JCA2007 4-6 June
The Japan-China Joint Conference of Acoustics 2007

Preprocessing effects on speech intelligibility in reverberation using mixed natural and electroacoustical sounds

Nao Hodoshima[1], Peter Svensson[2] and Takayuki Arai[1]
[1] Department of Electrical and Electronics Engineering, Sophia University
[2] Department of Electronics and Telecommunications,
Norwegian University of Science and Technology

ABSTRACT

This study evaluates a preprocessing approach for reducing reverberation effects when listeners hear both sounds from a talker and from a loudspeaker. Steady-state suppression, as described by Arai *et al*. [Proc. Autumn Meet. Acoust. Soc. Jpn., 2001; Acoust. Sci. Tech., 23, 229-232, 2002], was used as the preprocessing approach which suppressed steady-state portions of a speech signal before it is radiated from loudspeakers. We simulated two different halls in which public address systems were installed. Stimuli for a syllable identification test were prepared by convolving unprocessed and steady-state suppressed signals with calculated impulse responses, and were presented to 20 young adults with normal hearing. Results showed that a loudspeaker gain affected the performance of steady-state suppression. Results also showed that the effect of the mixture of sounds from a loudspeaker and from a talker on the performance of steady-state suppression was negligible when a direct-to-reverberation ratio is high at a talker microphone. The mixture of natural and electroacoustical sounds makes the study of steady-state suppression more realistic and tests its robustness.

INTRODUCTION

Reverberation makes speech perception difficult. Numbers of studies have shown that people with hearing impairments, elderly people and non-native listeners are much more affected by reverberation than young native listeners with normal hearing (e.g., [1]). Therefore, it is important to provide clear speech for various populations in public spaces by reducing the effect of reverberation in architectural acoustic or electroacoustic ways, that is, to build “barrier-free” listening environments.

There are several approaches to improve speech intelligibility in reverberant environments based on electroacoustics. Examples are the use of directional loudspeakers, a postprocessing approach and a preprocessing approach. Postprocessing is applied to speech signals that are captured in reverberation (e.g., [2]). Preprocessing processes speech signals before they are affected by reverberation (e.g., [3]).

“Overlap-masking” (i.e., reverberation tails that mask the following segments) is a contributor to the reduction of speech intelligibility in the presence of reverberation [4]. In order to reduce overlap-masking directly, steady-state suppression was proposed as a preprocessing technique [5, 6]. This technique automatically suppresses steady-state portions of speech (e.g., vowel nuclei) that are relatively less important for syllable perception compared to spectral transitions [7]. Several listening tests showed that steady-state suppression significantly improved consonant identification for young people with normal hearing [8, 9] and elderly people [10, 11] under diotic and dichotic listening environments at reverberation time (RT) of 0.7-1.3 s.

In the previous studies [8-11], speech signals were convolved with impulse responses of a room as a diotic condition. This condition simulates the situation where no talker was present in the same room as listeners, and dry speech was used as an input of steady-state suppression. However, when steady-state suppression is used in the situation where the talker is present in a room, the listeners in the room hear both the natural (unprocessed) sound from the talker and the amplified processed sound from a loudspeaker. Also, the input of a public address (PA) system is a speech signal picked up by a close microphone in front of the talker in the room, meaning that the input of PA system is not completely dry speech.

The purpose of this study was to investigate the effect of steady-state suppression in the situation where a listener hears a natural sound from a talker in addition to a processed sound from a loudspeaker in two questions. The first question was whether a loudspeaker gain affects the performance of steady-state suppression. The second question was whether the input of PA system affects the performance of steady-state suppression differently between dry speech and speech captured by a talker microphone, in other words, between a pre-recorded situation and a real situation. In order to test the two questions, a listening test was carried out for young people with normal hearing using unprocessed and steady-state suppressed speech under simulated reverberant environments in which public address systems were virtually installed.

LISTENING TEST

Participants

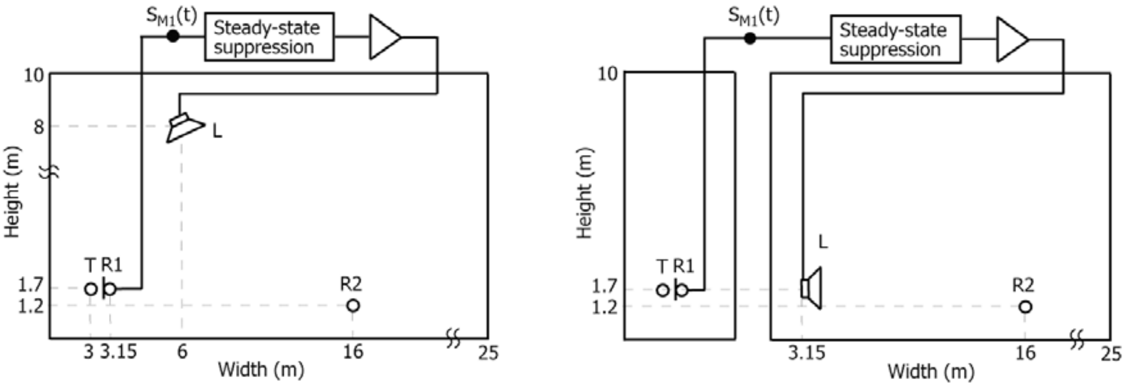
Twenty native speakers of Japanese (4 males and 16 females, aged 22 to 37 years old) participated in the listening test. None of them reported a history of unusual noise exposure or listening difficulties. They had normal hearing as the air-conduction thresholds were less than 25 dB HL from 125 to 8 kHz for both ears.

Room simulation

Two simple rooms (the small and the large rooms) that had shoebox shapes with 100% diffusion were calculated by CATT-Acoustic [12]. The small room had a volume of 500 m^3 and RT of 1.2 s, and the large room had a volume of 15750 m^3 and RT of 1.8 s. Two source positions corresponding to a talker (T) with a measured directivity from a mouth of a mannequin [13] and a loudspeaker (L) with vertical array-type directivity, and two receiver positions corresponding to a microphone in front of the talker (R1) with cardioid directivity and a listener (R2) with omnidirectional directivity were prepared in each room.

An electroacoustic path consisted of a steady-state suppression unit and a loudspeaker gain. Three gain conditions (gain 0, gain 1 and gain 2) were prepared. The loudspeaker gain was 0.0 dB in gain 0. The total level at R2 was 2.2 dB higher on average with gain 1 and 5.5 dB higher on average with gain 2 compared to the natural sound.

Source and receiver positions in a vertical section of the small room are shown in Fig. 1a for gains 1 and 2 and Fig. 1b for gain 0. For stimuli with gains 1 and 2, we simulated the situation where a listener attended a lecture in a hall and a speech signal was recorded by a clip-on microphone on the talker. R2 was a little bit away from the



a) gain 1 and gain 2 b) gain 0

Fig. 1. Source (T: talker, L: loudspeaker) and receiver (R1: microphone in front of the talker and R2: listener) positions in a vertical section of the small room for a) gains 1 and 2 and for b) gain 0. $S_{M1}(t)$ is the input of PA system.

main direction of the loudspeaker in order to avoid the situation where intelligibility at the listener's position is too high. For stimuli with gain 0, we simulated the situation where no talker was present in the same room as listeners.

Monaural impulse responses between T and R1 ($IR_{T\text{-}R1}$), T and R2 ($IR_{T\text{-}R2}$), L and R1 ($IR_{L\text{-}R1}$), L and R2 ($IR_{L\text{-}R2}$) were calculated in each room by CATT-Acoustic which used randomized tail-corrected cone-tracing [14]. After the impulse response calculations, the loudspeaker gains were added in $IR_{L\text{-}R1}$ and $IR_{L\text{-}R2}$, making four conditions (two rooms x two loudspeaker gains) for each of $IR_{L\text{-}R1}$ and $IR_{L\text{-}R2}$.

Stimuli

The speech materials consisted of nonsense 14 consonant-vowel syllables (vowel: /a/, and consonants: /p, t, k, b, d, g, s, ʃ, h, dz, dʒ, tʃ, m, n/) as targets embedded in a Japanese carrier phrase, which was the same as that used in [11-14]. All possible CV combinations were selected, excluding those that do not meet Japanese phonotactics. The speech materials were obtained from the ATR speech database of Japanese. The speaker was a 40-year-old male. The A-weighted energy was set equal for speech materials in each room and each gain condition.

Stimuli at R2 consisted of a natural sound from the talker with gain 0 as in Eq. (1):

$$T\{s(t)\}*IR_{T-R2} \tag{1}$$

and a natural sound from the talker plus an amplified speech sound from the loudspeaker with gains 1 and 2 as in Eq. (2):

$$s(t)*IR_{T-R2}+T\{s_{M1}(t)\}*IR_{L-R2} \tag{2}$$

$$T\{s(t)\}=\begin{cases} x\ (unprocessed\quad condition\ ) \\ P(x)\ (\ processed\quad condition\ ) \end{cases}$$

where * denotes convolution, $s(t)$ denotes a speech signal and $P(x)$ denotes that steady-state suppression which suppresses the amplitude of steady-state portions to 40% is applied to x. Therefore, 168 stimuli [two processing conditions (unprocessed/ processed) x two rooms (small and large rooms) x three gains (gain 0, gain 1 and gain 2) x 14 speech materials] were prepared.

Stimuli at the input of the PA system consisted of dry speech and speech captured by a talker microphone. The speech captured by the talker microphone is a natural sound from the talker plus an amplified speech sound from the loudspeaker in the large room with gain 2 as in Eq. (3):

$$s(t) * IR_{T-R1} + (s(t) * IR_{T-R1}) * IR_{L-R1}$$
$$= s_{M1}(t) * (1 + IR_{L-M1}) \quad (3)$$

where * denotes convolution and $s(t)$ denotes a speech signal. Therefore, 28 stimuli (14 dry speech and 14 speech captured by a talker microphone) were prepared.

Procedures

The computer-controlled listening test was conducted in a sound-treated room. The stimuli were presented at a sampling frequency of 16000Hz over headphones (STAX SR-303) through a digital audio amplifier (Onkyo MA-500U) that was connected to a computer. The sound level was adjusted to a comfortable level for the participants before the trials began, and the comfortable level was kept constant throughout the test. In each trial, a stimulus was presented, after which a computer monitor displayed the 14 syllables. The participants were instructed to use the mouse to click on a syllable on the monitor. Once a syllable was selected, the next trial was presented. During trials, 168 stimuli at R2 were randomly presented first, followed by 28 stimuli at the input of PA system that were presented randomly.

RESULTS AND DISCUSSION

Effect of loudspeaker gain

Figure 2 shows the mean percent correct at R2 for each room, gain, and processing condition. For the stimuli with gains 1 and 2, an ANOVA for repeated measures was carried out with room (small and large), gain (gain1 and gain2) and processing (unprocessed and processed). For the stimuli with gain 0, an ANOVA for repeated measures was carried out with room (small and large) and processing (unprocessed and processed).

For the stimuli with gains 1 and 2, results showed that the mean percent correct in the small room was significantly higher than in the large room [$p < 0.01$], that was consistent with the previous studies [11-14]. The mean percent correct with gain 2 was also significantly higher than that with gain 1 [$p = 0.02$]. The mean percent correct in the processed condition was also significantly higher than that in the unprocessed condition [$p = 0.01$]. Interactions between room and gain [$p = 0.01$] and between room and processing [$p = 0.02$] were also significant. Post-hoc analyses showed that the mean percent correct with gain 2 was significantly higher than that with gain 1 [$p < 0.01$] for the small room. For the large room, the mean percent correct in the processed condition was significantly higher than that in the unprocessed condition [$p < 0.01$]. These results showed that the mean percent corrects mainly increase due to higher loudspeaker gain in the small room. In the large room, on the other hand, the mean percent corrects mainly increase due to steady-state suppression rather than the loudspeaker gain. Post-hoc analyses also showed that processed speech significantly performed better than unprocessed speech at gain 1 [$p = 0.03$] and gain 2 [$p < 0.01$] in the large room, indicating that introducing a loudspeaker changed the performance of steady-state suppression.

For the stimuli with gain 0, results showed that the mean percent correct in the small room was significantly higher than that in the large room [$p < 0.01$], that was consistent with the previous studies [11-14]. The mean percent correct in the processed

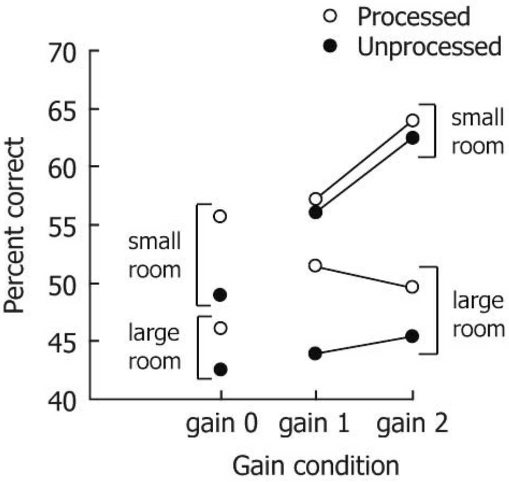


Fig. 2. Mean percent correct of 14 CV syllables at a listener position for each room, gain and processing condition.

condition was also significantly higher than that in the unprocessed condition [$p < 0.01$]. No interaction was significant. Post-hoc analyses showed no significant difference between mean percent corrects of unprocessed and processed conditions for both room conditions.

Processed speech had higher mean percent correct under all gain conditions in this study. This suggests that steady-state suppression improves syllable identification in both situations where a talker and a listener are in the same room and in different rooms.

Input of PA system

The *t*-test showed no statistically significant difference between dry speech (99.7%) and speech captured by a talker microphone (100%), indicating that the effect of steady-state suppression would be the same for the two cases of the input of a preprocessing approach: a pre-recorded voice which corresponds to the dry speech and a real voice which corresponds to speech captured by a talker microphone. This means that the same steady-state suppression may be applied to the two cases when a direct-to-reverberation ratio is high at the talker microphone. This would be useful for a practical use of steady-state suppression. When we think of amplifying a pre-recorded voice as well as a real voice in a same enclosure (e.g., an automatic announcement of the arrival and departure of trains in a station and a live announcement of a delayed train in a station), the maximum improvement in speech intelligibility would be obtained by steady-state suppression with the same parameters.

CONCLUSIONS

The effect of steady-state suppression was investigated in the situation where a listener hears both natural and electroacoustical sounds. Results from a listening test showed that introducing a loudspeaker changed the performance of steady-state suppression, implying that the effect of steady-state suppression also might be different in a direction and a level from a loudspeaker. The results also showed that the effect of steady-state suppression was the same for the two cases of the input of PA system (a pre-recorded voice and a real voice) when a direct-to-reverberation ratio is high at a talker microphone. For a practical application of steady-state suppression, future research would find which directivities and gains from loudspeaker(s) would be appropriate for steady-state suppression.

ACKNOWLEDGEMENTS

This research was supported by a Grant-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science.

REFERENCES

[1] A. K. Nábělek and P. K. Robinson, "Monaural and binaural speech perception in reverberation for listeners of various ages", J. Acoust. Soc. Am., 71(4), 1242-1248 (1982).

[2] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering", Proc. ICASSP, 7, 156-159 (1982).

[3] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments", Speech Comm., 45(2), 101-113 (2005).

[4] A. K. Nábělek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification", J. Acoust. Soc. Am., 86(4), 1259-1265 (1989).

[5] 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., 1, 449-450 (2001).

[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", Acoust. Sci. Tech., 23(4), 229-232 (2002).

[7] S. Furui, "On the role of spectral transition for speech perception", J. Acoust. Soc. Am., 80(4), 1016-1025 (1986).

[8] 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(6), 4055-4064 (2006).

[9] 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(2), 212-214 (2005).

[10] Y. Miyauchi, N. Hodoshima, K. Yasu, N. Hayashi, T. Arai and M. Shindo, "A preprocessing technique for improving speech intelligibility in reverberant environments: The effect of steady-state suppression on elderly people", Proc. Interspeech, 2769-2772 (2005).

[11] N. Hodoshima, Y. Miyauchi, K. Yasu and T. Arai, "Steady-state suppression for improving syllable identification in reverberant environments: A case study in an elderly person", Acoust. Sci. Tech., 28(1), 53-55 (2007).

[12] CATT, Bengt-Inge Dalenbäck, Gothenburg, Sweden (www.catt.se).

[13] J. L. Flanagan, "Analog measurements of sound radiation from the mouth", J. Acoust. Soc. Am., 32(12), 1613-1620 (1960).

[14] J. E. Summers, R. R. Torres, Y. Shimizu and B. L. Dalenbäck, "Adapting a randomized beam-axis-tracing algorithm to modeling of coupled rooms via late-part ray tracing", J. Acoust. Soc. Am., 118(3), 1491-1502 (2005).