Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

A Case Study of Bootstrap Masker Quality Assessment for Speech-Privacy Protection

Yosuke Kobayashi

Graduate School of Engineering
Muroran Institute of Technology
27-1 Mizumoto, Muroran, Hokkaido, Japan
ykobayashi@csse.muroran-it.ac.jp

Kazuhiro Kondo

Graduate School of Science and Engineering
Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata, Japan
kkondo@yz.yamagata-u.ac.jp

Received January 2017; revised June 2017

Abstract. *In this paper, we discuss the quality assessment of a new method for the generation of a masker for speech privacy protection. This masker includes speech characteristics that prevent eavesdroppers from overhearing conversations in public spaces. Previous research shows that maskers generated from the target speech perform better in interfering with the listening, than the other maskers. Therefore, we propose a bootstrap (BS) masker method that efficiently generates a masker from a small sample of the recorded speech. We evaluate the subjective speech intelligibility and establish that the BS masker can achieve the same level of intelligibility as that of the conventional additional masker at an approximately 4 dB lower target-to-masker ratio.*
Keywords: Speech-privacy, Masker, Intelligibility, Bootstrap

1. **Introduction.** Conversational speech is a fundamental aspect of human communication and varies richly in purpose. Conversations involving private information occur in public spaces such as hospitals, banks, and city offices. However, it is well known that the sound propagation range cannot be easily controlled. Sound and speech can leak from gaps such as doors or windows and may be overheard by a person eavesdropping in the corridor. Thus, there is a need to consider the speaker's privacy and ensure that conversations cannot be heard by a third party.

W. J. Cavanaugh *et al.* recognized two categories of speech privacy [1]: (1) Confidential privacy, in which, speech must not be intelligible to people in the space other than the conversation's participants, and (2) normal privacy, in which, speech must not interfere with the working environment of others. They also noted that the degree of privacy is related to an intelligibility indicator called the articulation index, now referred to as the speech intelligibility index [2]. Confidential privacy can be considered as speech privacy (SP) protection, because it prevents the leakage of the private information from speech. Therefore, increasing the confidential privacy in an architectural space can be considered as a method of information protection.

In recent years, a noise emitting method, called the masker, for enhancing confidential privacy has been implemented, mainly in North America. In some cases, it has been further improved; devices that emit pink noise and background music as maskers, have been used in open spaces for SP protection [3, 4]. We examined such techniques that use noise for masking conversational speech without sounding unnatural. In a prior research, the approaches proposed for generating SP protection maskers include time-domain reversal processing [5, 6] and speech synthesis processing [7]. Both approaches use signal processing technologies to delete the semantics from the source speech and generate maskers possessing the characteristics of the masked speech. With these methods, it is possible to generate more efficient maskers than the conventional pink noise, etc.

In our previous study, we had categorized the source-speech combinations used for generating the maskers into four types; when all the maskers were tested at the same sound levels, it was determined that the maskers generated from the speaker's own speech decrease the intelligibility the most [8]. However, numerous speech signals need to be obtained in advance to generate a masker from the speaker's speech in this method. For example, a new user of the masker system needs to speak a number of sentences beforehand to generate his or her personal masker; or, the administrator needs to edit and prepare multiple signals of the user beforehand from previous recordings. Assuming we have this recording, we generated a 10 s masker by preparing 16 different sequences of 10 s speeches of a speaker, and averaging all of these in our previous research [8]. However, it is often difficult to collect speech signals in advance to generate the maskers from speakers who have not previously participated in a conversation. We also confirmed that if we do not have the previous recordings of the speaker, the maximum performance cannot be obtained because maskers from speech other than their own speech need to be used to prepare the masker.

Bootstrapping is a type of Monte Carlo method, which uses random sampling and shuffling for estimating the distribution, based on a large volume of data from a small data set. In the research field of machine learning, it is reported that overtraining can be suppressed by improving the prediction accuracy of an unstable learning model by bootstrapping, which is also used for the statistical analysis of complex population parameters. This method can estimate the distribution of a large volume of data from a small data set. Previously, we proposed a masker generation method using an algorithm similar to the bootstrap method in machine learning [9]. We applied random sampling and shuffling to process the speech signals for masker generation from a small speech data set. We called this the bootstrap-type (BS) masker.

We had previously reported the bootstrapping masker's listening difficulty [10], with a subjective evaluation [11, 12]. It was determined that the proposed masker demonstrates the highest listening difficulty performance compared to the other masker types. However, the intelligibility of the BS masker was not evaluated and the absolute SP protection performance was unknown. In this research, intelligibility testing with the "Word Familiarity Controlled Word Lists 2007 (FW07) [13]" was conducted to further evaluate the performance of the BS masker in detail. This study was conducted as a case study with fixed synthesis parameters of the proposed BS method and conventionally used AD method maskers. As a result, we found a significant difference for the proposed BS masker compared to the conventional AD masker.

2. **Masker generation method.**

2.1. **AD Masker.**

D. Kobayashi *et al.* proposed a human speech-like noise (HSLN) [14]. The HSLN is a method for generating speech noise using multiple superposition of the

recorded utterance signals. The HSLN signal, $n_N[k]$, for the N-th superposition is defined as follows:

$$n_N[k] = \alpha \sum_{m=0}^{N-1} s[m f_s L + k], \quad 0 \leq k \leq f_s L - 1 \tag{1}$$

where N is the number of superpositions, k is the sample period, f_s is the sampling frequency, L is the total length of the HSLN signal, $s[\cdot]$ is the speech signal sample used for generating the maskers, and α is the normalization coefficient. This method assumes that the utterance continues for a sufficiently long interval (as shown in Fig. 1) and that the HSLN is created with $L = 1$ s and $N = 16$. L was set to 1 s to match the intelligibility evaluation word signal length, and N was set to 16 to directly compare the results to the previous study [8], which also used $N = 16$. The value of $N = 16$ was selected in the previous study as an appropriate value that meets the requirements of the masking efficiency and the natural quality of the masker. We also observed the same balance, and therefore decided to use the same parameters here. The HSLN is recognized as a signal with multiple speech superposition [14]. The normalization parameter of α was set to 1/16, the arithmetic mean. In this paper, we implemented the HSLN using a ring buffer to enable operation in real time; as in the example shown in Fig. 1, the segments were collected from the buffers four, five, and three from the left in the specified order and these were averaged. The HSLN, using a ring buffer, will henceforth be referred to as an additional (AD) masker.

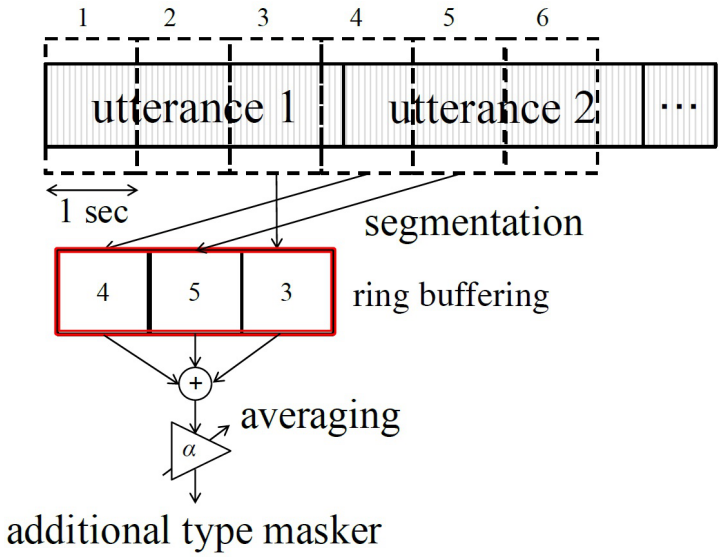

FIGURE 1. AD-type masker generation procedure (*e.g.* $L = 1$ s and $N = 3$).

2.2. **BS Masker.** For the AD masker, a speech cache of size, $L \times N$ s, is necessary. The superposition number should be reduced if only a few usable speech signals are available, leading to a risk of speech content leakage from the semantics of the source speech. Moreover, the speech needs to continue for a considerable duration for the superposition to be sufficient.

In order to solve this problem, we propose a BS-type masker that splits the cached speech into smaller sub-segments and then rejoins them, as shown in Fig. 2. We call this a BS masker because it is similar to a type of Monte Carlo method called bootstrap processing, which includes the resampling and shuffling processes.

In this research, the length of the segment is set to 1 s, and the length of the sub-segment is set to 125 ms, which is approximately equal to an average speech rate of 8 mora per second. In addition, the sub-segments do not possess any of the linguistic significance of the source speech on their own. Therefore, a 1 s masker is generated from a 1 s speech; even if the speaker changes during the dialogue, the masker generation continues. In this experiment, we set it to 1 s because it was empirically possible to follow the dialogue of this length in the preliminary experiments. Furthermore, multiple consonants and vowels are expected to be included in the masker for time lengths of approximately 1 s. Although not shown in Fig. 2, a Hamming window was applied over each sub-segment signal. Therefore, a quarter of the sub-segment length overlaps at each joint to suppress the masker generation in the low sound-pressure sections. A uniformly distributed random number sequence was used for resampling in the BS. Moreover, at the start and end of the coupled masker, there is a signal time with little signal power due to the windowing process. We removed this part and used the central one second as the BS masker.

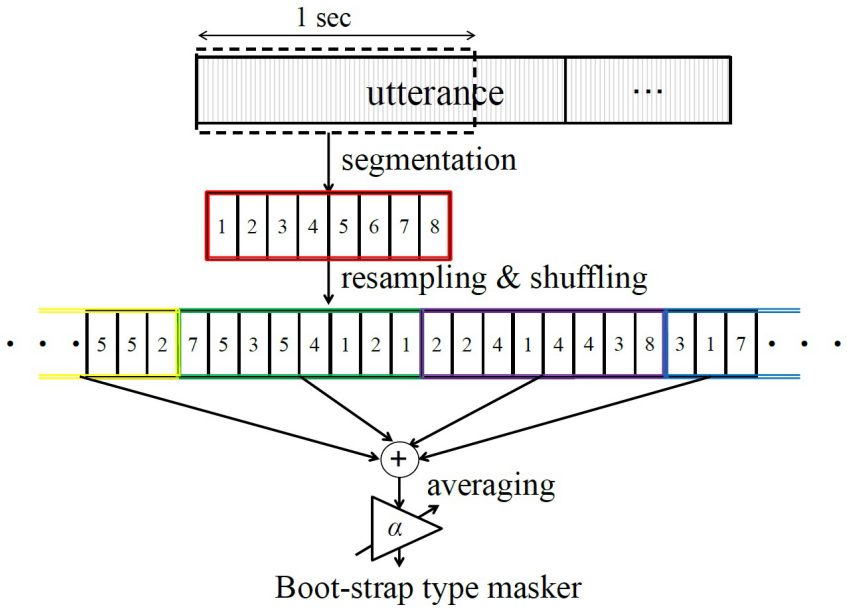

FIGURE 2. BS-type masker generation procedure.

The BS masker is better than the AD masker with respect to the following: The shuffling process suppresses the reverberation more effectively than the simple addition process. The efficiency of the resampling process enables the same sub-segment to be used several times. Therefore, a long speech signal is not necessary. Moreover, the generation of BS maskers that are longer than the cached speech length is also possible because shuffling permits duplication.

3. **Subjective Evaluation.**

3.1. **Approach.** W. J. Cavanaugh *et al.* reported that speech privacy is related to the speech intelligibility that invades one's own space [1]. In this research, we evaluate the intelligibility of the proposed BS and conventional AD maskers. The evaluation experiment was controlled using the energy ratio between the target speech (evaluated speech) and the masker was defined by the target to masker ratio (TMR). The TMR is similar to the signal to noise ratio (SNR) in signal processing. It is calculated using the target

speech energy, E_T and the masker energy, E_M, as depicted in Eq. (2).

$$\mathrm{TMR} = 10 \log_{10} \frac{E_T}{E_M} \tag{2}$$

3.2. **Intelligibility test method.** In this research, we compare the intelligibility of the maskers generated by the above methods using the FW07 [13]. This evaluation method defines the familiarity levels for Japanese words, which are based on a psychological evaluation. The FW07 consists of 20 lists of 20 evaluation words, categorized by different levels of familiarity, each spoken by two male and two female Japanese speakers. In this research, the intelligibility was evaluated by selecting one male and one female speaker from the list of sound sources ranked as, "highly familiar" in the FW07. The intelligibility of FW07 is defined in Eq. (3), in terms of the total number of words, T_W and the number of correct answers, C.

$$\text{Intelligibility} = \frac{C}{T_W} \tag{3}$$

3.3. **Speech recognition threshold.** The estimation function for the intelligibility is a logistic function, as shown in Eq. (4).

$$\text{Estimated intelligibility} = \frac{1}{1 + \exp\left(-\beta_0 - \beta_1 x\right)} \tag{4}$$

This function was calculated as a general linear model with the TMR as the explanatory variable. In the equation, x is the explanatory variable and β_n is the regression coefficient, calculated by maximum likelihood estimation. The value of x, at which the estimated intelligibility is 0.5, is the speech recognition threshold (SRT). In this research, we calculate the value of the explanatory variable as the SRT at which the discrimination rate of the evaluation words are 50%. The SRT can be calculated by Eq. (5), using the regression coefficients, β_n.

$$\mathrm{SRT} = -\frac{\beta_0}{\beta_1} \tag{5}$$

3.4. **Experimental settings.** Table 1 shows the evaluation conditions of the intelligibility test. For both the AD and BS maskers, 40 words from two lists in the FW07 were allocated, per TMR, ranging from −10 to 10 dB per speaker. Hence, the evaluation words per subject were assigned two methods, two speakers, five conditions, and 40 words each, for a total of 800 words.

Table 1. Speech intelligibility test condition.

Speaker	Female 1, Male 1
Masker	AD, BS
TMR	−10, −5, 0, 5, 10 dB
Word num.	40 words/condition
Subjects	10

A continuous-word speech signal without overlaps was developed from the unused words for each TMR condition, and utilized as the sound source for the maskers. Maskers using both the approaches were generated from the developed signal and a random segment was extracted for each test word. As the masker segments varied, the average frequency

response across 40 words was used to smooth the effect of the frequency response variation, in the used segments.

The evaluated speech was used in accordance with the standard configuration of the FW07. Subjective testers listened to random speech from a PC connected to an audio interface (Roland, UA-25) via headphones (Sennheiser, HDA-300). The subjects were ten men aged 18-22 and the experiments were conducted in a soundproof booth.

3.5. **Experimental Results.** Figure 3 shows the results of the evaluation and the overall average of the estimated function; Fig. 4 shows the results by speaker gender. The error bars in both figures are the 95% confidence intervals. Table 2 presents the list of SRTs calculated using the estimation function. Figure 3 demonstrates that the difference in the intelligibility of the maskers is 0.43 when the TMR is 0 dB. Therefore the BS masker exhibits considerably lower intelligibility. Thus, a significant decrease in the intelligibility of the BS masker was demonstrated, indicating that it is more effective than the AD masker. The SRT with the AD masker was -2.6 dB and that with the BS masker was 1.4 dB. Thus, the total difference in the SRT was 4 dB. Note that a higher SRT indicates a more effective masking.

Table 2. SRT for each method with respect to the TMR (dB).

	mean	female	male
BS	1.41	0.97	1.84
AD	-2.60	-2.01	-3.19

3.6. **Statistical Analysis.** In order to statistically analyze the two methods and the differences due to gender, we tried 3-way Repeated Measures ANOVA with three factors: TMR, method, and gender. The results are shown in Table 3. The factors with significant differences are marked with an asterisk. From the results, it can be seen that there is a significant difference between the AD and BS methods ($F(1, 9) = 5.721$, $p < .05$). Moreover, there is no significant difference due to the speaker's gender alone ($F(1, 9) = 2.811$, $p = 0.128$). On the other hand, all interactions involving the speaker's gender had significant differences. From the above results, it was shown that the masker method has a statistically significant effect on intelligibility.

3.7. **Discussion.** As shown in Fig. 4 and Table 2, the results for each generation method vary by speaker gender; however, the largest observed difference in the intelligibility, for the same TMR, was 0.2 at most for all the ranges. The SRT was lower for the males in the AD masker, whereas it was lower for the females, in the BS masker. The methods have varying effects based on the gender of the speaker, but the difference in the SRT by gender was less for the BS masker (0.9 dB) than for the AD masker (1.2 dB). This difference is less than that between the masker generation methods and is not a significant difference as mentioned in section 3.6. On the other hand, the effect of the masker method was greater for males, which is evidenced by the fact that the difference in SRT between the methods was 5.0 dB for males, whereas it was only 3.0 dB for females. This difference was significant, as shown in section 3.6. In summary, the BS masker performed more efficiently, than the AD masker, enabling the same level of intelligibility reduction at a masker level 4 dB less. Thus, the BS masker can be emitted 4 dB lower to expect the same level of SP protection as the AD masker.

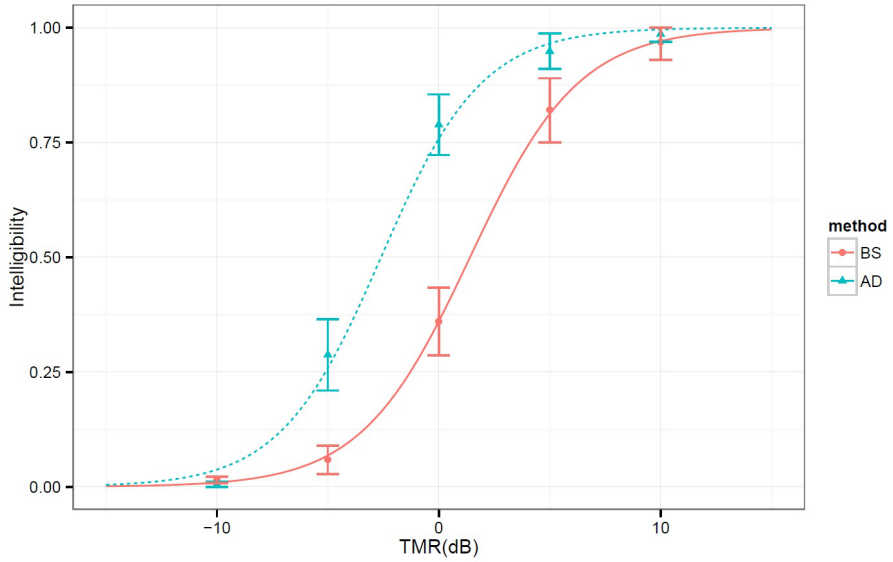

FIGURE 3. Intelligibility vs. the TMR with respect to the method; Error bars are 95% confidence intervals.

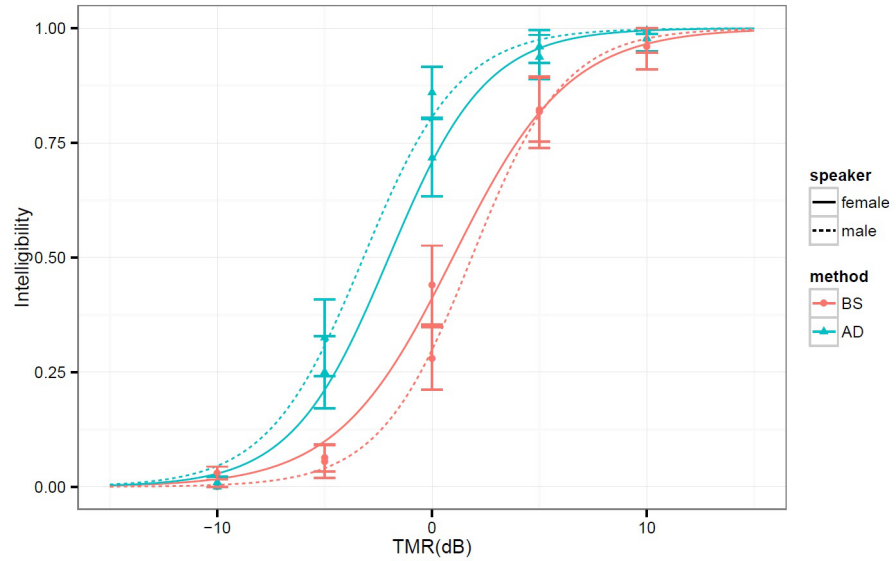

FIGURE 4. Intelligibility vs. the TMR with respect to the speaker; Error bars are 95% confidence intervals.

3.8. **Time–frequency analysis.** Figure 5 shows the spectrogram of a female speaker for a target sound (using the test word "kanzashi"), in which typical characteristics were observed. Figures 6, 7, and 8 display the spectrograms for the AD and BS maskers, respectively. The BS masker shows two examples generated from different speakers. These figures were calculated by resampling the evaluated speech, which had a TMR of 0 dB during the subjective evaluation in chapter 3 with a 16 kHz sampling frequency.

The target speech was a normal utterance and its speech characteristics were observed clearly for each speech mora. The AD masker's spectrum range at 4 kHz was considerable and had babble noise-like characteristics without demonstrating temporal change (flatter characteristics). However, the BS masker was more speech-like and exhibited temporal changes similar to those of the target speech. One of the critical results is that both the BS figures exhibit discontinuous time connections. This discontinuity is owing to the current generation algorithm that shuffles using a low-energy consonant segment.

TABLE 3. ANOVA results

Source	Sum of squares	DF	Mean Squares	F	Pr($>F$)
Subjects(S)	0.4389	9	0.04876		
TMR	28.993	4	7.248	1050	<2e-16 **
S*TMR	0.249	36	0.07		
Method	0.01280	1	0.012800	5.721	0.0404 *
S*Method	0.02014	9	0.002238		
Gender	0.001800	1	0.001800	2.811	0.128
S*Gender	0.005762	9	0.0006403		
TMR*Method	2.5190	4	0.6298	133.2	<2e-16 **
S*TMR*Method	0.249	36	0.07		
TMR*Gender	0.01776	4	0.004441	2.722	0.0446 *
S*TMR*Gender	0.05874	36	0.001632		
Method*Gender	0.05611	1	0.05611	107.4	2.65e-06 **
S*Method*Gender	0.00470	9	0.00052		
TMR*Method*Gender	0.1926	4	0.04814	39.57	1.04e-12 **
S*TMR*Method*Gender	0.0438	36	0.00122		
Total	32.424212	90			

** < 0.01, * < 0.05

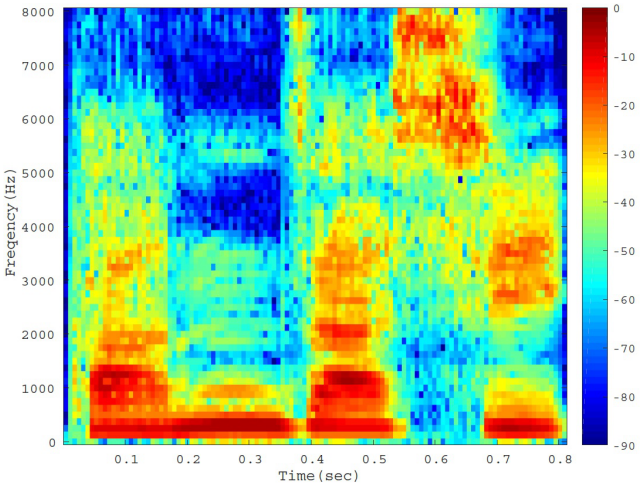

FIGURE 5. Spectrogram of the test word "kanzashi" from a female speaker; the color bar represents the magnitude(dB).

We consider that, in these results, because the BS masker was generated not by adding multiple speech segments but by adding smaller segments, it became speech-like, at least in the time-frequency level.

The above results indicate that the BS masker spectrum is more similar to the speech source spectrum containing semantics. This may explain the higher masking efficiency of the BS maskers. In addition, we established that with the BS masker, the TMR is more correlated with the spectrum distortion measures than with the AD masker. This suggests that we may be able to use the objective distance measure related to the spectrum distortion for estimating the TMR and even the intelligibility.

4. **Conclusions.** In this research, we conducted a subjective intelligibility evaluation of the proposed BS and conventional AD maskers. The results demonstrated that the BS

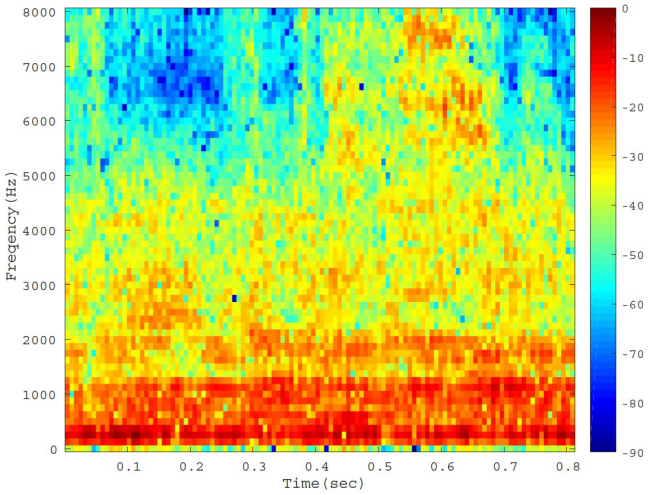

FIGURE 6. Spectrogram of the AD masker generated from a female speaker; the color bar represents the magnitude(dB).

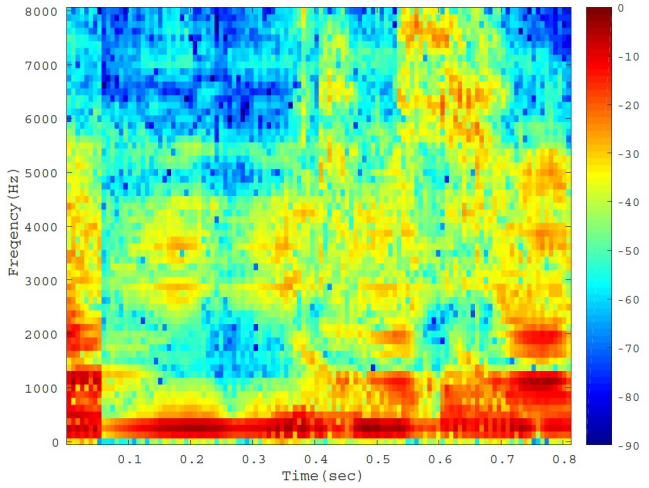

FIGURE 7. Spectrogram of the BS masker generated from a female speaker; the color bar represents the magnitude(dB).

masker could achieve the same level of SRT as the AD masker at 4 dB higher TMR (4 dB lower noise level). In addition, we also found a significant difference in the results of the statistical analysis for the masker generation method. Next, we analyzed the spectrogram of a female speaker for the target sound with both the types of maskers; it was determined that its similarity with the speech time–frequency characteristics may influence the subjective intelligibility. In addition, comparing the spectral distributions of the AD maskers and BS maskers, we observed that with the BS masker, the TMR is more correlated with the spectrum distortion from the original speech than with the AD masker. This result suggests that objective evaluation from the frequency characteristics is possible. In future, we intend to generate BS maskers based on the objective estimation results of its masking efficiency.

Acknowledgment. This work was supported in part by JSPS KAKENHI, Grant Number No. 16K21584, the Cooperative Research Project of the Research Institute of Electrical Communication, Tohoku University (H26/A14), the Artificial Intelligence Research

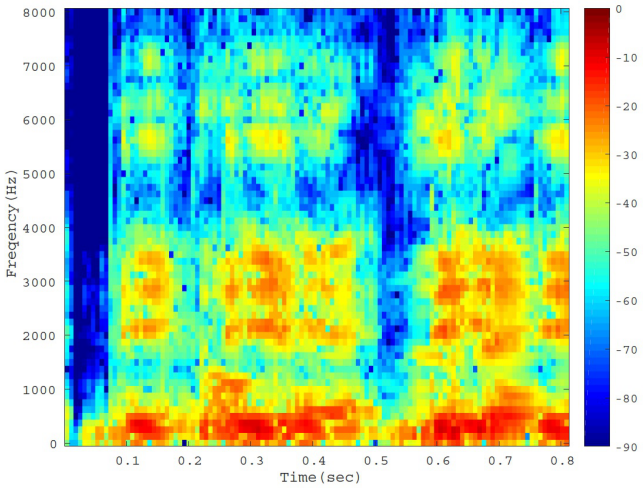

FIGURE 8. Spectrogram of the BS masker generated from a male speaker; the color bar represents the magnitude(dB).

Promotion Foundation, the Tateishi Science and Technology Foundation, and the Mazda Foundation.

REFERENCES

[1] W. J. Cavanaugh, W.R. Farrell, P. W. Hirtle and B. G. Watters, "Speech privacy in buildings.," *J. Acoust. Soc. Am.*, 34, pp. 475–492, 1962.

[2] ANSI S3.5-1997, "Methods for calculation of the speech intelligibility index," *American National Standards Institute*, New York, 1997.

[3] M. Fujiwara, Y. Shimizu, M. Hata, H. Lee, K. Ueno and S. Sakamoto, "Experimental study for speech privacy with a sound masking system in medical examination room," *Proc. of INTERNOISE 2009*, 2009.

[4] M. Fujiwara, M. Hata, T. Yamakawa and Y. Shimizu, "Experimental study of speech privacy with a sound-masking system in pharmacies," *Proc. of INTERNOISE 2011*, 2011.

[5] T. Arai, "Masking speech with its time-reversed signal," *Acoust. Sci. & Tech.*, vol. 31, no. 2, pp. 188–190, 2010.

[6] Y. Hara, M. Tohyama and K. Miyoshi, "Effects of temporal and spectral factors of maskers on speech intelligibility," *Applied Acoustics*, vol. 73, pp. 893–899, 2012.

[7] M. Akagi and Y. Irie, "Privacy protection for speech based on concepts of auditory scene analysis," *Proc. INTERNOISE 2012*, 2012.

[8] K. Kondo, H. Sakurai, S. Kashiwada and T. Komiyama, "Speaker and Gender-Dependent Maskers for Efficient Speech Privacy Protection," *Noise Control Eng. J.*, vol. 62, no. 6, pp. 411–421, Nov./Dec. 2014.

[9] T. Hesterberg, D.S. Moore, S. Monaghan, A. Clipson and R. Epstein, "Bootstrap Methods and Permutation Tests," *Introduction to the Practice of Statistics*, vol. 5, pp. 1–70, 2005.

[10] M. Morimoto, H. Sato and M. Kobayashi, "Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1607–1613, 2004.

[11] Y. Kobayashi and K. Kondo, "Bootstrap masker generation method for speech masking systems," *Proc. INTERNOISE 2014*, P-39, 8 pages, 2014.

[12] Y. Kobayashi and S. Kanemaru, ?Listening Difficulty Assessment for Real-time Bootstrap Masking System,??*Proc. of IEEE GCCE 2016*, pp. 413–415, 2016.

[13] S. Sakamoto, T. Yoshikawa, S. Amano, Y. Suzuki and T. Kondo, "New 20-word lists for word intelligibility test in Japanese," *Proc. of Inter-Speech 2006*, pp. 2158–2161, Sep. 2006.

[14] D. Kobayashi, S.Kajita, K.Takeda and F.Itakura, "Extracting speech features from human speech like noise," *Proc. ICSLP 1996*, pp. 418–421, 1996.