Acoust. Sci. & Tech. **28**, 6 (2007)

ACOUSTICAL LETTER

Inverse correlation of intelligibility of speech in reverberation with the amount of overlap-masking

Takayuki Arai[1,*], Yoshiaki Murakami[1], Nahoko Hayashi[1], Nao Hodoshima[1] and Kiyohiro Kurisu[2]

[1]*Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*
[2]*TOA Corporation, 2–1 Takamatsu-cho, Takarazuka, 665–0043 Japan*

(*Received 25 April 2007, Accepted for publication 17 May 2007*)

Keywords: Speech intelligibility, Reverberation, Overlap-masking, Steady-state suppression
PACS number: 43.55.Hy, 43.66.Dc, 43.71.Es, 43.72.Ew, 43.38.Tj [doi:10.1250/ast.28.438]

1. Introduction

Late reflections degrade speech intelligibility [1], whereas early reflections often help speech intelligibility, and this is called the Haas effect (e.g., [2]). It has been reported that the main cause of degradation in speech intelligibility in reverberant environments is overlap-masking [3–5]. Because of overlap-masking, reverberant components of prior speech segments mask successive segments. As a result, speech segments following reverberating segments are more difficult to understand. As the energy of the prior segments increases, the effect of overlap-masking also increases. This is particularly important when the preceding segment is a vowel, which has more power, and the subsequent segment is a consonant, which has less power [6,7].

A number of researches have proposed and discussed how the intelligibility of speech in reverberation can be estimated from an impulse response of a room. Reverberation time, such as T_{60}, is a simple objective parameter for estimating reverberation [8]. Speech intelligibility usually decreases as T_{60} becomes longer, but different rooms having the same T_{60} might yield different degrees of speech intelligibility. One example is the case where T_{60} is the same in different rooms, but the energy ratios of the direct-to-reverberated sounds are different. The *Deutlichkeit* value, such as D_{50} [9,10] and Clarity, such as C_{50} [9,11], take this direct-to-reverberation ratio into account. The speech transmission index (STI) is another parameter that is widely used to measure speech intelligibility objectively [12,13]. STI is based on the fact that the modulation transfer function depends on reverberation [14].

The intelligibility of speech also depends on the speech signal itself. To reduce overlap-masking, Arai *et al.* [6,7] proposed "steady-state suppression" as a preprocess for speech signals in reverberant environments. Strange *et al.* [15] showed that the information in steady-state portions of a speech signal was relatively insignificant compared with the information in transient portions. Additionally, steady-state portions usually have more energy compared with transient portions. In the "steady-state suppression" technique, overlap-masking is reduced by estimating and suppressing steady-state portions of speech that have high energy but are less important for speech perception, such as the nuclei of syllables. From the results of several experiments in simulated and real sound fields, we found that when we apply this process between a microphone and loudspeaker, it significantly improves speech intelligibility in reverberant environments (reverberation times of 0.7–1.3 s) [e.g., 6,7,16,17].

The conventional measures for estimating the intelligibility of speech in reverberation, which are based on the impulse response of a room, are independent of the speech signal itself. Therefore, they do not reflect the effect of any preprocesses, including nonlinear processing (e.g., the steady-state suppression technique), which are designed to be applied to the original speech signal. There are several empirical STI approaches for predicting the intelligibility of nonlinearly processed speech [18–21]. These approaches could handle nonlinearly processed speech, such as cochlear-implant processed speech.

In this study, we propose a new intelligibility measure that can take into account the effect of nonlinear preprocesses, with a view toward using the steady-state suppression technique to reduce the amount of overlap-masking (OLM) in reverberant environments. In particular, we show how this measure correlates with speech intelligibility.

2. Proposed measure for estimating intelligibility of speech

First, we focus on a target syllable within an arbitrary sentence. Figure 1 shows a conceptualized speech waveform. In this figure, $s(t)$ denotes a target syllable, whereas $p(t)$ denotes a sequence of pretarget syllables. We define $t = 0$ at the boundary between $p(t)$ and $s(t)$ on the horizontal (time) axis. Then, a new measure, the signal-to-OLM ratio *SOR*, is defined as

$$SOR = 10\log_{10}\frac{\int_0^T |s(t) * h_{50}(t)|^2 dt}{\int_0^T |p(t) * h(t)|^2 dt} \text{ [dB]},$$

where $h(t)$ is the impulse response of a room and $h_{50}(t)$ is the first 50 ms of the impulse response. In this case, the direct sound starts at $t = 0$ in the impulse response. The variable T is

*e-mail: arai@sophia.ac.jp

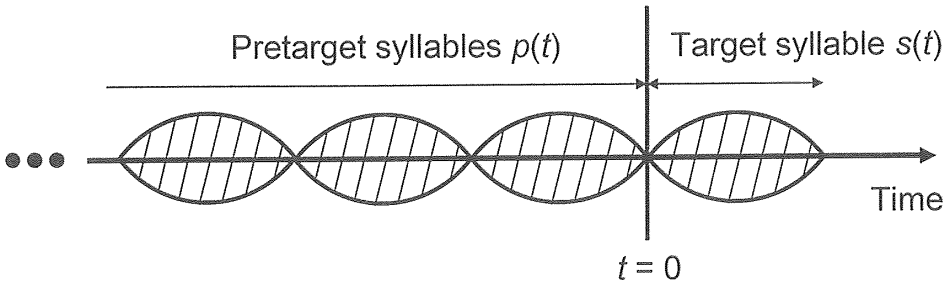

Fig. 1 Conceptualized speech waveform: target syllable $s(t)$ and pretarget syllables $p(t)$.

Table 1 Ten impulse responses with different combinations of the T_{60} and D_{50} values used in this study.

		T_{60} [s]			
		0.7	1.5	2.3	3.0
D_{50} [%]	30	Rev1	Rev2	Rev5	Rev8
	50		Rev3	Rev6	Rev9
	70		Rev4	Rev7	Rev10

the duration for calculating the measure. We used $T = 150$ ms in the following experiment because in this study, we focus particularly on the signal-to-OLM ratio within the initial consonant of a target syllable that is located at $t = 0$–150 ms of $s(t)$ in the experimental setup of this study.

3. Evaluation

To evaluate how the speech intelligibility correlates with the amount of overlap-masking, we compared the correct rate obtained from the following perceptual experiment with *SOR*.

3.1. Reverberant conditions

We conducted a perceptual experiment under artificial reverberant conditions by convolving speech samples with impulse responses. The impulse responses were artificially synthesized from white noise by multiplying temporal envelopes. Table 1 shows ten impulse responses with different combinations of T_{60} and D_{50}.

3.2. Speech samples

The original speech samples consisted of 14 nonsense consonant-vowel (CV) syllables (target syllables) embedded in a Japanese carrier phrase, "Daimoku to shite wa ____ to iimasu" (It is called ____ as a title). The vowel was /a/ and the consonants were /p, t, k, b, d, g, s, ʃ, h, dz, dʒ, tʃ, m, n/. The speech samples were obtained from the ATR Speech Database of Japanese. The same carrier sentence was used for all targets. The beginning position of the target vowel was adjusted to 150 ms from the offset of the pretarget carrier phrase. The ratio of the root-mean square (RMS) in the carrier phrase to that in the CVs was 1:0.7. Finally, we prepared original speech samples and processed speech samples by steady-state suppression following the method of Arai *et al.* [6,7] and further, added a step to avoid suppressing relatively longer continuants where the spectral moment is higher than 3,750 Hz [22], such as sibilant consonants.

3.3. Perceptual experiment

The stimuli were the speech samples convolved with each of the ten impulse responses used in this study.

3.3.1. Participants

Twenty-two young people with normal hearing (15 males and 7 females, aged 20 to 26 years) participated in the experiment. All were native speakers of Japanese.

3.3.2. Procedure

The experiment was conducted in a soundproof room. Stimuli were presented diotically through headphones (STAX SR-303) connected to a computer via the digital-to-analog (D/A) converter of a digital audio amplifier (MA-500U, Onkyo) that was connected to the computer via a USB interface. The sound level was adjusted to each participant's comfort level during a training session prior to the experiment. A stimulus was presented in each trial and the listeners were instructed to select one of 16 options, including 14 CVs, vowel /a/, and 'others,' displayed on the computer screen. The experiment was carried out at each listener's pace. For each listener, 280 stimuli were presented randomly (10 reverberation conditions × 14 CVs × 2 processing conditions).

3.4. Results and discussion

Table 2 shows the mean percentage of correct responses in the perceptual experiment (the second column) and the *SOR* value (the third column) for the ten reverberant conditions. The first and second rows of each cell show the speech intelligibility of nonprocessed speech samples and speech samples processed by steady-state suppression, respectively. We can see that steady-state suppression improves speech intelligibility as well as increases the *SOR* value under most of the experimental conditions used in this study.

Figure 2 shows the scatter plot of the speech intelligibility versus the *SOR* value for each pair of a nonprocessed condition (unfilled circle) and a processed condition (filled circle); a line connects the two conditions. The correlation coefficient among all of the points in this figure is 0.7025, which shows a high correlation between speech intelligibility and the *SOR* value.

4. Conclusions

In this study, we investigated the correlation between the intelligibility of speech in reverberation and the amount of overlap-masking (OLM) due to reverberation. There was a high correlation between the results of a perceptual experiment and the values of the newly proposed intelligibility measure, *SOR*, defined as the signal-to-OLM ratio. In other

Table 2 Mean percentages of correct responses of the perceptual experiment (the second column) and the *SOR* value (the third column) for the ten reverberant conditions. The first and second rows of each cell show speech intelligibility of nonprocessed speech samples and speech samples processed by steady-state suppression, respectively.

Condition	Intelligibility [%]	*SOR* [dB]
Rev1	69.2	−20.0
	72.7	−15.3
Rev2	48.4	−22.6
	53.9	−18.2
Rev3	58.1	−18.8
	61.4	−14.4
Rev4	58.8	−15.0
	62.3	−10.7
Rev5	39.3	−23.2
	45.1	−19.5
Rev6	49.0	−19.6
	54.9	−16.0
Rev7	58.1	−15.5
	61.4	−12.0
Rev8	38.8	−23.5
	38.6	−20.3
Rev9	47.1	−20.0
	52.3	−16.8
Rev10	58.4	−16.0
	60.4	−12.9

words, the intelligibility of speech in reverberation was inversely correlated with the amount of overlap-masking.

Two advantages of using our proposed measure are that: 1) it reflects the reverberation characteristics of a room, as contained in the impulse response of the room, and 2) it also reflects the characteristics of the speech signal itself, as well as the effect of any preprocesses, including nonlinear processing (e.g., steady-state suppression), applied to the original speech signal.

This time, we did not divide the speech signal into frequency bands but treated the signal as one band. However, we can also determine *SOR* by filter bank analysis based on the auditory filter. In this case, a more realistic amount of overlap-masking is estimated and the correlation between the intelligibility of speech and the amount of overlap-masking might be improved. The correlation can also be calculated for each consonant. In the future, we will use "model speech," such as amplitude-modulated white noise, instead of actual speech signals.

Acknowledgments

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

References

[1] A. K. Nábělek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, **17**, 724–739 (1974).

[2] H. Haas, "The influence of a single echo on the audibility of

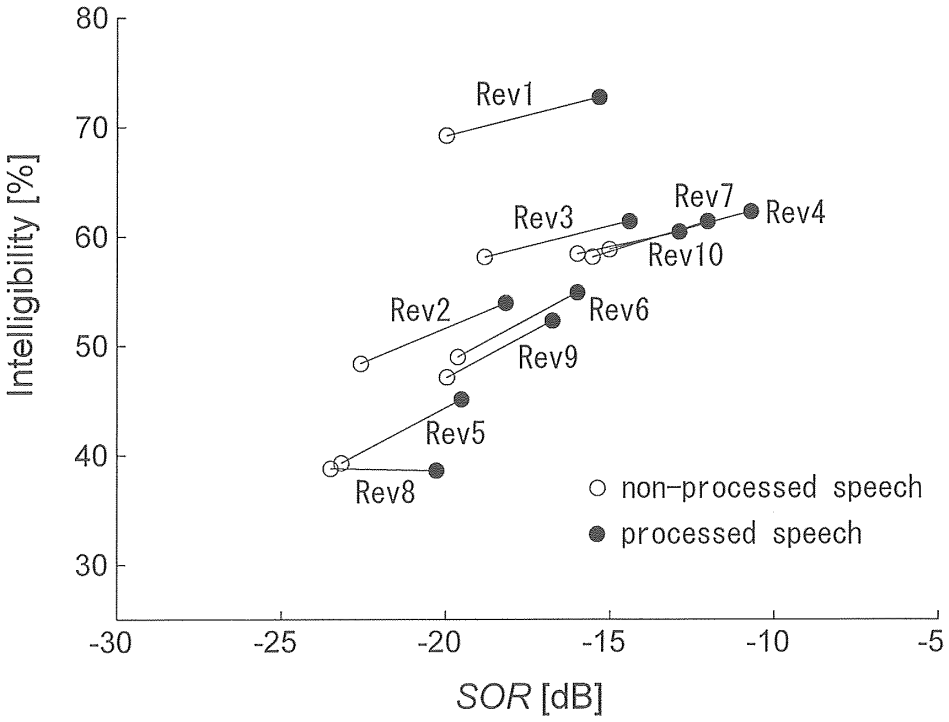

Fig. 2 Scatter plot of speech intelligibility versus the *SOR* value for each pair of a nonprocessed condition (unfilled circle) and a processed condition (filled circle); a line connects the two conditions.

speech," *J. Audio Eng. Soc.*, **20**, 145–159 (1972).
[3] V. O. Knudsen, "The hearing of speech in auditoriums," *J. Acoust. Soc. Am.*, **1**, 56–82 (1929).
[4] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).
[5] A. K. Nábělek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, **86**, 1259–1265 (1989).
[6] 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.*, Vol. 1, pp. 449–450 (2001).
[7] 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).
[8] W. C. Sabine, *Collected Papers on Acoustics* (Dover, New York, 1964).
[9] H. Kuttruff, *Room Acoustics*, 4th ed. (Spon Press, London, 2000).
[10] R. Thiele, "Richtungsverteilung und Zeitfolge der Schallrückwürfe in Räumen," *Acustica*, **3**, 291–302 (1953).
[11] W. Reichardt, O. A. Alim and W. Schmidt, "Definition und Messgrundlagen eines objektiven Masses zur Ermittlung der Grenze zwischen brauchbarer und unbrauchbarer Durchsichtigkeit bei Musikdarbietung," *Acustica*, **32**, 126–137 (1975).
[12] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," *J. Acoust. Soc. Am.*, **67**, 318–326 (1980).
[13] IEC 60268-16 Ed. 3.0, Sound system equipment — Part 16: Objective rating of speech intelligibility by speech transmission index (2003).
[14] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, **77**, 1069–1077 (1985).
[15] W. Strange, J. J. Jenkins and T. L. Johnson, "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.*, **74**, 695–705 (1983).
[16] 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**, 4055–4064 (2006).
[17] 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 binaural listening conditions," *Acoust. Sci. & Tech.*, **26**, 212–214 (2005).
[18] C. Ludvigsen, C. Elberrling, G. Keidser and T. Poulsen, "Prediction of intelligibility of non-linearly processed speech," *Acta Oto-Laryngol. Supple.*, **469**, 190–195 (1990).
[19] R. Drullman, J. M. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, **95**, 2670–2680 (1994).
[20] K. L. Payton, L. D. Braida, S. Chen, P. Rosengard and R. Goldsworthy, "Computing the STI using speech as a probe stimulus," *Past, Present and Future of the Speech Transmission Index* (TNO Human Factors, Soesterberg, 2002).
[21] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, **116**, 3679–3689 (2004).
[22] K. Kobayashi, Y. Hatta, K. Yasu, S. Minamihata, N. Hodoshima, T. Arai and M. Shindo, "Improving speech intelligibility for elderly listeners by steady-state suppression," *Tech. Rep. IEICE*, SP2005-168, pp. 31–36 (2006).