社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2003-53 (2003-06)

小・中規模ホール環境を想定した定常部抑圧による拡声音声の明瞭度改善

程島奈緒† 荒井隆行† 井上豪† 木下慶介† 楠本亜希子‡

† 上智大学理工学部電気電子工学科 〒102-8554 東京都千代田区紀尾井町 7-1

‡ Dept. of Veterans Affairs, Portland VA Medical Center, Portland, OR 97207, USA

E-mail: † n-hodosh@sophia.ac.jp

あらまし　残響により音声明瞭度が減少する原因は，主に先行する音素に付加される残響の尾が後続の音素に影響を与えるoverlap-maskingの影響によると考えられている(Bolt et al., 1949). 荒井らは, 残響によるoverlap-maskingの影響を軽減するため，エネルギは比較的大きいが音声知覚にそれほど重要ではないとされる定常部を抑圧する処理を行い，音声明瞭度の改善を得た(Arai et al., 2002). 本論文では0.4-1.0秒の残響条件における定常部抑圧処理の効果を調べた結果，0.8-0.9秒の残響時間において，またターゲットが破裂音の場合に処理による有意な改善が得られた．以上より，定常部抑圧処理は残響環境下において音声明瞭度を改善するための前処理として有効であると確認され，overlap-maskingの影響が実証された.

キーワード　音声強調，残響，音声明瞭度，マスキング，定常部抑圧

Improving intelligibility of speech by steady-state suppression as pre-processing in small to medium sized halls.

Nao HODOSHIMA† Takayuki ARAI† Tsuyoshi INOUE† Keisuke KINOSHITA†

and Akiko KUSUMOTO‡

† Dept. of Electrical and Electronics Eng., Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan

‡ Dept. of Veterans Affairs, Portland VA Medical Center, Portland, OR 97207, USA

E-mail: † n-hodosh@sophia.ac.jp

Abstract One of the reasons reverberation degrades speech intelligibility is the effect of overlap-masking, when segments of an acoustic signal are masked by reverberation components of previous segments (Bolt et al., 1949). To reduce overlap-masking, Arai et al. suppressed steady-state portions having more energy, but which are less crucial for speech perception, and confirmed promising results for improving speech intelligibility (Arai et al., 2002). We conducted a perceptual test with steady-state suppression under a reverberation time of 0.4s-1.0s. The results showed significant improvements for some reverberation conditions and especially for stop consonants. We certified that steady-state suppression is an effective pre-processing method for improving speech intelligibility under reverberant conditions and proved the effect of overlap-masking.

Keyword speech enhancement, reverberation, speech intelligibility, overlap-masking , steady-state suppression

1. Introduction

In large auditoriums, speech intelligibility is often degraded. This is due to reverberation caused by a superposition of reflected sounds with various delays and amplitudes. Although reverberation adds richness to music, it makes speech more difficult to understand. Because reverberation tails affect subsequent segments, an acoustic signal of one segment is masked by the reverberation components of the previous portion, and this effect of overlap-masking degrades intelligibility [1] [2].

There are several general approaches for improving speech intelligibility in reverberant environments: microphone array, post-processing and pre-processing. Microphone array takes advantage of spatial information about the sound source and assures the direction of the desired signal. Thus it can enhance the desired sound source by reducing noise and reverberation [3] [4].

A post-processing method is applied to speech signal already released into a room and affected by reverberation.

Inverse filtering [5] [6] and modulation filtering [7] [8] are examples. Minimum phase inverse filtering is a dereverberation technique that applies inverse filtering, supposing that the impulse response of a room has a minimum phase. Modulation filtering alters the modulation spectrum of a signal, which is derived from a frequency analysis of the temporal envelope of the band-passed signal.

In a pre-processing approach, a speech signal is processed between a microphone and loudspeaker, and can therefore be used with a Public Address (PA) system. Langhans and Strube applied the same technique in pre-processing as they used in post-processing, but no clear improvement was found [7].

It has been discovered that the important modulation frequency of a signal for speech perception is around 4Hz. It has also been confirmed that the peak of the modulation spectrum shifts to the lower modulation frequency and the modulation index is reduced as the acoustic signal is reverberated [9]. Thus, Kusumoto et al. enhanced this particular frequency region in their application of a modulation filter [10]. They showed promising results for improving speech intelligibility with pre-processing.

As a pre-processing approach, Arai et al. suppressed steady-state portions of speech in order to reduce the influence of overlap-masking caused by reverberation, and obtained clear improvements [11]. To decrease the effect of overlap-masking, one might think to lessen the energy of preceding portions beforehand so that the energy of reverberation components overlapping to a subsequent portion is attenuated. However, it is reported that while spectral transition is crucial for syllable perception, vowel nuclei are not necessary for either vowel or syllable perception [12]. Therefore, Arai et al. suppressed these steady-state portions with more energy but less crucial information for speech perception, in order to reduce the effect of overlap-masking caused by reverberation tails of previous portions [11]. They confirmed promising results for improving speech intelligibility.

The purpose of our study is to explore the effect of steady-state suppression under various reverberation conditions. Our ultimate goal is to provide a filter for pre-processing which is suitable for an individual auditorium. In order to achieve this, we need to better understand the relationship between the effect of a pre-processing filter and reverberation condition on speech intelligibility. Because it is difficult to examine the effects of both parameters simultaneously, in a previous study we varied reverberation time from 0.9 to 1.3s while applying a pre-processing filter and evaluated the effects on speech intelligibility [13]. The results show that steady-state suppression with a specific suppression rate prevented the degradation of speech intelligibility within a certain range of reverberation time. The results also show clear improvements for relatively shorter reverberation conditions within the range of 0.9 - 1.3s. In order to investigate the effect of steady-state suppression at shorter reverberation conditions than those in [13], we conduct a perceptual test with a reverberation time of 0.4 - 1.0s.

2. Perceptual Experiment

2.1. Reverberant conditions

The artificial impulse responses h_n were created as Eq. (1) to obtain the desired reverberation conditions [14]:

$$h_n(t) = e^{-t/\tau} h_o(t), \qquad (1)$$

where τ is a time constant. The original impulse response h_o used for this study was measured in the Hamming Hall, Higashi-Yamato City, Tokyo (A reflection board was not used.). Thus, we can obtain the desired reverberation time as a function of τ. Table 1 shows the set of reverberation conditions used in our experiment.

Reverberation time is defined as the time the decay curve of the impulse response decreases 60 dB from steady state. We used Early Decay Time (EDT), which is the time it took for 10 dB of reverberation decay, and we multiplied that by six to extrapolate the reverberation time. The value of reverberation time as seen in Table 1 is the average of 0.5, 1, and 2 kHz of reverberation time.

2.2. Steady-state Suppression

We applied the same method as in [11] to suppress steady-state portions of speech. This signal processing calculates the D parameter to measure a spectral transition [12] and defines a speech portion as steady-state when D is less than a certain threshold. D is essentially the same as the parameter proposed by Furui [12]; in this paper, we used D as the mean square of the regression coefficients for each time trajectory of the logarithmic envelope of a subband. Once a portion is

Table 1: Reverberation conditions used in the experiment

Impulse response	r1	r2	r3	r4	r5
RT (s)	0.4	0.5	0.7	0.9	1.0

considered steady-state, the amplitude of the portion is multiplied by the factor of 0.4 (a suppression rate of 40%). Fig. 1 shows waveforms of (a) original and (b) steady-state suppression (a suppression rate of 20%). The waveform is one of stimuli used in the experiment.

2.3. Stimuli

The original signals consisted of nonsense Consonant-Vowel (CV) syllables embedded in a Japanese carrier phrase. The twenty-four CVs (targets) used in the experiment are shown in Table 2. 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 a sentence data set. The beginning position of the target vowel was adjusted to 150ms from the end of the pre-target carrier phrase to control the amount of energy overlapping to the target from the previous portion. We used 150ms because mean durations of Japanese syllables are between 150-200ms.

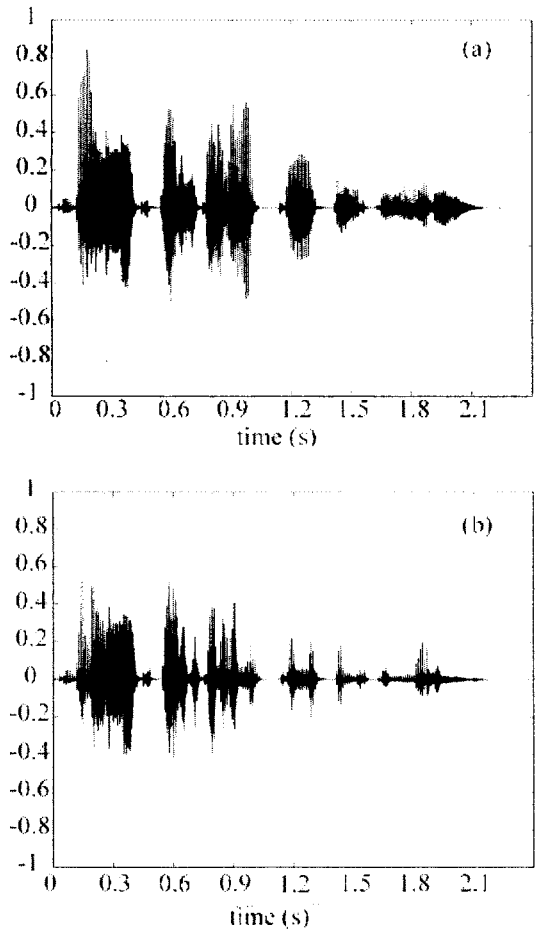

Fig. 1: The waveforms of (a) original and (b) steady-state suppression (a suppression rate of 20%)

Table 2: CVs used in the experiment

	Voiceless consonants +Vowels	Voiced consonants +Vowels
Stops	/pa/ /ta/ /ka/	/ba/ /da/ /ga/
+ Vowels	/pi/ /ki/	/bi/ /gi/
Fricatives	/sa/ /ʃa/ /ha/	
+ Vowels	/ʃi/ /hi/	
Affricates	/tʃa/	/dza/ /dʒa/
+ Vowels	/tʃi/	/dʒi/
Nasals		/ma/ /na/
+ Vowels		/mi/ /ni/

The stimuli consisted of four conditions: the original signals (Org), the processed signals (Proc), the original signals with reverberation (Org_rev) and the processed signals with reverberation (Proc_rev).

2.4. Subjects

Twenty-two normal hearing subjects (11 males and 11 females, ages 19 to 27) participated in the experiment. All were native speakers of Japanese.

2.5. Procedure

The experiment, controlled by a computer, was conducted in a soundproof room. The stimuli were presented with headphones (STAX SR-303), and the sound level was adjusted to each subject's comfort level. In the experiment, a stimulus was presented at each trial. Then 24 CVs in Kana orthography were shown on the PC screen. Subjects were forced to choose one of 24 CVs by clicking a button on the PC screen with a mouse. For each subject, 288 stimuli were presented randomly (5 reverberation conditions x 24 CVs x 2 processing conditions + 24 CVs x 2 processing conditions).

3. Experimental Results

The mean percent correct for each reverberation and processing condition is shown in Table 3. A 2 x 5 ANOVA for repeated measures was performed, confirming significant main effects of processing ($p<0.001$), impulse response ($p<0.001$) and interaction ($p=0.006$). For the comparison of means between processing, a t-test was performed for each impulse response. A significant difference was obtained for the r3-r4 conditions [r3: Org_rev (73.9%), Proc_rev (82.4%), $p<0.001$; r4: Org_rev (70.1%), Proc_rev (79.2%), $p<0.001$]. The mean percent correct for each processing condition without reverberation is shown in Table 4.

Table 3: Mean percent correct in each condition with reverberation [%]:

Impulse Responses	r1	r2	r3	r4	r5
Org_rev	90.7	84.3	73.9	70.1	69.5
Proc_rev	92.8	86.6	82.4	79.2	73.7

Table 4: Mean percent correct in each condition without reverberation [%]:

Org	97.0
Proc	97.2

Table 5 Mean percent correct in manner of articulation [%]:

	r1	r2	r3	r4	r5
Stops					
Org_rev	85.0	72.7	60.9	50.5	53.6
Proc_rev	91.4	85.9	80.9	74.5	64.1
Fricatives					
Org_rev	88.2	85.5	78.2	81.8	76.4
Proc_rev	88.2	76.4	80.0	79.1	79.1
Affricates					
Org_rev	100	95.5	86.4	83.6	84.5
Proc_rev	96.4	95.5	87.3	87.3	82.7
Nasals					
Org_rev	96.6	97.7	85.2	87.5	81.8
Proc_rev	97.7	89.8	83.0	80.7	79.5

Next, we classified the consonants into manner of articulation. The mean percent correct for each reverberation and processing condition is shown in Table 5. For stops, a 2x5 ANOVA for repeated measures confirmed significant main effects of processing [$F(1, 21)=94.87$, $p<0.001$] and impulse response [$F(4, 84)=54.82$, $p<0.001$]. Interaction [$F(4, 84)=4.48$, $p=0.003$] was also significant. For the comparison of means between processing, a t-test was performed for each impulse response. A significant difference was obtained for the r1-r5 conditions [r1: Org_rev (85.0%), Proc_rev (91.4%), $p=0.01$; r2: Org_rev (72.7%), Proc_rev (85.9%), $p<0.001$; r3: Org_rev (60.9%), Proc_rev (80.9%), $p<0.001$; r4: Org_rev (50.5%), Proc_rev (74.5%), $p<0.001$; r5: Org_rev (53.6%), Proc_rev (64.1%), $p=0.04$].

For fricatives, affricates and nasals, a 2x5 ANOVA for repeated measures confirmed that main effects of impulse response [fricatives: $F(4, 84)=4.62$, $p=0.002$; affricates: $F(4, 84)=22.45$, $p<0.001$; nasals: $F(1, 21)=9.26$*, $p=0.006$]. Main effects of processing [fricatives: $F(1, 21)=0.94$; affricates: $F(1, 21)=0.05$; nasals: $F(1, 21)=3.65$] and interaction [fricatives: $F(4, 84)=0.14$; affricates: $F(4, 84)=0.29$; nasals: $F(4, 84)=0.27$] were not significant.

4. Discussions

The significant main effect of reverberation confirms the correct rates decreased as reverberation time increased, regardless of processing type. The significant interaction for the mean percent correct and stops indicated that the effect of steady-state suppression depended on reverberation time. The interaction for the rest of manners of articulation showed that the effect of steady-state suppression was independent from reverberation time.

The significant main effect of processing is that steady-state suppression prevents degrading speech intelligibility in reverberant environments. The results show that correct rates for both Org and Proc were almost the same. This shows that speech intelligibility of processed stimuli was not degraded compared to ones in which steady-state portions were not suppressed. The results also show that the difference of correct rates between Proc_rev and Org_rev decreased as reverberation time shortened. The results indicate that 0.8 s is the lower limit of reverberation time in which we observe significant improvements with the suppression rate of 40 %. Therefore, our results confirm that ssteady-state suppression prevents degrading speech intelligibility as a pre-processing method and that the effect of steady-state suppression differs with respect to reverberation time.

In addition to the results where significant improvements were obtained with a reverberation time of 0.8 - 0.9 s, a close to significant difference was obtained for the r5 condition [Org_rev (69.5%), Proc_rev (73.7%), $p=0.057$]. The previous study [13] had tested the effect of steady-state suppression with a reverberation time of 0.9 - 1.3s, and significant improvements were observed with those of 0.9 - 1.2s. The combination of results in this

*As Mauchly's test of sphericity was significant, we adjusted the degrees of freedom.

study and in [13] indicate that steady-state suppression is effective for the r5 condition because the significant effect of processing may appear continuously in the certain range of reverberation time. Therefore, we concluded that the effect of steady-state suppression with the suppression rate of 40 % prevented degrading speech intelligibility for a reverberant time of 0.8 - 1.2s.

We classified the consonants into manner of articulation to determine if the effect of steady-state suppression depended on consonant type. When the targets were stops, it was showed that Proc_rev performed significantly better than Org_rev under all reverberation conditions. On the other hand, it was confirmed that correct rates did not differ with or without steady-state suppression for the rest of the manners of articulation. Thus, it can be said that the reason steady-state suppression prevented the degradation of average speech intelligibility is mainly due to significant improvement for stops.

We further investigated the reason steady-state suppression greatly prevents degrading of speech intelligibility for stops. When we examined the frequency regions of burst and spectral transition, which are cues for stop perception, we found that they coincided with strong energy regions of the previous vowel in our stimuli. Thus, when the stimuli are reverberated, these bursts and spectral transitions are seriously affected by the reverberation components of the previous vowel, decreasing speech intelligibility. We surmise that the degradation of speech intelligibility is prevented for stops because steady-state suppression reduced the amount of energy of the reverberation components overlapping to burst and spectral transition.

We determined that the effect of steady-state suppression could be expressed as a change of modulation spectrum of a speech signal. Thus, we compared the modulation spectra of signals with and without steady-state suppression following Hodoshima et al. [15]. (The modulation spectra are same as [15], in which figures are displayed).) The modulation spectrum is derived from a frequency analysis of the temporal envelope of a band-passed signal. First, speech signals are divided into four bands following Arai et al. [16] and modulation spectrum was calculated in each band. Then these modulation spectra were averaged over 24 sentences used in the perceptual experiment. When we compared the two modulation spectra, we observed that their modulation indices around 4Hz and above 10Hz are intensified by steady-state suppression in all bands.

It has been noted that the important modulation frequencies for speech perception lie between 1-16Hz, especially 4Hz, where the modulation spectrum usually reaches its maximum value [17] [18]. When the acoustic signal is reverberated, the peak of the modulation spectrum shifts to the lower modulation frequency and the modulation index is reduced [9]. Thus, there is a strong relationship between speech intelligibility and the modulation spectrum. It has also been shown that the modulation frequency around 4Hz reflects the syllabic rate of speech, and above 10Hz phoneme rate of speech [19].

These findings indicate that syllables and transitions are enhanced by steady-state suppression. In other words, steady-state suppression preemptively enhances modulation indices around frequencies important for speech perception so that the modulation index is prevented from being reduced by reverberation. Therefore, we have seen observed clear improvements by means of steady-state suppression in reverberant environments.

5. Conclusions

In this paper, we investigated the range of reverberation conditions for steady-state suppression by Arai et al. [11] that prevent degrading speech intelligibility by reducing the effect of overlap-masking. To explore the relationship between steady-state suppression and several reverberation conditions, we conducted a perceptual test with a set of artificial impulse responses.

The results showed that clear improvements were obtained with a reverberation time of 0.8 – 1.0 s. These show that steady-state suppression is an effective pre-processing method to prevent degrading speech intelligibility under reverberant conditions. It is considered that this effectiveness is observed because steady-state suppression reduces the effect of overlap-masking. Thus, this supports that speech intelligibility is degraded by the effect of overlap-masking caused by reverberation. We predict that the range of reverberation conditions in which clear improvements are observed may be different as we change the suppression rate of steady-state portions. Thus, we would like to investigate the upper limit of reverberation time which prevents degrading speech intelligibility by

steady-state suppression when we more and more suppress steady-state portions of speech. Eventually, we would like to explore the effect of steady-state suppression in an auditorium.

6. Acknowledgements

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

References

[1] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation", J. Acoust. Soc. Am., 21, pp. 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, pp. 187-194, 1978.

[3] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction", IEEE Trans. Acoust. Speech and Signal Process., ASSP-34 (6), pp. 1391-1400, 1986.

[4] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West and M. M. Sondhi, "Autodirective microphone systems", Acoustica, 73 (2), pp. 58-71, 1991.

[5] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response", J. Acoust. Soc. Am., 66(1), pp. 165-169, 1979.

[6] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics", IEEE Trans. Acoust. Speech and Signal Process., 36(2), pp. 145-152, 1988.

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

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

[9] T. Houtgast and H. J. M. Steeneken, "A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", J. Acoust. Soc. Am., 77(3), pp. 1069-1077, 1985.

[10] A. Kusumoto, T. Arai, M. Takahashi and Y. Murahara, "Modulation enhancement of speech as a preprocessing for reverberant chambers with the hearing-impaired", Proc. IEEE ICASSP, pp. 933-936, 2000.

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

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

[13] N. Hodoshima, T. Inoue, T. Arai and A. Kusumoto, "Suppressing steady-state portions of speech for improving intelligibility in various reverberant environments", Proc. China-Japan Joint Conference on Acoustics, pp. 199-202, 2002.

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

[15] N. Hodoshima, T. Inoue, T. Arai, K. Kinoshita and A. Kusumoto, "Suppressing steady-state portions of speech for improving intelligibility as pre-processing: Under various reverberant environments", Technical Report of IEICE, SP2002-65, pp.47-51, 2002.

[16] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," Proc. IEEE ICASSP, pp.933-936, 1998.

[17] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., 95(2), pp. 1053-1064, 1994.

[18] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," J. Acoust. Soc. Am., 105(5), pp. 2783-2791, 1999.

[19] A. J. Duquesnoy and R. Plomp, "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am., 68(2), pp. 537-544, 1980.