Speaking-Aid Systems Based on One-to-Many Eigenvoice Conversion for Total Laryngectomees

Hironori Doi*, Keigo Nakamura*, Tomoki Toda*, Hiroshi Saruwatari* and Kiyohiro Shikano*
* Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma, Nara Japan
E-mail: hironori-d@is.naist.jp Tel: +81-743-72-5288

***Abstract*—This paper proposes speaking-aid systems based on one-to-many eigenvoice conversion (EVC) for enhancing three types of alaryngeal speech: esophageal speech; electrolaryngeal speech; and body-conducted silent electrolaryngeal speech. Although alaryngeal speech allows laryngectomees to utter speech sounds, it suffers from lack of naturalness and speaker individuality. To improve the sound quality of alaryngeal speech, alaryngeal-speech-to-speech (AL-to-Speech) methods based on statistical voice conversion have been proposed. To recover the speaker individuality of alaryngeal speech, a one-to-many EVC capable of flexibly adapting the conversion model to given target natural voices was applied to the AL-to-Speech methods. The experimental results of objective and subjective evaluations demonstrate that the proposed methods yield significant improvements of speech quality and make the converted voice quality similar to the given target voice quality.**

I. Introduction

People who have undergone a total laryngectomy due to an accident or laryngeal cancer cannot produce speech sounds because their vocal folds have been removed. Therefore, they require an alternative speaking method to produce speech sounds without vibration of their vocal folds. This generated speech is called alaryngeal speech.

There are several types of alaryngeal speech. As typical alaryngeal speech, esophageal speech (ES) and electrolaryngeal speech (EL) are widely used in Japan. Moreover, we have proposed another type of alaryngeal speech, body-conducted silent EL (silent EL) [1], which is produced with small-powered sound source signals and detected with a Non Audible Murmur (NAM) microphone [2], to keep the sound source signals generated from EL less audible. Each of these three types of alaryngeal speech have their own advantages in terms of usability or quality. However, they suffer from lack of naturalness and speaker individuality.

To improve the sound quality of these three types of alaryngeal speech, conversion methods from each alaryngeal speech into normal speech [1], [3], [4] have been proposed based on statistical voice conversion (VC) [5], [6], [7]. These approaches are called alaryngeal-speech-to-speech (AL-to-Speech) based on VC in this paper. In AL-to-Speech based on VC, Gaussian mixture models (GMMs) of the joint probability densities of acoustic features between alaryngeal speech and normal speech are trained in advance using parallel data consisting of dozens of utterance-pairs of those two types of speech data. The trained models are capable of converting the acoustic features of alaryngeal speech to those of normal speech in a probabilistic manner while keeping linguistic information unchanged. This technique yields significant improvements of speech quality since the converted speech is basically generated according to the statistical properties of acoustic features of normal speech. However, the converted voice quality is determined by the target natural voices used in the training. To use a user's own original voice before undergoing total laryngectomy as the target voice, parallel data of alaryngeal speech and the original voices are needed. However, very few laryngectomy patients prepare such data.

To allow laryngectomees to flexibly control the converted voice quality, this paper applies one-to-many eigenvoice conversion (EVC) [8] to AL-to-Speech. The one-to-many EVC is a conversion method from a single source speaker's voice into an arbitrary target speaker's voice. This method allows us to control the speaker individuality of the converted speech by manipulating a small number of parameters or flexibly adapting the conversion model to an arbitrary target speaker using only a small number of given target speech samples in a text-independent manner. Therefore, the proposed method is expected to flexibly recover speaker individually of laryngectomees even if only a few speech samples of their original voices are available. In our previous work [9], one-to-many EVC has been applied to ES and its effectiveness has been demonstrated. In this paper, we additionally apply one-to-many EVC to EL and silent EL as well as ES and evaluate its effectiveness assuming that only a few original speech samples are available for adaptation.

II. Alaryngeal speech

Figure 1 shows the three types of alaryngeal speech dealt with in this paper. Each of them has their own advantages. However, their sounds are relatively unnatural compared with natural voices and they also suffer from lack of speaker individuality.

A. Esophageal speech (ES)

One of the biggest advantages of ES is that it can be produced without any equipment. Alternative excitation sounds are produced by releasing gases from or through the esophagus, and then they are articulated to generate ES. ES usually sounds more natural than the other types of alaryngeal speech. However, its sound quality is much lower than normal voices. It often includes specific sounds produced through the production mechanism of ES.

Proceedings of the Second APSIPA Annual Summit and Conference, pages 498–501,
Biopolis, Singapore, 14-17 December 2010.

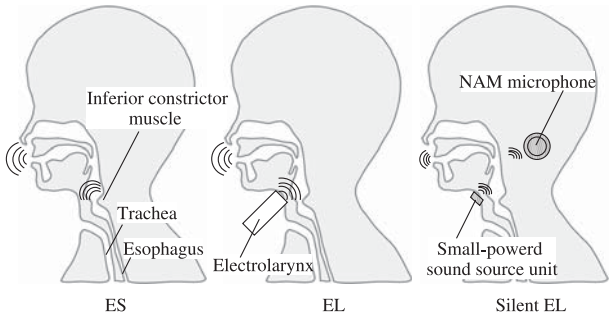

Fig. 1. Speaking methods for producing three types of alaryngeal speech (ES, EL, and silent EL).

B. Electrolaryngeal speech (EL)

EL is generated using an electrolarynx that is a medical device to mechanically generate the sound source signals, which are conducted into the oral cavity from the skin on the lower jaw. EL is easy to produce compared with ES. On the other hand, it is difficult to mechanically generate a naturally sounding F_0 contour, and therefore a monotone pitch is often generated. Consequently, it sounds very mechanical. Moreover, because the electrolarynx needs to generate enough loud sound source signals to make the produced speech sounds sufficiently audible, those signals are easily emitted outside and they would be perceived as noise by other people.

C. Body-conducted silent electrolaryngeal speech (silent EL)

To alleviate the issue of the noisy sound source signals caused by the electrolarynx, a novel device that generates extremely small-powered sound source signals has been proposed. People around the speaker are no longer annoyed by the generated sound source signals. On the other hand, the use of this device also makes the produced speech inaudible. Therefore, it is detected with NAM microphone, which is one of the body-conductive microphones capable of detecting extremely small signals in the vocal tract from the neck below the ear. The detected speech sounds are much more unnatural compared with EL due to the lower-powered excitation and body conduction.

III. One-to-many EVC

We describe one-to-many EVC as a technique for flexibly controlling voice quality of the converted speech. This method consists of a training process, an adaptation process, and a conversion process.

A. Training Process

As a conversion model, an eigenvoice GMM (EV-GMM) is trained using multiple parallel data sets consisting of a single source speech data set and many target speech data sets including various speakers' voices. Let us assume a source static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D_x)]^\top$ and a target static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ at frame t, where $\top$ denotes transposition of the vector. As a source speech parameter vector, we use a feature vector $\boldsymbol{X}_t$ to capture contextual features of the source speech, e.g., the joint static and dynamic feature vector or the concatenated feature vector from multiple frames. As a target speech feature vector, we use a feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of static and dynamic features. The EV-GMM models the joint probability density of the source and target parameter vectors as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}), \boldsymbol{\Sigma}_m^{(X,Y)}\right) \quad (1)$$

$$\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{A}_m\boldsymbol{w} + \boldsymbol{b}_m \end{bmatrix} \quad (2)$$

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (3)$$

where $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ is a target-speaker-dependent parameter for controlling target voice quality. $\boldsymbol{\lambda}^{(EV)}$ is a canonical EV-GMM parameter set consisting of α_m, $\boldsymbol{\mu}_m^{(X)}$, $\boldsymbol{\Sigma}_m^{(X,Y)}$, $\boldsymbol{A}_m$, and $\boldsymbol{b}_m$ for the m^{th} mixture component. $\boldsymbol{b}_m$ and $\boldsymbol{A}_m = [\boldsymbol{a}_m(1), \cdots, \boldsymbol{a}_m(j), \cdots, \boldsymbol{a}_m(J)]$ are a bias vector and eigenvectors, respectively. The number of eigenvectors is J.

Adaptive training [8] is used in the EV-GMM training. The canonical EV-GMM parameters and the target-speaker-dependent parameters are optimized by maximizing a total likelihood of the EV-GMM adapted to individual pre-stored target speakers as follows:

$$\{\hat{\boldsymbol{\lambda}}^{(EV)}, \hat{\boldsymbol{\omega}}_{(1:S)}\} = \underset{\boldsymbol{\lambda}^{(EV)}, \boldsymbol{\omega}_{(1:S)}}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{\omega}_s) \quad (4)$$

where $\hat{\boldsymbol{\lambda}}^{(EV)}$ denotes the updated canonical EV-GMM parameter set and $\hat{\boldsymbol{\omega}}_{(1:S)} = \{\hat{\boldsymbol{\omega}}_1, \cdots, \hat{\boldsymbol{\omega}}_S\}$ denotes a set of the updated weight vectors for the individual pre-stored target speakers. $\boldsymbol{Y}_t^{(s)}$ is the target feature vector of the s-th pre-stored target speaker at frame t.

B. Adaptation Process

The trained EV-GMM allows us to control the converted voice quality by manipulating the weight vector $\boldsymbol{w}$. If target speech data are available, the EV-GMM is flexibly adapted to the target speech by automatically determining the weight vector in a text-independent manner. The optimum weight vector is determined by

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(tar)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}_t \quad (5)$$

where $\{\boldsymbol{Y}_1^{(tar)}, \cdots, \boldsymbol{Y}_T^{(tar)}\}$ are a time sequence of the given target feature vectors. This adaptation process works well even if using only a few arbitrary utterances of the target speech.

C. Conversion Process

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ be a time sequence of the source and target feature vectors, respectively. The converted static feature vector sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing a likelihood

of the conditional probability density function of $\boldsymbol{Y}$ given $\boldsymbol{X}$ as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \qquad \text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{6}$$

where $\boldsymbol{W}$ is a window matrix to extend the static feature vector sequence into the joint static and dynamic feature vector sequence. Furthermore, the quality of the converted speech is dramatically improved by considering a global variance of the converted features [7].

IV. AL-to-Speech based on one-to-many EVC

To improve sound quality and recover speaker individuality of the three types of alaryngeal speech (ES, EL and silent EL), we propose the AL-to-Speech method based on one-to-many EVC. The proposed method allows the laryngectomees to utter speech sounds similar to their own normal voice quality even if only a few utterances of their previously recorded normal voices are available for the EV-GMM adaptation. Even if those voices are not available, the proposed method still allows the laryngectomees to change the converted voice quality by manipulating the weight parameter of the EV-GMM.

In AL-to-Speech, spectral segment feature vectors of alaryngeal speech are converted into multiple acoustic features of the target normal speech such as spectrum, aperiodic components [10], and F_0. The spectral segment feature vector is extracted with PCA from concatenated spectral feature vectors from multiple frames around a current analyzed frame. This feature is effective for improving conversion accuracy [4].

The proposed AL-to-Speech based on EVC also consists of training, adaptation and conversion processes. In the training process, two EV-GMMs for spectral estimation and aperiodic component estimation are independently trained using multiple parallel data sets consisting of alaryngeal speech data uttered by a laryngectomee and normal speech data uttered by many pre-stored non-laryngectomees. To build the GMM for F_0 estimation, first the target-speaker-dependent GMMs of joint probability density of the source spectral segment feature vectors and the target log-scaled F_0 are separately trained for all pre-stored target speakers. Then, the GMM yielding the best conversion accuracy in F_0 estimation is manually selected. In the adaptation process, the weight vectors of the EV-GMMs for the spectral and aperiodic component estimation are independently estimated using the spectral features and the aperiodic components extracted from the given target speech samples. In the conversion process, the converted spectral feature vectors and aperiodic components are independently estimated with the adapted EV-GMMs. On the other hand, an F_0 sequence is estimated with the selected target-speaker-dependent GMM. Then, it is further converted so that its mean μ_x and standard deviation σ_x on the log scale are equal to those of the adaptation speech data, μ_y and σ_y, as follows:

$$\log y_t = \frac{\sigma_y}{\sigma_x}(\log x_t - \mu_x) + \mu_y \tag{7}$$

where x_t and y_t denote the F_0 value estimated with the GMM and the converted F_0 value at frame t, respectively.

V. Experimental evaluations

To demonstrate the effectiveness of the proposed AL-to-Speech methods, we conducted experimental evaluations.

A. Experimental conditions

We recorded 50 phoneme-balanced sentences of ES uttered by one Japanese male laryngectomee and those of EL and silent EL uttered by another Japanese male laryngectomee, respectively. We also recorded the same sentences of normal speech uttered by 40 Japanese non-laryngectomees consisting of 27 male and 13 female speakers. Speech data of 30 non-laryngectomees consisting of 22 male and 8 female speakers were used for training and those of the other 10 non-laryngectomees consisting of 5 male and 5 female speakers were used as the target data for evaluation. From the recorded 50 sentences of each speaker, 40 sentences were used as training or adaptation data and the remaining 10 sentences were used as the test data. The sampling frequency was set to 16 kHz.

The 0-th through 24-th mel-cepstral coefficients were used as a spectral parameter. Mel-cepstrum analysis [11] was employed for alaryngeal speech and STRAIGHT analysis [12] was employed for normal speech. The shift length was 5 ms. To extract the spectral segment feature of ES, current and $\pm$ 8 frames were used for spectral and aperiodic estimation and current and $\pm$ 16 frames were used for F_0 estimation. For EL and silent EL, current and $\pm$ 8 frames were used for every parameter estimation. As the source excitation feature of normal speech, we used log-scaled F_0 values extracted with STRAIGHT F_0 extractor [13] and aperiodic components averaged on five frequency bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz) that were used for designing mixed excitation.

The EV-GMMs for spectral and aperiodic component estimation were trained in each type of alaryngeal speech. The number of eigenvectors was set to 29 in every EV-GMM. The number of mixture components was set to 64. The EV-GMMs were adapted to the target speakers using 1, 2, 4, 8, 16, or 32 utterances of their normal speech data. As a conventional approach (i.e., AL-to-Speech based on VC), we also trained the GMMs for spectral and aperiodic estimation using a parallel data set of each type of alaryngeal speech and normal speech of each target speaker. The number of utterance-pairs used in training was set to 1, 2, 4, 8, 16 or 32. The number of mixture components was optimized manually according to the training data size.

B. Objective evaluations

We evaluated the spectral estimation accuracy of each AL-to-Speech method with mel-cepstral distortion between the estimated and target mel-cepstra.

Figure 2 shows mel-cepstral distortion as a function of the number of adaptation utterances used in the proposed method or that of utterance-pairs used in the conventional method. The proposed method yields much better conversion accuracy compared with the conventional method when only a few utterances of the target normal speech are available. It is observed that the distortion caused by the proposed method using 16 adaptation utterances is almost equivalent to or

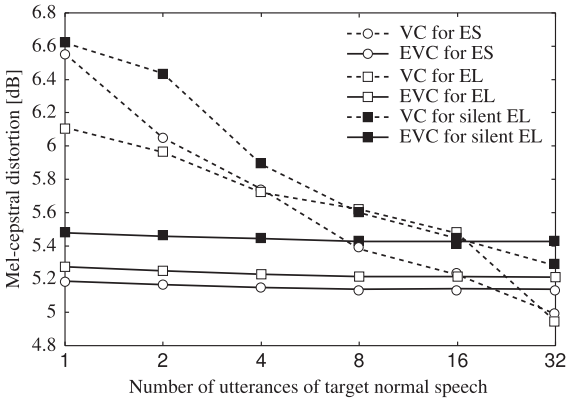

Fig. 2. Mel-cepstral distortion as a function of the number of utterances of target normal speech (i.e., utterance-pairs in VC or adaptation utterances in EVC).

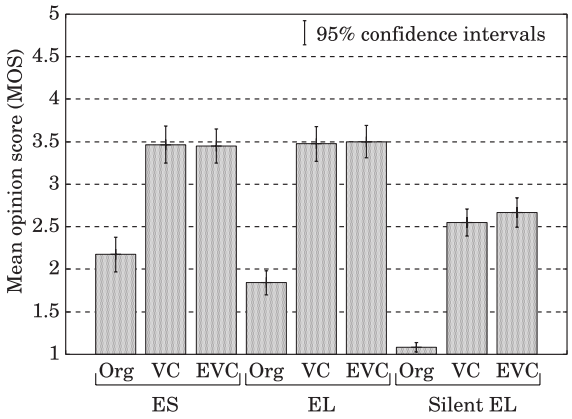

Fig. 3. Result of opinion test of speech quality. "Org", "VC", and "EVC" show original alaryngeal speech, converted speech by the conventional method, and converted speech by the proposed method, respectively.

smaller than that caused by the conventional method using 16 parallel utterance sets. Note that parallel data are not necessary in the proposed method since unsupervised adaptation using arbitrary utterances of the target normal speech is available.

C. Perceptual evaluations

We conducted an opinion test of speech quality. Ten listeners evaluated 9 types of speech samples including three types of alaryngeal speech (ES, EL, and silent EL), three types of converted speech with the conventional method (AL-to-Speech based on VC), and three types of converted speech with the proposed method (AL-to-Speech based on one-to-many EVC). The conventional method used 32 utterance-pairs for GMM training. On the other hand, only one utterance was used as adaptation data for the EV-GMMs in the proposed method. Each listener evaluated 135 speech samples.

Figure 3 shows the result of the test. All AL-to-Speech methods yield significant improvements of speech quality compared with the original alaryngeal speech. It is worthwhile to note that the proposed method achieves better speech quality than every type of original alaryngeal speech while keeping the external excitation sounds inaudible. The converted speech quality of the proposed method is equivalent to that of the conventional method in conversion of three types of alaryngeal speech. Note that the proposed method needs only one arbitrary utterance of the target normal speech while the conventional method needs 32 utterance-pairs of alaryngeal speech and the target normal speech.

VI. Conclusions

This paper has presented speaking-aid systems based on one-to-many EVC for three types of alaryngeal speech, esophageal speech, electrolaryngeal speech, and body-conducted silent electrolaryngeal speech. Our proposed methods are capable of converting alaryngeal speech into the target normal speech even if only one arbitrary utterance of the target speech is available. The experimental results have demonstrated that the proposed methods yield significant improvements of sound quality in every type of alaryngeal speech.

Acknowledgment

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

References

[1] K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari and K. Shikano, "Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments," *INTERSPEECH*, pp.2209–2212, Sep, 2008.

[2] Y. Nakajima, H. Kashioka, K. Shikano, N. Campbell, "Remodeling of the sensor for Non-Audible Murmur (NAM)," *INTERSPEECH*, pp.389–392, September 2005.

[3] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Electrolaryngeal speech enhancement based on statistical voice conversion," *INTERSPEECH*, pp. 1431–1434, Brighton, UK, Sep. 2009.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of esophageal speech using statistical voice conversion," *APSIPA*, pp. 805–808, Sapporo, Japan, Oct. 2009.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.

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

[7] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[8] Y. Ohtani, T. Toda, H. Saruwatari, K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Trans. Information and Systems*, vol. E93-D, no. 6, pp.1589–1598, June 2010.

[9] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "STATISTICAL APPROACH TO ENHANCING ESOPHAGEAL SPEECH BASED ON GAUSSIAN MIXTURE MODELS," *Proc. ICASSP*, pp. 4250–4253, Dallas, U.S.A., March 2010.

[10] 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 system STRAIGHT," *MAVEBA*, Florence, Italy, Sept. 2001.

[11] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.

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