# STATISTICAL APPROACH TO ENHANCING ESOPHAGEAL SPEECH BASED ON GAUSSIAN MIXTURE MODELS

*Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano*

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

## ABSTRACT

This paper presents a novel method of enhancing esophageal speech using statistical voice conversion. Esophageal speech is one of the alternative speaking methods for laryngectomees. Although it doesn't require any external devices, generated voices sound unnatural. To improve the intelligibility and naturalness of esophageal speech, we propose a voice conversion method from esophageal speech into normal speech. A spectral parameter and excitation parameters of target normal speech are separately estimated from a spectral parameter of the esophageal speech based on Gaussian mixture models. The experimental results demonstrate that the proposed method yields significant improvements in intelligibility and naturalness. We also apply one-to-many eigenvoice conversion to esophageal speech enhancement for flexibly controlling enhanced voice quality.

***Index Terms*—** laryngectomees, esophageal speech, speech enhancement, voice conversion, eigenvoice conversion.

## 1. 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. Esophageal speech is one of the alternative speaking methods for laryngectomees. Excitation sounds are produced by releasing gases from or through the esophagus, and then they are articulated for generating esophageal speech. Esophageal speech sounds more natural than speech generated by other alternative speaking methods such as electrolaryngeal speech. However, degradation of the intelligibility and naturalness in esophageal speech is caused by several factors such as specific noises and relatively low fundamental frequency. Consequently, esophageal speech sounds unnatural compared with normal speech.

Some approaches have been proposed to enhance esophageal speech based on the modification of its acoustic features, e.g., using comb filtering [1] or smoothing [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 those simple modification processes. More sophisticated and complicated processes are essential to dramatically enhance esophageal speech.

In this paper, we propose a statistical approach to enhancing esophageal speech using voice conversion (VC) [3, 4]. Our proposed method converts esophageal speech into normal speech in a probabilistic manner (Esophageal-Speech-to-Speech: ES-to-Speech). We train Gaussian mixture models (GMMs) of the joint probability densities between the acoustic features of esophageal speech and those of normal speech in advance using parallel data consisting of utterance-pairs of the esophageal speech and the target normal speech. Any esophageal speech sample is converted using the trained GMMs so that it sounds like normal speech. Because the converted speech is generated from statistics extracted from normal speech, this conversion process is expected to remove the specific noise sounds effectively and improve the $F_0$ pattern of the esophageal speech. However, the converted speech basically sounds like voices uttered by a different speaker from the laryngectomee. To make it possible to flexibly control the converted voice quality, we also apply one-to-many eigenvoice conversion (EVC) [5] to ES-to-Speech. The one-to-many EVC is a conversion method from a single source speaker into any arbitrary target speakers. This method allows us to control the converted voice quality by manipulating a small number of parameters or to flexibly adapt the conversion model for the given speech samples. This method seems very effective for helping laryngectomees to speak in their favorite voices or special voices sounding like their own voices that have already been lost.

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.

## 2. ESOPHAGEAL SPEECH

Figure 1 shows an example of speech waveforms, spectrograms, and $F_0$ contours of normal speech uttered by a non-laryngectomee and esophageal speech uttered by a laryngectomee in the same sentence. We can see that acoustic features of esophageal speech are considerably different from those of normal speech. Esophageal speech often includes some specific noisy sounds, which can be easily observed in the silence parts in the figure. These noises are produced through a process of generating excitation sounds, i.e., pumping air into the esophagus and the stomach and releasing air from them. Waveform envelope and spectral components of esophageal speech don't vary over an utterance as smoothly as those of normal speech. These unstable and unnatural variations cause the unnatural sounds of esophageal speech. Moreover, the pitch of esophageal speech is generally lower and less stable than that of normal speech. Consequently, a usual $F_0$ analysis process for normal speech often fails at $F_0$ extraction and the unvoiced/voiced decision in esophageal speech. These characteristics of esophageal speech cause severe degradation of analysis-synthesized speech quality.

The intelligibility and naturalness of esophageal speech strongly depend on the skill of individual laryngectomees in producing esophageal speech. However, some specific noises are essentially difficult to remove because they are caused by the production mechanism of esophageal speech.

## 3. VOICE CONVERSION ALGORITHM

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 VC methods. This method consists of a training and conversion process. Moreover, we also describe one-to-many EVC [5] as a technique for flexibly controlling the converted speech quality.

### 3.1. 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 $\boldsymbol{y}_t = [y_t(1), \cdots,$

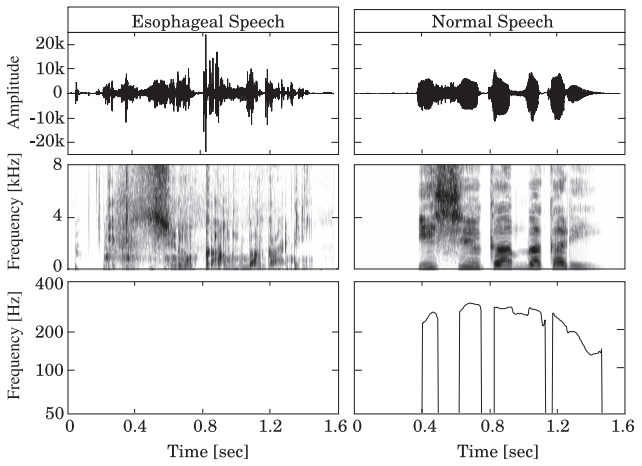


**Fig. 1**. Example of waveforms, spectrograms and $F_0$ contours of both esophageal and normal speech.

$y_t(D_y)]^\top$ at frame $t$, where $\top$ denotes transposition of the vector. As an input speech parameter vector, we use $\boldsymbol{X}_t$ to capture contextual features of source speech, e.g., the joint static and dynamic feature vector or the concatenated feature vector from multiple frames. As an output speech feature vector, we use $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of static and dynamic features.

Using a parallel training data set consisting of time-aligned input and output parameter vectors $[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top$, $[\boldsymbol{X}_2^\top, \boldsymbol{Y}_2^\top]^\top$, $\cdots$, $[\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]^\top$, where $T$ denotes the total number of frames, 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 $\boldsymbol{\Sigma}^{(v)}$.

### 3.2. 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 and the output feature 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 the following objective function,

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

where $\boldsymbol{W}$ is a window matrix to extend the static feature vector sequence into the joint feature vector sequence of static and dynamic features [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 $\omega$.

### 3.3. One-to-Many EVC

An eigenvoice GMM (EV-GMM) models the joint probability density in the same manner as shown in Eqs. (1) and (2), except for a definition of the target mean vector written as

$$\boldsymbol{\mu}_m^{(Y)} = \boldsymbol{A}_m \boldsymbol{w} + \boldsymbol{b}_m, \quad (6)$$

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

The EV-GMM is trained using multiple parallel data sets consisting of a single input speech data set and many output speech data sets including various speakers' voices. The trained EV-GMM allows us to control the converted voice quality by manipulating the weight vector $\boldsymbol{w}$. Moreover, the GMM for the input speech and new target speech are flexibly built by automatically determining the weight vector $\boldsymbol{w}$ using only a few arbitrary utterances of the target speech in a text-independent manner.

## 4. VOICE CONVERSION FROM ESOPHAGEAL SPEECH TO SPEECH (ES-TO-SPEECH)

In order to significantly enhance esophageal speech, it is essential to remove the specific noise sounds and to generate smoothly varying speech parameters over an utterance. Moreover, in order to synthesize enhanced speech with modified speech parameters, it is also essential to deal with difficulties of feature extraction of esophageal speech. To address these issues, we propose statistical voice conversion from esophageal speech into normal speech. Because converted speech parameters are generated from the statistics of the normal speech, the specific noise sounds and unstable variations are alleviated effectively by the conversion process. Furthermore, even if some speech parameters such as $F_0$ and unvoiced/voiced information are difficult to extract from esophageal speech, those parameters exhibiting properties similar to those of normal speech would be estimated by the conversion from other speech parameters robustly extracted from esophageal speech (e.g., spectral envelope). This estimation process may be regarded as a statistical feature extraction process.

### 4.1. Feature Extraction in ES-to-Speech [8]

The spectral components of esophageal speech vary unstably as mentioned in **Section 2**. Moreover, spectral features of some phonemes are often collapsed due to difficulties of producing them in esophageal speech. To alleviate these issues, we use a spectral segment feature extracted from multiple frames. At each frame, a spectral parameter vector at the current frame and those at several preceding and succeeding frames are concatenated. Then, dimension

reduction with principal component analysis (PCA) is performed to extract the spectral segment feature.

Although it is difficult to extract $F_0$ from esophageal speech (see Figure 1), we usually perceive pitch information of esophageal speech. Assuming that relevant information is included in spectral parameters, we use the spectral segment feature as an input feature for estimating $F_0$ in the conversion process. Moreover, in order to make the estimated $F_0$ correspond to the perceived pitch information of esophageal speech, as an output feature we use $F_0$ values extracted from normal speech uttered by a non-laryngectomee so as to make its prosody similar to that of esophageal speech.

### 4.2. Training and Conversion

In order to convert esophageal speech into normal speech, we use three different GMMs to estimate three speech parameters of the target normal speech, i.e., spectrum, $F_0$, and aperiodic components that capture noise strength of an excitation signal on each frequency band [9]. For the spectral estimation, we use a GMM to convert the input spectral segment features into the corresponding output spectral parameters. For the $F_0$ estimation, we use a GMM to convert the input spectral segment features into the output $F_0$ values and unvoiced/voiced information. For the aperiodic estimation, we use a GMM to convert the input spectral segment features into the output aperiodic components.

In synthesizing the converted speech, we design a mixed excitation based on the estimated $F_0$ and aperiodic components [9]. Then, we synthesize the converted speech by filtering the mixed excitation with the estimated spectral parameters.

### 4.3. Applying One-to-Many EVC to ES-to-Speech

We further apply one-to-many EVC to ES-to-Speech for flexibly controlling converted voice quality. The one-to-many EV-GMM for spectral conversion is trained using multiple parallel data sets of esophageal speech data uttered by the laryngectomee and normal speech data uttered by many non-laryngectomees. The trained EV-GMM allows laryngectomees to speak in their favorite voices, which are created by manipulating the weight vector for eigenvoices or by estimating proper weight values using only a small amount of those voices' data as adaptation if they are given.

## 5. EXPERIMENTAL EVALUATIONS

### 5.1. Experimental Conditions

We recorded 50 phoneme-balanced sentences of esophageal speech uttered by one Japanese male laryngectomee. We also recorded the same sentences of normal speech uttered by a Japanese male non-laryngectomee. He tried to imitate the prosody of the laryngectomee utterance-by-utterance as closely as possible. Sampling frequency was set to 16 kHz. We conducted a 5-fold cross validation test in which 40 utterance-pairs were used for training, and the remaining 10 utterance-pairs were used for evaluation.

The 0th through 24th mel-cepstral coefficients extracted with STRAIGHT analysis [10] were used as the spectral parameter. As the source excitation features of normal speech, we used log-scaled $F_0$ extracted with STRAIGHT $F_0$ analysis [11] and aperiodic components [9] 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 preliminarily 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 [8]. As a result, we set the number of mixture components to 32 for each of three GMMs. For the segment feature extraction, we used the current $\pm$ 8 frames in both the spectral estimation and the aperiodic estimation and the current $\pm$ 16 frames in the $F_0$ estimation, respectively.

**Table 1**. *Estimation accuracy of mel-cepstrum without power and aperiodic components. Mel-cepstral distortion with power (i.e., including the 0th coefficient) is shown in parentheses.*

| | Mel-cepstral distortion [dB] | Aperiodic distortion [dB] |
|---|---|---|
| Extracted | 8.46 (12.95) | 6.99 |
| Converted | 4.96 (6.26) | 3.71 |

**Table 2**. *$F_0$ correlation coefficient (Corr.) between the extracted/converted $F_0$ and the target $F_0$ extracted from normal speech and unvoiced/voiced decision error. For example, "VU" shows the rate of estimating a voiced frame as an unvoiced.*

| | Corr. | U/V decision error [%] |
|---|---|---|
| Extracted | 0.07 | 43.82 ($VU$: 42.60, $UV$: 1.22) |
| Converted | 0.68 | 8.36 ($VU$: 4.06, $UV$: 4.30) |

### 5.2. Objective Evaluations

Tables 1 and 2 show estimation accuracy of spectrum, aperiodic components, and $F_0$. It is observed that the acoustic features of esophageal speech are very different from those of normal speech. These large differences of the acoustic features are significantly reduced by ES-to-Speech. We can see that the proposed conversion method is very effective for estimating any of the three acoustic features.

### 5.3. Perceptual Evaluations

We conducted two opinion tests of intelligibility and naturalness. The following six types of speech samples were evaluated by 10 listeners.

**ES** recorded esophageal speech
**ES-AS** analysis-synthesized esophageal speech
**EstSpg** synthetic speech using converted mel-cepstrum, converted aperiodic components, and $F_0$ extracted from esophageal speech
**Est$F_0$** synthetic speech using extracted mel-cepstrum, extracted aperiodic components, and converted $F_0$
**CV** synthetic speech using converted mel-cepstrum, converted aperiodic components, and converted $F_0$
**NS-AS** analysis-synthesized normal speech

Each listener evaluated 120 samples[1] in each of the two tests.

Figures 2 and 3 show the result of the intelligibility test and that of the naturalness test, respectively. ES-AS causes significant intelligibility degradation compared to ES due to the difficulties of the acoustic feature extraction in esophageal speech. The specific noises and unstable variations on the spectrogram in esophageal speech are significantly alleviated by using the estimated spectral features. Moreover, the converted speech exhibiting pitch information similar to pitch perceived in esophageal speech is generated by using the estimated $F_0$ contour. Although significant improvements in intelligibility and naturalness are not observed when using only one of these estimated features, we can see that the ES-to-Speech (CV) estimating all acoustic features yields significantly more intelligible and natural speech than esophageal speech.

These results suggest that the proposed ES-to-Speech is very effective for improving both the naturalness and the intelligibility of esophageal speech.

[1] Several samples are available from http://spalab.naist.jp/ hironori-d/ICASSP/ES2SP/index.html

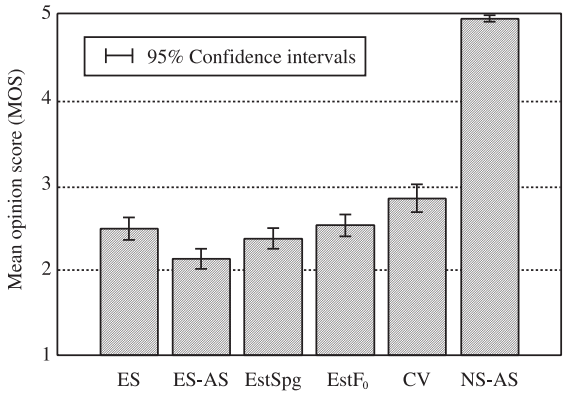


**Fig. 2**. Mean opinion score on intelligibility.

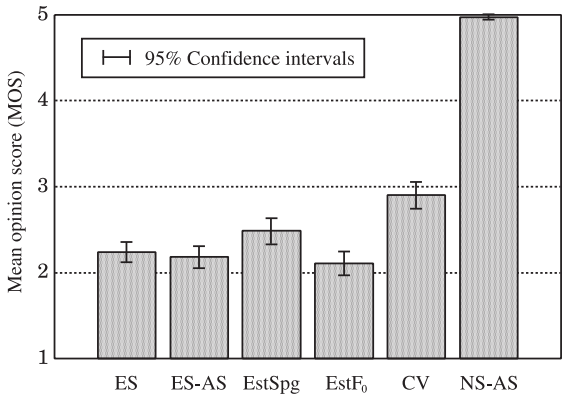


**Fig. 3**. Mean opinion score on naturalness.

### 5.4. Effectiveness of One-to-many VC

To make voice quality of the converted speech similar to the laryngectomee's own voice quality, we applied one-to-many EVC to ES-to-Speech. We trained the one-to-many EV-GMM using parallel data sets consisting of the esophageal speech and 30 speakers' normal speech. The EV-GMM was adapted to an esophageal speech sample shown in Figure 1, and then the esophageal speech sample was converted into normal speech using the adapted EV-GMM.

Figure 4 shows an example of waveform, spectrogram, and $F_0$ of the converted speech. We can see that 1) acoustic features of the converted speech vary more stably than those of the original esophageal speech sample and 2) the specific noise sounds are significantly alleviated by the conversion process. Namely, even if esophageal speech is used as the adaptation data, the adapted EV-GMM provides the converted speech of which properties are similar to those of normal speech. In addition, we have observed that the adapted EV-GMM makes the converted voice quality closer to the laryngectomee's voice quality compared with the GMM used in the previous evaluations, which were trained using the single non-laryngectomee's speech. Furthermore, we have also observed that the converted voice quality is flexibly changed by manipulating the weight values of the EV-GMM. The proposed ES-to-Speech with one-to-many EVC is expected to make possible a new speaking aid system that allows laryngectomees to control the converted voice quality as they want.

## 6. CONCLUSION

This paper has presented a novel method for enhancing esophageal speech using statistical voice conversion. The proposed method (ES-to-Speech) converts a spectral segment feature of esophageal speech into spectrum, $F_0$, and aperiodic components of normal speech independently using three different GMMs. The experimental results have demonstrated that ES-to-Speech yields significant improvements in intelligibility and naturalness of esophageal speech. Moreover, we have also applied one-to-many eigenvoice conversion to ES-to-Speech for flexibly controlling voice quality of the converted speech.

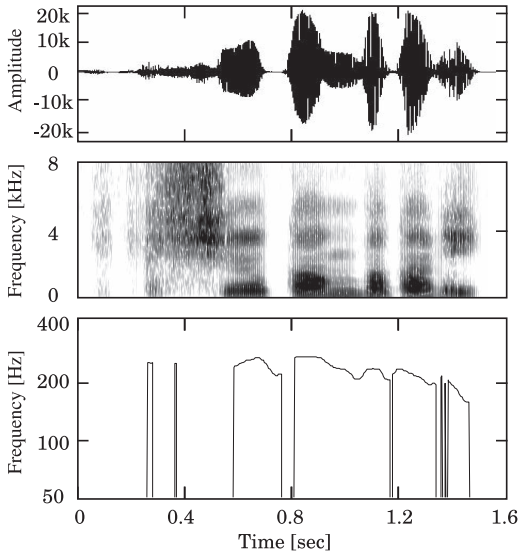


**Fig. 4**. Example of waveform, spectrogram, and $F_0$ of the converted speech by one-to-many EVC.

## 7. REFERENCES

[1] A. Hisada and 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, and 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, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.

[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. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Enhancement of Esophageal Speech Using Statistical Voice Conversion," *APSIPA 2009*, pp. 805-808, Sapporo, Japan, Oct. 2009.

[9] 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.

[10] 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.

[11] 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.