

ACOUSTICAL LETTER

Comparison of consonant identification improvements by steady-state suppression via a loudspeaker system between with and without natural sounds from a talker in reverberation

Nao Hodoshima[1,*], Peter Svensson[2,†] and Takayuki Arai[1,‡]

[1]*Department of Information and Communication Sciences, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim, NO-7491, Norway*

(*Received 30 September 2007, Accepted for publication 18 January 2008*)

Keywords: Speech enhancement, Reverberation, Speech intelligibility, Sound reinforcement systems
PACS number: 43.72.Ew, 43.55.Hy, 43.38.Tj [doi:10.1250/ast.30.59]

1. Introduction

Reverberation makes speech perception difficult for people with hearing impairments and for elderly people compared with young people having normal hearing (e.g. [1]). Therefore, reducing the effect of reverberation may help us to build "barrier-free listening environments" in public spaces where speech signals will be intelligible for various types of people.

One electroacoustical approach to improving speech intelligibility in reverberant environments is preprocessing (i.e. speech signals are emitted over a loudspeaker in a room) [2–5]. An example of preprocessing is steady-state suppression (SSS) [4,5], in which steady-state portions of speech (e.g. vowel nuclei) are suppressed. These portions have high energy and therefore cause a large amount of "overlap-masking" (i.e. reverberation tails mask the following phonemes) [6], but these portions are relatively less important for syllable perception compared with spectral transitions [7]. Therefore, SSS reduces the effect of overlap-masking with little degradation of the information needed for speech perception. Several listening tests have shown that SSS significantly improves syllable identification for young people with normal hearing [8,9] and for elderly people [10,11] in diotic and dichotic listening environments.

The previous studies on SSS [8–11] simulated the situations in which listeners hear processed electroacoustical sounds (e.g. recorded announcements over loudspeakers at a train station), and in which input to a public address (PA) system is dry speech. When a talker and listeners are present in the same enclosure (e.g. during lectures), the listeners hear both natural (unprocessed) sounds from the talker and processed electroacoustical sounds from the loudspeakers. Furthermore, the input to the PA system is not completely dry speech because speech signals picked up by a microphone close to the talker include electroacoustical (reverberant) sounds.

In this study, we investigated the effect of SSS in situations where listeners hear natural and electroacoustical sounds. We are interested in 1) how multiple-sound-source listeners hear (a single source as an electroacoutical path or two sources as natural and electroacoutical paths), and how 2) loudspeaker gain, and 3) the input to the PA system (dry or reverberant speech) affect the performance of SSS. These must be tested, to enable the practical application of SSS, from the following points of view. 1) The talker–listener distance (e.g. level differences between natural and electroacoutical sounds) would affect the performance of SSS. For example, when the sound level from a loudspeaker is much higher than that of natural sounds, a listener would benefit greatly from the processed sound. In contrast, when the sound level difference between natural and electroacoutical sounds is much smaller as the listener is closer to the talker, the effect of SSS would be much less compared with the previous situation because of the interaction of the two sounds. 2) The effect of SSS would increase nonlinearly as loudspeaker gain is increased. For example, the effect of SSS would increase when the gain is increased to a certain level, and then would remain the same when the gain is increased above that level. Studying the relationship between the gain and SSS allows us to find the appropriate gain and parameters of SSS in an enclosure. 3) The effect of the mixture of electroacoustical sounds with sounds from a talker at the talker microphone would be negligible on the performance of SSS because the level of electroacoustical sound is much larger compared with that of sound from a talker.

In order to verify these hypotheses, we carried out a listening test on young people with normal hearing using natural and steady-state suppressed speech under simulated reverberant environments in which PA systems were virtually installed on a computer. Diotic listening was used for simulated reverberant environments in order to compare the effect of SSS in this study with those in previous studies [8–11].

2. Listening test

2.1. Hypotheses

We investigated the effect of SSS on the listener side and the talker side in situations where a talker and listeners are present in the same enclosure. The hypotheses on the listener side were that 1) SSS improves speech intelligibility in situations where listeners receive sounds either from a single source (e.g. a talker) or from two sources (e.g. the talker and

*e-mail: n-hodosh@sophia.ac.jp
†e-mail: svensson@iet.ntnu.no
‡e-mail: arai@sophia.ac.jp

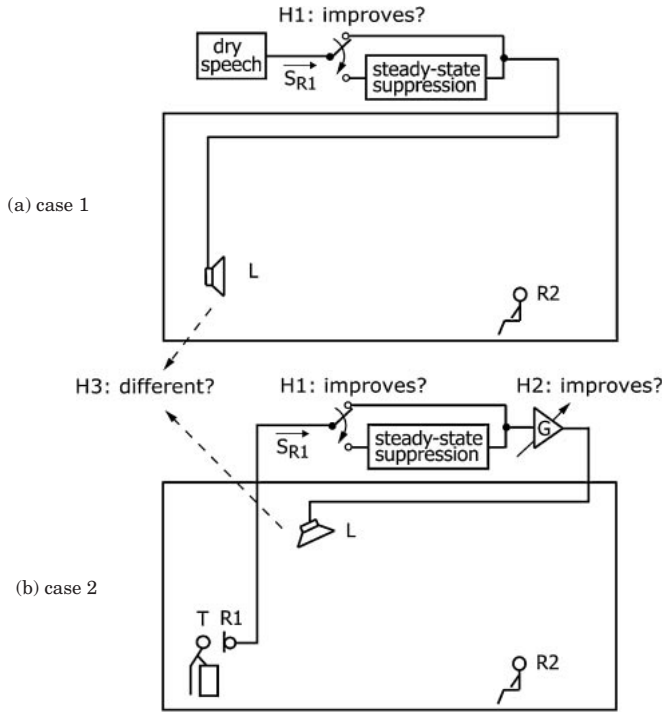

Fig. 1 Room simulation setting in a vertical section for case 1 (a) and case 2 (b). Sources (T: talker, L: loudspeaker) and receivers (R1: microphone close to T, R2: listener) are included. S_{R1} is input to the PA system. G is the loudspeaker gain. H1–H3 are hypotheses.

loudspeaker), the sound level from the loudspeaker being much higher than that from natural sounds, and 2) the improvement of speech intelligibility as a result of SSS increases as the loudspeaker gain increases. The hypothesis on the talker side was that 3) the effect of the mixture of electroacoustical sounds with sounds from a talker at the talker microphone is negligible when the talker-microphone distance is short (e.g. 15 cm). We simulated room conditions to test these hypotheses, as explained in the next section.

2.2. Room Simulation

Two simple rooms (one small and one large) of shoebox shape and 100% diffusion on the walls were simulated using the software CATT-Acoustic [12]. The small room had a volume of 500 m^3 and a reverberation time (RT) of 1.2 s, and the large room had a volume of 15,750 m^3 and RT of 1.8 s.

Figure 1 shows the room simulation settings in a vertical section. Source and receiver directivities were a measured one [13] for the talker (T), a vertical array for the loudspeaker (L), a cardioid for a clip-on microphone at a 15 cm distance from the talker (R1) and omnidirectional for the listener (R2). R2 was slightly deviated from the main direction of the loudspeaker in order to avoid the situation of the speech intelligibility at R2 being too high. The PA system consisted of a SSS unit and loudspeaker gain. The input to the PA system (S_{R1}) was either dry speech or speech captured by R1.

Table 1 summarizes how our hypotheses relate to two inputs to the PA system, the number of sources R2 receives and three gain conditions used in the room simulation. The gain is set as a relative level to natural sounds. For hypotheses

Table 1 Relationships between the hypotheses, the input of PA systems, the number of sources R2 receives and the gain conditions used in the room simulation.

Hypothesis	Input of PA system	Number of sources R2 receives	Increased level at R2 (re: natural sounds)
1 and 2 (listener side)	dry	single from T (case 1)	0.0 dB: gain 0
	talker microphone	two from T and L (case 2)	2.2 dB: gain 1 5.5 dB: gain 2
3 (talker side)	dry		
	talker microphone		5.5 dB: gain 2

1 and 2 on the listener side, we prepared two cases. Case 1 shown in Fig. 1(a) had a single source from T, dry speech as S_{R1} and gain 0 (no gain). Case 2 shown in Fig. 1(b) had two sources from T and L, speech from the talker microphone as S_{R1} and two gains (gain 1 and 2). Gain 0 was not included in case 2 since the sound level from L would be much larger than that from the natural sound when listeners simultaneously hear natural and electroacoustical sounds. For hypothesis 3 on the talker side, we used two inputs: dry speech and speech from the talker microphone.

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

2.3. Stimuli

Speech materials were 14 nonsense consonant-vowel syllables (vowel: /a/, and consonants: /p, t, k, b, d, g, s, ʃ, h, dz, dʒ, tʃ, m, n/) embedded in a Japanese carrier phrase obtained from the ATR speech database of Japanese, and were the same as those used in the previous studies [8–11]. The speech materials were recorded by a 40-year-old male at a sampling frequency of 16,000 Hz. Each sentence was scaled so that the A-weighted energy was equal for all speech materials under all gain conditions and in both rooms.

Stimuli for testing hypotheses 1 and 2 were sounds from T (case 1) as specified by Eq. (1), and sounds from T and L (case 2) as given by Eq. (2).

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

$$s(t) * IR_{\text{T-R2}} + G \cdot T\{s_{\text{R1}}(t)\} * IR_{\text{L-R2}} \tag{2}$$

$$\text{where} \quad T\{s(t)\} = \begin{cases} s(t) & (\textit{unprocessed condition}) \\ P(s(t)) & (\textit{processed condition}) \end{cases},$$

$s(t)$ indicates a speech signal, G indicates the gain of the loudspeaker, S_{R1} indicates input to the PA system, $P(s(t))$ indicates a processed signal whose amplitudes of steady-state portions of $s(t)$ are suppressed to 40% and $*$ indicates convolution. Note that no processing was applied for natural sounds in case 2. Also note that the processed condition in

case 1 indicates "a supertalker" (i.e. a talker speaking steady-state suppressed speech), and it seems somewhat unrealistic. However, this condition arises when we use recorded sounds, and this condition was adopted in order to compare the results of processed sounds with those of natural sounds (natural in case 1) as well as to compare the situation of this study with those of previous ones [8–11] by recreating the situation used in the previous studies [8–11] (i.e. the input to the PA system was dry speech and sounds were sent from a single source) in these simulated reverberant conditions.

For hypothesis 3, stimuli were dry speech, and speech captured by R1 in the large room with gain 2. Stimuli for speech captured by R1 are as given by

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

where $s(t)$, G, S_{R1} and $*$ are the same parameters as in Eqs. (1) and (2).

2.4. Participants

Participants were 20 native speakers of Japanese (4 males and 16 females, 22 to 37 years old). They had normal hearing with air-conduction thresholds of less than 25 dB HL from 125 to 8 kHz for both ears.

2.5. Procedures

The listening test was conducted in a sound-treated room. The stimuli were presented over headphones (STAX, SR-303) through a digital audio amplifier (Onkyo, MA-500U) that was connected to a computer. Each participant could adjust the playback level to a comfortable level. In each trial, a stimulus was presented, after which a computer monitor displayed the 14 syllables used in the listening test. The participants were instructed to use the mouse to click on the syllable heard on the monitor. Trials with 168 stimuli for hypotheses 1 and 2 (natural/processed × small/large rooms × three gains × 14 speech materials) were randomly presented first, followed by 28 stimuli for hypothesis 3 (two inputs × 14 speech materials) presented randomly.

3. Results and discussion

3.1. Listener side

Figure 2 shows the mean percent correct responses (mean scores) at the listener position for each room, gain, and processing condition. Separate repeated-measures ANOVAs were carried out for case 1, with room (small/large) and processing (natural/processed) as two factors, and for case 2 with room (small/large), processing (natural/processed) and gain (gain1/gain2) as three factors.

For case 1, the small room had significantly higher mean scores than the large room [$p < 0.01$]. Processed speech also had significantly higher mean scores than natural speech [$p < 0.01$]. These results were consistent with previous findings [8–11]. No interaction was significant. Post-hoc analyses showed no significant difference between mean scores of natural and processed speech for the two room conditions.

For case 2, the small room had significantly higher mean scores than the large room [$p < 0.01$]. Processed speech also had significantly higher mean scores than natural speech [$p = 0.01$]. These results were consistent with case 1. Gain 2 also had significantly higher mean scores than gain 1 [$p = 0.02$]. Interactions between room and gain [$p = 0.01$] and between room and processing [$p = 0.02$] were also significant. Post-hoc analyses showed that gain 2 had significantly higher mean scores than gain 1 in the small room [$p = 0.02$ for natural and $p = 0.01$ for processed], while no significant difference was found between gain 1 and gain 2 in the large room. These interactions and further analyses showed that the mean scores increase mainly owing to loudspeaker gain in the small room, while the mean scores increase mainly as a result of SSS in the large room. Post-hoc analyses also showed that processed speech performed significantly better than natural 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 SSS.

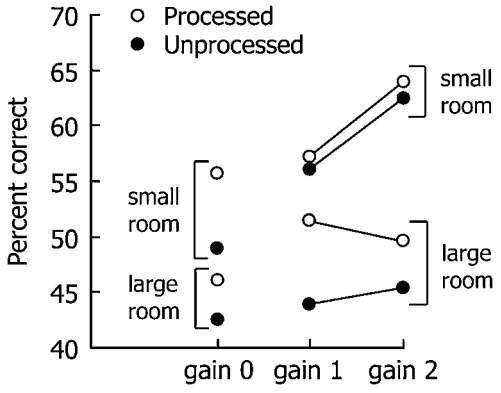

Fig. 2 Syllable identification scores at the listener position for each room, gain and processing condition.

Processed speech had higher mean scores under all gain conditions in this study. This suggests that SSS improves syllable identification in both situations where a talker and a listener are either in the same room or in different rooms.

3.2. Talker side

A t-test showed no statistically significant difference in the mean percent correct responses between dry speech (100.0%) and speech captured by a talker microphone (99.7%). This indicates that the effect of SSS would be the same for the two inputs when the direct-to-reverberation ratio is high at the talker microphone. This would be useful for a practical use of SSS. For a recorded announcement as well as a live one in the same enclosure (e.g. an automatic announcement of the arrival of trains and a live announcement of a delayed train in a station), the maximum improvement in speech intelligibility would be obtained by SSS with the same parameters.

4. Conclusions

The effect of SSS was studied in the situation where a listener hears both natural and electroacoustical sounds. Results from the listening test showed that 1) SSS improved syllable identification in cases when listeners receive sounds either from a single source (e.g. an electroacoustical path) or from two sources (e.g. natural and electroacoustical paths), 2) increasing the loudspeaker gain changed the performance of SSS, and 3) the effect of the mixture of electroacoustical sounds with sounds from a talker at the talker microphone

is negligible when the direct-to-reverberation ratio is high. Previous studies [10,11] indicated that SSS also might be effective for elderly people and people with hearing impairments, and therefore, future research should include tests on the effect of SSS under the current conditions with these populations in order to build "barrier-free listening environments". For a practical application of SSS, future research should be on the binaural listening condition as well as finding directivities and loudspeaker gains that are suitable for SSS.

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 and by the Research Council of Norway through the Acoustic Research Centre project.

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**, 1242–1248 (1982).

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

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

[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.*, Vol. 1, pp. 449–450 (2001).

[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," *Acoust. Sci. & Tech.*, **23**, 229–232 (2002).

[6] 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).

[7] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, **80**, 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**, 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**, 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*, pp. 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**, 53–55 (2007).

[12] CATT, B. L. Dalenbäck, Gothenburg, Sweden (www.catt.se).

[13] J. L. Flanagan, "Analog measurements of sound radiation from the mouth," *J. Acoust. Soc. Am.*, **32**, 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**, 1491–1502 (2005).