ACOUSTICAL LETTER

The effect of pre-processing approach for improving speech intelligibility in a hall: Comparison between diotic and dichotic listening conditions

Nao Hodoshima*, Takahito Goto, Noriko Ohata, Tsuyoshi Inoue and Takayuki Arai

Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 30 August 2004, Accepted for publication 4 October 2004*)

Keywords: Speech enhancement, Reverberation, Speech intelligibility, Overlap-masking, Steady-state suppression
PACS number: 43.72.Ew, 43.66.Dc, 43.71.Es, 43.55.Hy, 43.38.Tj [DOI: 10.1250/ast.26.212]

1. Introduction

In large auditoriums, understanding speech may become difficult. One reason is that reverberation causes 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, where reverberation tails of previous portions of a sound affect subsequent segments [1,2].

To improve speech intelligibility in reverberant environments, there are three general approaches: microphone-array, post-processing and pre-processing. Post-processing methods such as inverse filtering (e.g., [3,4]) and modulation filtering (e.g., [5,6]) are applied to speech signals already released into a room and affected by reverberation. Pre-processing approaches, on the other hand, processes a speech signal before it is affected by reverberation. It is a method that reduces the influence of reverberation on the transmission path. Kusumoto *et al.* proposed the modulation filtering which enhances the important modulation frequency region for speech perception [7]. Since pre-processing operates on a speech signal between a microphone and loudspeaker, this method can be used with a Public Address (PA) system.

Spectral transitions in a speech signal play an important role for speech perception. Furui [8] 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 [9,10]. This is because the information in the steady-state portions of the speech signal is redundant with that in transient segments. Also, both "delta" processing of cepstral features [8] and RelAtive SpecTrAl (RASTA) processing [11] enhance transitions of speech, and they have also been shown to increase recognition rates in automatic speech recognition.

Arai's pre-processing method [12] that suppresses steady-state portions of speech obtained clear improvements by reducing the masking influence caused by the reverberation components of the previous portion as described in [12]. Hodoshima *et al.* [13,14] conducted perceptual tests in a diotic environment (the same stimulus was presented simultaneously to both the right and left earphones) to confirm the effectiveness of Arai's technique [12] with a set of artificial reverberation conditions, in which reverberation times were 0.4–1.3 s. Clear improvements were obtained with reverberation times of 0.8–1.2 s.

Binaural listening enhances the ability to understand speech in reverberation [15]. Helfer [16] 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 (69.5%) was higher than that of diotic hearing (63.4%).

The purpose of this study is to compare the effect of steady-state suppression in two listening environments: dichotic and diotic. In the dichotic case, different stimuli are presented to listener's right and left ears in the largest lecture hall at Sophia University in Tokyo, Japan. In the diotic case, a speech signal is convolved with the monaural impulse response measured in the same lecture hall and is presented simultaneously to both the right and left earphones. Section 2 presents reverberation conditions used in our perceptual experiments. Section 3 describes Experiment I (a dichotic environment), and Section 4, Experiment II (a diotic environment). Finally, discussions and conclusions are detailed in Section 5.

2. Reverberation condition

The perceptual test was conducted in the largest lecture hall at Sophia University in Experiment I and the impulse response of the hall was used in Experiment II. Using the Time Stretched Pulse (TSP) signal [17], we measured the impulse response of the hall, which seats 822 people (see Fig. 1).

To calculate 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 derived from the average over the reverberation times at the 0.5-, 1.0- and 2.0-kHz of the 1-octave bandpassed impulse response was 1.3 s.

3. Experiment I

3.1. Steady-state suppression

We used the steady-state suppression method proposed by Arai *et al.* [12], to suppress steady-state portions of speech. In this study, the steady-state suppression was applied to whole sentences (as in [12] and labeled "whole-proc" in that work).

3.2. Stimuli

We used the same speech samples as Hodoshima *et al.*

*e-mail: n-hodosh@sophia.ac.jp

Fig. 1 The largest lecture hall at Sophia University (Tokyo, Japan).

used in [13]. 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/. They 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 types: the original (unprocessed) signals (Org) and the processed signals (Proc). Fifty-six stimuli were prepared in total (14 CVs $\times$ with/without processing $\times$ 2 repetitions). 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.

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 always an empty seat between adjacent two subjects. 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.

3.5. Experimental Results

The left part of Fig. 2 shows the mean value of the correct rate of the processed signals (Proc) and the unprocessed ones (Org) for the 24 subjects in this experiment. A *t*-test confirmed that the correct rate of the processed signals was significantly higher than for the unprocessed signals ($p < 0.01$).

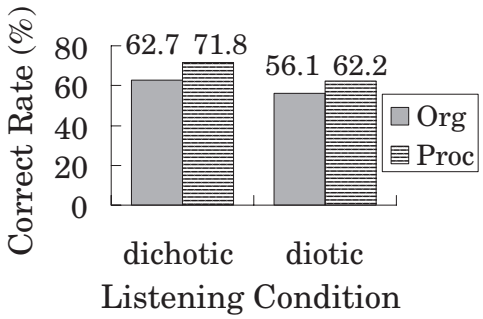

Fig. 2 The correct rate of the processed signals (Proc) and the unprocessed ones (Org) of the perceptual experiments in two listening conditions (dichotic in a hall and diotic simulating a reverberation environment).

4. Experiment II

4.1. Stimuli

We created the stimuli that simulated a reverberation environment. The stimuli were made by convoluting the same stimuli as in Experiment I with the impulse response measured from a single microphone (monaural impulse response) in the same hall as in Experiment I. These consisted of two types: the original signals (Org) and the steady-state suppressed signals (Proc). Therefore, 28 stimuli were prepared, i.e., 14 CVs $\times$ 2 (with or without) processing conditions in total.

4.2. Subjects

Twenty-one normal hearing subjects (17 males and 4 females with an average age of 22 years) participated in the experiment. All were native speakers of Japanese.

4.3. Procedure

The experiment, controlled by a computer, was conducted in a soundproof room. The stimuli were presented through headphones (STAX SR-303). The sound level was adjusted to each subject's comfort level during the practice session. In the main session, a stimulus was presented at each trial. Then 14 CVs in Kana orthography were shown on a PC screen. Subjects were forced to choose one of the 14 CVs by clicking a button on the screen with a mouse. When they selected a CV, the next stimulus was presented. For each subject, the stimuli were presented randomly.

4.4. Experimental Results

The right part of Fig. 2 shows the mean value of the correct rate of the processed signals with reverberation (Proc) and the unprocessed ones with reverberation (Org) for the 21 subjects in this experiment. A *t*-test confirmed that the correct rate of the processed signals was significantly higher than for the unprocessed signals ($p < 0.01$).

5. Discussions

Our results show that steady-state suppression yielded significant improvements not only in a diotic environment (Experiment II) as in [12–14] but in a dichotic environment as well (Experiment I). Steady-state suppression also shows more improvement in the dichotic environment than the diotic environment.

Results from our two experiments are consistent with

Helfer [16]. The correct rate of Proc in Experiment I is higher than that in Experiment II. And the correct rate of Org in Experiment I is higher than that in Experiment II. That means the correct rate of dichotic hearing was higher than that of diotic hearing. This effect may be caused by binaural advantage [15].

6. Conclusions

This study examined the effect of steady-state suppression in dichotic and diotic environments. The results showed that steady-state suppression significantly prevents degrading speech intelligibility in both of the listening conditions used in this study. We have seen the effectiveness of steady-state suppression in a diotic listening environment from the results of perceptual tests in various reverberation conditions [12–14]. This study also shows an efficacy of steady-state suppression in a dichotic listening environment. Further investigations should provide a more complete picture of its effectiveness.

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

References

[1] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).

[2] A. K. Nabelek 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).

[3] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, **66**, 165–169 (1979).

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 145–152 (1988).

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

[6] C. Avendano and H. Hermansky, "Study on the reverberation of speech based on temporal envelope filtering," *Proc. ICSLP*, Vol. 2, 889–892 (1996).

[7] A. Kusumoto, T. Arai, K. Kisnoshita, N. Hodoshima and N. Vaughan, "Modulation enhancement of speech by a preprocessing algorithm for improving intelligibility in reverberant environments," *Speech Commun.* (To be published).

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

[9] D. Kewley-Port, D. B. Pisoni and M. Studdert-Kennedy, "Perception of static and dynamics acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Am.*, **73**, 1779–1793 (1983).

[10] W. Strange, J. J. Jenkins and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, **74**, 695–705 (1983).

[11] H. Hermansky, and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, **2**, 578–589 (1994).

[12] 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).

[13] N. Hodoshima, T. Inoue, T. Arai and A. Kusumoto, "Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments," *Acoust. Sci. & Tech.*, **25**, 58–60 (2004).

[14] N. Hodoshima, T. Arai, T. Inoue, K. Kinoshita and A. Kusumoto, "Improving speech intelligibility by steady-state suppression as pre-processing in small to medium sized halls," *Proc. Eurospeech*, pp. 1365–1368 (2003).

[15] W. Koenig, "Subjective effects on binaural hearing," *J. Acoust. Soc. Am.*, **22**, 61–62 (1950).

[16] K. S. Helfer, "Binaural Cues and Consonant Perception in Reverberation and Noise," *J. Speech Hear. Res.*, **37**, 429–438 (1994).

[17] N. Aoshima, "Computer-generation pulse signal applied for sound measurement," *J. Acoust. Soc. Am.*, **69**, 1484–1488 (1981).