Impact of Various Small Sound Source Signals on Voice Conversion Accuracy in Speech Communication Aid for Laryngectomees

Keigo Nakamura[1], *Tomoki Toda*[1], *Hiroshi Saruwatari*[1], *Kiyohiro Shikano*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan

{kei-naka, tomoki, sawatari, shikano}@is.naist.jp

Abstract

We proposed a speaking aid system using statistical voice conversion for laryngectomees, whose vocal folds have been removed. This paper investigates the influence of various small sound sources on the voice conversion accuracy. Spectral envelopes and power of sound sources are controlled independently. In total 8 different kinds of sound source signals, e.g. pulse train, sierra wave and so on, are examined. Results of objective and subjective evaluations demonstrate that for voice conversion, sound sources with various spectral envelopes and power in a large degree are acceptable unless the power of them is comparable to that of silence parts.

Index Terms: speaking aid, laryngectomees, sound sources, voice conversion

1. Introduction

There is an operation called laryngectomy as one of typical medical cure to laryngeal cancer. This operation completely removes the patients' vocal folds. Therefore, people who have undergone laryngectomy (laryngectomees) have serious problems in phonation.

Because laryngectomees cannot generate vocal folds vibrations, they need another method to be able to speak [1]. This paper focuses on a method to speak with a medical device such as an electrolarynx. The benefits of using such a medical device are a fast learnablity and easy usability for people with poor physical power. However, this method to speak has some problems; generated voice (artificial speech) sounds mechanical because actual vocal folds vibration is too complex to be completely simulated by the medical device, and the sound source signal itself is too noisy for people around the speaker. As a solution to these problems, we proposed a speaking aid system for laryngectomees using statistical voice conversion [2].

In our previous work [2], only pulse train is used as the sound source signal. If our speaking aid system allows various sound sources, users can use any kinds of signals they like, e.g., speakable sound sources supporting clean auditory feedback or quiet sound sources not annoying to people around the user. However, this wouldn't make sense if voice conversion couldn't work for different kinds of sound source signals. Therefore, this paper investigates how various sound source signals affect the voice conversion accuracy. Experimental results demonstrate that for voice conversion sound sources with various spectral envelopes and power in a large degree are acceptable.

This paper is organized as follows. In section 2, our speaking aid system for laryngectomees is described. In section 3, several small sound source signals are designed. In section 4, objective and subjective evaluations are conducted. Finally, we summarize this paper in section 5.

2. Speaking Aid System

Figure 1 shows an overview image of the proposed speaking aid system for laryngectomees [2]. This system consists of the following four parts:

1. Articulating extremely small source signals: A sound source unit used in this system can output extremely small and arbitrary signals [3].
2. Recording artificial speech with Non-Audible Murmur (NAM) microphone [4] [5], i.e., body transmitted artificial speech: NAM microphone, which is a microphone for detecting extremely small signals from the skin directly with high quality. Detection works even if the signal cannot be catched by people around the speaker.
3. Converting the body transmitted artificial speech into a much more natural voice by statistical voice conversion [6]: Voice conversion used in this system is based on statistical spectral conversion including training and conversion parts. In the training part, joint probability density of input and output features are modeled with a Gaussian Mixture Model (GMM). In the conversion part, output features are determined based on maximum likelihood estimation on conditional probability density of output features given input features [6].
4. Output the converted speech.

Although it would be ideal to convert body transmitted artificial speech into ordinary speech, such conversion is very difficult because an appropriate F_0 counter has to be estimated from speech signals that do not have natural F_0 information like unvoiced speech. To avoid this problem, a method of converting body transmitted unvoiced speech into whisper that is familiar to unvoiced speech has been proposed [7], and it has been reported that the voice conversion of body transmitted unvoiced speech into whisper significantly works better than that into ordinary speech. Therefore, this paper also discusses how to convert body transmitted artificial speech into whisper.

We have conducted a subjective evaluation to demonstrate the effectiveness of our speaking aid system. The experimental conditions are the same as described in [2]. Five kinds of speech samples including 1) original body transmitted artificial speech, 2) converted one, 3) target whisper, 4) air transmitted artificial speech with an existing electrolarynx and 5) air transmitted ordinary speech were evaluated in terms of naturalness. Target whisper is used as the limit of the converted whisper. The fourth speech is evaluated as the conventional method. The last stimulus is given as the final limit of our system. As a result, voice conversion significantly improves the naturalness of original speech as shown in Table 1. Moreover, naturalness of

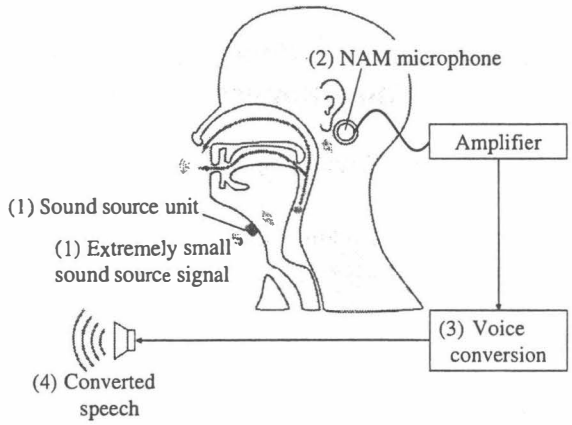

Figure 1: *Overview of our speaking aid system.*

Table 1: Result of opinion test on naturalness using a 5-point scale opinion score (1: Bad - 5: Excellent).

Speech samples	MOS
Artificial speech with an electrolarynx	1.6
Body transmitted artificial speech	1.0
Converted body transmitted artificial speech	2.8
Target whisper	3.9
Ordinary speech	5.0

converted speech of our speaking aid system is much better than air conducted artificial speech using the electrolarynx which is a typical conventional speaking aid system.

3. Sound Source Signals

3.1. Designing Ideas

Many electrolarynxes are designed so that expired artificial speech sounds natural as much as possible [8] [9]. On the other hand, because our speaking aid system employs voice conversion and the sound source unit generates extremely small signals, we can design sound sources based on some other views as follows:

- Designing sound sources so that the voice conversion accuracy is not degraded.
- Designing sound sources so that the speaker can clearly hear auditory feedback of his or her own body transmitted artificial speech.
- Designing sound sources so that the signals cannot be heard by people around the speaker.

As the first step, this paper investigates the impact of different sound sources on the voice conversion accuracy. Various sound sources are designed by controlling the spectral envelope and power independently. Specifically, three different kinds of spectral envelopes and five different kinds of power variations are designed as described in the following subsections.

3.2. Spectral Changes

In case of spectral changes, pulse train, sierra wave and compensation wave into target speech, i.e., whisper in this paper, are

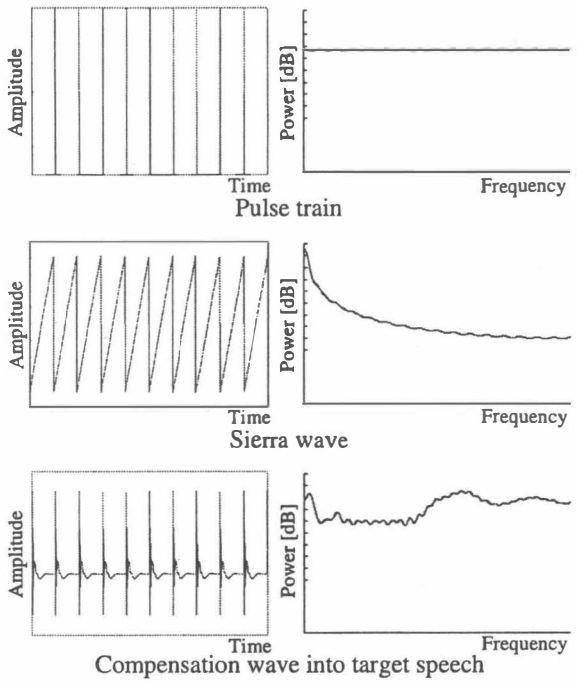

Figure 2: *Waveforms and power spectra of three sound source signals. Every fundamental frequency is set to 100 [Hz].*

designed. Figure 2 shows waveforms and spectral envelopes of individual sound source signals.

- Pulse train:
  Pulse train is used as one of the most basic signals.
- Sierra wave:
  It is said that vocal folds vibrations can be approximated with asymmetry triangle wave. Sierra wave is used as the asymmetry triangle wave in this paper.
- Compensation wave into target speech:
  The voice conversion accuracy might be improved by getting input features close to target ones. We designed a sound source so that a long-term spectrum of body transmitted artificial speech would be close to that of the target speech. First, averaged mel-cepstral coefficients of input body transmitted artificial speech using the pulse train as the sound source signal and those of the target speech in training data are calculated. Next, the former mel-cepstral coefficients are subtracted from the latter ones. Finally, intended sound source based on the result of mel-cepstral coefficients are available by convolving the acquired impulse response with pulse train. Consequently, as shown in the figure 2, high frequency components are enhanced.

Figure 3 shows an example of spectrograms of body transmitted artificial speech using three kinds of sound sources and that of target whisper detected with a head-set microphone. One characteristic of body transmitted artificial speech is that high-frequency components are not observed because it does not include radiation characteristics from lips. The compensation wave enhances higher frequency components because it simulates a sort of the radiation characteristics. On the other

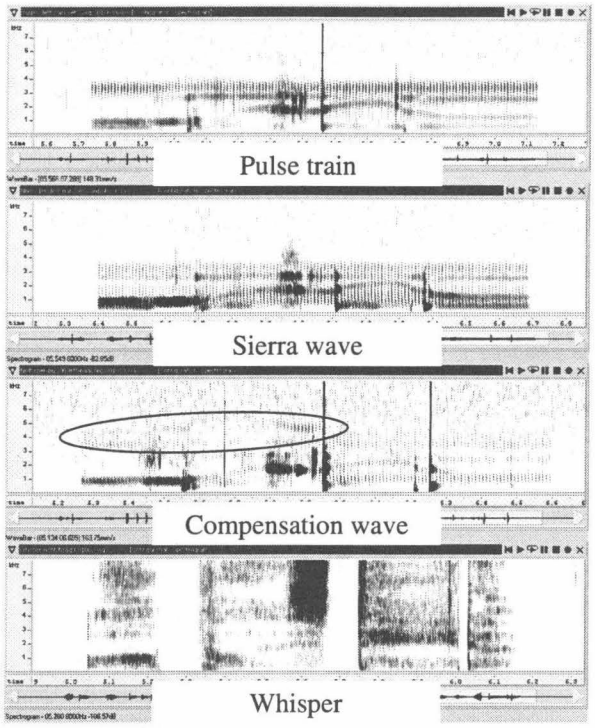

Figure 3: *An example of spectrograms of body transmitted artificial speech using pulse train, sierra wave compensation wave and target whisper.*

hand, power in lower frequencies is relatively suppressed. Consequently, some information around 4 - 5 [kHz] can be seen as indicated with an ellipsoid in the figure. Formants of the body transmitted artificial speech using sierra wave can be seen clearly than others because much power of the signal is in lower frequencies.

3.3. Power Changes

In case of power changes, spectral envelope is fixed only to sierra wave. Figure 4 shows power histograms of 50 utterances using minimum, basic, and maximum power of sierra wave. Power of the minimum signal is -27 [dB] and that of the maximum one is +18 [dB] compared with the basic one. Left distributions show power histograms of silence parts and right ones show those of speaking parts.

In the histograms when using minimum signals, most of the right distribution overlaps the left one. This shows that most of the frames have almost the same power as silence parts.

4. Experimental Evaluations

In order to compare voice conversion accuracy for various sound sources, objective and subjective evaluations are conducted.

4.1. Conditions

One Japanese male, who learned the way to speak with a medical device for 21 days with an existing electrolarynx, imitated the artificial speech. When recording speech, auditory feedback of the detected body transmitted artificial speech was given to the speaker with a headphone to allow him to check his artic-

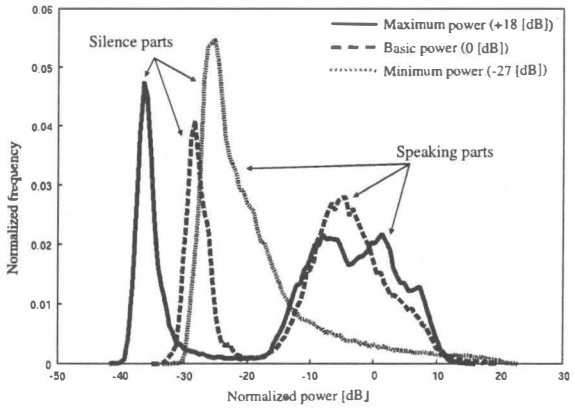

Figure 4: *Power histograms of 50 utterances using sierra wave with minimum, basic, and maximum power. The horizontal axis is frame normalized power.*

ulation. The same vibration unit from our previous work [2] is employed in the experiment. Fundamental frequency of all sound source signals were set to 100 [Hz]. The target speech was set to whisper of the speaker. We used 70 sentences of read Japanese newspaper articles for cross validation. Fifty sentences out of them were used as training data for voice conversion, and the remaining 20 sentences were used as evaluation data. The number of mixtures of the GMM was set to 32 in all conversions. Mel-cepstral distortion was used as an objective evaluation measure.

For changing the spectral envelope, 3 kinds of signals of pulse train, sierra wave and compensation wave into whisper were used as the sound sources. Average power of each sound source was adjusted to that of the pulse train. When average power of signals was changed, 6 kinds of different sierra waves of which powers were set to -27 [dB], -18 [dB], -9 [dB], 0 [dB], +9 [dB], and +18 [dB] were designed. Only articulation without sound sources from the device was evaluated as the lower limit of the power. In this case, only sound sources generated when pronouncing several consonants were articulated. Moreover, the electrolarynx was also evaluated as the upper limit of the power. Note that the power of electrolarynx was still much larger than that of the maximum sierra wave.

In subjective evaluation, 5 persons evaluated the naturalness of converted whisper from 7 kinds of sound sources; only articulation, the electrolarynx, the minimum and the maximum power of sierra wave, the basic power of sierra wave, compensation wave into whisper, and pulse train. An opinion score was set to a 5-point scale (1: Bad - 5: Excellent). Twenty evaluation speech samples were selected randomly from every cross validation set. In total 140 speech samples were evaluated by each subject.

4.2. Results and Discussion

Table 2 shows the result of objective evaluation. The mel-cepstral distortion before conversion when using compensation wave is smaller than when using other different spectral envelopes. After conversion, differences of spectral envelope do not significantly affect the accuracy of voice conversion (5.1-5.2 [dB]). When changing power, the accuracy of voice conversion varies between 5.1 [dB] and 5.5 [dB]. However, there is no apparent tendency that a power increase or decrease causes

Table 2: Mel-cepstral distortion [dB] when using several sound sources

Sound sources (numbers show power [dB])	Before conversion	After conversion
Articulation (-∞)	10.1	6.0
Sierra wave (-27)	9.9	5.5
Sierra wave (-18)	9.9	5.2
Sierra wave (-9)	10.2	5.5
Pulse train (0)	11.0	5.1
Sierra wave (0)	11.7	5.2
Compensation wave (0)	10.3	5.1
Sierra wave (+9)	12.0	5.1
Sierra wave (+18)	11.3	5.5
Electrolarynx	13.0	5.1

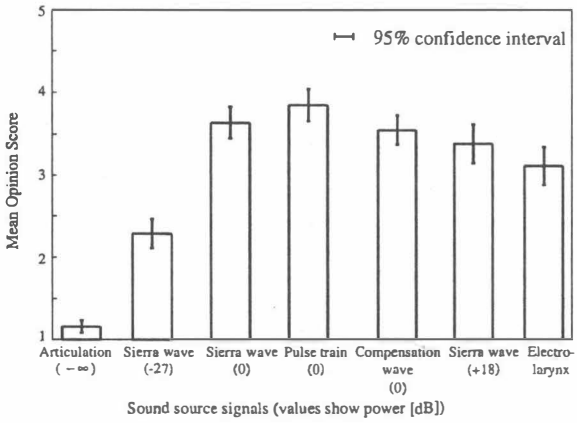

Figure 5: *Result of opinion test on naturalness using a 5-point scale opinion score (1: Bad - 5: Excellent).*

significant degradation of the voice conversion accuracy. It is expected that other factors such as variations of articulation utterance-by-utterance, the location of the NAM microphone and the location of the sound source unit rather give much influence on the voice conversion accuracy. Consequently, voice conversion works for various sound sources. Note that a sound source signal is necessary because the distortion of converted voice of articulation is large (6.0 [dB]).

Figure 5 shows the result of the subjective evaluation. MOS of converted speech of articulation gives only 1.2 and that of the minimum power of sierra wave gives 2.3. Too small power of sound sources results in a poor quality of the converted voice. Consequently, such sound sources are inappropriate in the speaking aid system. On the other hand, when increasing the power, the quality of the converted speech does not vary so much for the other sound sources (3.1-3.9). Moreover, there are no significant degradations when changing the spectral envelope. Consequently, voice conversion works with various sound source signals unless their power is very close to that of silence parts.

5. Conclusions

This paper examined the influence of various small sound source signals on voice conversion used in our speaking aid system for laryngectomees. Sound sources are designed so that the voice conversion accuracy would not be degraded by changing spectral envelopes and power independently. We designed 8 different kinds of signals and conducted objective and subjective evaluations. As a result, there were no significant degradations when changing the spectral envelope. Consequently, this paper showed that voice conversion worked well for sound sources with various spectral envelopes and power in a large degree unless the power of them is comparable to that of silence parts.

We will design new sound sources, e.g., speakable sound sources supporting clean auditory feedback or quiet sound sources not annoying people around the user. Although we evaluated naturalness, intelligibility of converted speech signals will be evaluated in our future work. We will also record speech data of a real laryngectomee and use them as input data of the voice conversion to extend results to real-life conditions.

6. Acknowledgements

This research is partially supported by SCOPE-S.

7. References

[1] L. Coltart, "Voice restoration after laryngectomy," Nursing Standard, Vol.13, Num.12, Dec. 1998.

[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, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition," IEICE Trans. Information and Systems, Vol.E89-D, No. 1, pp. 1-8, 2006.

[5] 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.

[6] 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.

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

[8] M. Hashiba, N. Uemi, M. Oikawa, Y. Yamaguchi, Y. Sugai and T. Ifukube, "Industrialization of the Electrolarynx with a Pitch Control Function and Its Evaluation," IEICE D-II, Vol. J84, No.6, pp.1240-1247, June 2001.

[9] N. Uemi, T. Ifukube, M. Takahashi and J. Matsushima, "Design of a new electrolarynx having a pitch control function," Proc. 3rd IEEE International Workshop on, pp. 198-203, July 1994.