequencies for speech perception lie between 1-16Hz, especially 4Hz, where the modulation spectrum usually reaches its maximum value [2], [3]. When the acoustic signal is reverberated, the peak of the modulation spectrum shifts to the lower modulation

frequency and the modulation index is reduced [1]. Thus, there is a strong relationship between speech intelligibility and the modulation spectrum.

Modulation filtering, which alters the modulation spectrum of a signal, has been proposed for improving speech intelligibility in reverberant environments. There are two general approaches to modulation filtering: post-processing and pre-processing. Post-processing is a dereverberation technique applied to a signal already released into a room and affected by reverberation. Langhans et al. proposed the theoretical inverse MTF (IMTF) filter. To improve speech intelligibility, they 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 [4]. Avendano et al. also artificially increased the modulation depth, but they employed an IMTF filter derived from their own training data [5].

In the pre-processing approach, a speech signal is processed between the microphone and loudspeaker. Langhans et al. applied the same technique to pre-processing as they used in with post-processing [4], but no clear improvement was found. In light of the discovery that the important modulation frequency of a signal for speech perception is around 4Hz, Kusumoto et al. enhanced this particular frequency region in their application of the modulation filter, and found promising results for improving speech intelligibility [6].

Our ultimate goal is to provide a suitable filter for each auditorium having a distinct reverberation time. In order to achieve this, we need to better understand the relationship between the effect of a modulation filter and reverberation time. In the previous studies [4], [6], the reverberation time was fixed and various approaches were tested. In this study, we fixed the modulation filter and varied reverberation time to explore the relationship between the degree of enhancement and reverberation. Finally, using the same modulation filter as in [6], we conducted a perceptual test with a set of artificial reverberations to further explore this same relationship.

PERCEPTUAL EXPERIMENT

Preparing Impulse Response for Desired Reverberation Time

The artificial impulse responses were adapted to create our reverberation conditions. We used the reverberation time T_{60}, defined as time that the decay curve decreased 60 dB from the steady state. A decay curve is obtained by integration over a reversed-time scale of a squared impulse response [7]. In the absence of discrete echoes, an impulse response of a room is approximated as a product of an exponential decay with time constant τ_o and a stationary noise $w(t)$ as in Eq. (2.1).

$$h_o(t) = e^{-t/\tau_o} w(t) \qquad (2.1)$$

Reverberation time changes as τ_o varies. Therefore, we can obtain the desired reverberation time as a function of τ_o. From Eq. (2.1), the new impulse response h_n (with the new time constant τ_n) was derived from Eq. (2.2). The original impulse response h_o used for this

study was measured in the Hamming Hall in Higashi Yamato City, Tokyo (a reflection board was not used).

$$h_n(t) = e^{-t/\tau} h_o(t) \qquad (1/\tau = 1/\tau_n - 1/\tau_o) \qquad (2.2)$$

Table 1 shows a set of reverberation times and time constants of the impulse response used in our experiment (h_o is identical to h3 in Table 1). τ is adjusted so that an interval of reverberation time changed 0.1 second.

Table 1.- Characteristics of the impulse responses used in our experiment

Impulse response	h1	h2	h3	h4	h5	h6
Rev. time (s)	0.90	1.00	1.11	1.21	1.30	1.40
Time constant τ (s)	0.19	0.22	0.25	0.28	0.31	0.38

Modulation Filtering

First, an original signal was split into 1/3-octave bands. In each band the envelope was extracted by the Hilbert transform. Then modulation filtering was applied to the envelope. We used the same modulation filter as was used in [6], which emphasizes the modulation frequency around 4Hz. After applying half-wave rectification, the filtered signal was obtained by multiplying the filtered envelope by the carrier signal in each band. We then applied the band-bass filter to remove frequency components outside the range of the band. Finally, the processed signal was obtained by summing up the filtered signals for each band.

Stimuli

The stimuli consisted of nonsense Vowel-Consonant-Vowel (VCV) syllables embedded in the Japanese carrier phrase, “Watashi no namae wa __ desu” (my name is __). The vowels were /a/ and the consonants were /p/, /t/, /k/, /b/, /d/, /g/, /s/, /ʃ/, /h/, /z/, /ɾ/, /ʧ/, /m/ and /n/. Each stimulus was spoken by a male native speaker of Japanese. The stimuli were grouped into three conditions: the original signals without reverberation (Clean), the original signals with reverberation (Rev_org) and the processed signals with reverberation (Rev_proc).

Subjects

Thirty-one normal hearing subjects (15 males and 16 females, ages 18 to 23) and one profoundly hearing-impaired subject (male, age 32) participated in the experiment. All were native speakers of Japanese. The hearing-impaired subject wore hearing aids during the perceptual test. His hearing level without hearing aids was 96dB (right ear) and 98dB (left ear).

Procedure

The experiment, controlled by a computer, was conducted in the soundproof room. The stimuli were presented from the headphones (STAX SR-303), and the sound level was adjusted to each subject’s comfort level. Before the main session, six stimuli were presented to the subjects for practice. In the main session a stimulus was presented twice. Then 14 CVs in Kana orthography were shown on the screen. Subjects were forced to choose one of 14 CVs by clicking a button on the screen with a mouse. After they selected, the next stimulus was presented. For each subject, 182 stimuli were presented randomly (6 reverberation conditions x 14 CVs x 2 processing conditions + 14 clean conditions).

EXPERINMENTAL RESULTS

The mean percent correct (MPC) for each reverberation and processing condition is shown in Tables 2 and 3. We analyzed the results for 30 normal-hearing subjects (and on that basis excluded one outlier) and one hearing-impaired subject. Table 2 shows the results with normal hearing subjects and the hearing-impaired subject. For the normal-hearing subjects, the stimuli in the clean condition (not shown) were clear enough so that the MPC was above 95%. A 2 x 6 ANOVA for repeated measures was performed, confirming significant main effects of processing types ($p=.003$) and impulse response types ($p<.001$). For the comparison of means between processing types, a t-test was performed for each impulse response type. A significant difference was obtained for the h1 condition ($p=.008$). For the hearing-impaired subject, MPC in the clean condition was 64.3%.

Table 2.- Mean percent correct with normal hearing subjects and the hearing-impaired subject

		h1	h2	h3	h4	h5	h6
Normal hearing subjects	Rev_org	88.1	84.5	83.8	81.7	81.7	80.5
	Rev_proc	83.1	81.7	81.4	79.1	79.3	77.9
Hearing-impaired subject	Rev_org	21.4	14.3	21.4	14.3	7.1	28.6
	Rev_proc	21.4	21.4	28.6	14.3	14.3	21.4

Next, we classified the targets into manner of articulation with the normal hearing subjects. The results for fricatives and stops are shown in Table 3. For fricatives, A 2 x 6 ANOVA for repeated measures confirmed that the main effect of processing types ($p=.017$) was significant. A close-to-significant main effect of the impulse response types was present ($p=.058$). A t-test was performed for each impulse response type. Significant differences were obtained for the h3 condition ($p=.005$). For stops, A 2 x 6 ANOVA for repeated measures confirmed that main effect of processing ($p<.001$) and impulse response types ($p<.001$) were significant factors. A t-test was done for each impulse response type. Significant differences were obtained for the h1-h5 conditions (h1: $p=.011$; h2: $p=.005$; h3: $p=.021$; h4: $p=.002$ and h5: $p=.030$).

Table 3.- Mean percent correct in fricatives and stops with normal hearing subjects

		h1	h2	h3	h4	h5	h6
Fricatives	Rev_org	90.0	88.0	85.3	84.0	83.3	85.3
	Rev_proc	89.3	93.3	92.0	89.3	88.0	86.0
Stops	Rev_org	83.9	77.8	77.8	76.1	76.1	68.9
	Rev_proc	75.7	69.4	70.6	67.8	68.9	68.3

DISCUSSION

For the normal hearing subjects, it was found that Rev_org had a higher MPC than Rev_proc in the h1 condition and that the MPCs did not differ between processing types at h2-h6 conditions. The MPCs of both processing conditions decreased linearly as reverberation time increased. We classified the targets into manner of articulation to know whether the effect of the modulation filtering depended on types of consonants. When the targets were fricatives,

Rev_proc performed better than Rev_org under all reverberant conditions, and especially clear improvement was found under the h3 condition. The MPCs of both processing conditions changed quadratically as reverberation time was longer. When the targets were stops, we found that Rev_org had higher MPCs than Rev_proc for the h1-h5 conditions and that the MPC did not differ between processing types under the h6 condition. The MPCs of both processing conditions decreased linearly as reverberation time increased. Therefore, we confirmed that the effect of the modulation filter differed with respect to reverberation time, and the relationship between our filter and reverberation time varied with respect to manner of articulation.

For the hearing-impaired subject, the MPCs of Rev_proc were higher than that of Rev_org under the h2, h3 and h4 conditions. This indicates that the modulation filtering might improve speech intelligibility for the hearing-impaired.

We examined a confusion matrix to identify the reason for the increased intelligibility when the targets were fricatives in Rev_proc. In Rev_proc, fewer subjects confused alveolar fricatives (/s/, /z/) with palato-alveolar fricatives (/ʃ/, /ʒ/) as compared to Rev_org. For the fricatives /s/ and/z/, strong energy regions in transition in Rev_proc were better preserved than they were in Rev_org. We believe that the strong energy regions were preserved in transition because the modulation filtering essentially suppressed the steady-state portions of speech [8]. Arai et al. hypothesized that the acoustic signal of a certain segment is masked by the reverberation components of the previous portion, and this masking effect degrades speech intelligibility [8]. Arai et al. suppressed the steady-state portions that have more energy but which are less crucial for speech perception. A perceptual test was conducted to reduce the masking effect and the results showed a possibility that the steady-state suppression prevented the degradation of speech intelligibility. Because the fricatives in general contain more steady-state portions than stops, much energy can be suppressed in fricatives, and the masking effect caused by reverberation can be reduced more than for stops.

CONCLUSION

In this paper we enhanced some temporal dynamics of speech in advance for improving speech intelligibility in reverberant environments. To explore the relationship between the degree of enhancement and reverberation time, we conducted a perceptual test with artificial reverberation. The results showed that the effect of our filter depended on reverberation time and manner of articulation of the targets. Clear improvement was obtained especially when the targets were fricatives at reverberation time 1.1 seconds. The results also showed the possibility that the modulation filter might improve speech intelligibility for the hearing-impaired. We concluded that our proposed method was effective under specific conditions.

ACKNOWLEDGEMENT

We appreciate Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for offering to use the impulse response data. Also, we thank the speakers and the subjects who participated in our experiment.

BIBLIOGRAPHICAL REFERENCES

[1] 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(6), pp. 1069-1077, 1985.

[2] R. Drullman, J. M. Festen and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *J. Acoust. Soc. Am*., 95(2), pp. 1053-1064, 1994.

[3] T. Arai, M. Pavel, H. Hermansky and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” *J. Acoust. Soc. Am*., 105(5), pp. 2783-2791, 1999.

[4] C. Avendano and H. Hermansky, “Study on the dereverberation of speech based on temporal envelope filtering,” *Proc. ICSLP*, pp. 889-892, 1996.

[5] T. Langhans and H. W. Strube, “Speech enhancement by nonlinear multiband envelope filtering,” *Proc. IEEE ICASSP*, pp. 156-159, 1982.

[6] A. Kusumoto, T. Arai, M. Takahashi and Y. Murahara, “Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired,” *Proc. IEEE ICASSP*, pp. 933-936, 2000.

[7] M. R. Schroeder, “New method of measuring reverberation time,” *J. Acoust. Soc. Am*., 37, pp. 409-412, 1965.

[8] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoustical Science and Technology*, 23, 2002 (to be published).