Active Speech Obscuration with Speaker-dependent Human Speech-like Noise for Speech Privacy

Yoshitaka Ohshio*, Haruka Adachi†, Kenta Iwai†, Takanobu Nishiura†, and Yoichi Yamashita†
* Graduate School of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan.
E-mail: is0057iv@ed.ritsumei.ac.jp, Tel: +81-77-561-5075
† College of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan.
E-mail: {is0244fi@ed, iwai18sp@fc, nishiura@is, yyama@is}.ritsumei.ac.jp, Tel: +81-77-561-5075

***Abstract*—This paper introduces a new active speech obscuration with speaker-dependent human speech-like noise (HSLN) for speech privacy. Recently, speech privacy is regarded as an important issue in open public spaces such as hospitals, pharmacies, banks, and so on. To protect speech privacy, speech obscuration methods utilizing HSLN have been studied. HSLNs are designed by superposing various speech signals and speech obscuration is achieved by hearing the target speech and HSLN at the same time. Conventionally, HSLN is designed with the pitch of the target speech as the sole speaker-dependent characteristic. However, additional speaker-dependent characteristics are required because the performance of speech obscuration is still insufficient. Therefore, we propose a speaker-dependent HSLN design method for effective speech obscuration that uses the third formant frequency of the target speech in addition to pitch as speaker-dependent characteristics. The third formant frequency is related to voice quality, which depends on the shape and length of the vocal tract. It follows that the proposed method can effectively mask the target speech by the HSLN considering the pitch and third formant frequency, which are analyzed from the speech. Experimental results demonstrate the effectiveness of the proposed method.**

I. Introduction

In recent years, private information leakage has become an increasingly serious social problem [1]. It is particularly problematic in open public spaces such as hospitals, pharmacies, banks, and so on because other people might overhear private information. Sound absorption and sound insulation walls have been used to protect against speech leakage, but these approaches are often not practical because they require putting thick walls without gaps between the speaker and listeners.

To solve this problem, human speech-like noise (HSLN) [2], [3] for speech obscuration has been attracting attention. HSLNs are designed by superposing various speech signals and are used as a masker signal for informational masking [4], [5]. Informational masking is defined as the degradation of the speech intelligibility of a maskee embedded in a context with the masker. Informational masking often occurs in cases where the characteristics of the HSLNs are similar to those of a speech signal with speaker-dependent characteristics. The optimal condition for masking a speech signal is when the HSLN is designed by a target speech signal. However, it may be uncomfortable for the target speaker to design an HSLN with his or her own voices. Therefore, HSLNs are generally designed by the speech signals of other people. In the conventional method [3], the HSLN is designed considering the speaker's pitch and using the speech signals of other people to effectively mask the target speech. However, the performance of the conventional method is insufficient because the conventional HSLN is designed considering only the pitch as a speaker-dependent characteristic.

In this paper, we propose a speaker-dependent HSLN design method for effective speech obscuration that considers the speaker's third formant frequency [6] in addition to the pitch. The third formant frequency is related to voice quality, which depends on the shape and length of the vocal tract. The proposed method analyzes the pitch and third formant frequency of the target speech as speaker-dependent characteristics. Then, an HSLN with characteristics similar to those of the target speech is selected from an HSLN database. This database is designed in advance by classifying the speech signals of other people for each characteristic. In this way, the proposed method can effectively mask the target speech by using the HSLN. Evaluation experiments were performed to determine the effectiveness of the proposed method.

II. Related research for speech obscuration

Auditory masking is the phenomenon where listeners find it difficult to perceive an affected sound due to the presence of other sounds [4]. The affected sound is called "maskee" and the other sounds are called "masker". Informational masking, which is one type of auditory masking, is defined as the degradation of the speech intelligibility of the maskee embedded in a context with the masker [5]. Informational masking often occurs in cases where the speaker-dependent characteristics of the maskee and masker, e.g., the speech contents, pitch, and voice quality, are similar to each other [6]. The optimal condition for masking a speech signal is that an individual's own voice is used as the masker.

There have been various studies on the designs of human speech-like noise (HSLN) [2], [3], which is the masker for informational masking, for use with speech obscuration. Figure 1 shows an example of speech obscuration using HSLN.

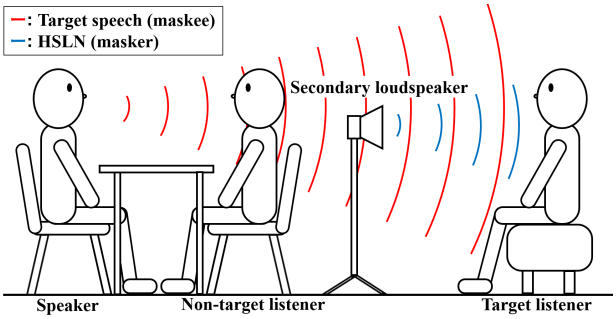

Fig. 1. Speech obscuration by HSLN.

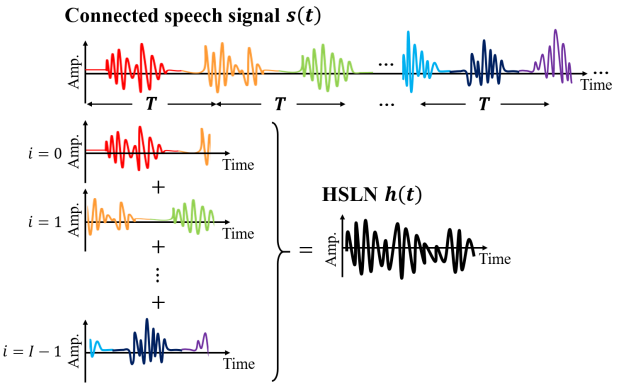

Fig. 2. Design of HSLN.

As shown, the HSLN is emitted to the target listener from a secondary loudspeaker and speech obscuration is achieved by hearing the target speech (maskee) and HSLN (masker) at the same time. On the other hand, the HSLN might not interfere with conversation for the non-target listener since this is emitted to only the target listener. Figure 2 shows the design of an HSLN. The HSLN $h(t)$ is designed by superposing various speeches as

$$h(t) \quad = \quad \sum_{i=0}^{I-1} s(iT+t), \tag{1}$$

where $t(0 \leq t < T)$ is the time index, T is the length of the HSLN signal, $s(t)$ is the connected speech signal, and $i(0 \leq i < I)$ is the index of the superposing numbers. To enable more informational masking, the conventional method [3] designs the HSLN considering the speaker's pitch by superposing the speech signals of other people. However, it is difficult for the conventional HSLN to mask the target speech signal because it considers only the pitch. It is necessary to consider other speaker-dependent characteristics for more effective speech obscuration.

III. Proposed Speech Obscuration Method with Speaker-Dependent Human Speech-Like Noise

In this paper, we propose a speech obscuration method with speaker-dependent HSLN for speech privacy. The proposed HSLN is designed considering the speaker's third formant frequency [6] in addition to the pitch, which is the sole speaker-dependent characteristic used in the conventional method [3]. The third formant frequency is related to voice quality, which depends on the shape and length of the vocal tract. Hence, the proposed method can effectively mask the target speech by using HSLN with a consideration of additional speaker-dependent characteristics of the speech signal. We assume that the target speech is a single worker in an open public space and that the condition of the speaker does not greatly change (i.e., the speaker-dependent characteristics barely change.)

Figure 3 gives an overview of the proposed method. To analyze the target speech $x(t)$, which is captured from the microphone, the frame-divided target speech $\tilde{x}(m;n)$ is obtained as

$$\tilde{x}(m;n) \quad = \quad x(mN+n), \tag{2}$$

where m is the frame index, $n(0 \leq n < 2N)$ is the time index in the frame, and N is the shift length. The proposed method designs the HSLN $\tilde{y}(m;n)$ corresponding to the target speech signal $\tilde{x}(m;n)$ frame by frame. In the first step, the pitch $f_{\mathrm{o}}^{[\tilde{x}]}(m)$ and third formant frequency $f_{3}^{[\tilde{x}]}(m)$ are computed by cepstral analysis and LPC analysis of the target speech $\tilde{x}(m;n)$, respectively. Then, the proposed frame-divided HSLN $\tilde{y}(m;n)$ is selected using $f_{\mathrm{o}}^{[\tilde{x}]}(m)$ and $f_{3}^{[\tilde{x}]}(m)$ from the HSLN database. This database, which consists of the speaker-dependent HSLN $\tilde{h}_{\alpha,\beta}(m;n)$, is designed in advance by classifying the frame-divided speech signals of other people $\tilde{s}_j(m;n)$ into separate classes, where j is the index of frame-divided speech signal and α and β are the indexes of the pitch and third formant frequency, respectively. The proposed HSLN $y(t)$, which is designed by overlap-adding $\tilde{y}(m;n)$, is emitted from the secondary loudspeaker to the target listener. In the proposed method, speech obscuration is achieved by hearing $x(t)$ and $y(t)$ simultaneously, the same as the conventional method [3].

The details of the design and the selection of the speaker-dependent HSLN are discussed below.

A. Design of database for speaker-dependent HSLN

First, the pitch $f_{\mathrm{o}}^{[\tilde{s}_j]}(m)$ and third formant frequency $f_{3}^{[\tilde{s}_j]}(m)$ are computed in advance by cepstral analysis and LPC analysis of the frame-divided speech signal of other people $\tilde{s}_j(m;n)$, respectively. Then, the frame-divided speech signal $\tilde{s}_j^{[\alpha,\beta]}(m;n)$ is classified into $C_{\alpha,\beta}$ according to

$$L_\alpha \leq f_{\mathrm{o}}^{[\tilde{s}_j]}(m) < H_\alpha, \tag{3}$$

$$L_\beta \leq f_{3}^{[\tilde{s}_j]}(m) < H_\beta, \tag{4}$$

where L and H are the lower and upper limit frequencies of each class $C_{\alpha,\beta}$, respectively. Then, the speaker-dependent HSLN $\tilde{h}_{\alpha,\beta}(m;n)$ is designed as

$$\tilde{h}_{\alpha,\beta}(m;n) \quad = \quad \sum_{j=0}^{J_{\alpha,\beta}-1} \tilde{s}_j^{[\alpha,\beta]}(m;n), \tag{5}$$

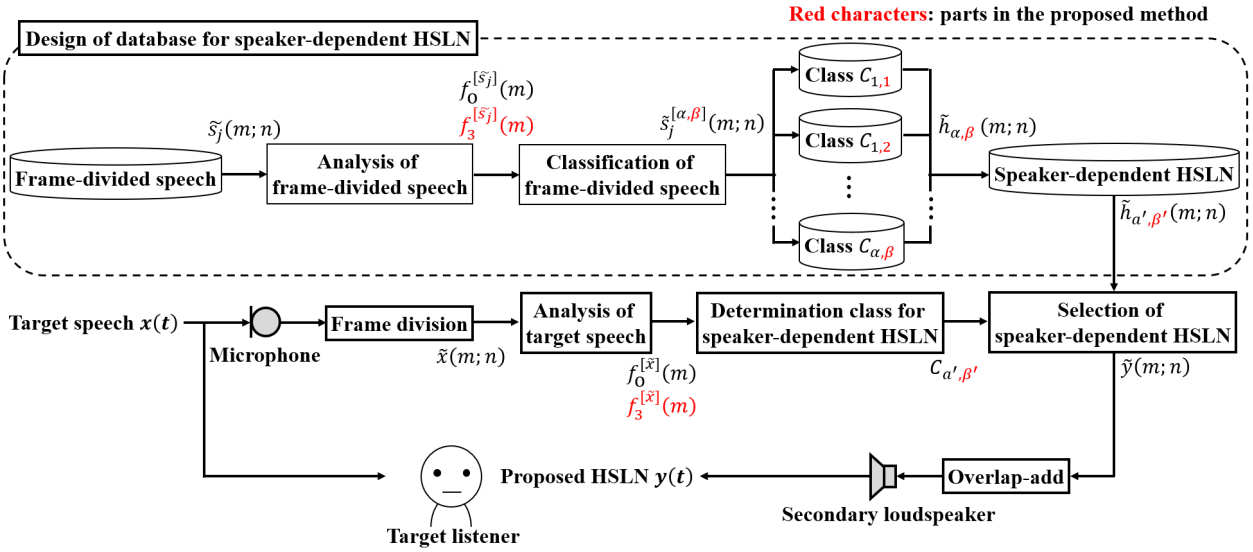

Fig. 3. Overview of the proposed method.

where $J_{\alpha,\beta}$ is the superposing numbers of the frame-divided speech signal $\tilde{x}(m;n)$ in the class $C_{\alpha,\beta}$. A database for speaker-dependent HSLN that consists of the speaker-dependent HSLN $\tilde{h}_{\alpha,\beta}(m;n)$ is designed in advance.

B. Selection of speaker-dependent HSLN

The proposed method determines the class $C_{\alpha',\beta'}$ from the speaker's pitch $f_0^{[\tilde{x}]}(m)$ and third formant frequency $f_3^{[\tilde{x}]}(m)$ according to

$$L_{\alpha'} \leq f_0^{[\tilde{x}]}(m) < H_{\alpha'}, \quad (6)$$

$$L_{\beta'} \leq f_3^{[\tilde{x}]}(m) < H_{\beta'}, \quad (7)$$

where α' and β' are the indexes of the class for the pitch and third formant frequency, respectively. Then, the proposed frame-divided HSLN $\tilde{y}(m;n)$ is designed by selecting the speaker-dependent HSLN $\tilde{h}_{\alpha,\beta}(m;n)$ in class $C_{\alpha',\beta'}$ as

$$\tilde{y}(m;n) = \tilde{h}_{\alpha',\beta'}(m;n). \quad (8)$$

The computational cost for the proposed method is $O(n) = n\log n$ per frame and the proposed method realizes real-time processing.

IV. Evaluation Experiment

We performed objective and subjective evaluation experiments to determine the effectiveness of the proposed method.

A. Experimental conditions

In the objective experiment, we evaluate the perceptual evaluation of speech quality (PESQ) [7] of the evaluation sounds. The PESQ represents the evaluation index of speech quality. In the subjective experiment, participants answer the utterance content of the evaluation sounds and we evaluate the word intelligibility rates of those sounds. Evaluation sounds consist of the following:

- w/o f_0, f_3 [2]: The conventional HSLN without the speaker-dependent characteristics
- w/ f_3: The HSLN with the third formant frequency
- w/ f_0 [3]: The conventional HSLN with the pitch
- The proposed method
- w/o HSLN

Table I lists the classes for each HSLN design. Evaluation sounds were emitted two times from a secondary loudspeaker (FOSTEX, FE83Fn) driven by a power amplifier (BOSE, 1705H). Participants were two females and five males in their twenties. The HSLN was designed using the sound sources from the database [8]. We utilized ten words from the database [9] as the target speech. The experimental environment was an ordinary office (L_A =45.8 dB). The sampling rate, quantization, and shift length were set to 16 kHz, 16 bits, and 960 points, respectively. The gain ratio G between the target speech and HSLN was set to -3 dB calculated as

$$G = 20\log_{10}\frac{\sum_{n=0}^{2N-1}|\tilde{x}(m;n)|}{\sum_{n=0}^{2N-1}|\tilde{h}(m;n)|}. \quad (9)$$

B. Experimental results

Figure 4 shows the results of the objective evaluation experiment for the PESQ. Horizontal axis represents each condition for speech obscuration and vertical axis represents the average of the PESQ. Lower values mean higher performance in speech obscuration. As shown in the figure, the proposed method has the lowest PESQ compared with the other methods. When only f_3 or f_0 is used to design the HSLN, the PESQ is almost the same as that for the conventional method [2]. This demonstrates that f_3 and f_0 should be used to design the HSLN for effective speech obscuration.

Figure 5 shows the results of the subjective evaluation experiment for word intelligibility rates. Horizontal axis represents

TABLE I
CLASS FOR EACH HSLN DESIGN

		Third formant frequency f_3 Hz		
		$f_3 < 2700$	$2700 \leq f_3 < 3100$	$3100 \leq f_3$
Pitch f_o Hz	$f_o < 125$	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$
	$125 \leq f_o < 160$	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$
	$160 \leq f_o < 200$	$C_{3,1}$	$C_{3,2}$	$C_{3,3}$
	$200 \leq f_o < 250$	$C_{4,1}$	$C_{4,2}$	$C_{4,3}$
	$250 \leq f_o < 315$	$C_{5,1}$	$C_{5,2}$	$C_{5,3}$
	$315 \leq f_o$	$C_{6,1}$	$C_{6,2}$	$C_{6,3}$

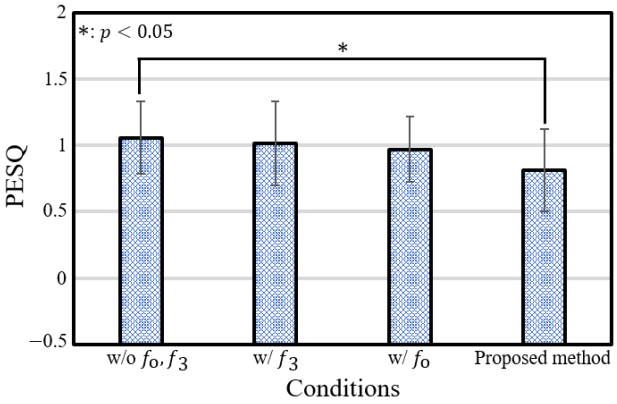

Fig. 4. Objective evaluation results for PESQ.

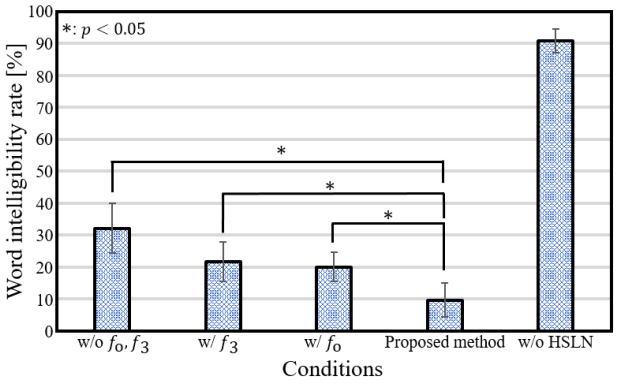

Fig. 5. Subjective evaluation results for word intelligibility.

each condition for speech obscuration and vertical axis represents the average word intelligibility rates. Lower values mean higher performance in speech obscuration. As shown in the figure, each method using HSLNs has a higher performance than those without HSLN. In particular, the proposed method has the highest performance and its differences are significant ($p < 0.05$). In addition, the proposed method unquestionably achieves speech obscuration because its word intelligibility rate is lower than 20 % [8]. This demonstrates the necessity of HSLN design that considers both pitch and third formant frequency for speech obscuration.

The results of these experiments demonstrate the effectiveness of the proposed HSLN design method. On the other hand, it is important to accurately estimate both the pitch f_o and the third formant frequency f_3 for high performance of speech obscuration in these methods. Thus, in order to apply the methods to real environments, it is necessary to independently estimate the pitch and the third formant frequency for multiple speakers.

V. CONCLUSION

In this work, we proposed a speaker-dependent HSLN design method for speech privacy in open public spaces. As speaker-dependent characteristics, the proposed method considers the third formant frequency, which is related to voice quality, in addition to the pitch. Experimental results demonstrated the effectiveness of the proposed method. In future work, we will expand the proposed method for use with multiple speakers.

ACKNOWLEDGMENTS

This work was partly supported by JST COI, JST SCORE and JSPS KAKENHI Grant Number JP18K19829.

REFERENCES

[1] W.J. Cavanauch and W.R. Farrel, “Speech privacy in buildings,” *J. Acoust. Soc. Am,* 34(4), pp. 475-492, 1962.
[2] D. Kobayashi, S. Kajita, and K. Takeda, “Extracting speech features from human speech like noise,” *Proc. ICSLP 96,* pp. 418-421, 1996.
[3] A. Ito, A. Miki,Y.Shimizu, K. Ueno, H.J. Lee, and S. Sakamoto, “Oral information masking considering room environmental condition, Part 1: Synthesis of maskers and examination on their masking efficiency,” *Proc. InterNoise2007,* pp. 419-428, 2007.
[4] R.L. Wegel and C.E. Lane, “The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear,” *Phys. Rev.,* 23, pp. 266-285, 1924.
[5] C.S. Watson, W.J. Kelly, and H.W. Wroton, “Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty,” *J. Acoust. Soc. Am.,* 60(5), pp. 1176-1186, 1976
[6] G. Kidd, C.R. Mason, V.M. Richards, F.J. Gallun, and N.I. Durlach, “Informational masking,” in *Springer Handbook of Auditory Perception of Sound Sources,* edited by W.A. Yost, 29, pp. 143-190, 2008.
[7] “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” *ITU-T Recommendation,* P. 862, 2001.
[8] K. Ueno, H. Lee, S. Sakamoto, A. Ito, M. Fujiwara, and Y. Shimizu, “Experimental study on applicability of sound masking system in medical examination room,” *Proc. Acoustics’08,* pp. 1331-1336, 2008.
[9] S. Amano, S. Sakamoto, T. Kondo, and Y. Suzuki, “Development of familiarity-controlled word lists 2003 (FW03) to assess spoken-word intelligibility in Japanese,” *Speech Commun.,* 51, pp. 76-82, 2009.