Enhancement of Esophageal Speech Using Statistical Voice Conversion

Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192 Japan
E-mail: {hironori-d, kei-naka, tomoki, sawatari, shikano}@is.naist.jp Tel: +81-743-72-5288

***Abstract*—This paper presents a novel method of enhancing esophageal speech based on statistical voice conversion. Esophageal speech is one of the speaking methods for total laryngectomees. Although it allows laryngectomees to speak by generating a sound source and articulating it to produce audible speech sounds using their esophagus and vocal organs, the generated voices sound unnatural. To improve the naturalness of esophageal speech, we propose a voice conversion method from esophageal speech into normal speech (ES-to-Speech). A spectral parameter and excitation parameters, such as F_0 and aperiodic components, of normal speech are separately estimated from the spectral parameter of the esophageal speech in the sense of maximum likelihood using different Gaussian mixture models. We conduct objective and subjective evaluations of the proposed method. The experimental results demonstrate that the proposed method yields significant improvements in naturalness of esophageal speech while maintaining its intelligibility.**

I. Introduction

Speech has been used as the ordinary way for most people to communicate with each other. Unfortunately, it is not always available to everyone. For instance, people who have undergone a total laryngectomy because of an accident or laryngeal cancer, cannot produce speech sounds because their vocal cords have been removed. Thus, they require another method to produce speech sounds.

Esophageal speech is one of the speaking methods for laryngectomees. Alternative sounds are produced by releasing gases from or through the esophagus. Esophageal speech enables laryngectomees to speak without any equipment. Moreover, esophageal speech sounds closer to a natural voice compared with speech generated by other speaking methods such as using an electrolarynx. However, degradation of naturalness is caused by several factors such as specific sounds of the esophageal speech and relatively low fundamental frequency, F_0. Consequently, esophageal speech sounds unnatural compared with normal speech.

There have been proposed some attempts at enhancing esophageal speech by modifying its acoustic features, e.g., using the comb filtering process [1] or a smoothing process [2]. However, since the acoustic features of esophageal speech exhibit quite different properties from those of normal speech, it is basically difficult to compensate for the acoustic differences between them using simple modification processes. Therefore, their effectiveness is limited.

In this paper, we propose a method of enhancing esophageal speech based on statistical voice conversion [3], [4] from esophageal speech into normal speech (Esophageal-Speech-to-Speech: ES-to-Speech). In the proposed method, we train Gaussian mixture models (GMMs) of the joint probability densities between the acoustic features of esophageal speech and those of normal speech using parallel data consisting of utterance-pairs of esophageal speech and target normal speech. The enhanced speech exhibiting properties of the target normal speech is generated by converting the acoustic features of esophageal speech into those of the target in the sense of maximum likelihood using the trained GMMs without any linguistic features. We conduct objective and subjective evaluations of the proposed method. The experimental results demonstrate that naturalness of esophageal speech is significantly improved by the proposed method.

This paper is organized as follows. In section II, we describe characteristics of the esophageal speech. In section III, we describe the statistical voice conversion algorithm used in this paper. In section IV, we describe the proposed enhanced method. In section V, experimental evaluations are presented. Finally, we summarize this paper in the section VI.

II. Esophageal Speech

Figure 1 shows air flows from lungs in non-laryngectomees and total laryngectomees, respectively. In laryngectomees, the trachea and the oral cavity connecting the esophagus are completely separated from each other to prevent food from entering the trachea. Therefore, laryngectomees not only cannot generate vocal fold vibration but also expire air through the oral cavity. Consequently, they have to produce speech sounds in alternative ways.

Esophageal speech is generated as follows. First, speakers pump a certain amount of air from the mouth into the esophagus and the stomach, which play the role of the respiratory organs. Next, releasing the air from them, sound excitations are generated by vibrating tissues around the entrance of the esophagus. Finally, the esophageal speech is produced by articulating the generated sounds in the same manner as by non-laryngectomees.

Figure 2 shows an example of speech waveform, spectrogram, and F_0 of normal speech and esophageal speech in the same sentence. We can see that acoustic features of the esophageal speech are considerably different from those of normal speech. Some specific sounds of esophageal speech are shown in silent parts. They are often produced through preparation for generating sounds, e.g., pumping air into the

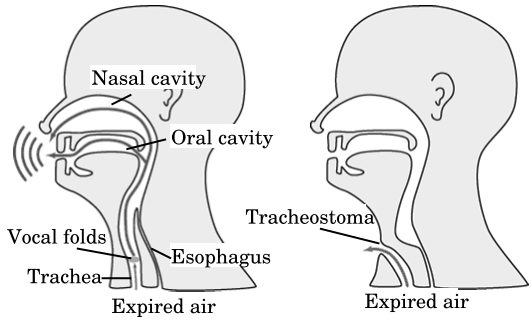

Fig. 1. Air flows from lungs in non-laryngectomee (left) and total laryngectomee (right).

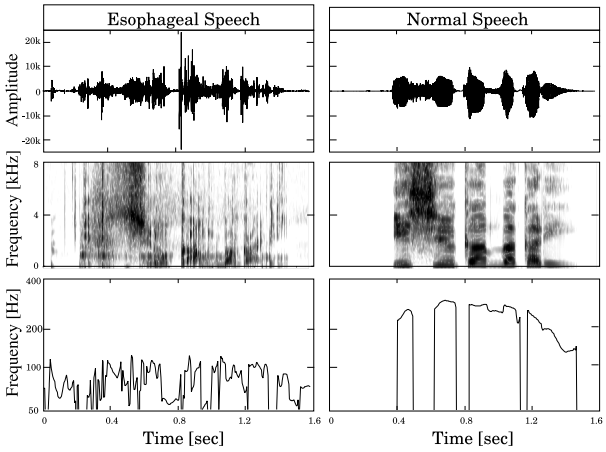

Fig. 2. Example of waveforms, spectrograms and F_0 counters of both esophageal and normal speech. F_0 values of esophageal speech have been extracted as described in section IV.

esophagus. Waveform envelope, spectral components, and F_0 of esophageal speech don't vary over an utterance as smoothly as those of normal speech. These unstable and unnatural variations cause specific noisy sounds of esophageal speech. These characteristics of esophageal speech make a feature extraction process, e.g., F_0 extraction and voiced/unvoiced decision, much more difficult compared with normal speech, and therefore, they also cause severe quality degradation of analysis-synthesized esophageal speech. In addition, some phonemes (e.g., glottal fricatives such as /h/) are difficult to produce in esophageal speech. Moreover, F_0 of esophageal speech is much lower than that of normal speech.

It depends on the level of skill of each laryngectomee as to how large the acoustic differences are between normal speech and esophageal speech. However, some differences are very critical because they are caused by the production mechanism of esophageal speech.

III. Voice Conversion Algorithm Based on Maximum Likelihood Estimation

Here we describe a conversion method based on maximum likelihood estimation of speech parameter trajectories considering a global variance (GV) [4] as one of the state-of-the-art statistical voice conversion methods.

A. Training Process

Let us assume an input static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D_x)]^\top$ and an output static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ at frame t, respectively. As an input speech parameter vector, we use $\boldsymbol{X}_t$ capturing segmental features of input speech, e.g., the joint static and dynamic feature vector or the concatenated feature vector from multiple frames [5]. As an output speech parameter vector, we use $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of both static and dynamic feature vectors.

Using a parallel training data set consisting of time-aligned input and output parameter vectors $[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top$, $[\boldsymbol{X}_2^\top, \boldsymbol{Y}_2^\top]^\top$, $\cdots$, $[\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]^\top$, the joint probability density of the input and output parameter vectors is modeled by a GMM [6] as follows:

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

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. A parameter set of the GMM is $\boldsymbol{\lambda}$, 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.

The probability density of the GV $\boldsymbol{v}(\boldsymbol{y})$ of the output static feature vectors $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ over an utterance is also modeled by a Gaussian distribution,

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}) \quad (3)$$

where the GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D_y)]^\top$ is calculated by

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \quad (4)$$

A parameter set $\boldsymbol{\lambda}^{(v)}$ consists of a mean vector $\boldsymbol{\mu}^{(v)}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}^{(v)}$.

B. Conversion Process

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ be a time sequence of the input parameter vectors and that of the output parameter vectors, respectively. The converted static feature sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing a product of the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the GV probability density,

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \quad \text{subject to}\, \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \quad (5)$$

where $\boldsymbol{W}$ is a window matrix to extend the static feature vector sequence into the joint static and dynamic feature vector sequence [7]. A balance between $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ and $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ is controlled by the weight ω.

IV. Voice Conversion from Esophageal Speech to Speech (ES-to-Speech)

In esophageal speech, the spectral components vary unstably and acoustic characteristics of some phonemes are not observed well as mentioned in Section II. To alleviate these fluctuations and compensate for missing acoustic characteristics, we extract a spectral segment feature from multiple frames. At each frame, spectral parameter vectors at current, preceding and succeeding frames are concatenated, and then, dimension reduction with principal component analysis (PCA) is performed to extract the spectral segment feature. Moreover, it is also difficult to extract F_0 from the esophageal speech because most frames are often determined as unvoiced and the extracted F_0 values vary widely and discontinuously when using a usual F_0 extractor. To extract informative F_0 as an input feature for the conversion process, we extract several F_0 candidates at each frame, and then we select the F_0 candidate closest to the average F_0 value calculated over F_0 values previously determined at several preceding frames (5 frames in this paper).

In order to convert input esophageal speech into normal speech, we use three types of GMMs for estimating output spectrum, F_0, and aperiodic components capturing noise strength on each frequency band of an excitation signal [10], respectively. For the spectral estimation, we use a GMM for converting the input spectral segment feature into the output spectral parameter. For the aperiodic estimation, we use a GMM for converting the input spectral segment feature into the output aperiodic components. For the F_0 estimation, we use a GMM for converting the input spectral segment feature into the output F_0. As an alternative method, we also use a GMM for converting the segment feature extracted from joint vectors of the input spectral parameter and F_0 at multiple frames into the output F_0. In synthesizing the converted speech, first we design a mixed excitation based on the estimated F_0 and aperiodic components [10]. Then, we synthesize the converted speech by filtering the mixed excitation with the estimated spectral feature.

V. Experimental Evaluations

In order to demonstrate the effectiveness of ES-to-Speech, we conducted experimental evaluations.

A. Experimental Conditions

We recorded 50 sentences of esophageal speech uttered by one Japanese male laryngectomee. Then, we recorded the same sentences of normal speech uttered by another Japanese male speaker. In order to make the target normal speech acoustically correlate with the source esophageal speech, the speaker tried imitating the prosody of the laryngectomee as well as he could by carefully listening to the recorded laryngectomee's speech samples utterance by utterance. Sampling frequency was set to 16 kHz.

We conducted a 5-fold cross validation test in which 40 utterance-pairs out of the recorded data were used for training, and the other 10 utterances were used for the test.

TABLE I
Estimation accuracy of F_0. Correlation coefficient between F_0 extracted from esophageal speech and that from normal speech is 0.12.

Input Feature	Correlation	Voiced/Unvoiced Error [%]
Spectrum	0.68	8.36 ($V \rightarrow U : 4.30, U \rightarrow V : 4.05$)
Spectrum & F_0	0.68	8.39 ($V \rightarrow U : 4.35, U \rightarrow V : 4.04$)

The 0th through 24th mel-cepstral coefficients extracted with STRAIGHT analysis [8] were used as the spectral parameter. As the source excitation features, we used a log-scaled F_0 extracted with STRAIGHT F_0 extractor [9] and aperiodic components [10] on five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, which were used for designing mixed excitation. The shift length was 5 ms.

We optimized several parameters such as the number of mixture components of each GMM and the number of frames used for extracting the spectral segment feature so that estimation accuracy was improved. As a result, we set the number of mixture components to 32 for each of the spectral estimation, the aperiodic estimation, and the F_0 estimation and set the number of input frames to current ± 8 for both the spectral estimation and the aperiodic estimation and to current ± 16 for the F_0 estimation, respectively.

B. Objective Evaluations

We evaluated input features for the F_0 estimation and an impact of the segment features on the spectral estimation.

Table I shows correlation coefficients of F_0 between normal speech and the converted speech and errors of voiced/unvoiced decision. Even if using both spectral and F_0 features, estimation accuracy for F_0 and voiced/unvoiced decision is almost equal to that when using only spectral features. Therefore, we used only the spectral feature for the F_0 estimation in the latter experimental evaluations.

Figure 3 shows mel-cepstral distortion at each phoneme category. The use of the spectral segment features is effective for improving spectral estimation accuracy. We can see that estimation accuracy is considerably improved for some phonemes such as liquid, unvoiced plosive, and voiced fricative. Although an increase of the number of input frames tends to make mel-cepstral distortion decrease, the use of too many input frames causes the degradation of mel-cepstral distortion due to the information loss by dimension reduction with PCA.

C. Perceptual Evaluations

We conducted two opinion tests on intelligibility and naturalness, respectively. The following five kinds of speech were evaluated.

$\boldsymbol{ES}$: recorded esophageal speech.
$\boldsymbol{NS}$: recorded normal speech.
$\boldsymbol{EstSpq}$: synthetic speech using the converted mel-cepstrum and the extracted F_0 of esophageal speech.
$\boldsymbol{EstF_0}$: synthetic speech using the extracted mel-cepstrum of esophageal speech and the converted F_0.
$\boldsymbol{EstSpq - EstF_0}$: synthetic speech using the converted mel-cepstrum and the converted F_0.
Note that the estimated aperiodic components were used for

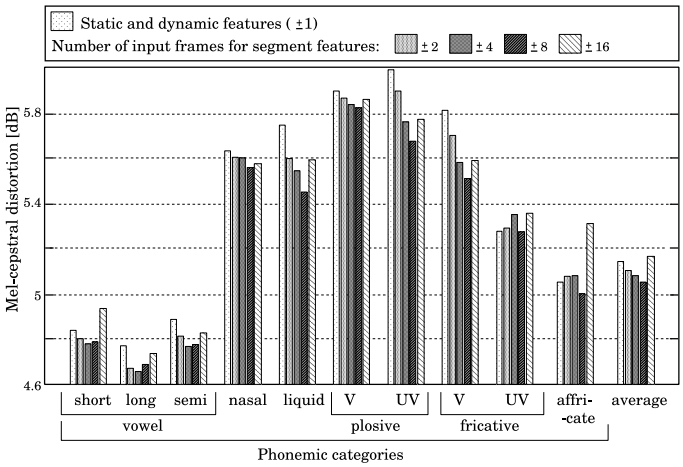

Fig. 3. Estimation accuracy of mel-cepstrum. The notation "V" denotes voiced phonemes, and "UV" denotes unvoiced phonemes.

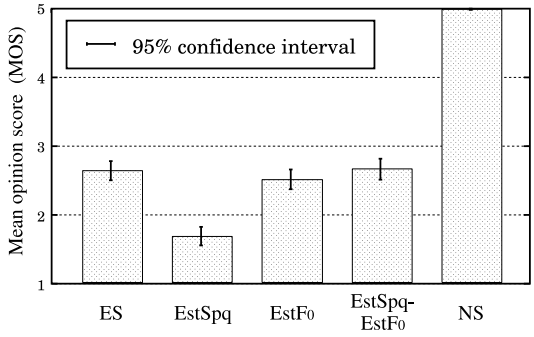

Fig. 4. Mean opinion score on intelligibility.

synthesizing all three kinds of synthetic speech using the mixed excitation.

1) Experimental Result on Intelligibility: Figure 4 shows the result of the intelligibility test. The proposed method (EstSpq-EstF_0) maintains intelligibility of the recorded esophageal speech (ES). It is found that the spectral conversion doesn't significantly affect intelligibility from a comparison between EstSpq-EstF_0 and EstF_0, although we have observed that the specific noisy sounds of esophageal speech are removed well by the spectral conversion. On the other hand, we can see that significant degradation of intelligibility is caused when directly using F_0 extracted from esophageal speech for synthesizing speech (EstSpq). This degradation would be caused by the difficulty of F_0 extraction for esophageal speech.

2) Experimental Result on Naturalness: Figure5 shows the result of the naturalness test. Naturalness is slightly improved by the F_0 estimation (EstF_0). If the F_0 estimation is not performed, severe degradation of naturalness is caused by the use of F_0 extracted from esophageal speech (EstSpq) as observed in the intelligibility test. We can see that a large improvement in naturalness is yielded by further performing the spectral conversion as well as the F_0 estimation (EstSpq-EstF_0) in order to remove the specific noisy sounds of esophageal speech.

These results suggest that the proposed method (EstSpq-EstF_0) is very effective for improving the naturalness of esophageal speech while keeping intelligibility equal.

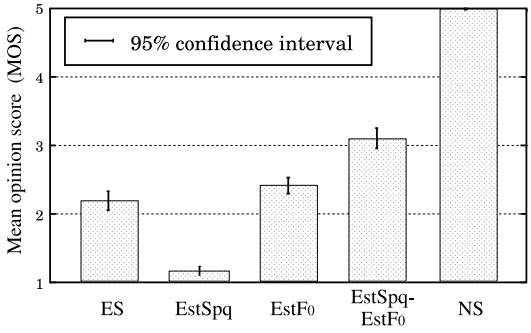

Fig. 5. Mean opinion score on naturalness.

VI. Conclusions

This paper has presented a novel method for enhancing esophageal speech based on statistical voice conversion from esophageal speech into normal speech (ES-to-Speech). Spectrum, aperiodic components, and F_0 of normal speech are separately estimated from a spectral segment feature of the esophageal speech in the sense of maximum likelihood using individual Gaussian mixture models. We have conducted objective evaluations and subjective evaluations. The experimental results have demonstrated that ES-to-Speech yields significant improvements in naturalness of esophageal speech while maintaining its intelligibility.

Acknowledgments

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] A. Hisada, H. Sawada. "Real-time clarification of esophageal speech using a comb filter," International Conference on Disability, Virtual Reality and Associated Technologies, pp.39–46, 2002.

[2] K. Matui, N. Hara, N. Kobayashi, H. Hirose. "Enhancement of esophageal speech using formant synthesis," *Proc. ICASSP*, pp. 1831–1834, Phoenix, Arizona, May. 1999

[3] 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.

[4] 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.

[5] T. Toda, A.W. Black, and K. Tokuda, "Statistical Mapping between Articulatory Movements and Acoustic Spectrum with a Gaussian Mixture Model," *Speech Communication*, Vol. 50, No. 3, pp. 215–227, Mar. 2008.

[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] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[8] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne. "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*, Vol. 27, No. 3-4, pp. 187–207, 1999.

[9] H. Kawahara, H. Katayose, A. Cheveigne and R. D. Patterson. "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.

[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 2001*, Florence, Italy, Sept. 2001.