19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

ENHANCING SPEECH IN REVERBERATION BY STEADY-STATE SUPPRESSION

PACS: 43.55.Jz

Kei Takahashi[1]; Keiichi Yasu[1]; Nao Hodoshima[1]; Takayuki Arai[1]; Kiyohiro Kurisu[2];
[1]Sophia University; 7-1 Kioi-cho, Chiyoda-ku, Tokyo, Japan; kei-taka@sophia.ac.jp
[2]TOA Corporation; 2-1 Takamatsu-cho, Takarazuka, Hyogo, Japan;

ABSTRACT
In order to improve speech intelligibility in reverberation, Arai *et al.* [Autumn Meet. Acoust. Soc. Jpn., 2001; Acoust. Sci. Tech. 23(8), 2002] proposed steady-state suppression. Our ultimate goal is to implement steady-state suppression in real time by a digital signal processor. For quasi-real-time processing, a consonant enhancement technique in order to compensate for the loudness recruitment was proposed by Yasutake *et al.* [Tech. Rep., of IEICE Japan, Vol. HIP2005-94, 2005], These two techniques in Arai *et al.* and Yasutake *et al.* are essentially the same in terms of enhancing consonants. Therefore, we simplified the algorithm of steady-state suppression by including the consonant detection method used in Yasutake *et al.* for quasi-real-time processing. We found the high concordance rate of steady-state portions of both the proposed and the previous methods of steady-state suppression.

INTRODUCTION
One reason that reverberation degrades speech intelligibility is overlap-masking, which occurs when reverberant tails of previous portions of a sound mask subsequent segments [1, 2]. For example, Figure 1 shows an illustration of overlap-masking on the utterance “October” [3]. The left part of Figure 1 shows the original speech signal without reverberation, and the right part of Figure 1 shows reverberant signals which were obtained by taking the convolution of the original speech signals with an impulse response of a room having a 1.1 s reverberation time. As you can see in Figure 1, consonants that have weak energy (e.g., /k/, /t/, /b/) can be masked by the reverberation tails that have strong energy (e.g., /o/).

Arai *et al.* [4, 5] proposed the steady-state suppression approach in order to reduce overlap-masking. This approach splits a speech signal into 1/3-octave bands for detecting steady-state portions, and suppresses steady-state portions of speech which have more energy but which are less crucial for speech perception. Steady-state suppression improved speech intelligibility, in particular reverberation time (e.g., Hodoshima *et al.* [6] and Goto *et al.* [7]).

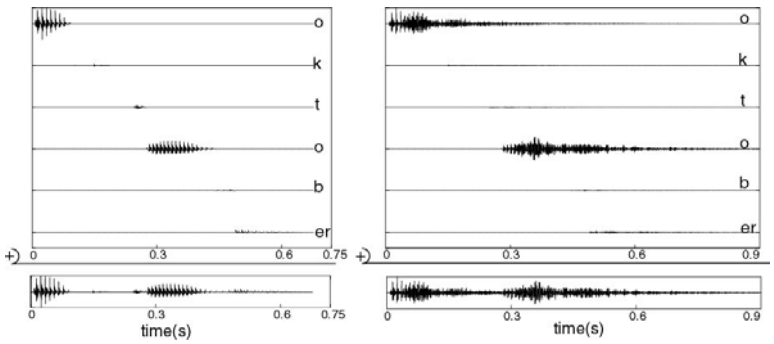


Figure 1 An illustration of overlap-masking on the utterance “October” [3]

In order to compensate for the loudness recruitment, Yasutake *et al.* [8] proposed a consonant enhancement technique which is simple and easy to realize in quasi-real-time processing since the consonant detection method [8] used temporal changes of energy in two adjacent frames of a signal without dividing it into a number of frequency bands. Our ultimate goal is to build steady-state suppression into a public address (PA) system. This study aims to implement steady-state suppression in quasi-real time by a digital signal processor (DSP). To do this, we planed to simplify the algorithm of steady-state suppression while detecting steady-state portions in a high accuracy. Since the consonant enhancement technique [8] can easily be implemented in real time, we included our modified consonant detection method used in [8] in steady-state suppression for realizing real-time processing.

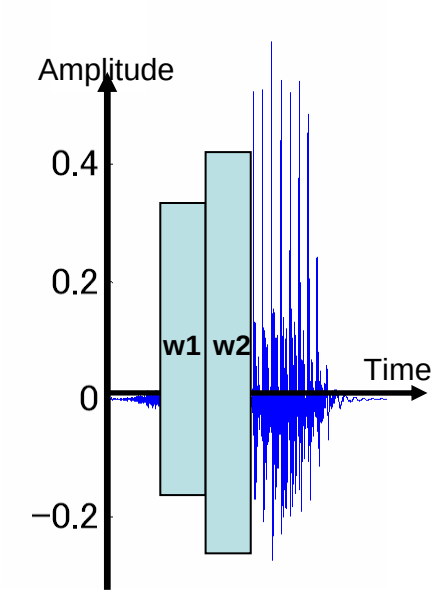
CONSONANT ENHANCEMENT

The consonant enhancement technique proposed by Yasutake *et al.* [8] detects consonant portions by comparing energies in two adjacent frames (w1, w2) on temporal axis (see Figure 2). If the energy in w2 is larger than that in w1, w1 is considered as a consonant. If the energy in w1 is larger than that in w2, w1 is considered as a vowel. On the other hand, steady-state suppression proposed by Arai *et al.* [4, 5] detects steady-state portions by temporal change of energy in five adjacent frames. See more details in references [4, 5].

These two methods [4, 5, 8] are identical in detecting temporal changes of energy, but are different in the following four points:

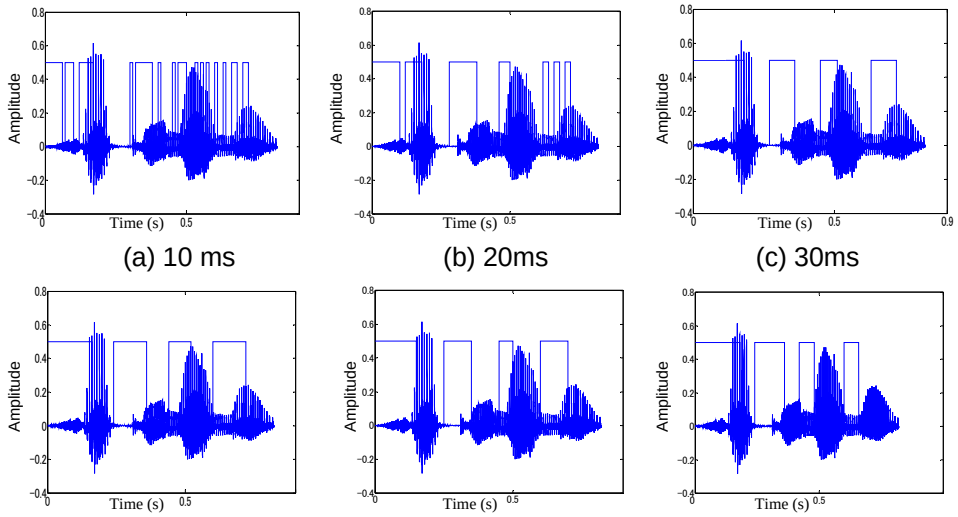
1) **Frame length and number of frames** – steady-state suppression used five frames of a speech signal while the consonant enhancement technique used two frames of a speech signal.
2) **Frequency bands** – steady-state suppression used 1/3-octave band while the consonant enhancement technique used a full band.
3) **Purpose** – the purpose of steady-state suppression is decreasing overlap-masking while the purpose of the consonant enhancement technique is a compensation for the loudness recruitment.
4) **Target of processing** – steady-state suppression processes the onset and coda of a syllable while the consonant enhancement technique processes only the onset of a syllable.

1) and 2) are related to performance of a real-time processing. Delay time by processing is decided by 1). The effect of frame length is seen in the section “Frame length”. 2) is related to the computational complexity by processing. If this amount is small, the processing is easily realized with DSP. Therefore, the proposed steady-state suppression (the proposed approach) described in the section “Function which compensates for the loudness recruitment” uses the consonant detection which deals with full band. 3) and 4) are related to configuration of the function which compensates for the loudness recruitment. This is explained in the section “Function which compensates for the loudness recruitment”.



	w1 > w2	w1 < w2
w1 is…	vowel	consonant
Enhancement	No	Yes

Figure 2 Consonant determination algorithms by Yasutake *et al.* [8]



(a) 10 ms (b) 20ms (c) 30ms

(d) 40ms (e) 50ms (f) 60ms

Figure 3 Consonant detection when the frame length is varied from 10 to 60 ms

Frame length

Figure 3 shows how consonant detection changes when the frame length is varied from 10 to 60 ms for a Japanese word. For the frame lengths of 10 and 20 ms, detection error of consonants occur while above the frame length of 30 ms, detection does not change much.Therefore, we decided to use a frame length of 30 ms to prevent detection error of consonant portions as well as to minimize the delay time by processing.

Function which compensates for the loudness recruitment

Yasutake *et al.* [8] defined the knee points as switching points of amplification characteristic of consonants on the function which compensates for the loudness recruitment while we defined knee points as the boundary of steady-state portions and transitions on the same function. We set knee points by not suppressing consonant portions with 14 syllables (/p/, /t/, /k/, /b/, /d/, /g/, /s/, /h/, /ʃ/, /tʃ/, /dʒ/, /dz/, /m/, /n/) used in Hodoshima *et al*. [3] by checking manually. Knee points were set in two parts which are the beginning and ending points of steady-state portions. Figure 4 shows steady-state suppression on the function which used to compensate for the loudness recruitment. The beginning point was set at -0.72 dB and ending point was set at 6 dB by setting knee points. The portion from -0.72 to 6 dB is regarded as the steady-state portion, and is suppressed by 7.95 dB. Suppressing of steady-state portions by 7.95 dB corresponds to suppressing the amplitude of steady-state portions to 40%. We used the suppression rate of 40% by which steady-state suppression improved speech intelligibility in [6]. A speech signal with the proposed approach and an original speech signal are shown in Figure 5.

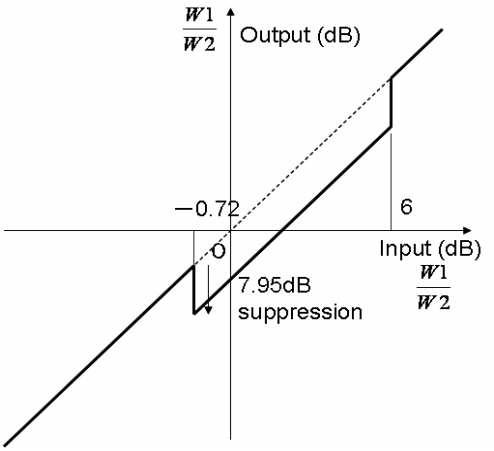


Figure 4 Steady-state suppression on the function which originally used to compensate for the loudness recruitment

RESULT AND DISCUSSION

Comparison of the proposed approach with previously proposed steady-state suppression

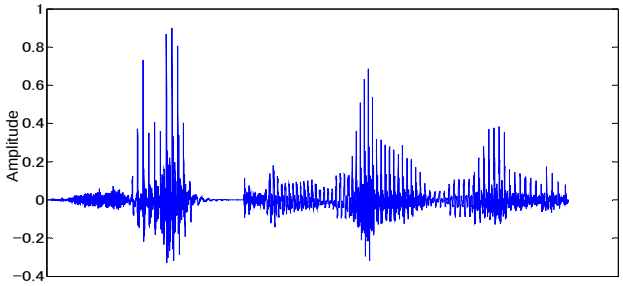
We investigated the concordance of steady-state portions between the proposed and the previously proposed approach of steady-state suppression. We used the previous approach used in [9, 10] because steady-state suppression in [9, 10] avoided suppressing relatively longer continuants such as sibilant consonants. In order to compare the two approaches, we selected six syllables (/p/, /d/, /ʃ/, /tʃ/, /dʒ/, /m/) from the 14 syllables which are used at determination of knee points.

Table 1 shows the concordance rates in percentage of detection of steady-state portions between the proposed and the previous approaches. The concordance rate is 91.8% on average in Table 1. Figure 6 shows detection of steady-state portions for the syllable /ma/. In Figure 6 (a), the right part of a portion which was detected as the steady-state portion is silence. Therefore, the silence is not counted as steady-state portions and we calculated the concordance rates. When portions have the vertical value of 1.0 in Figure 6, they correspond to transitions and the others correspond to steady-state portions. In addition, the previous approach prevents suppressing the amplitude of steady-state portions drastically by tapering the boundary between steady-state portions and transitions.

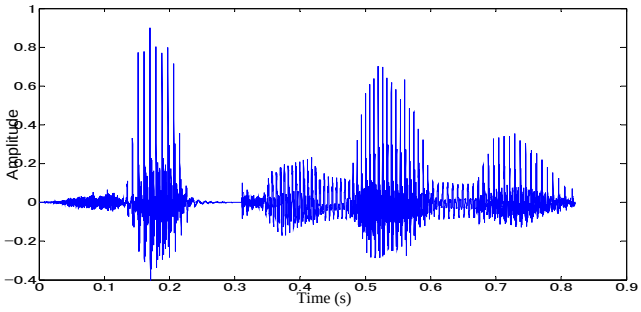
In Figure 6, the beginning or ending part of steady-state portions delays or precedes by 14.9 ms (difference of beginning or ending of steady-state portions between proposed and previous approaches, respectively) compared to previous approach.

Table 1 Concordance rate of steady-state portion between proposed and previous approaches

Syllable	pa	da	ʃa	tʃa	dʒa	ma	average
Concordance rate (%)	92.0	91.4	97.2	94.3	84.3	91.5	91.8

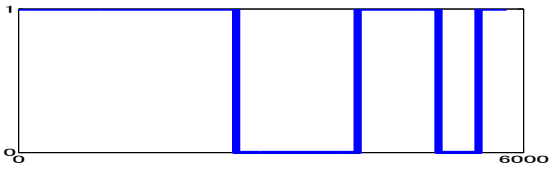


(a) Proposed steady-state suppression (proposed approach)

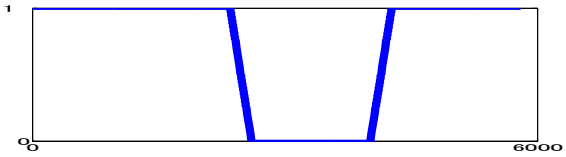


(b) Original speech signal

Figure 5 Speech with steady-state suppression (proposed approach) and original one



(a) Proposed approach



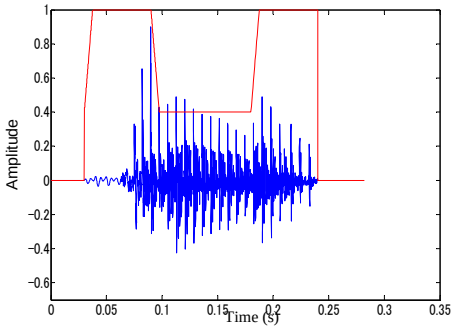
(b) Previous approach

Figure 6 Detection of steady-state portions by proposed and previous approaches

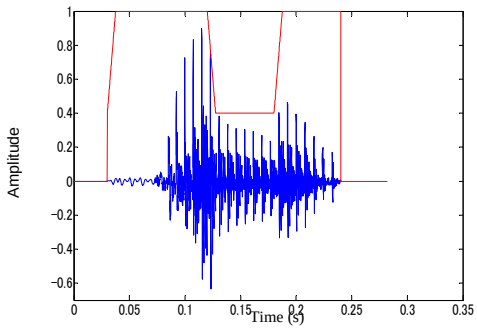
Detection of steady-state portions and the starting point of each frame

Steady-state suppression [3, 4] used five frames for detecting steady-state portions while Yasutake *et al.* used two frames for detecting consonants. Delay time by processing is decided by the frame length and the number of frames. In the proposed approach, when we used two frames with a frame length as 30 ms for detecting steady-state portions and processed a speech signal frame-by-frame, delay time would be up to 30 ms. On the other hand, in the previous approach, when we use five frames with a frame shift as 10 ms for detecting steady-state portions and processed a speech signal, delay time would be up to the same 30 ms.

If the frame shift decreases and the number of frames increases, accuracy in detection of steady-state portions would increase without increasing delay time. Changing the starting point of each frame also can cause to increase detection error of steady-state portions. Figure 7 shows how detection of steady-state portions changes when we add silence to the beginning of a speech signal. Decreasing a frame shift would improve the accuracy of detection of steady-state portion.



(a) 10 ms



(b) 20 ms

Figure 7 Detection of steady-state portions when we add silence to the beginning of a speech signal. The silence is (a) 10 ms and (b) 20 ms.

Relationship between loudness recruitment and steady-state suppression

It has been reported that temporal masking was increased for people with hearing impairments compared with people with normal hearing [11]. Plomp indicated that the automatic gain control (AGC) such as compensating for the temporal masking attenuates envelope variations in particular the attack/release times [13]. By suppressing amplitude of a speech signal as well as enhancing envelope curve of the speech signal, Kurosawa *et al*. improved word intelligibility in the condition with a simulated hearing impairment [14].

Arai *et al.* [15] focused on the increased temporal masking of people with hearing impairments, and proposed applying steady-state suppression to hearing aids. Kobayashi *et al.* carried out listening tests for elderly people, and speech intelligibility was improved by steady-state suppression [9, 10]. On the other hand, the consonant enhancement technique by Yasutake *et*

al. [8] aimed at compensating for the loudness recruitment.Not only the technique proposed by Yasutake *et al.* [8] but also steady-state suppression enhances consonants. Therefore, steady-state suppression can be considered as compensating both for the loudness recruitment and the increase in the temporal masking.

CONCLUSIONS

We realized steady-state suppression on the function which used to compensate for the loudness recruitment in order to apply steady-state suppression for real-time processing in reverberation. We found the high concordance rate of steady-state portions of both the proposed and the previous steady-state suppression. In order to investigate how the proposed steady-state suppression improves speech intelligibility in reverberation, we would like to carry out listening tests as a subjective evaluation and compare the amounts of overlap-masking with or without steady-state suppression in the future.

ACKNOWLEDGEMENT

A part of this research was supported by Open Research Center Project from MEXT and Grants-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science.

References:

[1] R. H. Bolt, A. D. MacDonald: Theory of speech masking by reverberation. Journal of the Acoustical Society of America **21, No.6** (1949) 577–580

[2] A. K. Nábělek, T. R. Letowski, F. M. Tucker: Reverberant overlap- and self-masking in consonant identification. Journal of the Acoustical Society of America **86** (1989) 1259-1265

[3] N. Hodoshima, T. Goto, N. Ohata, T. Inoue, T. Arai: The effect of pre-processing approach for improving speech intelligibility in a hall: Comparison between diotic and dichotic listening conditions. Acoustical Science and Technology **26** (2002) 212-214

[4] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto, T. Kitamura: Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. Proceedings of Autumn Meeting of the Acoustical Society of Japan (2001) 449-450 (in Japanese)

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

[6] N. Hodoshima, T. Arai, A. Kusumoto, K. Kinoshita: Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments. Journal of the Acoustical Society of America **119 No.6** (2006) 4055-4064

[7] T. Goto, T. Inoue, N. Ohata, N. Hodoshima and T. Arai: The effect of pre-processing for improving speech intelligibility in the Sophia University lecture hall. Proceedings of Autumn Meeting of the Acoustical Society of Japan (2003) 613-614 (in Japanese)

[8] T.Yasutake, Y. Nakamura: Quasi-real-time Consonant Enhancement System. Technical report of IEICE Japan HIP 2005-94 (2005) (in Japanese)

[9] K. Kobayashi, Y. Hatta, K. Yasu, N. Hodoshima, T. Arai, M. Shindo: Consonant enhancement of monosyllable for elderly listeners by steady-state suppression. Proceedings of Spring Meeting of the Acoustical Society of Japan (2005) 321-322 (in Japanese)

[10] K.Kobayashi, Y. Hatta, K. Yasu, S. Minamihata, N. Hodoshima, T. Arai, M. Shindo: Improving speech intelligibility for elderly listeners by steady-state suppression. Technical report of IEICE Japan, SP2005-168 (2006) 31-36

[11] S. E. Gehr, M. S. Sommers: Age differences in backward masking. Journal of the Acoustical Society of America, **106, No.5** (1999) 2793-2799

[12] B. C. J. Moore and B. R. Glasberg: A comparison of four methods implementing automatic gain control (AGC) in hearing aids. British Journal of Audiology **22, No.2** (1988) 93-104

[13] R. Plomp: The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function. Journal of the Acoustical Society of America **83, Issue 6** (1988) 2322-2327

[14] T. Kurosawa, R. Nisimura, Y. Suzuki: A study on validity of envelope emphasis for amplitude compression hearing aids. Proceedings of Spring Meeting of the Acoustical Society of Japan (2003) 459-460

[15] T. Arai, K. Yasu, N. Hodoshima: Effective speech processing for various impaired listeners. Proceedings of the International Congress on Acoustics **II** (2004) 1389-1392