Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech

Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science,
Nara Institute of Science and Technology, Japan
{kei-naka, tomoki, sawatari, shikano}@is.naist.jp

Abstract

The aim of this paper is to improve the naturalness of speech using a medical device such as an electrolarynx. There are several problems associated with using existing electrolarynxes, such as the fact the loud volume of the electrolarynx itself might disturb smooth interpersonal communication, and that the generated speech is unnatural. We propose a novel speaking-aid system for total laryngectomees using a new sound source as an alternative to the existing electrolarynx and a statistical voice-conversion technique. The new sound-source unit outputs extremely small signals that cannot be heard by people around the speaker. Artificial speech is recorded with a NAM microphone through soft tissues of the head. From the result of voice conversion, the body-transmitted artificial speech is consistently converted to a more natural voice. We also demonstrate that the speech recognition performance of the proposed system substantially increases in terms of objective evaluation.

Index Terms: total laryngectomees, voice conversion, Non-Audible Murmur (NAM) microphone, electrolarynx.

1. Introduction

Speech is one of our fundamental methods for communication. However, it is not available for everybody. A total laryngectomy is the most common surgery for laryngeal cancer, in which the vocal folds are completely removed, and people who have undergone a total laryngectomy (total laryngectomees), cannot make utterances using vibration of their own vocal folds. Instead, they require some other methods to generate sound. There are some alternative methods for total laryngectomees to generate speech without vocal folds vibration. One of them used in total laryngectomees' daily life is called esophageal speech. Releasing gases from or through the esophagus produces the sound of esophageal speech. Another method for speech production is to use a medical device.

We focused on the latter method using a sound source unit, with which it is easy to master the method of uttering compared with esophageal speech. Moreover, it also requires less physical energy. The aim of this paper is to improve the naturalness of artificial speech generated by the sound source unit as typified by the electrolarynx for total laryngectomees. A novel speaking-aid system is proposed, in which we use a new sound-source unit that outputs an extremely small signal. The artificial speech generated by this unit is recorded with a NAM microphone [1][2] through soft tissues of the head. The system converts the artificial speech to a text or to a natural voice to make human or man-machine communication smoother. Results of objective evaluations revealed that, word accuracy of the body-transmitted speech substantially increased, and this speech was converted to a more natural voice on the mel-cepstral distortion between the converted speech and the target speech.

This paper organized as follows. In Section 2, we describe relevant issues regarding total laryngectomees. Section 3 explains the proposed speaking-aid system, and in Section 4 we present an evaluation of the experimental results. Finally, we summarize this paper in Section 5.

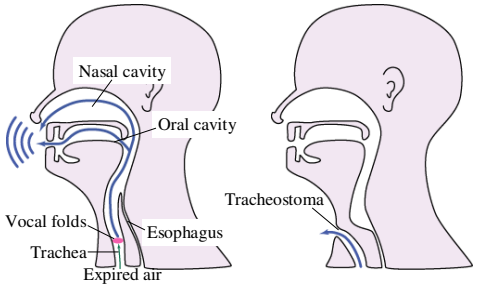

Figure 1: *The route of the air flowing from the lungs. Left figure shows the case of non-disabled people, right figure shows the case of the total laryngectomees.*

2. Total Laryngectomees

2.1. Conditions of Total Laryngectomees

Figure 1 shows how people produce speech sounds depending on the way air flows from the lungs. The left figure shows the case of non-disabled people, and the right one shows the case of total laryngectomees. Non-disabled people generate speech sounds by making the air flow from their lungs to the vocal tract and vibrating the vocal folds. However, total laryngectomees do not have vocal folds. Instead, they have a tracheostoma, which is a hole created at the end of the trachea to allow air for breathing. As Figure 1 shows, the throat and the esophagus of total laryngectomees are completely separated. Because the air flowing from the lungs does not pass through the oral cavity, they cannot generate the speech sounds in the same way as non-disabled people. Consequently, they have to produce speech sounds in an alternative way.

2.2. Problems of the Existing Electrolarynx

One of the most common medical devices for generating sound sources is the electrolarynx, which is a hand-held, battery-driven device operated by pressing it under the mandible. Figure 2 shows

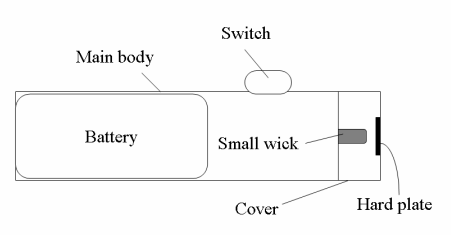

Figure 2: *Basic components of an electrolarynx.*

Figure 3: *A scene of using an electrolarynx.*

the basic components of an electrolarynx, and Figure 3 depicts a scene of using it. The small wick vibrates, and this vibration is amplified to strike a hard plate, and this is used as the sound-source signal. The electrolarynx's, vibration is conducted through the mandible to the esophagus and the resulting sound is emitted from the mouth as artificial speech. Although an electrolarynx is easy to use, there are some problems. We focus primarily on two problems, i.e., the loudness of the electrolarynx itself and the unnaturalness of the artificial speech. It is known empirically that especially in a calm environment, the signal of the electrolarynx floats out from the pressed location even if the user sets the volume to minimum. This might disturb smooth human communication. Moreover, the artificial speech generated with existing electrolarynxes is mechanical and unnatural.

3. Proposed Speaking-Aid System for Total Laryngectomees

Figure 4 shows the proposed speaking-aid system for total laryngectomees. We use a new sound-source unit that generates an extremely small signal as an alternative to the existing electrolarynx. The artificial low-energy speech is detected through the soft tissues of the head with a Non-Audible Murmur (NAM) microphone [1][2]. The body-transmitted artificial speech is then converted to natural speech, such as a whisper using a statistical voice-conversion technique to facilitate smoother human-human communication. For man-machine communication, the body-transmitted signal is directly recognized with an automatic speech recognition engine.

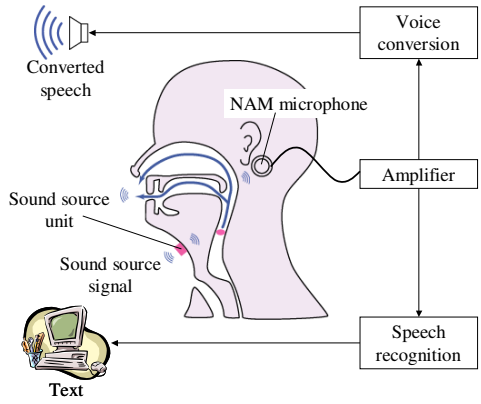

Figure 4: *Proposed speaking-aid system for total laryngectomees.*

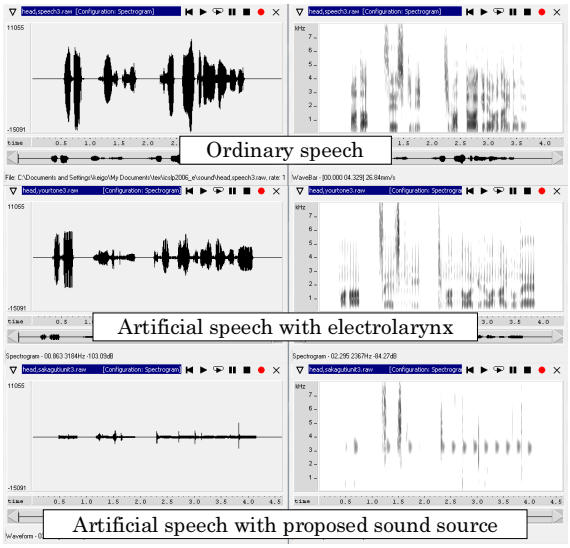

Figure 5: *An example of waveforms and their spectrograms recorded with a conventional microphone for sampling air-conducted sound. The content of the utterance is "M a d a s s e: sh i k i n i k i m a q t a w a k e d e w a n a i n o d e".*

3.1. Sound-Source Unit

We apply a new sound-source unit as an alternative to the existing electrolarynx to address the problem that the volume of the electrolarynx itself might inhibit smooth communication. This sound-source unit is a wave transducer that outputs extremely small sound-source signals compared with the existing electrolarynx. In this paper, the simplest signal, i.e., a pulse train, is used as the sound-source signal with a fundamental frequency of 100 Hz. Figure 5 depicts an example of waveforms and their spectrograms recorded with a conventional microphone for sampling air-conducted sound. The left column shows the waveform and the right one shows the spectrogram. It is clear that the existing air-conducted microphone cannot detect the artificial speech from the proposed sound-source unit with enough quality, especially under noisy conditions, although it can detect that of the existing electrolarynx.

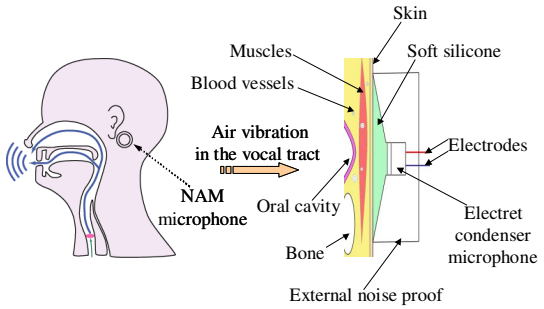

Figure 6: *The structure of a Soft-Silicone Type NAM microphone, and the location of the NAM microphone attachment.*

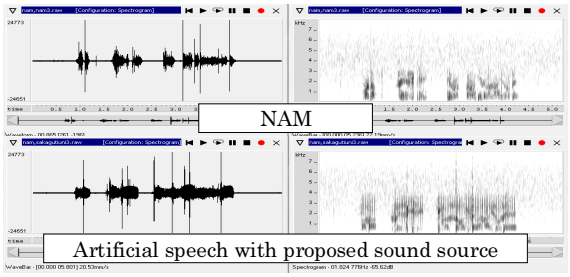

Figure 7: *An example of waveforms and their spectrograms recorded with a NAM microphone. The content of the utterance is "M a d a s e: s h i k i n i k i m a q t a w a k e d e w a n a i n o d e".*

3.2. Non-Audible Murmur (NAM) and NAM Microphone

Nakajima et al. proposed an acoustic sensor called the Non-Audible Murmur (NAM) microphone, which detects speech signals directly on the skins [1][2]. NAM is defined as articulated respiratory sound without vocal folds vibration transmitted through the soft tissues of the head. Because NAM has low power, it is very difficult to be detected with conventional microphones for sampling air-conducted sound. To solve this problem, the NAM microphone, which is attached to the top of the neck skin low behind the earlobe, is proposed. Figure 6 displays the structure of a NAM microphone and the attachment location. Figure 7 shows an example of waveforms and their spectrograms detected with a NAM microphone. This figure indicates that the NAM microphone can detect speech with low power such as NAM with sufficient quality. Moreover, the NAM microphone is very robust against external noises. NAM has special acoustic characteristics. Unvoiced plosives are detected very strongly, and frequency components over 4 kHz are almost lost. As shown in Figure 7, since acoustic characteristics of body-transmitted artificial speech are similar to those of NAM, it is expected that the techniques of automatic speech recognition [7] and voice conversion [5] are useful for detecting body-transmitted artificial speech as well as NAM.

3.3. Voice Conversion

Artificial body-transmitted speech not only has low intelligibility, but it also sounds very mechanical. To address these problems, we adopt a technique called voice conversion [3] that converts one voice into another. In this paper, the body-transmitted artificial speech is in fact converted to a whisper [8]. To do this we used the statistical spectral conversion method based on the maximum likelihood estimation (MLE). This method uses a Gaussian Mixture Model (GMM) to model the joint probability density of the source and target features. To train the GMM, we apply the same sentence pairs of source and target voices.

It has been reported that the conversion from NAM to ordinary speech (NAM-to-Speech) significantly improves the intelligibility and voice quality of NAM [5]. It is expected that the body-transmitted artificial speech could also be able to be converted to more natural speech by using this voice conversion method.

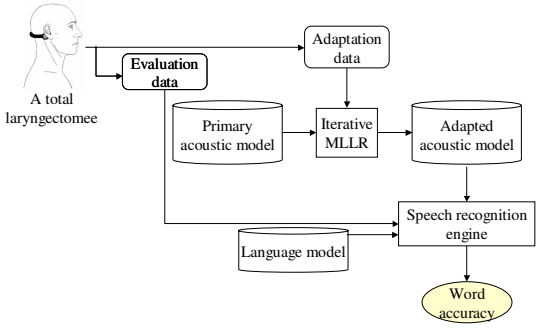

Figure 8: *The setup of the speech recognition experiment*

4. Experimental Evaluations

To demonstrate the effectiveness of the proposed speaking-aid system for total laryngectomees, we performed experiments on speech recognition and voice conversion.

4.1. Experimental Conditions

Figure 8 shows the setup of the speech recognition experiment. In this work, we used simulated voices of a non-disabled Japanese male speaker. To correctly interrupt the air flowing from the lungs, the subject made utterances with the new sound source while stopping breathing. The male speaker read 50 sentences from newspaper articles, which were used as training data for speaker adaptation, and 20 sentences for evaluation. All of them were recorded at 48 kHz. We used 16 kHz down-sampled data. We created an adapted acoustic model for body-transmitted artificial speech by iteratively performing Maximum Likelihood Linear Regression (MLLR) adaptation. MLLR is one of the speaker adaptation techniques for modifying a large number of parameters with a small amount of adaptation data [4]. The mean vectors of a mixture Gaussian HMM system are estimated with a set of linear transformations to maximize the likelihood for the adaptation data. The proposed system's recognition performance was evaluated for a newspaper dictation task. We used Julius [6] as the decoder for large-vocabulary continuous speech recognition (LVCSR). The language model was a 20K-word trigram model, and the acoustic model was a phonetic tied-mixture (PTM) model with 64 mixtures and 3,000 states. The number of leaves on a regression tree for MLLR was set to 128.

Figure 9 shows the setup of the voice conversion experiment. We employed natural whispers uttered by the male speaker as the target speech, and used the artificial speech data used in the recognition experiment as the source speech. We trained a Gaussian Mixture Model (GMM) to convert the source speech to the target speech. Following that, the evaluation data of the source speech were converted to that of the target speaker through the conver-

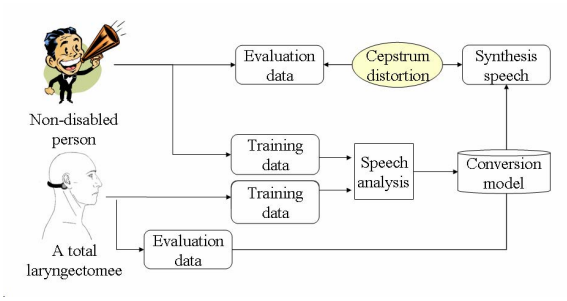

Figure 9: *The setup of the voice conversion experiment*

Table 1: *Improvement of word accuracy by iterative MLLR adaptation in LVCSR and that of mel-cepstral distortion by voice conversion from body-transmitted artificial speech to whispers*

word accuracy [%]	2.4 → 64.7
mel-cepstrum distortion [dB]	6.7 → 5.0

sion model. The mel-cepstral distortion between the converted speech and the target whisper was then calculated, with the 0th through 24th mel-cepstral coefficients being used as spectral features at each frame. Full covariance matrices were employed for the GMM, in which the number of mixtures was set to 32.

4.2. Experimental Results

Table 1 shows the experimental results. The improvement in the word accuracy indicates the possibility for man-machine communication by total laryngectomees. Moreover, the decrease in the mel-cepstral distortion between the synthesized speech and the target speech indicates the possibility of making human-human communication smoother. Figure 10 shows an example of waveforms and spectrograms of the body-transmitted artificial speech, the converted speech, and the target whispers. It is clear that the acoustic characteristics of the converted speech are physically much closer to the target whispers than the original speech. These results indicate the strong possibility of realizing the proposed speaking-aid system.

5. Conclusions

This paper proposed a speaking-aid system for total laryngectomees. To realize the system, we primarily adopted two novel techniques, i.e., a NAM microphone as a device to detect signals with only low power, and voice conversion to improve the naturalness of the body-transmitted speech. We employed extremely small sound-source signals as an alternative to the existing electrolarynx to make human-human communication smoother. The body-transmitted speech was evaluated based on the mel-cepstral distortion between the converted speech of a simulated total laryngectomee and the target whispers of a non-disabled person. The mel-cepstral distortion decreased from 6.7 dB to 5.0 dB. The speech was also recognized with a speaker-adapted acoustic model created using Maximum Likelihood Linear Regression (MLLR). The word accuracy for a 20 K-word dictation task improved from 6.4% to 64.7%.

As future work, we will investigate the conversion from the body-transmitted artificial speech not only to whispers but also to ordinary speech.

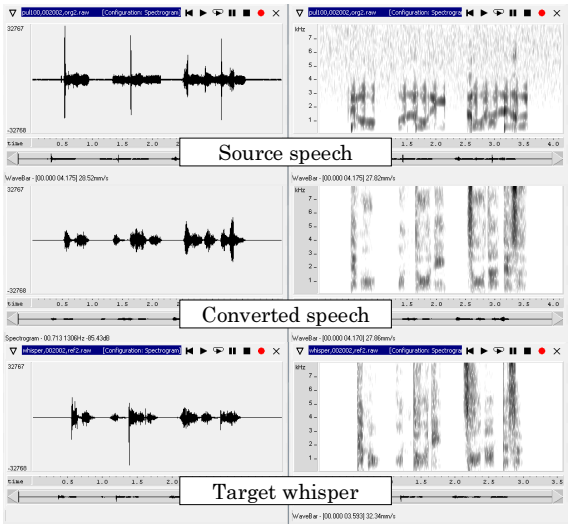

Figure 10: *An example of waveforms and spectrograms of body-transmitted speech, the converted speech, and the target whisper*

6. Acknowledgements

This research is supported by SCOPE-S.

7. References

[1] 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.

[2] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, “Non-Audible Murmur (NAM) Recognition,” IEICE Trans. Information and Systems, Vol.E89-D, No. 1, pp. 1-8, 2006.

[3] T. Toda, A.W. Black, and K. Tokuda, “Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005), Vol. 1, pp. 9-12, 2005.

[4] M.J.F. Gales, and P.C. Woodland, “Mean and Variance Adaptation within the MLLR Framework,” Computer Speech & Language, Vol. 10, pp. 249-264, 1996.

[5] T. Toda, and K. Shikano, “NAM-to-Speech Conversion with Gaussian Mixture Models,” Proceedings of Interspeech 2005, pp. 1957-1960, 2005.

[6] A. Lee, T. Kawahara, and K. Shikano, “Julius — An Open Source Real-Time Large Vocabulary Recognition Engine,” Proc. 7th European Conference on Speech Communication and Technology, pp. 1691-1694, 2001.

[7] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, “Accurate Hidden Markov Models for Non-Audible Murmur (NAM) Recognition Based on Iterative Supervised Adaptation,” IEEE Automatic Speech Recognition and Understanding Workshop, pp. 73-76, 2003.

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