An Evaluation of Target Speech for a Nonaudible Murmur Enhancement System in Noisy Environments

Sakura Tsuruta, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
E-mail: [tsuruta.sakura.tl0, ko-t, tomoki, neubig, ssakti, s-nakamura]@is.naist.jp Tel: +81-743-72-5265

***Abstract*—Nonaudible murmur (NAM) is a soft whispered voice recorded with NAM microphone through body conduction. NAM allows for silent speech communication as it makes it possible for the speaker to convey their message in a nonaudible voice. However, its intelligibility and naturalness are significantly degraded compared to those of natural speech owing to acoustic changes caused by body conduction. To address this issue, statistical voice conversion (VC) methods from NAM to normal speech (NAM-to-Speech) and to a whispered voice (NAM-to-Whisper) have been proposed. It has been reported that these NAM enhancement methods significantly improve speech quality and intelligibility of NAM, and NAM-to-Whisper is more effective than NAM-to-Speech. However, it is still not obvious which method is more effective if a listener listens to the enhanced speech in noisy environments, a situation that often happens in silent speech communication. In this paper, assuming a typical situation in which NAM is uttered by a speaker in a quiet environment and conveyed to a listener in noisy environments, we investigate what kinds of target speech are more effective for NAM enhancement. We also propose NAM enhancement methods for converting NAM to other types of target voiced speech. Experiments show that the conversion process into voiced speech is more effective than that into unvoiced speech for generating more intelligible speech in noisy environments.**

I. INTRODUCTION

Speech communication is one of the most fundamental communication methods in our daily life, and the advancement of technology such as cell phones has enabled us to communicate with each other whenever and wherever. However, it has also created the awareness of existence of some situations where we hesitate to talk; e.g., we would have trouble maintaining privacy in a crowd, or speaking itself can annoy others in quiet environments such as a library or an office.

In recent years, silent speech interfaces have attracted attention as a new speech communication style [1] that allows us to talk while keeping silent. One of the promising silent speech media is nonaudible murmur (NAM) [2]. NAM is a soft whispered voice recorded with NAM microphone through body conduction. It is produced by articulating respiratory sounds without using vocal-fold vibration. One of the biggest drawbacks of NAM is that its acoustic properties are very different from those of natural voices because of the essential properties of body-conductive recording, i.e., lack of radiation characteristics from the lips and effect of the low-pass characteristics of soft tissues. Consequently, its naturalness and intelligibility are significantly degraded.

To address this issue, statistical NAM enhancement methods have been proposed [3]. In these methods, acoustic parameters of NAM are converted into those of air conducted normal speech (NAM-to-Speech) or a whispered voice (NAM-to-Whisper) by using statistical voice conversion (VC) techniques [4][5]. It has been reported in [3] that NAM-to-Whisper outperforms NAM-to-Speech because 1) NAM-to-Whisper can avoid the F_0 and unvoiced/voiced information prediction from NAM, which is an essentially difficult prediction process, and 2) spectral conversion accuracy in NAM-to-Whisper is higher than that in NAM-to-Speech because spectral parameters of NAM are more similar to those of a whispered voice than those of normal speech. However, because they were compared to each other through listening tests under only a quiet environment, it is still not obvious which method is more effective if a listener listens to the enhanced speech in noisy environments, a situation that often happens in real silent speech communication.

In this paper, we investigate what kinds of target speech are more effective for enhancing intelligibility of NAM. Silent speech communication is often used to support speech telecommunication between persons in very different environments. NAM will be used by a person in a relatively quiet environment to speak to the other while keeping silent. On the other hand, the other person is not always in such a special environment. In this paper, as one of the typical situations in silent speech communication, we assume the situation where NAM uttered by a speaker in a quiet environment needs to be conveyed to a listener in noisy environments.

II. NAM ENHANCEMENT METHODS BASED ON STATISTICAL VOICE CONVERSION

In statistical NAM enhancement methods, acoustic features of NAM uttered by a speaker are converted into those of more intelligible and natural voices, such as normal speech or a whispered voice, uttered by the same speaker. These methods consist of training and conversion processes. In the training process, we train a Gaussian mixture model (GMM) to model the joint probability density of the source and target acoustic features [6] using a corresponding joint feature vector set generated by performing automatic time frame alignment for a parallel data set consisting of NAM and target speech. In the conversion process, the source acoustic features are converted into the target acoustic features based on maximum likelihood estimation of speech parameter trajectories considering global variance (GV) [5].

978-616-361-823-8

A. Conversion from NAM to Normal Speech (NAM-to-Speech)

Spectral segment features [7] of NAM are used as the source features. On the other hand, spectral and excitation features of normal speech are used as the target features, where F_0 including unvoiced/voiced (U/V) information and aperiodic components [8] are used as the excitation features. We separately train three GMMs for converting the spectral segment features of NAM to the individual target features of normal speech, i.e., the spectral, F_0, and aperiodic features. In the conversion process, the spectral segment features of NAM are converted into each of the target features of normal speech using the corresponding GMM. After a mixed excitation signal [9] is designed according to the converted F_0 and aperiodic features, the converted speech is generated by filtering the mixed excitation signal with the converted spectral features.

B. Conversion from NAM to Whisper (NAM-to-Whisper)

The whispered voice is totally unvoiced speech. It sounds much more natural and intelligible compared to NAM. Therefore, naturalness and intelligibility of NAM are significantly improved by NAM-to-Whisper conversion. As the source features, the spectral segment features of NAM are used as in NAM-to-Speech. On the other hand, as the target feature, only the spectral features of the whispered voice are used because white Gaussian noise can be always used as an excitation signal in synthesizing a whispered voice. Therefore, we need to train a single GMM for converting the spectral segment features of NAM into the spectral features of the whispered voice. In the conversion process, we perform the spectral conversion in the same manner as in NAM-to-Speech.

III. INVESTIGATION OF TARGET SPEECH TO IMPROVE INTELLIGIBILITY IN NOISY ENVIRONMENTS

The improvements of intelligibility of NAM through NAM enhancement systems is the most important factor to make it possible to use NAM in silent speech communication. Moreover, because NAM is a special speaking style, environmental situations where a speaker uses NAM will be relatively limited, e.g., the speaker utters in NAM to avoid annoying others in quiet environments such as in a library. On the other hand, the environmental situation of the listener is not limited at all. It is highly possible that the listener calls a speaker in noisy environments. Therefore, it is worthwhile to investigate intelligibility of the enhanced speech in various environments.

In this paper, we investigate what kinds of target speech yield larger improvements in intelligibility of NAM in some noisy environments. Although NAM-to-Whisper outperforms NAM-to-Speech in a very quiet environment such as in soundproof room as reported in [3], it is possible that NAM-to-Speech conversely outperforms NAM-to-Whisper in noisy environments because F_0 patterns of voiced speech could be useful as an acoustic cue to perceive speech sounds. We also propose voicing of a whispered voice and electrolaryngeal (EL) speech [10] as new forms of target voiced speech.

A. Voicing of Whispered Voice

As reported in [3], spectral conversion accuracy of NAM-to-Whisper is higher than that of NAM-to-Speech. Therefore, we propose another NAM enhancement method to convert NAM to a new target voice generated by the spectral feature of the whispered voice and the excitation features of the normal speech as shown in Fig. 1. The spectral features are determined with the GMM for NAM-to-Whisper. On the other hand, the excitation features are predicted with GMMs in NAM-to-Speech. We also implement continuous F_0 patterns (CF_0) [10] as the F_0 features because it yields significant improvements in prediction accuracy of F_0 patterns. The enhanced speech is denoted as "converted whisper+CF_0."

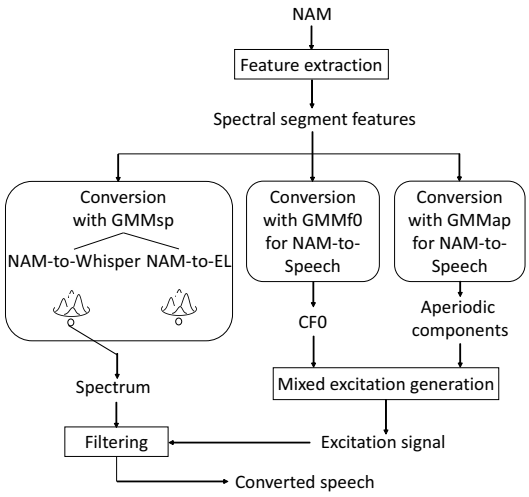

Fig. 1. Conversion process of proposed methods

B. Electrolaryngeal Speech

Although "converted whiper+CF_0" is totally voiced speech, a combination of the spectral feature of unvoiced speech (i.e., the whispered voice) and the excitation features of voiced speech (i.e., the normal speech) causes a mismatch of unvoiced/voiced information between the spectral and excitation features. This mismatch might cause adverse effects on intelligibility because we never use a voice suffering from such mismatch in actual speech communication.

To avoid this mismatch, we also propose a conversion method from NAM to another new target voice based on EL speech [10]. The EL speech is produced by one of the alternative speaking methods for laryngectomees who have removed their vocal codes. It is produced by articulating voiced excitation signals artificially generated from an electrolarynx. Consequently, EL speech is totally voiced speech. Moreover, it has been reported that intelligibility of EL speech is relatively high [11]. The conversion process of this method is also shown in Fig. 1. In the training process, we record EL speech produced by the same speaker as in NAM recording. Then, we train another GMM for converting the spectral segment features of NAM into the spectral features of EL speech. In the conversion process, the spectral features are determined with the trained GMM. On the other hand, the excitation features are determined with the same GMMs as used in generating "converted whisper+CF_0." The enhanced speech is denoted as "converted EL+CF_0."

IV. Experimental Evaluations

A. Experimental Conditions

We recorded normal speech, a whispered voice, and EL speech using an air-conductive microphone, and NAM using a NAM microphone as well as an air-conductive microphone. The speaker was one Japanese male. He uttered 50 sentences included in the ATR phonetically balanced sentence set [12]. He also uttered the most unfamiliar 160 words included in the familiarity controlled word lists 2007 [13] in NAM. All of these words consisted of 4 Japanese mora. The sampling frequency was set to 16 kHz.

In the NAM enhancement methods, the 0th through 24th mel-cepstral coefficients were used as the spectral parameters. STRAIGHT analysis [14] was used for normal speech and EL speech. Mel-cepstral analysis was used for the whispered voice and NAM. As the excitation parameters, F_0 and aperiodic components were extracted from normal speech with STRAIGHT analysis. The shift length was set to 5 ms. Current $\pm$ 4 frames were used for extracting the spectral segment feature of NAM.

We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance pairs were used for evaluation. We train three GMMs for spectral conversion, 1) GMM_SP for converting into the spectral features of the normal speech, 2) GMM_WH for converting into those of the whispered voice, and 3) GMM_EL for converting into those of the EL speech. We also train three GMMs for the excitation prediction, 1) GMM_F_0 for converting into the F_0 features of the normal speech, 2) GMM_CF_0 for converting into the continuous F_0 features of the normal speech, and 3) GMM_AP for converting into the aperiodic features of the normal speech. The numbers of mixture components was set to 64 for the spectral conversion, 32 for the F_0 and CF_0 prediction, and 16 for the aperiodic prediction. Note that the air-conducted NAM simultaneously recorded with (body-conducted) NAM was used to improve accuracy of time frame alignment between NAM and the target voices. Table 1 and Table 2 show conversion accuracy of the trained GMMs evaluated by the cross-validation test.

We conducted an opinion test on listenability using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent) as in [11]. The following 8 kinds of speech samples were evaluated:

- NAM: the original NAM
- WH: the original whisper
- EL: the original EL speech
- SP: the original normal speech
- CVSP: the enhanced speech by NAM-to-Speech using GMM_SP, GMM_F_0, and GMM_AP
- CVWH: the enhanced speech by NAM-to-Whisper using GMM_WH
- CVWH+CF_0: the enhanced speech "converted whisper+CF_0" using GMM_WH, GMM_CF_0, and GMM_AP
- CVEL+CF_0: the enhanced speech "converted EL+CF_0" using GMM_EL, GMM_CF_0, and GMM_AP

These samples were presented to listeners using a headphone in a soundproof room assuming three conditions: 1) a quiet environment without external noise, 2) a noisy environment by adding office noise so that SNR was set to 0 dB, and 3) another noisy environment by adding office noise so that SNR was set to -10 dB. The SNR was set to these values considering that A-weighted sound pressure level of speech presented from cell phones (55 dB(A)), that of the office noise (55 dB(A)), and that of the babble noise (65 dB(A)). The number of listeners was 12. Each listener evaluated 240 samples consisting of 10 samples from each kind of speech and each condition.

We also conducted a dictation test on intelligibility. We evaluated speech samples of NAM, CVSP, CVWH, and CVWH+CF_0 using utterances of the unfamiliar words. We considered two environments, the quiet environment without external noise and the noisy environment by adding 0 dB SNR of the office noise. The number of listeners was 4. Each listener evaluated 160 samples consisting of 20 samples from each kind of speech and environment.

TABLE I
SPECTRAL CONVERSION ACCURACY OF EACH MODEL. MEL-CEPSTRAL DISTORTION IS CALCULATED WITHOUT POWER INFORMATION

	GMM_SP	GMM_WH	GMM_EL
Mel-cepstral distortion	4.2 dB	4.5 dB	5.4 dB

TABLE II
EXCITATION FEATURE PREDICTION ACCURACY OF EACH MODEL

	GMM_F_0	GMM_CF_0	GMM_AP
Aperiodic distortion	N/A	N/A	3.6 dB
U/V error rate	8.9 %	13.5 %	N/A
F_0 correlation coefficient	0.43	0.53	N/A

B. Experimental Results

The result of the opinion test on listenability is shown in Fig. 2. For the original speech, SP has the highest score and NAM has the lowest score in every environment. On the other hand, the score of WH is higher than that of EL in the quiet environment, they become almost equal as the external noise level increases. This result shows that listenability of unvoiced speech (i.e., the whispered voice) is more sensitive to the external noise level than that of the voiced speech.

For the converted speech, we can see that most of the NAM enhancement methods yields significant improvements in listenability in every environment. The score of CVSP is almost equal to or better than that of CVWH. Moreover, in a comparison between CVWH and CVWH+CF_0, we can see a tendency that the score of CVWH is higher than that of CVWH+CF_0 in the quiet environment but they are reversed as the noise level increases. Namely, the voiced speech is more robust against external noise in terms of listenability. It is interesting that the enhanced speech CVWH+CF_0 outperforms the original whispered voice WH in the noisy environment with -10 dB SNR. On the other hand, the score of CVEL+CF_0 is much lower than the other types of enhanced speech. This is caused by poor spectral conversion accuracy as shown in Table 1.

Table 3 shows the result of the dictation test on intelligibility. The score shows mora correct rate. It can be observed that

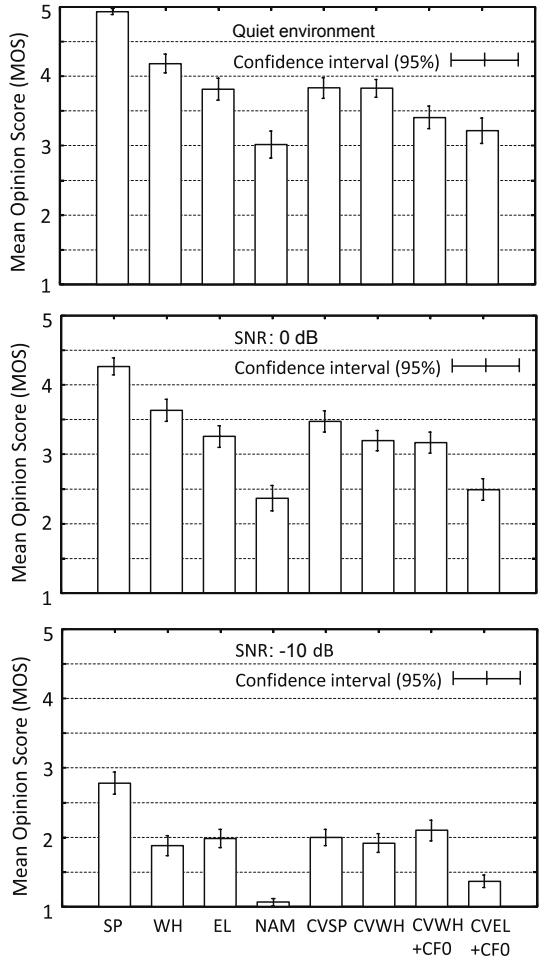

Fig. 2. Result of opinion test on listenability (top figure shows quiet environment, middle one shows SNR = 0 dB, bottom one shows SNR = -10 dB)

NAM is less intelligible but the NAM enhancement methods are capable of effectively improving its intelligibility. The intelligibility of CVWH is greatly degraded in the noisy environment. On the other hand, degradation of the intelligibility of CVSP and CVWH+CF_0 caused by the external noise is much smaller than that of CVWH. This tendency is the same as observed in the previous test on listenability; i.e., voiced speech is more robust to external noise than unvoiced speech. Moreover, we can also see that CVSP is again better than CVWH, in cotrast to the results reported in [3]. It is possible that this inconsistency is caused by speaker differences or differences of utterances to be evaluated. Investigating this more thoroughly is future work.

From these results, it is demonstrated that the NAM enhancement methods converting NAM to voiced speech such as normal speech or voicing of the whispered voice are effective for generating intelligible enhanced speech in noisy conditions.

TABLE III
RESULT OF DICTATION TEST ON INTELLIGIBILITY

	Mora correct rate (Quiet environment)	Mora correct rate (SNR = 0 dB)
NAM	55.0%	42.5%
CVSP	67.8%	63.1%
CVWH	61.6%	55.0%
CVWH+CF_0	65.3%	62.5%

V. CONCLUSIONS

In this paper, we have investigated what kinds of target speech are more effective for generating intelligible speech by nonaudible murmur (NAM) enhancement assuming one of the typical situations of silent speech communication, where NAM is uttered by a speaker in a quiet environment and the enhanced speech is conveyed to a listener in noisy environments. The experimental result has demonstrated that the conversion process from NAM into voiced speech such as normal speech or voicing of a whispered voice is more robust to external noise than that into unvoiced speech such as the whispered voice. We plan to investigate more about target speech for making the enhanced speech more intelligible in NAM enhancement. We will also implement several techniques to improve intelligibility of normal speech in noisy environments for NAM enhancement.

ACKNOWLEDGMENT

Part of this work was supported by JSPS KAKENHI Grant Numbers: 26280060 and 23240023.

REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun*, vol. 52, no. 4, pp. 270-287, 2010.

[2] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano, "Non-audible murmur(NAM) recognition," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 1, pp. 1-8, 2006.

[3] T. Toda, M. Nakagiri, K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2505-2517, Sep. 2012.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131-142, Mar. 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trafectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.

[6] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," *in Proc. ICASSP*, Seattle, WA, pp. 285-288, May.1998

[7] T. Toda, A.W. Black, K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215-227, 2008.

[8] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. MAVEBA*, Firenze, Italy, pp. 2266-2269, Sep. 2006.

[9] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *in Proc. INTERSPEECH*, Pittsburgh, PA, Sep.2006, pp. 2266-2269.

[10] K. Tanaka, T. Toda, G. Neubig, S. Sakti, S. Nakamura, "A Hybrid Approach to Electrolaryngeal Speech Enhancement Based on Noise Reduction and Statistical Excitation Generation," *IEICE Transactions on Information and Systems*, Vol. E97-D, No. 6, pp. 1429-1437, Jun. 2014.

[11] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, K. Shikano, "Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion.," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 172-183, Jan. 2014.

[12] Y. Sugisaka, K. Takeda, M. Abe, S. katagiri, T. Umeda, and H. Kuwabata, "A large-scale Japanese speech database," in*Proc. ICSLP90*, Kobe, Japan, Nov. 1990, pp.1089-1092.

[13] T. Kondo, S. Amano, S. Sakamoto, and Y. Suzuki, "Spoken Word Intelligibility of Young and Old Adults with Familiarity-Controlled Word Lists 2007 (FW07)," *Proc. ICP 2008*, PS-Tue-pm-157, 2008.

[14] H. Kawahara, H .Katayose, A. de Cheveigne, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity," *in Proc. EUROSPEECH*, Budapest, Hungary, pp. 2781-2784, Sep. 1999.