Steady-state Suppression in Reverberation: A Comparison of Native and Nonnative Speech Perception

Nao Hodoshima†, Dawn Behne‡ and Takayuki Arai†

† Department of Electrical and Electronics Engineering
Sophia University, Tokyo, Japan
n-hodosh@sophia.ac.jp

‡ Department of Psychology
Norwegian University of Science and Technology, Trondheim, Norway

Abstract

This study investigated whether the steady-state suppression method proposed by Arai *et al*. (2001, 2002) improved consonant identification for nonnative listeners in reverberation. It also compared the effect of steady-state suppression on consonant identification by native and nonnative listeners in reverberation. We used steady-state suppression as a preprocessing technique which processes speech signals before they are radiated from loudspeakers in order to reduce the amount of overlap-masking. Participants were 24 native English (native listeners) and 24 Japanese speakers (nonnative listeners), both with normal hearing. A diotic Modified Rhyme Test was conducted with and without steady-state suppression for reverberation times of 0.4, 0.7 and 1.1 s and a non-reverberant condition. The results showed that native listeners performed better than nonnative listeners, and that the mean percentage of correct answers in initial consonants was higher than in final consonants. The results also showed that processed and unprocessed speech was comparable for word initial and final consonants. These findings indicate that parameters of steady-state suppression would need adjustment to accommodate speech materials and reverberant conditions. They also suggest that the difficulties that nonnative listeners have might not be due to the actual acoustic-phonetic information from the signal.

Index Terms: speech enhancement, nonnative listeners, reverberation, steady-state suppression

1. Introduction

With the increase in internationalization, more and more opportunities for listening or speaking a foreign language arise. It has been reported that nonnative listeners [1][2] have more difficulty understanding speech under noisy and/or reverberant environments than native listeners. Low noise and reverberation levels are therefore preferable in public spaces where nonnative languages are commonly in use, such as international airports, stations, and conference rooms. Care must also be taken in classrooms, lecture halls, etc for educational purposes.

Hazan and Simpson [3][4] reported a speech enhancement method for nonnative listeners in noise. They enhanced consonantal regions of VCV stimuli by processing stimuli before adding noise, and presenting them in a background of noise with the long-term average spectrum corresponding to the speech signal. Nonnative listeners groups obtained significantly higher intelligibility scores for the enhanced stimuli compared to natural stimuli.

Reduced speech intelligibility from reverberation is primarily due to overlap-masking [5][6]. Arai *et al*. [7][8] proposed reducing the effect of overlap-masking by steady-state suppression as a pre-processing approach. This technique suppresses steady-state portions of speech that are not necessary for syllable perception [9], such as vowel nuclei. Steady-state suppression statistically improved consonant identification for young normal hearing listeners [10][11] and older listeners [12] under diotic listening conditions and a hall at reverberation times (RTs) of 0.7 to 1.3 s. To our knowledge, steady-state suppression has not previously been suggested for nonnative listeners in reverberant environments.

The purpose of the current study was 1) to investigate whether steady-state suppression proposed by Arai *et al*. [7] [8] improved consonant identification for nonnative listeners in reverberation and 2) to compare the effect of steady-state suppression on consonant identification by native and nonnative listeners in reverberant environments. This was done by testing native English and Japanese listeners in an English diotic Modified Rhyme Test (MRT) conducted with and without steady-state suppression for three reverberant conditions and a non-reverberant condition.

2. Experiment

2.1. Participants

Table 1 shows two listener groups included in the study: "native listeners" had English as a first language, whereas "nonnative listeners" all had Japanese as their first language. Native listeners were all living in Norway when the experiment was run and 16 of them had used Norwegian as a foreign language daily or weekly for an average of 5.7 years. Six of them had not learned any foreign languages. Nonnative listeners were considered to have an average level of English proficiency in Japan based on the following criteria: 1) they had English as a second language, 2) they never lived abroad, 3) they attained a C ranking on the TOEIC (Test of English for International Communication) proficiency scale, or attained a intermediate or primary ranking on the English test which all first year students take at Sophia University, where the

Table 1 Summary of 'native listeners' and 'nonnative listeners' group characteristics.

	Native listeners	Nonnative listeners
Native language	English	Japanese
Numbers	14 males, 10 females	6 males, 18 females
Age	18 to 50 years (average: 31 years)	20 to 30 years (average: 22 years)
Thresholds	less than 25 dB HL from 250 to 4 kHz	less than 25 dB HL from 125 to 8 kHz

experiment was conducted. Nonnative listeners began learning English when they were, on average, 11 years old, and had studied English for an average of 11 years. The different threshold frequency ranges for native and nonnative listeners was due to the use of different audiometers at the two testingsites. None of the subjects reported a history of unusual noise exposure or listening difficulties.

2.2. Speech materials

The version of the MRT used in Kusumoto *et al*. [13] was also used in the current study for direct comparison with previous findings [1][2]. The target words developed by Kruel *et al*. [14] were embedded within the carrier phrase "You will mark the ___, please." All 6 lists of sentences, each composed of 50 monosyllabic words, were used. The speaker was a 34-year-old male native speaker of standard American English. The average intensity of the stimuli was normalized across sentences.

Two processing conditions were used in this study: original (unprocessed) speech and steady-state suppression as was used by Arai *et al*. [7][8]. The steady-state suppression method calculates the *D* parameter to detect spectral transitions of a speech signal [9], and defines speech portions as steady-state when *D* is less than a specified threshold. Once a portion is considered steady-state, the amplitude of the portion is multiplied by a factor of 0.4, giving a 40% suppression rate.

Speech materials in both processing conditions were reproduced with three reverberant conditions: RT of 0.4, 0.7 and 1.1 s. The impulse responses were obtained by multiplying exponential decays by the impulse response measured in Hamming Hall in Tokyo as described in [15]. The RT values are the average RTs derived from the Early Decay Time at the center frequencies of 0.5, 1, and 2 kHz of the 1-octave bandpass filtered impulse response.

Each of the six conditions (2 processing conditions x 3 reverberant conditions) was assigned to one of the 6 MRT lists and counterbalanced across subjects. An Additional 50 sentences were randomly selected from across the six lists and used as stimuli in a non-reverberant condition. This gave a total of seven conditions.

2.3. Procedure

Speech materials were presented over headphones in sound treated rooms at Sophia University and the Norwegian University of Science and Technology. Headphones were STAX SR-303 (electrostatic, open-back, push-pull type at frequency ranges of 7-41000 Hz) for nonnative listeners and AKG K271 (dynamic, closed-back type at frequency ranges of 16-28000Hz) for native listeners. The sound level was adjusted to a comfortable level for each subject beforehand, and maintained throughout the experiment.

Each participant was tested in all seven conditions in a consonant identification task. Half of the trials were word-initial identification tasks, while the rest was word-final identification tasks. In any given trial, a test sentence was presented over headphones, after which a PC monitor displayed six words which differed in a single initial or final consonant depending on tasks. Participants were instructed to mouse-click the word they heard on the monitor. Once they had selected a word, the next trial was presented. Trials with the 300 reverberant stimuli were randomly presented first, followed by the 50 randomly presented stimuli in the non-reverberant condition. The reverberant condition preceded the non-reverberant condition, rather than randomizing the conditions, so that familiarity with the stimuli from the non-reverberant condition would not affect the reverberant condition. Before starting the experiment, listeners had five practice trials to become familiar with the procedure.

2.4. Results

The mean percentage of correct answers (scores) for each reverberant and processing condition was calculated. The scores for the different conditions are an average of the 24 participants. Figure 1 shows the mean scores by native and nonnative listeners for each of the reverberant and processing conditions collapsed across word position. Figure 2 presents the results with initial and final word position shown separately. Both figures show native (nl) and nonnative (nnl) listeners' mean scores presented in the original (org), steady-state suppression (proc) conditions for non-reverberant stimuli and three different reverberation times. For example, "nl_org" means native listeners' mean score in the original condition. A 2 x 2 x 3 x 2 mixed ANOVA was carried out with listener group as a nonrepeated factor, consonant position, RT and processing as repeated variables, and scores as the dependent variable. Results show that native listeners had a higher score than nonnative listeners [$F(1,46) = 119.00$, $p < 0.01$], and scores was higher for initial consonants than for final consonants [$F(1,46) = 1881.06$, $p < 0.01$]. The score also reliably differed across RTs [$F(2,92) = 68.47$, $p < 0.01$]. RT means shown in Figure 1 show a general pattern of decreasing score as RT increases. No reliable difference in score was observed between the steady-state and unprocessed conditions. In addition to these main effects, significant interactions were observed between consonant position and listener group [$F(1,46) = 111.32$, $p < 0.01$] and between consonant position and reverberant condition [$F(2,92) = 16.20$, $p < 0.01$]. Other interactions were not significant.

In the non-reverberant condition, a 2 x 2 x 2 mixed ANOVA was carried out with listener group as a nonrepeated factor, consonant position and processing as a repeated variable, and score as the dependent variable. Results show that native listeners had a higher score than nonnative listeners [$F(1,46) = 96.73$, $p < 0.01$]. The score was also higher for initial consonants than for final consonants [$F(1,46) = 57.16$, $p < 0.01$]. In addition to these main effects, the interaction between listener group and consonant position was significant [$F(1,46) = 29.63$, $p < 0.01$].

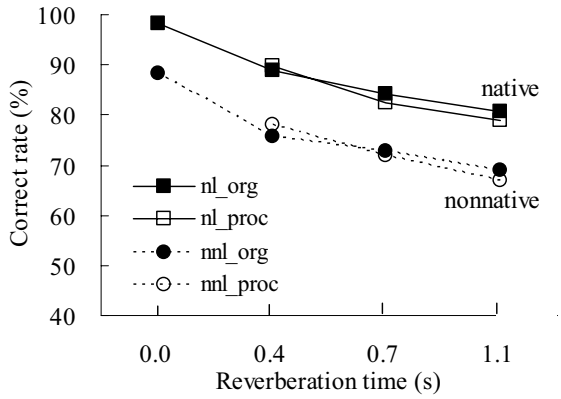

Figure 1 Native (nl) and nonnative (nnl) listeners' mean score presented in the original (org), steady-state suppression (proc) conditions for non-reverberant stimuli and three different reverberation times.

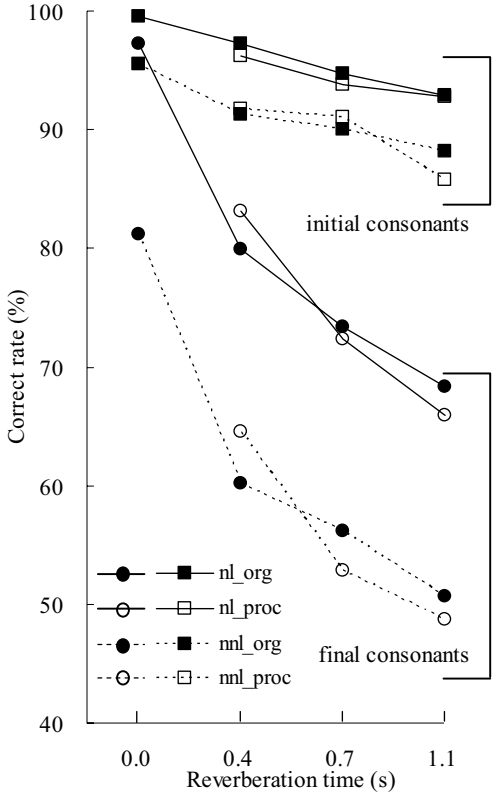

Figure 2 Native (nl) and nonnative (nnl) listeners' mean score for word initial and final consonants presented in the original (org), steady-state suppression (proc) conditions for non-reverberant stimuli and three different reverberation times.

Table 2 The difference between the mean score (%) for native and nonnative listeners in the non-reverberant, unprocessed and processed conditions for initial and final consonants.

	Non-reverberant	Unprocessed	Processed
Initial consonants	4.0	5.1	4.8
Final consonants	16.0	18.2	18.4

3. Discussion

3.1. Native and nonnative listeners

Consistent with previous findings [1][2], the mean score was higher for native listeners than nonnative listeners in the reverberant conditions. However, there was no significant interaction between listener group and reverberant condition.

For non-reverberant speech, this conflicts with findings for word identification in non-reverberant speech [2]. This inconsistency might be due to a difference between the participants' English proficiency level. Nonnative listeners in the word identification study [2] lived abroad for a couple of years, while nonnative listeners in the current study had not been living abroad. Interestingly, in the current study, the difference between the mean score for native and nonnative listeners in non-reverberant speech was close to that in the reverberant conditions (see Table 2).

3.2. Scores in consonant positions

In both word initial and final consonants the mean score decreased as RT increased, but at a different rate: final consonants had a greater decrease than initial consonants. In the reverberant conditions, as well as in the non-reverberant condition, the mean score in initial consonants was nevertheless higher than in final consonants, as was also observed in [16]. One reason for this might be production differences: initial consonants generally have higher intensities than final consonants [17]. The focus for future research could instead be on differences between initial and final consonants with different RTs. The basis for this is yet to be investigated and is a topic for future research.

In addition, the difference between the mean score for initial and final consonants was larger for nonnative listeners than for native listeners under the reverberant conditions (20.8% difference for native listeners and 34.2% for nonnative listeners). That nonnative listeners have more difficulty identifying final consonants than initial consonants is consistent with previous findings (e.g., [2]).

3.3. The effect of steady-state suppression

Under the conditions tested in this study, the results for processed and unprocessed speech were comparable for word initial and final consonants, and for native and nonnative listeners. The increased magnitude of the modulation spectrum from steady-state suppression is nevertheless known to reduce the effect of reverberation [10]. Steady-state suppression preemptively enhances the modulation spectrum important for

speech perception around 4 Hz and above 10 Hz, which may have lead to reverberation not reducing the modulation index; that is, speech intelligibility [10]. A possible reason for not observing a reliable improvement in consonant identification with steady-state suppression in this study might be that parameters such as the suppression rate would need adjustment to accommodate the speech material used in this study. Future research may lead to appropriate parameters for speech materials and reverberant conditions.

That steady-state suppression did not improve consonant identification for native listeners in this study suggests that the native listeners perceptual processing in this task might not have been dependent on the acoustic signal alone, and that their experience with the language gave them an adequate knowledge base to manage the task. That no difference between native and nonnative listeners was observed for steady-state suppression suggests that the difficulties that nonnative listeners have may be at a deeper level of processing than extracting the actual acoustic-phonetic information from the signal. Although steady-state suppression might help listeners (native and nonnative, alike) to extract acoustic-phonetic information by reducing the amount of reverberation, they still might not be able to use that information given due to their lower level of proficiency with the language. A direction for future research would be to follow up on this and compare responses by native listeners and nonnative listeners as well as children, who like nonnative listeners, have less speech and language experience.

4. Conclusions

The current study 1) investigated whether steady-state suppression proposed by Arai *et al*. [7][8] improved consonant identification for nonnative listeners of English in reverberation and 2) compared the effect of steady-state suppression on consonant identification by native English listeners and nonnative listeners of English in reverberant environments. Contrary to expectations, results showed no difference between native and nonnative listeners for steady-state suppression under the current conditions. These results indicate that parameters such as the suppression rate would need adjustment to accommodate speech materials and reverberant conditions. These findings also suggest that the difficulties nonnative listeners have in reverberant environments might be due to other aspects of speech perception than extracting acoustic-phonetic information from the signal.

5. Acknowledgements

This research was supported by Grants-in-Aid for Scientific Research (A-2, 16203041) and Grants-in-Aid for JSPS Fellows (176911), both from the Japan Society for the Promotion of Science. The authors would like to thank Hideki Tachibana, Kanako Ueno and Sakae Yokoyama for offering the use of the impulse response data.

6. References

[1] A. K. Nábělek and A. M. Donahue, "Perception of consonants in reverberation by native and non-native listeners", *J. Acoust. Soc. Am*., 75(2):632-634, 1984.

[2] Y. Takata and A. K. Nábělek "English consonant recognition in noise and in reverberation by Japanese and American listeners", *J. Acoust. Soc. Am*., 88:663-666, 1990.

[3] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant perception by non-native listeners: Preliminary results", *Proc. StiLL Workshop*, 119-122, 1998.

[4] V. Hazan and A. Simpson, "The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects", *Language and Speech*, 43(3):273-294, 2000.

[5] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation", *J. Acoust. Soc. Am*., 21:577-580, 1949.

[6] A. K. Nábělek 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.

[7] 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 (in Japanese).

[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", *Acoust. Sci. Tech*., 23:229-232, 2002.

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

[10] 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.

[11] 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.

[12] 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. Eurospeech*, 2769-2772, 2005.

[13] A. Kusumoto, T. Arai, K. Kinoshita and N. Hodoshima, "Modulation enhancement of speech by preprocessing for improving intelligibility in reverberant environment", *Speech Com*., 45(2):101-113, 2005.

[14] E. J. Kreul, J. C. Nixon and K. D. Kryter, "A proposed clinical test of speech discrimination", *J. Speech Hear. Res*., 11:536-552, 1968.

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

[16] V. O. Knudsen, "The hearing of speech in auditoriums", *J. Acoust. Soc. Am*., 1(1):56-82, 1929.

[17] M. A. Redfold and R. L. Diehl, "The relative perceptual distinctiveness of initial and final consonants in CVC syllables", *J. Acoust. Soc. Am*., 106(3):1555-1565, 1999.