

The Use of Air-Pressure Sensor in Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion

Keigo Nakamura¹, Tomoki Toda¹, Hiroshi Saruwatari¹, and Kiyohiro Shikano¹

¹Graduate School of Information Science, Nara Institute of Science and Technology, Japan
 {kei-naka, tomoki, sawatari, shikano}@is.naist.jp

Abstract

In our previous work, we proposed a speaking-aid system converting electrolaryngeal speech (EL speech) to normal speech using a statistical voice conversion technique. The main weakness of our system is the difficulty of estimating natural contours of the fundamental frequency (F_0) from EL speech including only built-in F_0 contours. This paper proposes another speaking-aid system with an air-pressure sensor to enable laryngectomees to control F_0 contours of the EL speech using their breathing air. The experimental result demonstrates that 1) the correlation coefficient of F_0 contours between the converted and the target speech is improved from 0.58 to 0.78 by the use of the air-pressure sensor and 2) the synthetic speech converted by the proposed system sounds more natural and is more preferred to that by our conventional aid system.

Index Terms: Electrolarynx, Air-pressure sensor, Laryngectomee, Voice conversion, Speaking-aid

1. Introduction

Voice rehabilitation of laryngectomees, whose vocal folds have been removed due to laryngectomy, is an important research topic. An electrolarynx (EL) is a major medical device that enables laryngectomees to obtain their speech again. When a user speaks with an EL, it is held by the user's one hand and attached to his or her lower jaw. The vibrations are transmitted through the skin and the electrolaryngeal speech (EL speech) is produced by the user's articulation. The problem focused on in this paper is unnaturalness of EL speech due to the built-in sound source signals.

Many researchers like [1] or [2] have tried to address unnatural EL speech. Uemi *et al.* have developed an EL [1] that enables laryngectomees to control the F_0 of the EL using the breathing air from the tracheostoma through an air-pressure sensor as Fig. 1 shows. Murakami *et al.* have enhanced the EL speech [2]. In their approach, many conversion rules are stored in a dictionary from training data. Those rules are then applied to test data. In this approach, the input utterances will not be accepted if no suitable conversion rule is not found.

We have proposed a speaking-aid system [3] which enhances EL speech based on a statistical voice conversion (VC) technique using Gaussian mixture models (GMMs) [4]. This system consists of four parts: 1) generating the sound source signals, 2) recording the produced EL speech, 3) converting the recorded EL speech, and 4) presenting the converted normal speech. Our previous work has demonstrated that smooth F_0 contours can be estimated from spectral features of the EL speech. On the other hand, the correlation of F_0 contours between the converted and the target speech has been 0.38 which still leaves room for improvement. This is because this system

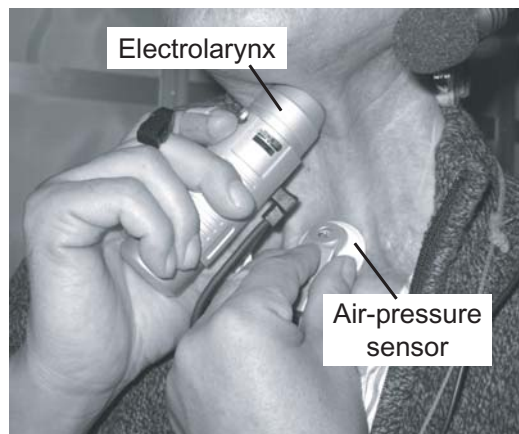


Figure 1: Scene of uttering EL speech with air-pressure sensor.

estimates target F_0 contours from only the spectral information of the EL speech.

In order to estimate more natural F_0 contours, this paper proposes the use of the air-pressure sensor in our speaking-aid system. The proposed system is shown in Fig. 2. This system allows laryngectomees to control F_0 of the EL speech by their breathing air. Therefore, more useful information for estimating natural F_0 contours is available when converting the EL speech generated with the air-pressure sensor (EL(air) speech) to normal speech. This paper estimates F_0 contours using both the spectral and the F_0 information of the EL(air) speech.

This paper is organized as follows. In **Section 2**, the used GMM-based VC method is overviewed. Our conventional speaking-aid systems are described in **Section 3**, and the proposed system is described in **Section 4**. These systems are experimentally evaluated in **Section 5**. This paper is concluded in **Section 6**.

2. GMM-based statistical VC [4, 5, 6]

GMM-based statistical VC consists of a training part and a conversion part.

2.1. Training part

Two speech signals of the source and the target speech are decided, and the training data consisting of same utterance pairs are recorded. These two speech signals are automatically aligned in advance using the dynamic time warping procedure.

Here, let $\mathbf{x}_t = [x_t(1), \dots, x_t(d_x)]^\top$ and $\mathbf{y}_t = [y_t(1), \dots, y_t(d_y)]^\top$ be a static source and target feature vector at frame t , respectively, where d_x and d_y denote the dimensions of \mathbf{x}_t and \mathbf{y}_t , respectively. \top denotes transposition. For the

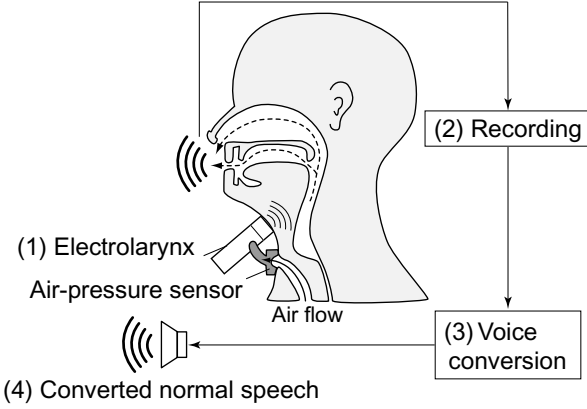


Figure 2: Overview of proposed speaking-aid system that enhances EL speech using air-pressure sensor.

training data of the source speech at frame t , a feature vector capturing dynamic movement is used, which is denoted as \mathbf{X}_t . For the training data of the target speech, a joint feature vector $\mathbf{Y}_t = [\mathbf{y}_t^\top, \Delta \mathbf{y}_t^\top]^\top$ is used. After preparing these training data, a GMM is trained to describe the joint probability density of the source and the target feature vectors as follows:

$$P(\mathbf{X}_t, \mathbf{Y}_t | \lambda) = \sum_{m=1}^M w_m \mathcal{N}([\mathbf{X}_t^\top, \mathbf{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}),$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix},$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. m denotes the mixture component index, and M denotes the total number of the mixture components. A parameter set of the GMM is denoted by λ , which consists of weights w_m , mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components. $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$ represent the mean vectors of the m th mixture component for the source and the target features, respectively. $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ represent the covariance matrices and $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$ represent the cross-covariance matrices of the m th mixture component for the source and the target features, respectively.

2.2. Conversion part

Let $\mathbf{X} = [\mathbf{X}_1^\top, \dots, \mathbf{X}_T^\top]^\top$ and $\mathbf{Y} = [\mathbf{Y}_1^\top, \dots, \mathbf{Y}_T^\top]^\top$ be a time sequence of the source and the target feature vectors, respectively, where T denotes the number of frames. The converted static feature sequence $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^\top, \dots, \hat{\mathbf{y}}_T^\top]^\top$ is determined to maximize the likelihood of the conditional probability density of \mathbf{Y} given \mathbf{X} as follows:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{Y} | \mathbf{X}, \lambda) \quad \text{subject to } \mathbf{Y} = \mathbf{W} \mathbf{y},$$

where \mathbf{W} is a matrix to extend the static feature sequence to the parameter vector sequence consisting of the static and the dynamic features [7].

The converted speech quality can be more enhanced by considering the global variance (GV) parameters [4].

3. Conventional speaking-aid systems using GMM-based VC

We have so far proposed two types of speaking-aid systems converting different types of EL speech signals to normal speech.

The main difference of these systems is the sound sources.

3.1. Speaking-aid system for EL speech [3]

This system converting EL speech into normal speech is the most basic system among our proposed systems. The four components of this system are the same as Fig. 2. This system is supposed to be used in situations in which only converted speech is mainly presented to listeners, such as telecommunication or lectures.

To present normal speech, this system estimates target spectral, F_0 , and aperiodic components of the excitation signal from only source spectral information, since the number of vibrations of the EL is fixed, and therefore, the produced EL speech does not have effective F_0 information. In the VC part, spectral segmental feature vectors of the source EL speech are used to compensate for lost information in the user's articulation. These feature vectors are constructed by following procedures. First, static feature vectors at frames $t \pm L$ are concatenated as $\mathbf{c}_t = [\mathbf{x}_{t-L}^\top, \dots, \mathbf{x}_t^\top, \dots, \mathbf{x}_{t+L}^\top]^\top$. Then, a low dimensional feature vector \mathbf{X}_t is extracted from \mathbf{c}_t by PCA procedure.

3.2. Speaking-aid system for EL speech using extremely small sound source signals [8]

The four components of this system are also the same as Fig. 2. This system employs another sound source unit generating extremely small signals [9] to address not only unnaturally sounding EL speech but also the noisy sound sources generated from the EL. EL speech using the small signal is captured using a Non-Audible Murmur (NAM) microphone [10], which is attached on the muscle at the back of the user's neck. NAM microphone captures the small-powered EL speech through the soft tissues of the head. The captured body-conducted EL speech using the small sound source signals (EL(small) speech) is converted to normal speech. This system is supposed to be used in users' daily conversations as well as telecommunication, since the sound sources and the produced EL(small) speech is assumed not to be heard by listeners. The VC method of this system is the same as the system described in Section 3.1.

4. Proposed speaking-aid system with the air-pressure sensor

In order to estimate more natural F_0 contours, this paper introduces the use of the air-pressure sensor [1] described in Section 1 to obtain intentionally controlled F_0 contours of the EL speech used in our aid system. This paper proposes the other speaking-aid system as shown in Fig. 2, which converts EL speech produced using this air-pressure sensor (EL(air) speech) to normal speech. This system is supposed to be used in the same situations as the system described in Section 3.1.

Although F_0 contours of EL speech are intentionally controlled by the speaker, these contours do not vary smoothly as shown in Fig. 4. Moreover, all phonemes are basically produced as voiced phonemes in EL speech because the EL always generates the sound source signals during speaking. Consequently, it is essentially difficult for laryngectomees to produce a natural F_0 contour with the EL even if they use the air-pressure sensor. Therefore, we use both spectral and F_0 information of the EL(air) speech for estimating more natural F_0 contours. In the VC part, the source data for the F_0 estimation are prepared as follows. Segmental feature vectors of the spectrum and the F_0 are independently constructed in the same manner as described

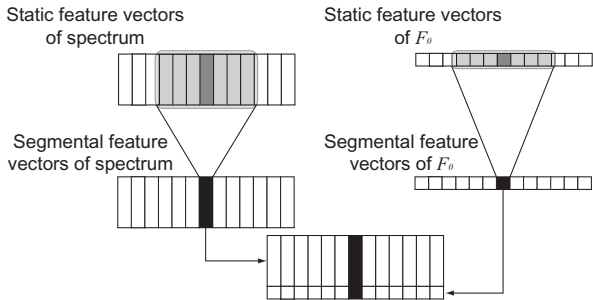


Figure 3: Flow chart of constructing segmental feature vectors using spectral and F_0 feature vectors.

in Section 3.1, and the source data are obtained by concatenating these vectors as Fig. 3 shows.

It is essential in VC to use source and target features that are correlating with each other. To obtain these data, a laryngectomee has trained how to control F_0 using the air-pressure sensor for one month. The laryngectomee has further trained to control F_0 for more three weeks so that the pitch of the EL(air) speech sounds similar to that of the target normal speech. After this training, EL(air) speech was recorded. However, we have noticed that it is too difficult to mimic the target pitch pattern by controlling F_0 with breathing air. Moreover, F_0 patterns of the recorded EL(air) speech are significantly different from those of the target normal speech. Therefore, we have additionally recorded target normal speech for the recorded EL(air) speech. In this recording, a target speaker has been asked to utter normal speech while mimicking the pitch patterns of the recorded EL(air) speech as naturally as possible. Note that the F_0 contours of the recorded EL(air) speech are still different from those of the re-recorded target normal speech as shown in Fig. 4. For example, an F_0 contour of the recorded EL(air) speech varies discontinuously; on the other hand, that of the target normal speech varies smoothly and naturally. These differences are removed by the VC in our proposed system.

5. Experimental evaluations

5.1. Experimental conditions

The source speaker was one laryngectomee (Japanese male), who was proficient in speaking with an EL. The target speaker was one non-laryngectomee (also a Japanese male). Both speakers recorded 50 phoneme-balanced sentences which served as our training data and 30 utterances of newspaper articles which served as our test data. The source speaker recorded three kinds of alaryngeal speech, which were EL speech, EL(small) speech using pulse train with 100 Hz, and EL(air) speech. EL(air) speech and normal speech were recorded using the method described in Section 4.

The number of mixture components of the GMMs to estimate spectrum, aperiodic components, and F_0 was set to 32, 32, and 64, respectively. The 0th through 24th mel-cepstral coefficients, which were extracted by mel-cepstral analysis [11], were used as the source spectral parameters in which the 0th coefficient captured power information. The concatenating frame length for the source segmental feature vectors was set to 8. After the concatenation of frames, 50- and 2-dimensional components were extracted frame by frame to construct spectral and F_0 segmental feature vectors, respectively. Acoustic features of the target speech were extracted by STRAIGHT analysis [12].

Mel-cepstral distortion measured the spectral conversion

Table 1: Mel-cepstral distortions without power information

Source speech	Source - Target [dB]	Converted - Target [dB]
EL(small) speech	11.42	4.55
EL speech	8.96	4.25
EL(air) speech	9.51	4.12

Table 2: Results of F_0 estimation accuracy

Source speech	Correlation \pm standard deviation	U/V decision error rate [%]
EL(small) speech	0.66 \pm 0.11	6.13
EL speech	0.58 \pm 0.15	5.44
EL(air) speech	0.78 \pm 0.08	4.73

accuracy. The F_0 accuracies were evaluated by unvoiced or voiced (U/V) decision error rates and the correlation coefficient between target and converted F_0 contours.

Six non-laryngectomees subjectively evaluated 1) intelligibility, 2) naturalness, and 3) preference, which were all rated using a five-point-scaled opinion score (1: Bad - 5: Excellent). Seven kinds of stimuli were evaluated: analysis-synthesized target normal speech, three kinds of recorded source speech signals (EL speech, EL(small) speech, and EL(air) speech), and three kinds of the converted speech signals from each source speech. When synthesizing the speech waveforms, the GV parameters of only the converted spectra were taken into account.

5.2. Experimental results

5.2.1. Objective results

Table 1 shows the results of mel-cepstral distortion. As the table shows, VC powerfully enhances spectral performance of the source speech. The results of EL and EL(air) speech conversion are much better than the results of the EL(small) speech conversion. This is because EL and EL(air) speech contain much more information than EL(small) speech.

As Table 2 shows, EL(air) speech conversion achieves higher correlation and less U/V decision errors than other results. These results demonstrate that the use of the air-pressure sensor effectively improves the F_0 estimation accuracy.

Fig. 4 shows an example of F_0 of the EL(air) speech, the converted speech, and the target speech, respectively. As this figure shows, VC powerfully works to make the F_0 contours of the EL(air) speech smoothly and continuously varying while suitably switching voiced or unvoiced decisions.

5.2.2. Subjective results

Fig. 5 shows the mean opinion score (MOS) for each test.

A Intelligibility

The intelligibility of source EL speech is higher scored than that of other converted speech signals. This is because the source speaker well knows how to produce intelligible EL speech. The degradation of the intelligibility by VC is future work. On the other hand, the scores of EL and EL(air) speech conversion stay at almost 3.5, and therefore, we believe that these results are acceptable.

B Naturalness

The naturalness of each converted speech is scored higher than each source speech. Moreover, the rating for converted

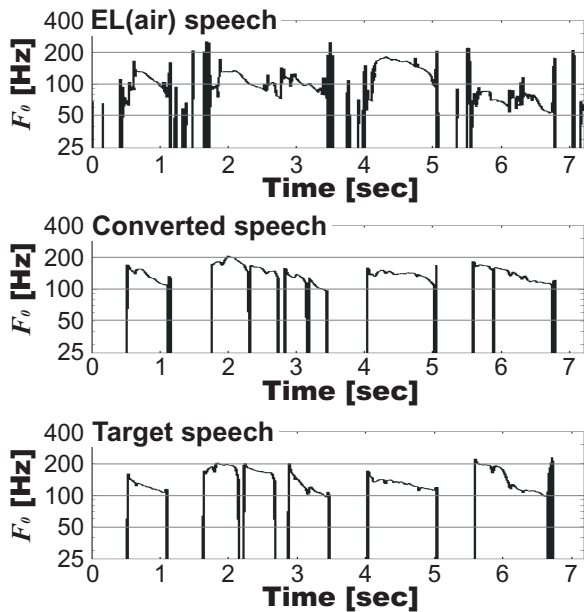


Figure 4: Example of F_0 contours.

speech from EL(air) speech has a higher score than that from EL speech. Therefore, the use of the air-pressure sensor is effective for improving the naturalness.

C Preference

The converted speech from the EL(air) speech is scored higher than that from the EL speech. This is because the improvement of the naturalness of the converted speech has affected the subjects much more than the degradation of intelligibility. From this result, we conclude that the use of the air-pressure sensor effectively improves the voice quality of the converted speech signals.

6. Conclusion

This paper introduced an air-pressure sensor and proposed a speaking-aid system to enhance EL(air) speech using a GMM-based VC method. It was shown that the results of this EL(air) speech conversion were better than those of EL speech conversion. Experimental results demonstrated the effectiveness of the use of the air-pressure sensor.

7. Acknowledgements

The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method. This research was supported in part by MIC SCOPE. This research was also supported in part by Grant-in-Aid for JSPS Fellows.

8. References

- [1] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Design of a New Electrolarynx Having a Pitch Control Function", Proceedings of 3rd IEEE International Workshop of Robot and Human Communication:198–203, Nagoya, Japan, July 1994.
- [2] K. Murakami, K. Araki, M. Hiroshige, and K. Tochinnai, "A Method for Speech Transform from Electrolaryngeal Speech to Normal Speech (in Japanese)", IEICE Trans. J87-D-I(11):1030–1040, Nov. 2004.

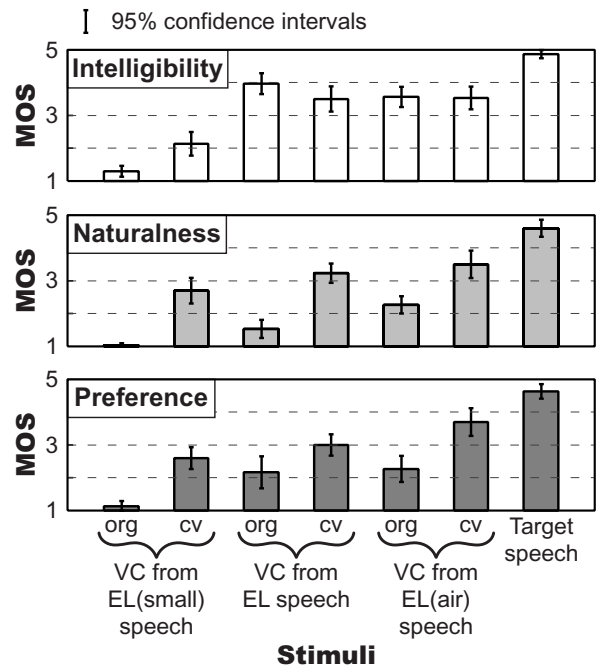


Figure 5: Subjective result by six non-laryngectomees. 'org' and 'cv' denote original and converted speech, respectively.

- [3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion", Proc. Interspeech 2009 - Eurospeech:1431–1434, Brighton, U.K., Sep. 2009.
- [4] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory", IEEE Trans. Audio, Speech and Language Proc. 15(8):2222–2235, Nov. 2007.
- [5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Transaction on Speech and Audio Proc. (SAP), 6(2):131–142, 1998.
- [6] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis", Proc. ICASSP 1998:285–288, Seattle, U.S.A., May 1998.
- [7] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. ICASSP 2000:1315–1318, Istanbul, Turkey, June 2000.
- [8] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "A Speech Communication Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech", IEICE Trans. Information and Systems, J90-D(3):780–787, 2007 (in Japanese).
- [9] Y. Hosoi and T. Sakaguchi, "Silent voice input system without exhalation -theory and applications-", Technical Report of IEICE, SP2003-105:13–16, 2003.
- [10] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Re-modeling of the Sensor for Non-Audible Murmur (NAM)", Proc. of Interspeech 2005:293–296, Lisbon, Portugal, Sept. 2005.
- [11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 137–140, San Francisco, USA, March 1992.
- [12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, 27(3–4):187–207, Apr. 1999.