Evaluation of Speaking-Aid System with Voice Conversion for Laryngectomees Toward Its Use in Practical Environments

Keigo Nakamura[1], *Tomoki Toda*[1], *Yoshitaka Nakajima*[2],
Hiroshi Saruwatari[1] *and Kiyohiro Shikano*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Department of Psychiatry, Osaka University Graduate School of Medicine, Japan
{kei-naka, tomoki, sawatari and shikano}@is.naist.jp, yoshitaka-nakajima@umin.ac.jp

Abstract

In this paper, we evaluate our previously proposed speaking-aid system with voice conversion for laryngectomees. The proposed system employs a sound source unit generating extremely small signals to keep them from annoying other persons, and then it statistically converts articulated signals captured with a body-attached microphone into natural speech. We have so far shown the effectiveness of the proposed system using speech data imitated by a non-laryngectomee, which have recorded in a sound proof room. In this paper, we further investigate 1) whether such small sound source signals cause the lack of auditory feedback under noisy environments and 2) whether the proposed system is effective for real laryngectomees. Experimental results demonstrate that 1) an explicit auditory feedback is useful to keep the speaker's articulation stable and 2) the voice conversion dramatically improves the naturalness of the laryngectomee's speech but it slightly degrades its intelligibility.

Index Terms: Speaking-aid, Laryngectomee, Voice conversion, Auditory feedback, Non-Audible Murmur microphone

1. Introduction

Laryngectomy is a surgical procedure usually conducted for patients with advanced laryngeal cancer or a similar problem. Losing the larynx means that the patients (so-called laryngectomees) lose their original voice, and therefore, they evidently have difficulty communicating with other persons. An electrolarynx is one of the easiest alternative methods for laryngectomees to speak without vocal fold vibration [1]. We focus on two problems in using an electrolarynx, namely 1) the leakage of sound source signals and 2) unnaturalness of the artificial speech resulting from mechanical excitation signals. Conventional electrolarynxes need to generate enough loud sound source signals to make artificial speech sufficiently audible. Therefore, the sound source signals themselves might be noisy for other people around the laryngectomee especially in a quiet environment. Furthermore, electrical imitation of our vocal fold vibration causes awful degradation of the generated speech quality.

We have proposed a speaking-aid system using a voice conversion technique for laryngectomees to provide much natural speech communication [2]. First, a user attaches a sound source unit under the lower jaw and articulates sound source signals. The feature of the sound source signals is that the power is extremely small so that people around the user cannot hear it [3]. To capture the small artificial speech sufficiently, we use Non-Audible Murmur (NAM) microphone [4], which has a large dynamic range. Because artificial speech detected with NAM microphone is still mechanical and unnatural, the captured data are converted into more natural speech using a statistical voice conversion technique [5]. Finally, the converted speech is presented as the user's voice.

In our previous work, the proposed system has been evaluated using speech data imitated by a non-laryngectomee, which was recorded in a sound-proof room [2]. Experimental results have demonstrated that both the naturalness and the intelligibility of the converted speech are much better than those of artificial speech using a conventional electrolarynx. However, it is still doubtful whether this system is practically effective. We wonder whether external noises might mask the auditory feedback generated by small excitation signals. The difficulty of acquiring the auditory feedback would cause a negative spiral of instability of the articulation, the degradation of the converted speech quality and more instability of the articulation. Furthermore, imitational speech cannot completely copy artificial speech uttered by real laryngectomees. Therefore, it is necessary to investigate the effectiveness of the proposed system using real laryngectomees' data.

Our aim in this paper is to investigate 1) effectiveness of explicit auditory feedback and 2) performance of the proposed system for real laryngectomee's data. Although it is ideal for the user to acquire the auditory feedback from the presented converted speech, high quality voice conversion techniques that can work in real time haven't been attained yet. Therefore, we propose a method to provide the detected signals with NAM microphone directly to the speaker as the explicit auditory feedback. Moreover, we record speech utterances of one real laryngectomee and evaluate the performance of the proposed system using them. Experimental evaluations demonstrate that 1) the explicit auditory feedback is effective for making the speaker's articulation stable and 2) the naturalness of converted speech utterances is dramatically improved, but their intelligibility is slightly degraded.

This paper is organized as follows. In section 2, the proposed speaking-aid system is explained. Auditory feedback in our proposed system is described in section 3, and speech data uttered by a laryngectomee is described in section 4. The effectiveness of the explicit auditory feedback and the performance of the proposed system using the laryngectomee's data are experimentally evaluated in section 5. Finally, this paper is concluded in section 6.

2. Speaking-Aid System with Voice Conversion for Laryngectomees

2.1. Structure of Our Proposed System

Figure 1 shows our proposed speaking-aid system with a voice conversion technique for laryngectomees [2].

1 Sound Source Unit and Signals

First, the user attaches a sound source unit under the lower jaw where is the same position as the conventional electrolarynx. The sound source unit used in the system can output extremely small signals that are difficult to be heard by other people around the user [3]. Small excitation signals also make it difficult for those people to hear low-volume artificial speech.

2 NAM microphone

NAM microphone [4], which captures signals through soft tissues of the head, is employed as the recording device. NAM microphone has a large dynamic range compared to other microphones. Consequently, it can detect various types of body-transmitted speech such as extremely low-volume artificial speech in the proposed system and ordinary speech.

3 Statistical Voice Conversion

To improve the quality of detected body-transmitted artificial speech, this system adopts a statistical voice conversion technique, which consists of training and conversion parts. In the training part, correspondences between source and target features are modeled. In the conversion part, source features are converted into target ones using maximum likelihood estimation of spectral parameter trajectory based on Gaussian Mixture Model (GMM) [5]. We set whisper as the target data to avoid the problem that the target F_0 frequencies are estimated from the source features not including F_0 information [6].

4 Output the Converted Speech

People around the user basically hear the converted speech.

2.2. Applications of the Proposed System

Our proposed system has some applications as shown in Table 1 by using different source data in the voice conversion. Although the proposed system has benefits of presenting natural speech, noise robustness and silent excitations, it requires quite complex voice conversion techniques because the proposed sound source unit and NAM microphone cause significant quality degradation of the source data. Other system constructions, which make the recorded source data more informative for the voice conversion process, are also sufficiently useful in limited environments. For example, in the use of telephones, it is not necessarily the case that the speaker must articulate the proposed small excitation signals because it is possible to provide only converted speech of body-transmitted artificial speech to the listener. Furthermore, if the user speaks in a quiet environment, an air-conductive microphone can be used as the recording sensor. Therefore, the voice conversion of the artificial speech using a conventional electrolarynx recorded with NAM/air-conductive microphone is also meaningful.

3. Auditory Feedback in Our System

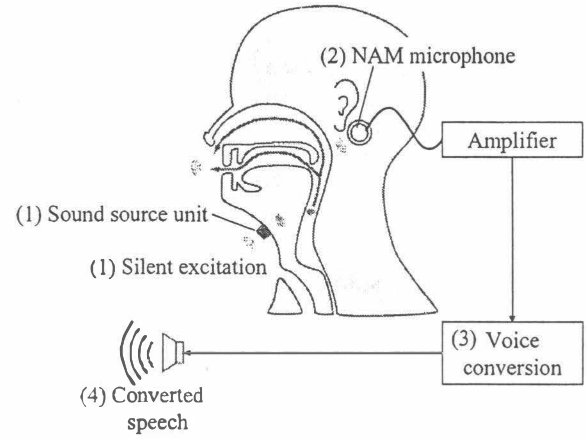


Figure 1: Image of the proposed speaking-aid system for laryngectomees.

Table 1: An example of applications of our speaking-aid system

Index	Sound source unit	Transmission micropohone	Benefits
A	Proposed unit	NAM microphone	1. Natural speech 2. Noise robustness 3. Silent excitations
B	Electrolarynx	NAM microphone	1. Natural speech 2. Noise robustness
C	Electrolarynx	Air-conducted microphone	1. Natural speech

Auditory feedback is a value-laden term that carries the implicit idea that speakers listen to the sound of their voice and send the result of this processing back through the brain to a level where this information can be compared with the production the speaker intended to produce [7]. Because our speaking-aid system uses extremely small excitation signals, the user may not be able to listen to the auditory feedback. Artificial speech has so far been recorded in a sound-proof room, and the speaker has been able to acquire the auditory feedback from the small artificial speech. However, the artificial speech with onl y small power may tend to be easily masked by external noises, and the user would have difficulty gaining auditory feedback from such masked artificial speech. This difficulty may cause a negative spiral of instability of the user's articulation and the degradation of the converted speech.

To prohibit such risks, we propose a method to provide detected signals with NAM microphone to the user's ear directly as explicit auditory feedback signals. The proposed method is evaluated by the stability of the user's articulation resulting from the presence of the explicit auditory feedback under noisy environments.

4. Artificial Speech Data of a Real Laryngectomee

4.1. Speech Recording

The speaker is a Japanese male in his 50s. This male underwent the laryngectomy around ten years ago and he has used an electrolarynx in his daily life.

We record artificial speech using the same sound source unit used in our previous work [2, 8]. As shown in Figure 2, the sound source unit, which outputs small excitation signals,

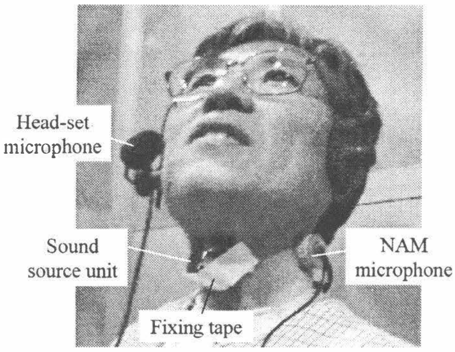


Figure 2: Scene of the speech recording of a laryngectomee.

is attached just under the lower jaw and is affixed by tape. The reason why the unit is affixed by tape is that it is expected to achieve hands-free communication because the proposed sound source unit is smaller than a conventional electrolarynx. 100 phoneme-balanced sentences are recorded. Small excitation signals are constructed with a pulse train of which the fundamental frequency is 100 [Hz] [8]. All utterances are recorded with NAM microphone and a head-set microphone simultaneously.

Moreover, we record artificial speech using a conventional electrolarynx uttered by the same laryngectomee on another day. These data are supposed to be used as the source data of applications B and C as shown in Table 1. 50 phoneme-balanced sentences are recorded. All utterances are recorded with NAM microphone and a tie-clip microphone simultaneously.

4.2. Setting of the Target Speech

In the conversion model training, we need to prepare the target speech samples for constructing the parallel data. One way to prepare the target speech is to use speech data from the existing speech corpora. However, the speaking speed and the pause insertions of the data in our speech corpora are different from those of the laryngectomee's data recorded above. Average speaking speed of the laryngectomee is almost 1.3 times slower than that of a part of our speech corpora. Furthermore, the laryngectomee tends to insert a pause more frequently compared with non-laryngectomees. These differences confuse alignments between source and target data in the voice conversion of our proposed system. Although these problems should be addressed, it is beyond the scope of this paper. To avoid the problems, we use speech data in which one non-laryngectomee speaks so that the speaking speed and pause insertion are as close to the source data as the speaker can accomplish.

5. Experimental Evaluations

To investigate 1) the effectiveness of explicit auditory feedback and 2) the performance of the proposed system for a real laryngectomee's data, we respectively ran two experimental evaluations.

5.1. Effectiveness of Explicit Auditory Feedback

5.1.1. Experimental Conditions

The speaker was a non-laryngectomee Japanese male who was trained to speak with an electrolarynx for 21 days. Speech data were recorded in a sound-proof room. The noise of an air-conditioner in a noise database [9] was used as the external noise. The usual noise level in the recording room was almost 31 [dBA]. External noise was output so that the noise levels at a place almost 50 [cm] away from the loud speakers were 50 [dBA], 55 [dBA], 60 [dBA], 65 [dBA] respectively. These sound levels were expected to replicate indoor noise environments of daily life.

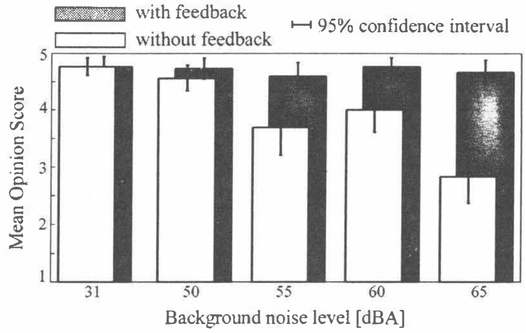


Figure 3: Result of opinion test about the stability of articulation under noisy environments.

Small excitation signals were constructed with a simple pulse train, sierra wave of which power was aligned to that of the pulse train and another sierra wave of which power was much larger than those of the others. Fundamental frequencies of these small excitation signals were all set to 100 [Hz]. As the explicit auditory feedback, body-transmitted artificial speech detected with NAM microphone was directly provided to one of the speaker's ears by a closed-type earphone. The volume of the explicit auditory feedback was controlled to each noise level so that the speaker felt it most easy to hear. In the case of not giving explicit auditory feedback, the speaker only gained his auditory feedback from expired artificial speech. Recorded utterances were nine newspaper articles in each recording condition.

The stability of the speaker's articulation was subjectively investigated by the speaker himself. 270 recorded utterances were randomly evaluated with a five point scaled opinion score (1: Bad - 5: Excellent) in the sound-proof room.

5.1.2. Experimental Result

Figure 3 indicates that if the explicit auditory feedback is not provided to the speaker, the articulation evidently becomes unstable especially in scenes in which the noise level is over 55 [dBA]. On the other hand, by providing the explicit auditory feedback, the speaker's articulation is kept stable in all situations in this experiment.

Figure 4 shows an example of waveforms with or without the explicit auditory feedback. It is demonstrated that in the case of not providing the explicit auditory feedback, bursts of some consonants such as /t/ and /k/ are extremely suppressed. However, such consonants can be uttered clearly even in loud background noise by providing the explicit auditory feedback with an earphone. As a result, explicit auditory feedback is quite effective for the speaker to keep his articulation stable under noisy environments.

5.2. System Evaluation Using Real Laryngectomee's Data

5.2.1. Experimental Conditions

Three kinds of artificial speech as shown in Table 1 were used as the source data of the voice conversion. The Whisper of a non-laryngectomee who recorded as described in section 4.2 was set as the target speech. In the evaluation of system A in Ta-

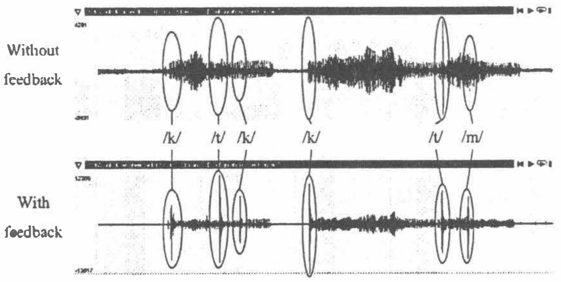


Figure 4: Example of waveforms of body-transmitted artificial speech using pulse train under 55 [dBA] of background noise level. The content is ’kihoNtekinikonoho: aNwai: toomou’.

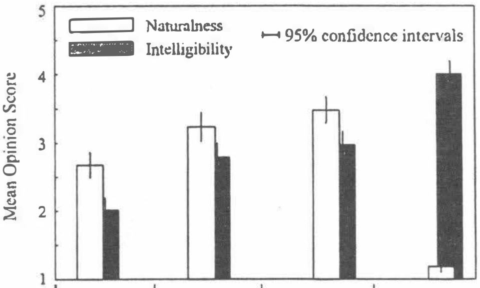
ble 1, a cross validation test using 100 newspaper articles was conducted, in which 90 utterances were for the training and the other ten utterances were for the evaluation. About the other sets B and C in Table 1, cross validation tests using 49 newspaper articles included in set A were conducted, in which 42 utterances were for training and the other seven utterances were for the evaluation. The number of mixture components was set to 64 in all tests.

Subjective evaluation was performed in which the intelligibility and the naturalness were evaluated independently with a five point scaled opinion score (1: Bad - 5: Excellent). Subjects were three non-laryngectomees. The evaluated stimuli were converted artificial speech of A, B and C, and source artificial speech of C (temporarily re-labeled as D). Stimulus A was evaluated as the output of the proposed system. Stimuli B and C were evaluated as the output of applications described in subsection 2.2. Stimulus D was evaluated as conventional speech. 35 utterances were selected from each cross validation set and a total of 140 utterances were evaluated randomly.

5.2.2. *Experimental Result*

Figure 5 brings to light that although the intelligibility of converted speech doesn't exceed that of conventional speech, the naturalness of the converted speech dramatically rises above that of conventional speech. This result displays two differences from our previous results using the non-laryngectomee's data [2, 8], namely, 1) the difference in the intelligibility can be seen by using different sound source units and excitation signals (A versus B) and 2) although the intelligibility of the converted speech is still acceptable, it is degraded (C versus D). About problem 1), the difference might be caused by the difference in the way of fixing the unit rather than that of the sound source units and excitation signals. Although the conventional electrolarynx was held by the speaker's hand in the recording, the sound source unit that outputs small excitation signals was affixed by tape. It is worthwhile to evaluate the effectiveness of the voice conversion using the laryngectomee's data recorded while holding the sound source unit with the speaker's own hand.

Difference 2) is a more essential problem. We think that this difference is caused by many factors such as the proficiency of the method of utterance using an external device, conversion into a different target speaker from the source speaker and so on. Further investigations are required.



Index	A	B	C	D
Unit	Small	EL	EL	EL
Microphone	NAM	NAM	Head-set	Head-set
Conversion	Done	Done	Done	None

Figure 5: Result of subjective evaluation of voice conversion using real laryngectomee's data.

6. Conclusion

In this paper, we evaluated 1) the effectiveness of explicit auditory feedback and 2) the performance of the proposed speaking-aid system using a real laryngectomee's data. With carefully recording target data to the source data, we dramatically succeeded at improving the naturalness of the converted data, but the voice conversion slightly degraded the intelligibility. In our future work, we will continue to improve the converted artificial speech uttered by the real laryngectomee.

7. Acknowledgements

This work was supported in part by MEXT Grant-in-Aid for Young Scientists (A) and MIC SCOPE.

8. References

[1] S. E. Williams and J. B. Watson, “Differences in Speaking Proficiencies in Three Laryngectomee Groups,” Arch Otolaryngol Vol. 111, pp. 216–219, April 1985.

[2] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, “Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech,” Proc. Interspeech 2006, pp.1395–1398, Sept. 2006.

[3] Y. Hosoi and T. Sakaguchi, “Silent voice input system without exhalation -theory and applications-,” Technical Report of IEICE, SP2003-105, pp. 13-16, 2003.

[4] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, “Remodeling of the Sensor for Non-Audible Murmur (NAM),“ Proceedings of Interspeech 2005, pp. 293-296, 2005.

[5] T. Toda, A. W. Black, and K. Tokuda, “Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 8, pp. 2222-2235, Nov. 2007.

[6] M. Nakagiri, T. Toda, H. Saruwatari, and K. Shikano, “Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion,” Proceedings of Interspeech 2006.

[7] P. Howell, “Effects of delayed auditory feedback and frequency-shifted feedback on speech control and some potentials for future development of prosthetic aids for stammering,” Stammering Res. 1(1): 31–46, April, 2004.

[8] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, “Impact of Various Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees,” Proceedings of the 10th European Conference on Speech Communication and Technology (Interspeech 2007 - Eurospeech), pp. 2517-2520, Aug. 2007.

[9] NOISE DATABASE, JEIDA, NOS-9601, Mar., 1996.