The Effect of Pre-processing for Improving Speech Intelligibility in the Sophia University Lecture Hall

Nao Hodoshima, Takahito Goto, Noriko Ohata, Tsuyoshi Inoue, Takayuki Arai
Sophia University, Tokyo, JAPAN
n-hodosh@sophia.ac.jp

Abstract

One reason reverberation degrades speech intelligibility in an auditorium is the effect of overlap-masking, which occurs when segments of an acoustic signal are affected by reverberation components of previous segments (Bolt and MacDonald, JASA, 1949). Arai et al. suggested a pre-processing technique to prevent the effect of overlap-masking. Their technique champions the suppression of steady-state portions of speech that have more energy but that are less crucial for speech perception than are transitional portions of speech (Arai et al., Acoust. Sci. and Tech., 2002). Arai et al. confirmed promising results for improving speech intelligibility (Arai et al., AST, 2002). Hodoshima et al. explored the effect of steady-state suppression with a set of artificial impulse responses and showed that steady-state suppression is an effective pre-processing method to prevent degradation of speech intelligibility under specific reverberant conditions (Hodoshima et al., China-Japan Joint Conf. on Acoust, 2002).

In this study, we conducted a perceptual test using Arai's pre-processing technique in an actual environment: the largest lecture hall in Sophia University in Tokyo, Japan, which has a reverberation time of approximately 1.0s. This study confirmed that steady-state suppression is an effective pre-processing method for improving speech intelligibility not only in simulated reverberant conditions but in an actual hall.

1. Introduction

In a large auditorium, understanding speech may be difficult. One reason is reverberation caused by a superposition of reflected sounds with various delays and amplitudes. Although reverberation adds richness of sound for music, it makes speech more difficult to understand. One of the reasons reverberation degrades speech intelligibility is overlap-masking, which occurs when reverberation tails of previous portions of a sound affect subsequent segments [1, 4].

To improve speech intelligibility in reverberant environments, there are three general approaches: microphone-array, post-processing and pre-processing. A post-processing method such as inverse filtering (e.g. [5,6]) and modulation filtering (e.g. [7,8]) is applied to a speech signal already released into a room and affected by reverberation. The pre-processing approach, which processes a speech signal before it is affected by reverberation, is a method that reduces the influence of reverberation on the transmission path. Since pre-processing operates on a speech signal between a microphone and loudspeaker, this method can be used with a Public Address (PA) system.

Furui showed that spectral transitions are crucial for syllable and vowel perception, where as vowel nuclei (i.e. steady state portions) are not necessary for either vowel or syllable perception [9]. This is because the information in the steady-state portions of the speech signal is redundant with that in transient segments [10]. Also, both "delta" processing of cepstral features [9] and RelAtive SpecTrAl (RASTA) processing [10] may be used to enhance transitions of speech, and they have also been shown to increase recognition rates in automatic speech recognition.

Arai's pre-processing method obtained clear improvements by reducing the masking influence caused by the reverberation components of the previous portion as described in [2]. Hodoshima et al., [3] and Inoue et al., [11] conducted perceptual tests in a soundproof room to confirm the effectiveness of Arai's technique [2] with a set of artificial reverberation conditions, in which reverberation times were 0.4-1.3s. Clear improvements were obtained with reverberation times of 0.8-1.2 s.

The purpose of this paper is to investigate whether steady-state suppression improves speech intelligibility in an actual environment. The actual environment used in this study was the largest lecture hall at Sophia University in Tokyo, Japan. In Section 2, we describe the signal processing of the steady-state suppression technique proposed by Arai [2]. Section 3 presents a perceptual experiment. Section 4 contains result of this experiment. Finally, discussion and conclusions are detailed in Section 5.

2. Signal Processing

2.1. The effect of overlap masking

Overlap-masking is the main reason reverberation degrades speech intelligibility [1, 4]. Fig. 1 shows an illustration of overlap-masking. We used the English word "October" as a speech sample, from the Multi-Lingual Speech Corpus created by the Department of Eastern and Western Linguistic Culture, Tsukuba University in Japan. The speaker we used was an English male. The left part of Fig. 1 shows the original speech signal, with no reverberation. The left bottom panel shows the original utterance, while the left top five panels are the waveforms for each phoneme /o/, /k/, /t/,

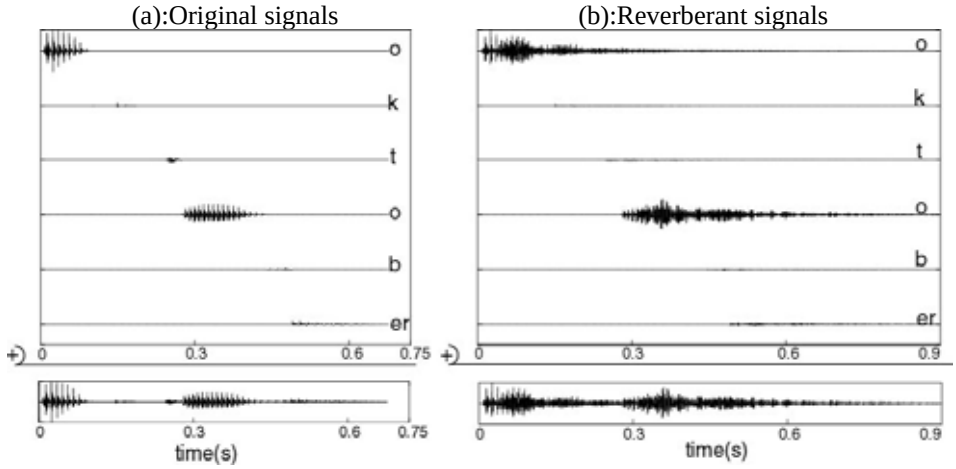

Figure 1: An illustration of overlap-masking for the word "October"

/o/, /b/ and /er/, manually segmented. The right part of Figure 1 shows the reverberant speech signals corresponding to the original signals in the left part of this figure. The reverberant signals were obtained by taking the convolution of the original speech signals with an impulse response of a room having a 1.1s reverberation time. The entire waveform of this utterance shown in the bottom panel is a summation of the top five waveforms, each of which corresponds to a constituent. As you can see in Figure 1, consonants that have weak energy such as /k/, /t/, /b/ are masked by the reverberation tails of the previous vowels. This indicates that when the previous phoneme has strong energy, such as with a vowel, the following phoneme is much more affected by the preceding reverberation tails.

2.2. Steady-state suppression

We used the steady-state suppression method proposed by Arai et al. [2], to suppress steady-state portions of the speech. Figure 2 is a diagram of our steady-state suppression technique. First, an original signal was split into 1/3-octave bands. Then, the temporal envelope was extracted from each band. After down-sampling, 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 [5]. 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 a factor of 0.4 (a suppression rate of 40%). We used this factor because a pilot study confirmed it is a reasonable degree of suppression [2] when the reverberation time is 1.1s.

3. Perceptual Experiment

3.1. Reverberation condition

Using the TSP signal, we measured the impulse response of the largest lecture hall in Sophia University, which seats 822 people (see Figure 3).

To calculate the reverberation time, we used Early Decay Time (EDT), which is the time it takes for 10 dB of reverberation decay, and we multiplied it by six to estimate the reverberation time. The average reverberation time in the hall at the center frequency of 0.5, 1, and 2 kHz of the 1-oct bandpassed impulse response is about 1.0 s.

3.2. Stimuli

We used the same stimuli as Hodoshima et al. [3] used. The original speech samples consisted of 14 nonsense Consonant-Vowel (CV) syllables embedded in a Japanese carrier phrase. The vowel was /a/ and the consonants were /p/, /t/, /k/, /b/, /d/, /g/, /s/, /ʃ/, /h/, /dz/, /dʒ/, /tʃ/, /m/ and /n/. 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 the sentence data set. We normalized the root-mean-square (RMS) energy in the CVs that have the same vowel, and then normalized the ratio of RMS in the carrier phrase relative to RMS energy in the CVs.

The stimuli have two conditions: the original (unprocessed) signals (Org) and the processed signals (Proc). Fifty-six stimuli were prepared (14 CVs x with/without processing × 2 repetitions) in total. The stimuli were arranged randomly.

3.3. Subjects

Twenty-four normal hearing subjects (12 males and 12 females with an average age of 23.5 years) participated in the experiment. All were native speakers of Japanese.

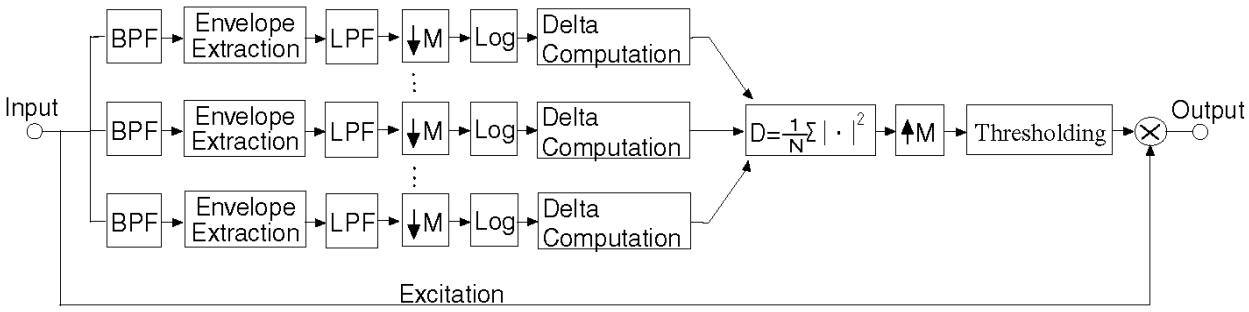

Figure 2: Block diagram of steady-state suppression

Figure3: The largest lecture hall at Sophia University

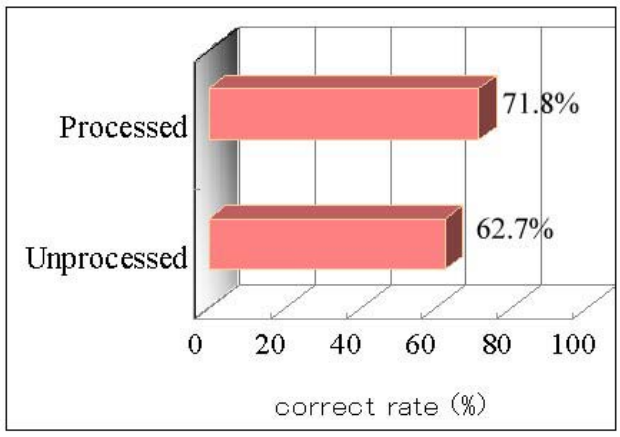

Figure4: The correct rate of perceptual experiment

3.4. Procedure

The experiment was conducted in the largest lecture hall in Sophia University. Each subject sat in the back-center portion of the hall. There was an empty seat between each subject. The stimuli were presented through two loudspeakers in the center of the stage. The sound level was adjusted to subjects' comfort level before the perceptual test began. In the perceptual test, a stimulus was presented once for each trial. After listening to each stimulus one time, subjects were expected to choose one of 14 CVs in Kana orthography appearing on the answer sheet provided to them. They were given 5 seconds to record their selection.

4. Experimental Results

Figure 4 shows the mean value of the correct rate for the 24 subjects in this experiment, in terms of processed or unprocessed signals, respectively. A t-test confirmed that the correct rate of the processed signals was significantly higher than for the unprocessed signals [$p<0.001$].

5. Discussions and Conclusions

Our results show that speech intelligibility is improved in an actual environment when we apply our pre-processing technique, which suppresses the steady-state portions of the signal.

In fact, steady-state suppression has yielded improvements not only in our current study, but also in a previous study [3], both of which involved a reverberation time of 1.0s.

Table 1 shows the correct rate of processed and unprocessed signals for both the binaural and diotic conditions. The correct rate of processed signals in a binaural environment was 71.8% and that of unprocessed signals was 62.7%. In the diotic environment, the correct rate of processed signals was 68.3% and that of unprocessed signals was 63.4%. Therefore we can confirm that steady-state suppression improves speech intelligibility in both binaural and diotic environments.

Helfer [12] compared the correct rate of binaural hearing with that of diotic hearing at a 1.6 s reverberation time and confirmed that the correct rate of binaural hearing was significantly higher (2.7%) than that of diotic hearing. By comparing the result of the present study with that of our previous study [3], we found a similar tendency in the correct rates of the processed signals in the binaural and diotic environments.

We found the gap between correct rates for processed and unprocessed signals to be 9.1% in a binaural environment (Processed: 71.8%, Unprocessed: 62.7%). The gap was 4.9% in a diotic environment (Processed: 68.3%, Unprocessed: 63.4%). This indicates that steady-state suppression in a binaural environment could

Table 1: The correct rate of processed signals in the previous study [3] and present study (reverberation time was 1.0s for both)

Experiment	Processed	Unprocessed
Binaural (Present study)	71.8%	62.7%
Diotic (Previous study)	68.3%	63.4%

improve speech intelligibility much better than in a diotic environment.

6. Summary

We confirm that the pre-processing method of steady-state suppression decreases the degradation of speech intelligibility. In our future work, we would like to investigate more specifically the best suppression rate for steady-state portions and an accurate range of reverberation times where this technique yields the most improvement in speech intelligibility. At the same time, we would like to promote this technique for practical use in our future work.

Acknowledgements

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

References

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

[2] Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A. and Kitamura, T., “Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments,” *Acoustical Science and Technology*, 23: 229-232, 2002.

[3] Hodoshima, N., Inoue,T., Arai, T., and Kusumoto, A., “Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments,” *Proc. China-Japan Joint Conference on Acoustics*, 199-202, 2002 (also, *Acoustical Science and Technology*, 25(1): 58-60, 2004).

[4] Nabelek, A. K. and Robinette, L., “Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects,” *J. Acoust. Soc. Am.*, 63: 187-194, 1978.

[5] Neely, S. T., and Allen, J. B., "Invertibility of a room impulse response." *J. Acoust. Soc. Am.* 66: 165-169, 1979.

[6] Miyoshi, M., and Kaneda, Y., "Inverse filtering of room acoustics." *IEEE Trans. on Acoustics Speech and Signal Processing* 36(2): 145-152, 1988.

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

[8] Avendano, C., and Hermansky, H., "Study on the reverberation of speech based on temporal envelope filtering," *Proc. ICSLP* 2, 889-892, 1996.

[9] Furui, S., “On the role of spectral transition for speech perception,” *J. Acoust. Soc. Am*., 80(4): 1016-1025, 1986.

[10] Hermansky, H., and Morgan, N., "RASTA processing of speech." *IEEE Trans. Speech Audio Process.* 2: 578-589, 1994.

[11] Inoue, T., Hodoshima, N., Arai,T., Kinoshita, K. and Kusumoto, A., “Improvement of speech intelligibility under various reverberant environments by the steady-state suppression” *Proc. Autumn Meet. Acoust. Soc. Jpn*,1: 377-378, 2002 (in Japanese).

[12] Helfer, K. S., “Binaural Cues and Consonant Perception in Reverberation and Noise”, *J. Speech and Hearing Research*, 37: 429-438, 1994.