Bootstrap masker generation method for speech masking systems

Yosuke KOBAYASHI [1]; Kazuhiro KONDO[2]

[1] National Institute of Technology, Miyakonojo College, Japan

[2] Yamagata University, Japan

ABSTRACT

Currently, speech masking systems make use of pre-recorded speech signals to generate maskers, which we call offline generated maskers. The offline masker includes no feedback from the speech environment, and only the overall averaged masking effect of the speech is gained, not the environment. In our prior study, a speaker-dependent masker that is generated by mixing pre-recorded speech of the speaker to be masked has been found to be the most effective, and can mask at low sound levels. Accordingly, the real time acquisition of the masked speaker speech signal is required to create the masker, which we shall call online generated maskers.

In the online masker generation, there is a problem that sufficient amount of sound material may not be available from the cached memory in real time. Therefore, we have applied the bootstrap method used in machine learning techniques, and generated a masker as if many speech samples from a small amount of speech is available. We tested the proposed masker using two subjective indexes, *i.e.*, annoyance and listening difficulty. We used sentence speech signals recorded with a dummy head. We compared the bootstrap type online masker (BS), the ring buffer addition type online masker (RA), and 3 types of offline maskers. These maskers were played at three signal noise ratio (5, 0, -5 dB). As a result, the annoyance scores of the online maskers were about the same as the offline maskers. However, the listening difficulty scores improved, and the BS type online masker was the most effective masker when the SNR is higher, at 5 and 0 dB. In addition, the masking effect of the speaker dependent condition (masker created from the target speaker speech) using the BS-Human Speech-Like Noise (HSLN) was found to be significantly higher than others, especially at the Target to Masker Ratio (TMR) of 5 dB.

Keywords: Speech privacy, masking system, Bootstrapping, Online signal processing,
I-INCE Classification of Subjects Number(s): 74.9

1. INTRODUCTION

In recent years, masking system for speech privacy protection that presents electro-acoustic maskers, which may commonly Back Ground Music (BGM) or pink noise, are used in open spaces, such as a bank or a pharmacy. In the latest products, human speech-like noise (1) maskers (HSLN maskers) are also available. For example, Fujiwara *et al.* have proposed a masker generated from instantaneous speech signals, where the recorded speech signal is time-reversed and played out to mask the speech (2). This masker seems to be efficient, and can lower the speech intelligibility effects at low levels. But its masker is the pre-recorded sound signal which includes no feedback from the current speech environment, and thus the ambient characteristics is not reflected in the masker. So far, we have studied the sound sources to be used for the generation of HSLN maskers that aims to protect speech privacy at the smallest possible level (3). Accordingly, we compared the four masker generation methods that use the speaker's own speech. The results showed that the HSLN masker generated using the speaker's own speech, as well as speech with same gender were found to significantly reduce the speech intelligibility, *i.e.*, effective for speech privacy protection. On the other hand, Akagi and Irie have proposed a similar masker that is generated from recorded instantaneous speech samples (4). They preserve the fine structure of the speech spectrum, but

[1] ykobayashi@cc.miyakonojo-nct.ac.jp
[2] kkondo@yz.yamagata-u.ac.jp

scramble the spectrum envelope. This masker seems to be even more efficient, but again seems to present some unnatural characteristics, which can be annoying. These problems seem to be due to the manner the speech signals are scrambled.

In this paper, we propose a method for generating a HSLN masker with efficient use of speaker's own speech signal, using only a small cache of about 1 s speech. In the proposed method, cached speech is segmented into small segments, and is slightly duplicated and shuffled. This is similar to the bootstrap method which is used in the Monte Carlo method in the machine learning field. We have developed a GUI for the subjective test of the proposed method. This GUI provides a psychological 4-point scale evaluation measure for each of the test signal as "listening difficulty (5)" and "annoyance (6)". As the result of the subjective evaluation, the proposed method is higher than others for the listening difficulty, but the annoyance is comparable. From the above results, we have confirmed that the bootstrap type HSLN masker is more efficient compared to the conventional methods.

2. COMPEARED MASKER GENERATION METHODS

2.1 Overview

Synthesis of HSLN masker used in speech masking system has been proposed in (1, 3). In this paper, we compare the addition type HSLN masker (AD-HSLN masker), the proposed bootstrapped HSLN masker (BS-HSLN masker), and the classical maskers.

2.2 Classical Maskers

We select pink noise, babble noise available from SPIB (7), and multi-talker noise from the TY-89 CD (8), all of which are widely used in current products. In this paper, we named these thre maskers the classic maskers.

2.3 Addition type Human Speech-Like Noise Masker (AD-HSLN masker)

D. Kobayashi *et al.* proposed the N times addition type HSLN signal as follows.

$$\mathrm{HSLN}[k] = \alpha \sum_{m=0}^{N-1} s[mK+k], 0 \le k \le K-1 \quad (1)$$

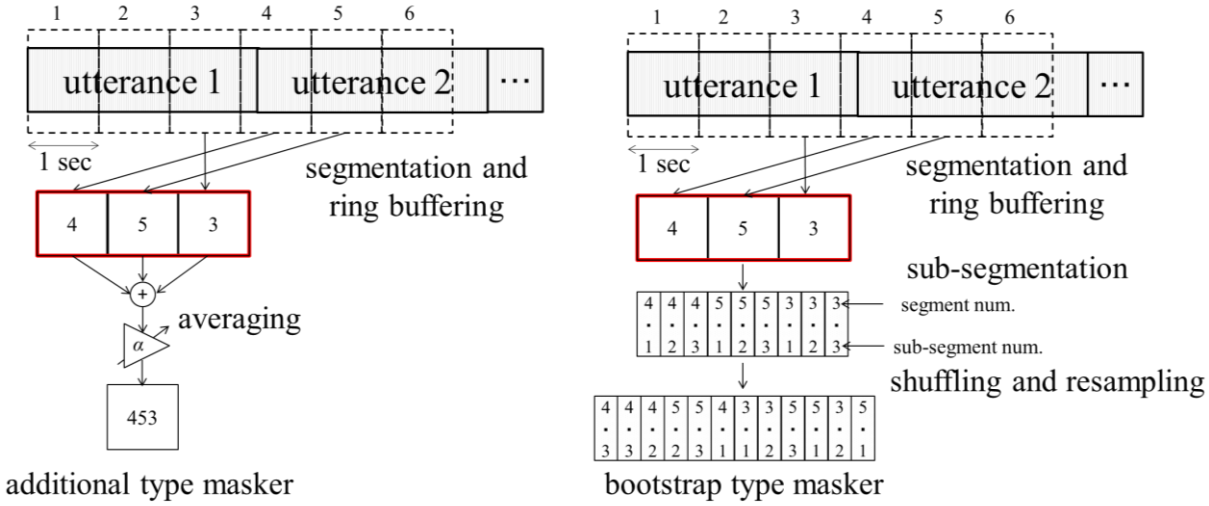
Where N is the number of additions, k is the sample number, K is HSLN signal time length, S [] is source speech signal and α is normalization coefficient. Accordingly, equation (1) is the process of adding cyclically segmented signal of length K. In this paper, we improve the synthesis of this addition type HSLN masker to suit online processing. We set the synthesis parameters of the HSLN masker to N = 16 and K = 1 s, as shown in Figure 1 (a). However, it is necessary to respond to sound longer than 16 seconds for online processing. This can be accomplished using a ring buffer as shown in Formula 2 when the number of additions, N is 17 or more. The normalization factor α, was so that the averaging factor. That is, it is set to 1 / N when N = 16 or below, and N=1/16 when N is more than 16.

$$\mathrm{HSLN}[k] = \frac{1}{16} \sum_{m=N-16}^{N-1} s[mK+k], 0 \le k \le K-1 \quad (2)$$

2.4 Bootstrap type Human Speech-Like Noise Masker (BS-HSLN masker)

In the AD-HSLN masker, there is a possibility that the efficiency of the masking is reduced when the number of additions is small. The effect of the low-level segments due to the time variation of the speech may become apparent if not enough samples are averaged. Thus, with the AD-HSLN masker, there is a possibility that the speech contents may leak when N is small, because the language information may remain audible in the speech materials. Therefore, in order to increase the performance of the AD-HSLN masker, it is necessary to use long speech samples to increase the number of additions. To solve this issue, we propose a method to re-divide segments into smaller sub-segments of speech in random order. The concept of this proposal is shown Figure 1 (b). In the

proposed method, duplicates of the sub-segments are shuffled and added. We named this proposed masker the BS-HSLN masker because this operation is similar to the bootstrap which is one of the Monte Carlo methods. In this paper, we have set the sub-segment length to 1/8 s, which is roughly the duration of one mora at average speech rate. These are recombined by overlapping 1/4 frame length, and applying the Hamming window to the boundary of each sub-segment. This process is not shown in Figure 1 (b). It is possible to suppress the effect of a low-level segments in HSLN masker using this process. Moreover, the synthesis of maskers longer than the cached speech length is also possible since the shuffling allows duplicates.



(a) Addition type (b) Bootstrap type

Fig. 1 Human Speech-Like Noise Masker Generation Procedure (*e.g.* $N = 3$).

3. SOUND SOURCE

3.1 Sound Corpus

In subjective test, we used 2 male and 2 female speaker's voice signal. Each speaker spoke 60 sentences from the Acoustical Society of Japan Continuous Speech Corpus for Research (ASJ-JIPDEC) 503 ATR phonetically balanced sentences. Text examples are as follows.

- ARAYURU GEN'JITSUO SUBETE JIBUN'NO HO-E NEJIMAGETANODA
- RYOKAN'YA HOTERUNI TSUKUTO HIJO-GUCHIO TAZUNERU
- NIQPON'NO ESUPERAN'TOTOSHITE YAHARI HYO-JUN'GOWA HITSUYO-DA

We select 20 speech samples for the synthesis of the masker, and all 60 speech samples were used as target speech for the subjective test. Since there is a limit to the number of sentences of the same speaker in the corpus, we allowed duplicates in the target speech sources and synthesized speech source.

3.2 Seed Speech Source for Masker Generation

As shown in our previous work (3), speech privacy protection effect of the HSLN masker synthesized from the own speaker's speech was the best, and the same gender's speech was the second. Therefore, we compare the BS-HSLN masker and AS-HSLN masker in the combinations shown in Table 1. In this table, we have indexed the two male speakers as m1 and m2, and the female speakers as f1 and f2. We have named as follows the speaker-dependency type, in accordance with the combination of the target speech talker and the seed speech talker. The speaker dependent (SD) type designates the same speaker for target and seed speech. The gender dependent (GD) type designates the same gender speech for the target and the seed speech, but different speaker. In the speaker-independent (SD) type, there is no match in the speaker in the target and seed.

Table 1 Combination of the experimental speaker.

Seed	Target	Dependent type	Seed	Target	Dependent type
m1	m1	SD	f1	m1	SI
m1	f1	SI	f1	f1	SD
m2	m1	GD	f2	m1	SI
m2	f1	SI	f2	f1	GD

3.3 Target Speech Source

We performed a subjective test for a total of 24 conditions. The conditions consist of one control condition, 3 classical masker conditions, the BS-HSLN masker and the AD-HSLN masker which includes 10 kinds of combinations (see Table 1) of synthetic base speech source. The 3 sentences were assigned to each of the 23 conditions to be added to the masker. The required number of sentences is 69 sentences by speaker gender. The subjective evaluations by speech signal generated for each gender speaker (F1 and M1) were performed for every condition. Total test signals with the added maskers are 138 sentences. The remaining 12 sentences were used for training and control conditions.

4. SUBJECTIVE TEST SET UP

4.1 Psychological Attribute Scale

W.J. Cavanaugh *et al.* reported that speech privacy level is closely associated with speech intelligibility (9). For this reason, many of speech masking systems have been tested for subjective speech intelligibility or subjective voice articulation. However, we would like to evaluate many test parameters in the prototype development of the new masker synthesis methods. Thus, the evaluation time of these subjective intelligibility tests becomes enormous.

Therefore, we considered testing using psychological attribute scale only to evaluate the speech masking system. We select the two psychological attribute scales, the listening difficulty (LD) scale (5) and the annoyance (AN) scale (6). Table 2 shows the LD score and its corresponding criteria. This scale is determined by the percentage of score other than "not difficult to listen". The LD rate is calculated by the following equation. T is the total number of responses. The $Count(L1)$ is the sum of "Not difficult" responses.

$$LD\ rate = \frac{T - Count(L1)}{T} \quad (3)$$

Annoyance scale is a measure that has been discussed by the team6 (Community Response to Noise) of the International Commission on Biological Effect of Noise (ICBEN). A maximum score of 5 was used in its definition. However, maskers that are too annoying are not practical. Therefore, we merged the "Extremely annoying" and "Very annoying" score, resulting in a 4 point scale shown in Table 3. The AN ratio is to be calculated in the same way as the LD ratio. The AN rate is calculated by the following equation. The attributes of the A1 and A2 categories both end with “not” in the Japanese translation (“not at all annoying” and “relatively not annoying”). Therefore, we deded to use A1 and A2 responses together as positive responses on the AN scale. T is the total number of responses. $Count(L1)$ and $Count(L2)$ are sum of "Not at all annoying" and "Slightly annoying" responses.

$$AN\ rate = \frac{T - (Count(A1) + Count(A2))}{T} \quad (4)$$

Table 2 List of the listening difficulty scale.

Score	Attributes
L1	Not difficult
L2	Slightly difficult
L3	Fairly difficult
L4	Extremely difficult

Table 3 List of the annoyance scale.

Score	Attributes
A1	Not at all annoying
A2	Slightly annoying
A3	Moderately annoying
A4	Extremely annoying, Very annoying

4.2 Subjective Test GUI

Figure 2 shows the GUI of subjective test for Windows PC. Subjects operate this GUI as follows. The vertical axis of the GUI window is set to the annoyance scale, and the horizontal axis is set to the listening difficulty scale. The subjects select their ratings from this plane. Playback of the same speech source is allowed only once.

[1] Play the sound source by clicking the play button.
[2] Radio button is selected which corresponds to the evaluation value of the 2 scales.
[3] Click the button to go to the next sample.

4.3 Recording Condition

To eliminate differences in the presentation of sound between the subjects, we pre-recorded the emitted masker and target speech from two loudspeakers using a dummy head (Southern Acoustics, SAMURA) with binaural earphone type microphone (Roland, CS-10EM) placed in the ears, located in the soundproof rooms at Yamagata University. Figure 3 shows the recording condition. We set the target emission loudspeakers at a height of 2.0 m from the floor directly in front of the dummy head, and a masker emission loudspeaker at the same distance and the height position of the 0.2 m from the floor (both of the loud speakers were BOSE MMS-1). For calibration between the loudspeakers, we emitted 1/1 octave band noise with a center frequency 1 kHz from each speaker, and set the level emmitted from each speaker to 60 dBeq at the dummy head around the pinna. This level is defined as 0 dB TMR (Target to Masker Ratio). We recorded speech and maskers with TMR at -5, 0, 5 dB by adjusting the gain of the masker loud-speaker. We recordeda total of 138 sentences in each condition of the TMR. The samples for the control condition is recorded by emitting sound from only target loudspeaker.

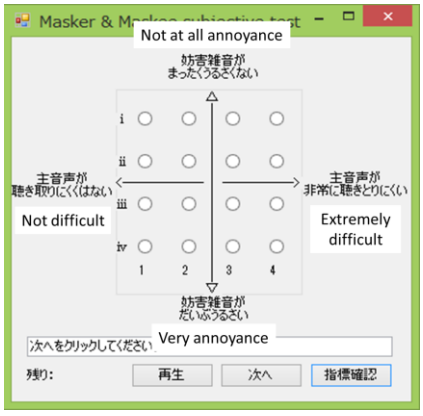


Fig. 2 GUI screenshot.

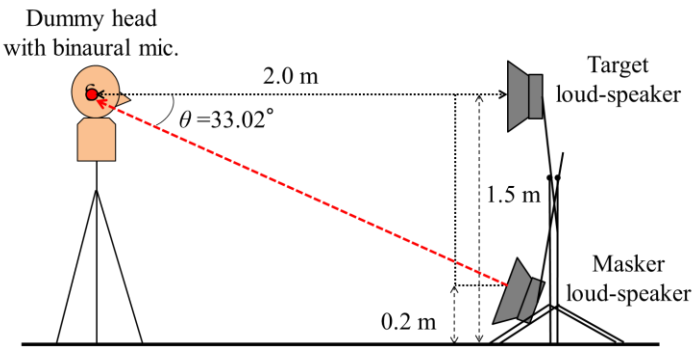


Fig. 3 Recording condition.

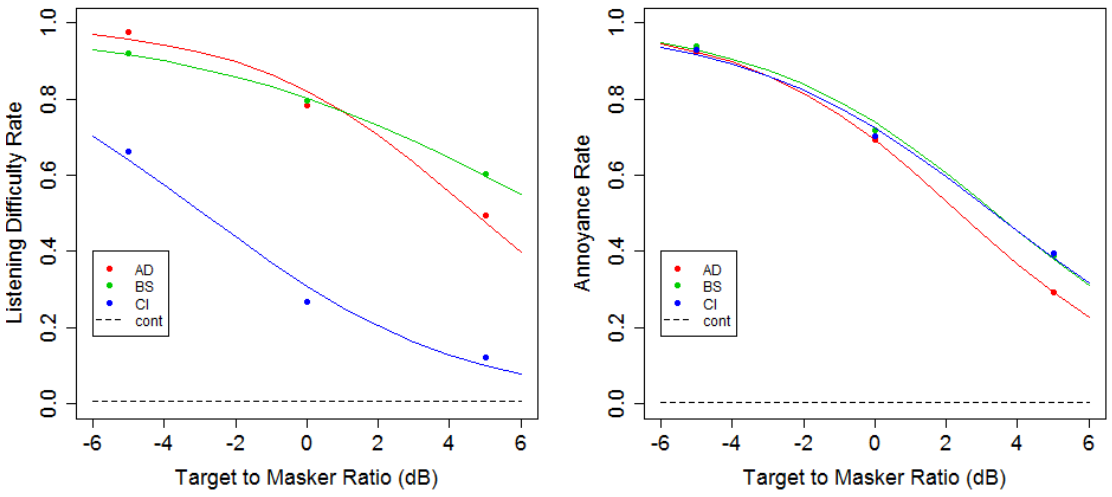
4.4 Experimental Conditions

The subjects were eleven men and women, age was 28 years old from 17 years old, and no abnormality in hearing was reported from the subjects. We trained all subjects for the target speech characteristic before experiments using non-masked target speech. Recorded speech using settings described in section 4.3 is presented randomly using headphones (Sennheiser, HD-25II) from the computer that connects the audio interface (Roland, UA-25EX).

4.5 RESULTS AND DISCUTTIONS

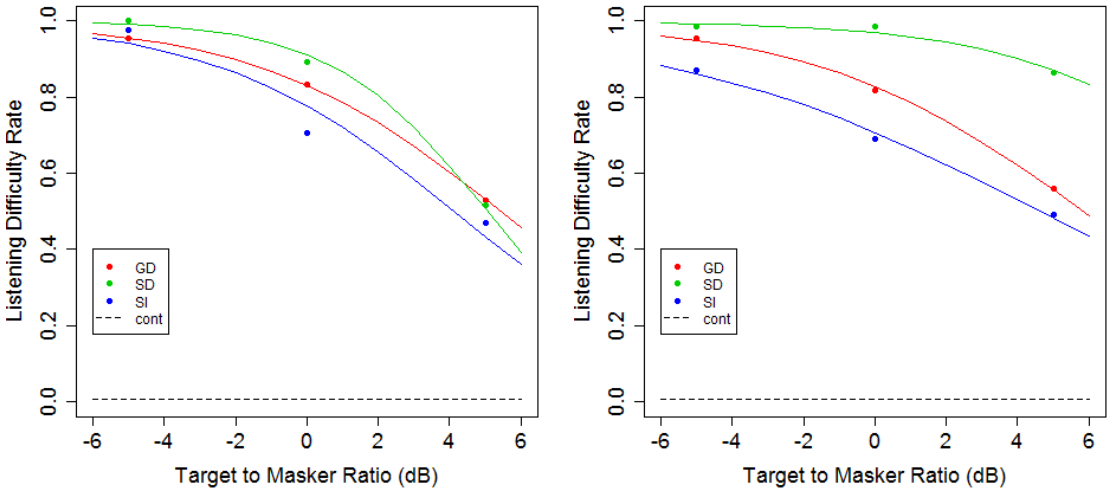
Fig.5 show relationship between the LD ratio and the AN ratio to the TMR by masker type. Plots in the figure are average of all conditions for each of Classic maskers (CL), AD-HSLN masker (AD) and BS-HSLN masker the (BS), respectively. Regression lines are sigmoid functions which were determined by the general linear model (GLM). The dotted line shows the results of the control condition, of which AN ratio and LD ratio are also virtually zero. The LD ratio is high for AD and BS HSLN masker. The LD score is virtually same for the HSLN masker at TMR 5 dB, and CL at TMR -5 dB, as can be seen in Fig. 5 (a). This means that the HSLN maskers can mask at 10 dB smaller levels than the CL maskers for the same LD score. On the other hand, there is no difference between the three systems in the AN ratio, as can be seen from Fig. 5 (b).

Fig.6 and 7 shows the relationship between the LD ratio and the AN ratio vs. the TMR by the dependency type of the HSLN maskers. The LD scores are SD > GD > SI, similar to the observations in previous research (3). The masking effect shown with SD conditions with BS-HSLN is significantly higher than others, especially at TMR of 5 dB. In the AD-HSLN maskers, there is no difference by the speaker dependency of the maskers. However, in the BS-HSLN masker, there is a significant difference between SD and others. The SD BS-HSLN masker shows the highest annoyance scores, but also highest listening difficulty from the above results. In other words, this masker is very effective as a masker.



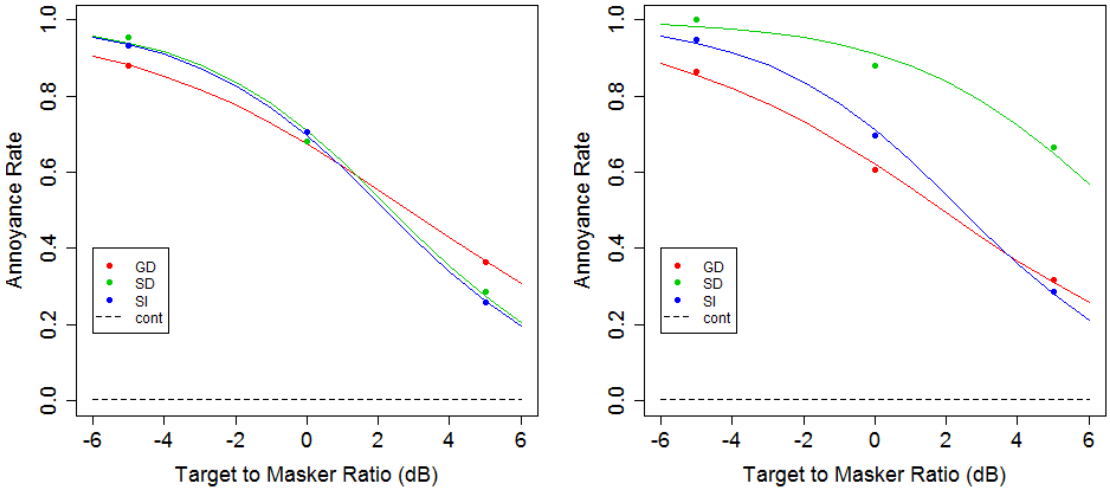
(a) LD rate (b) AN rate

Fig. 5 Relationship between psychological attribute rates to TMR by generation methods.



(a) AD-HSLN masker (b) BS-HSLN masker

Fig. 6 Relationship between LD rate rates to TMR by dependent type.



(a) AD-HSLN masker (b) BS-HSLN masker

Fig. 5 Relationship between AD rate rates to TMR by dependent type.

5. CONCLUSIONS

We proposed a masker synthesis method that uses bootstrapped speech samples for speech masking system. The proposed method shows significantly higher "Listening difficulty" compared to the classical maskers and addition type HSLN masker, and about the same "annoyance" at the same target-to-masker ratio level. Moreover, under the SD conditions, in which the same the target speaker and seed speaker is used, higher listening difficulty and annoyance was shown compared to the GD and SI maskers. Although not compared in this paper, BS-HSLN maskers can also be synthesized using much less speech samples compared to the AD-HSLN masker. For this reason, with the BS-HSLN masker synthesis method, it is much easier to utilize the target speaker's own speech considering real-time operation. We plan to do a thorough comparison of seed speech length required for both the AD-HSLN and the BS-HSLN masker synthesis the future.

REFERENCES

1. Kobayashi D, Kajita S, Takeda K and Itakura F, EXTRACTING SPEECH FEATURES FROM HUMAN SPEECH LIKE NOISE. Proc. International Conference on Spoken Language Processing 1996; 3-6 Oct. 1996; Philadelphia, PA 1996; p.418-421.
2. Fujiwara M, Shimizu Y, Hata M, Lee H, Ueno K, and Sakamoto S. Experimental study for speech privacy with a sound masking system in medical examination room, Proc. Inter-noise 2009; Ottawa, Canada 2009; p.
3. Kondo K, Komiyama T and Kashiwada S. Towards Gender-Dependent Babble Maskers for Speech Privacy Protection. Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing 2013; 16-18 Oct. 2013; Beijing, China 2013; p
4. Akagi M. and Irie Y. Privacy protection for speech based on concepts of auditory scene analysis, Proc. Inter-noise 2012; 19-22 Aug. 2012; New York, NY 2012; 485.
5. Morimoto M, Sato H. and Kobayashi M, Listening difficulty as a subjective measure for evaluation of speech transmission performance in public spaces J. Acoust. Soc. Am. 2004; 116(3):1607-1613.
6. Masden K. and Yano T. Evaluating the equivalence of verbal scales and question stems to be used in English and Japanese noise annoyance questions. J. Sound and Vibration 2004; 277:589–601.
7. Don J and Shami P. N. The Signal Processing Information Base IEEE Signal Processing Magazine 1993; 10(4): 36-43.
8. Yonemoto K, Tateishi T, Koba K. and Kurauchi N. Hearing aid evaluation CD (TY-89) and 57S list from monosyllabic articulation test as well as sound level Audiology Japan 1989; 32(5):429-430. In Japanese.
9. Cavanaugh W.J, Farrell W.R. and Hirtle P.W. Speech privacy in buildings. J. Acoust. Soc. Am. 1962; 34:475-492.