Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion

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

***Abstract*—In this paper, we present novel speaking-aid systems based on one-to-many eigenvoice conversion (EVC) to enhance 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 the lack of speech quality and speaker individuality. To improve the speech quality of alaryngeal speech, alaryngeal-speech-to-speech (AL-to-Speech) methods based on statistical voice conversion have been proposed. In this paper, one-to-many EVC capable of flexibly controlling the converted voice quality by adapting the conversion model to given target natural voices is further implemented for the AL-to-Speech methods to effectively recover speaker individuality of each type of alaryngeal speech. These proposed systems are compared with each other from various perspectives. The experimental results demonstrate that our proposed systems are capable of effectively addressing the issues of alaryngeal speech, e.g., yielding significant improvements in speech quality of each type of alaryngeal speech.**

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

I. Introduction

PATIENTS who suffer from laryngeal cancer require total laryngectomy, which is a surgical operation to remove the larynx and tissues around the larynx such as the vocal folds. People who have undergone total laryngectomy, called laryngectomees, cannot speak in the usual manner owing to the removal of their vocal folds. Because speech is one of the most popular methods of human communication, laryngectomees experience inconvenience in many situations of their daily life. Therefore, they desire to be able to speak using medical devices or following rehabilitation in order to reintegrate into their individual, social, and regular activities. To accomplish their wish, alternative speaking methods to produce speech sounds using residual organs or medical devices instead of vocal cords have been used. Speech sounds generated by alternative speaking methods without vocal fold vibration are called alaryngeal speech.

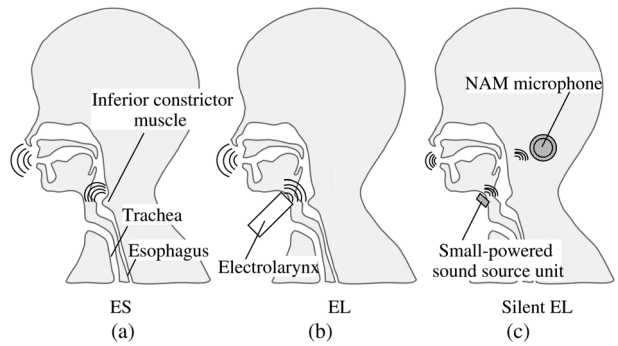


Fig. 1. Alternative speaking methods for laryngectomees: (a) to produce esophageal (ES) speech, (b) to produce electrolaryngeal (EL) speech, and (c) to produce silent electrolaryngeal (silent EL) speech.

There are various alternative speaking methods. In this paper, three speaking methods shown in Fig. 1 are focused on. Among them, the speaking methods for esophageal speech (ES speech) and electrolaryngeal speech (EL speech) are the most popular methods in Japan. The speaking method for ES speech is one of the alternative speaking methods that generate alaryngeal speech with residual organs. ES speech is generated by modulating alternative excitation sounds that are produced by releasing gases from or through the esophagus by articulatory movement. This speaking method allows laryngectomees to speak without any equipment and ES speech sounds more natural than the other types of alaryngeal speech such as EL speech, but its sound quality is not comparable to normal speech uttered by non laryngectomees. Although it generally takes a long time to learn the speaking method for ES speech, support for learning it is provided by many volunteers in Japan.

The speaking method for EL speech is one of the most popular alternative speaking methods using medical devices. Alternative excitation sounds are produced using an electrolarynx, which is a medical device to mechanically generate the sound source signals. The generated sound source signals are conducted as alternative excitation sounds into the oral cavity from the skin on the lower jaw. Then, the alternative excitation sounds are articulated to produce EL speech sounds. It is much easier to learn how to speak using the electrolarynx than to learn how to produce ES speech. Moreover, users need less physical power to produce EL speech compared with other alaryngeal speech, such as ES speech. However, the EL speech sound is mechanical and artificial because the generated fundamental frequency (F_0) contour is totally unnatural owing to the pre-defined frequency of the vibration (i.e., F_0 of the

Manuscript received March 27, 2013; revised July 10, 2013; accepted October 03, 2013. Date of publication October 23, 2013; date of current version November 22, 2013. This work was supported in part by MIC SCOPE and MEXT Grant-in-Aid for Young Scientists (A) and in part by a JSPS Research Fellowships for Young Scientists. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark J. F. Gales.

H. Doi, T. Toda, H. Saruwatari, and K. Shikano are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan (e-mail: hironori-d@is.naist.jp; tomoki@is.naist.jp; sawatari@is.naist.jp; shikano@is.naist.jp).

K. Nakamura was with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan. He is now with Rakuten, Inc., Tokyo 140-0002, Japan.

Digital Object Identifier 10.1109/TASLP.2013.2286917

sound source signals). Additionally, because the electrolarynx needs to generate sufficiently loud sound source signals to make the produced EL speech sufficiently audible, the sound source signals are readily emitted outside, disturbing speech communication.

To resolve the issue of the emitted sound source signals in the speaking method for EL speech, a new speaking method for silent EL speech has been proposed [1]. A new sound source unit is used to generate less audible sound source signals. Since the produced speech also becomes less audible, it is detected with a nonaudible murmur (NAM) microphone [2], which is a body-conductive microphone capable of detecting extremely soft speech from the neck below the ear. The detected speech signals are presented outside as silent EL speech while keeping the external sound source signals sufficiently silent.

Although these three types of alaryngeal speech allow laryngectomees to speak again, their sound quality and intelligibility are severely degraded compared with those of normal speech. Moreover, alaryngeal speech sounds are of similar voice quality[1] regardless of the speaker differences because the production mechanism of the sound source signals in each type of alaryngeal speech strongly affects the voice quality of the produced speech. Consequently, the alaryngeal speech suffers from the degradation of speaker individuality.

Several attempts at improving alaryngeal speech quality have been carried out. A new electrolarynx using an air pressure sensor has been developed to control F_0 of the sound source signals by an expiratory pressure. Although it is not easy to accurately control F_0 by adjusting the expiratory pressure, this device makes it possible for laryngectomees to produce more naturally sounding EL speech, the F_0 of which effectively varies over an utterance [3]. One weakness of this device is that both hands are needed to hold an electrolarynx and air pressure sensor while speaking. Moreover, it is still difficult to mechanically generate sound source signals similar to those naturally generated by vocal fold vibrations. Consequently, the produced EL speech quality is still different from the natural voices produced by non laryngectomees.

As another attempt, speech enhancement methods based on the modifications of acoustic features of ES speech using signal processing, such as comb filtering [4], smoothing of acoustic parameters [5], formant manipulation [6], and noise reduction based on auditory masking [7] have been proposed. Although they are useful in the alaryngeal speech enhancement, quality improvements are still limited since the acoustic features of alaryngeal speech exhibit quite different properties from those of normal speech, and therefore, it is basically difficult to compensate for those acoustic differences using such a simple modification process.

Recently, statistical approaches to alaryngeal speech enhancement have been proposed [8], [9] to convert alaryngeal speech into target normal speech while keeping linguistic information unchanged. The statistical enhancement framework consists of training and conversion processes. In the training process, a conversion function from acoustic features of alaryngeal speech into those of target normal speech is modeled using training data including utterance pairs of alaryngeal speech and normal speech. In the conversion process, any utterance of alaryngeal speech is converted to that of target normal speech on the basis of the conversion function. This data-driven approach is capable of more complicated acoustic modifications to compensate for the large acoustic differences between alaryngeal speech and normal speech. As typical conventional methods, a codebook mapping method [10] and a probabilistic conversion method based on Gaussian mixture models (GMMs) [11] have been applied to alaryngeal speech enhancement [8], [9]. The GMM-based conversion method is one of the most popular voice conversion methods. It is well defined mathematically and its conversion performance is relatively high. It has been reported that the alaryngeal speech enhancement method based on the GMM-based voice conversion, which is called Alaryngeal-Speech-to-Speech (AL-to-Speech), is highly effective for improving the naturalness and intelligibility of individual types of alaryngeal speech [1], [3], [9].

Although these conventional enhancement methods allow laryngectomees to speak in more natural voices than alaryngeal speech, recovering speaker individuality is minimally considered. In fact, it is essentially difficult to flexibly control the voice quality of enhanced alaryngeal speech by these methods. In the statistical voice conversion approaches, it is possible to change the converted voice quality using different target voices but it is necessary to prepare training data consisting of utterance pairs of the alaryngeal speech and each target voice, which is very laborious. To flexibly change the converted voice quality to recover speaker individuality or provide a unique voice for laryngectomees, one-to-many eigenvoice conversion (EVC) [12] has been applied to ES speech enhancement (called ES-to-Speech) in our previous work [13]. One-to-many EVC is a technique for converting a specific 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 to flexibly adapt the conversion model to an arbitrary target speaker on the basis of a small number of given target speech samples in a text-independent manner. ES-to-Speech based on EVC helps laryngectomees speak in their favorite voices or in their own voices that have already been lost but a few recorded samples are available.

In this study, we develop AL-to-Speech systems capable of flexibly controlling the enhanced voice quality based on one-to-many EVC for not only ES speech but also EL speech and silent EL speech. The effectiveness of the proposed AL-to-Speech systems based on VC/EVC for the three types of alaryngeal speech is evaluated from various perspectives. The features of each AL-to-Speech system are demonstrated through various comparisons among the different AL-to-Speech systems. In this paper, we present further details of the proposed systems, more discussions, and more evaluations than those in our previous work [9].

The paper is organized as follows. In Section II, the characteristics of three types of alaryngeal speech, namely, ES speech, EL speech and silent EL speech are described. In Section III, the statistical VC algorithm and a one-to-many EVC algorithm are described. In Section IV, the proposed method of enhancing ala-

[1]In this paper, the term "voice quality" is used to represent speech characteristics on speaker individuality affected by both glottal excitation parameters and vocal tract spectral parameters.

ryngeal speech is discussed. In Section V, our proposed method is experimentally evaluated. Finally, this paper is summarized in Section VI.

II. Alaryngeal Speech

The larynx including the vocal fold has to be removed by laryngectomy, if the larynx has severe trouble such as cancer. Because the larynx prevents food from entering the trachea, the trachea and the oral cavity connecting the esophagus are completely separated from each other when this function is lost by total laryngectomy. Therefore, laryngectomees cannot generate vocal fold vibrations nor expire air through the oral cavity. They have to produce speech sounds in an alternative manner.

In this section, we describe the three types of alaryngeal speech generated by an alternative speaking method for laryngectomees. Fig. 2 shows an example of speech waveforms, spectrograms, F_0 contours, and aperiodic components of (a) normal speech, (b) ES speech, (c) EL speech, and (d) silent EL speech in the same sentence.

F_0 contours are extracted with STRAIGHT analysis [14] by manually optimizing an F_0 search range to reduce F_0 extraction errors as much as possible. Spectrograms and 5-band aperiodic components (averaged on 0-1, 1-2, 2-4, 4-6, and 6-8 kHz frequency bands) [15] are also extracted with STRAIGHT analysis [16], [17].

A. Esophageal Speech (ES speech)

ES speech sounds more natural than the other types of alaryngeal speech. Although a speaker skilled in producing ES speech can control prosody using residual organs, the produced sound is constantly low in tone regardless the speaker. Moreover, specific unnatural sounds caused by producing the excitation sounds in the manner mentioned above are often observed. The difficulty of producing ES speech makes a spectral envelope vary unstably but it still captures phoneme-dependent acoustic characteristics and is capable of conveying linguistic information. Aperiodic components are constantly noisy in all frequency bands.

Even if we can perceive pitch information (i.e., perceptual tone related to prosody) in ES speech, it is difficult to directly extract F_0 patterns corresponding to pitch information from ES speech because the excitation signals are less periodic. This is similar to pitch perception in a whispered voice, which is unvoiced speech. We have found that pitch information is also perceived in an ES speech sample resynthesized from a mel-cepstrum sequence including power coefficients and noise excitation [13]. Therefore, we expect that an acoustic cue of pitch information in ES speech is included in its spectral envelope and power.

B. Electrolaryngeal Speech (EL Speech)

It is easier to stably speak using EL speech than using ES speech because stable excitation signals are generated using medical device. However, EL speech sounds mechanical owing to artificial excitation signals. Although the spectral envelope stably varies according to each phoneme, it is distorted by the sound source signals leaked from the electrolarynx. The electrolarynx used in this study generates sound source signals with almost constant F_0 values and a high periodicity. Excitation parameters such as F_0 and aperiodic components are easily extracted from EL speech but are less informative since they capture only the acoustic characteristics of the artificial excitation signals.

C. Body-Conducted Silent Electrolaryngeal Speech (Silent EL Speech)

Silent EL speech allows the user to speak without leaked excitation sounds even with an electrolarynx. However, silent EL speech sounds much more unnatural than EL speech owing to its lower-powered sound source signals and body conduction. It basically has similar acoustic characteristics to EL speech except that (1) the signal-to-noise ratio of silent EL speech is much lower than that of EL speech and (2) high-frequency components over 3 or 4 kHz are severely attenuated by the lack of radiation characteristics from the lips and by the effect of the low-pass characteristics of the soft tissue [18].

III. Voice Conversion

A. Basic Voice Conversion (VC)

VC is the method that converts the source speaker's voice into the target speaker's voice in a probabilistic manner. In this section, we describe a conversion method based on maximum likelihood estimation of speech parameter trajectories considering global variance (GV) [19] as one of the state-of-the-art VC methods. This method consists of a training and a conversion process.

1) Training Process: Let us assume a D_x dimensional input static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D_x)]^\top$ and a D_y dimensional output static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ at frame t, respectively, 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, such as the joint feature vector of static and dynamic feature vectors 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 feature vector $\boldsymbol{y}_t$ and a dynamic feature vector $\Delta\boldsymbol{y}_t$.

By using a parallel training data set consisting of time-aligned input and output parameter vectors $\boldsymbol{Z}_1 = [\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top$, $\boldsymbol{Z}_2 = [\boldsymbol{X}_2^\top, \boldsymbol{Y}_2^\top]^\top, \cdots, \boldsymbol{Z}_T = [\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]^\top$ determined by Dynamic Time Warping (DTW), where T denotes the total number of frames, the joint probability density of the input and output parameter vectors is modeled with a GMM [20] as follows:

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

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

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a 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 α_m, mean vectors $\boldsymbol{\mu}_m^{(Z)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(Z)}$ for individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(Z)}$ consists

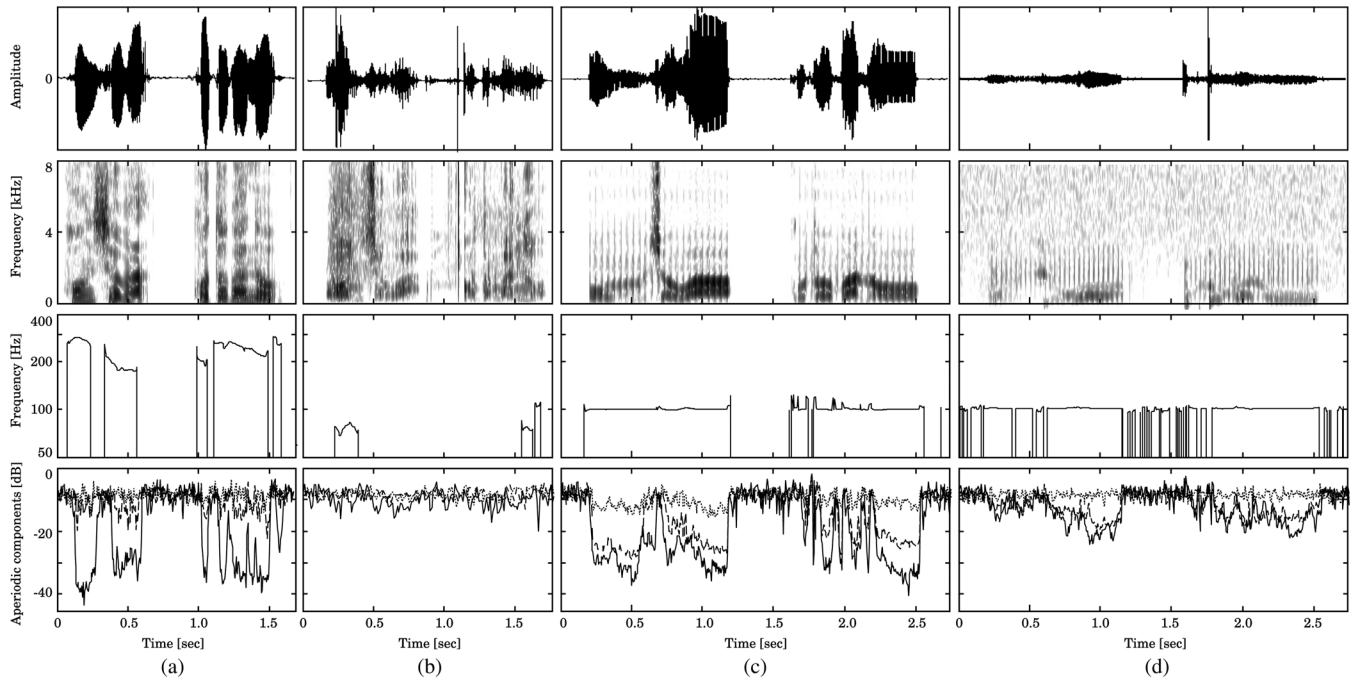


Fig. 2. Example of acoustic features, i.e., waveforms, spectrograms, F_0 contours, and aperiodic components of (a) normal speech, (b) ES speech, (c) EL speech, and (d) silent EL speech in the same sentence fragment /h o N sy o w a k o t o b a n o/. In aperiodic components, the solid line, coarse broken line, and fine broken line represent averaged aperiodic components in low frequency band, middle frequency band, and high frequency band, respectively.

of an input mean vector $\boldsymbol{\mu}_m^{(X)}$ and an output mean vector $\boldsymbol{\mu}_m^{(Y)}$. The covariance matrix $\boldsymbol{\Sigma}_m^{(Z)}$ consists of input and output covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ and cross-covariance matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$.

The GV is defined as the variance of features over one utterance. To consider the GV in the conversion, 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 with a Gaussian distribution,

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

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

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

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

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 output feature vectors, respectively. The converted static feature vector 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 an objective function defined as a product of the GV probability density function given by Eq. (3) and the conditional probability density function $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$, which is analytically derived from the joint probability density given (1), as follows:

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

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 [21]. The balance between $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ and $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ is controlled by the weight ω.

B. One-to-Many Eigenvoice Coversion (EVC)

We also describe one-to-many EVC [12] as a technique for flexibly controlling the voice quality of the converted speech. This method consists of training, adaptation, and conversion process.

Training Process: In one-to-many EVC, Eigenvoice GMM (EV-GMM) is used as a conversion model. 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. EV-GMM models the joint probability density of the input and output parameter vectors as follows:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}), \boldsymbol{\Sigma}_m^{(Z)}\right), \tag{6}$$

$$\boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{A}_m\boldsymbol{w} + \boldsymbol{b}_m \end{bmatrix}, \tag{7}$$

$$\boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \tag{8}$$

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

Adaptation Process: The trained EV-GMM allows users to control the converted voice quality by manipulating the weight

TABLE I
CHARACTERISTICS OF ACOUSTIC FEATURES OF ALARYNGEAL SPEECH

	ES	EL and silent EL
Spectrum	Unstably varying according to phonemes (still informative)	Varying according to phonemes (still informative)
Aperiodic components	Constantly noisy (less informative)	Depending on mechanical excitation (less informative)
F_0	Hard to be extracted (less informative)	Constant due to mechanical excitation (less informative)

vector $\boldsymbol{w}$. If users have target speech data, the GMM for the source 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. The optimal weight vector $\hat{\boldsymbol{w}}$ is estimated by maximizing the likelihood of the marginal distribution as follows [22]:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}, \tag{9}$$

where $\boldsymbol{Y}^{(tar)}$ is a time sequence of the target features for the adaptation.

Conversion Process: In the conversion process, the user gives a target-speaker-dependent parameter $\hat{\boldsymbol{w}}$ by entering data manually or by estimating from target speech samples. The adapted EV-GMM is generated with $\hat{w}$ as shown in (7). Then, the converted speech is estimated in the same manner as basic VC shown in (5).

IV. VOICE CONVERSION FROM ALARYNGEAL SPEECH TO SPEECH (AL-TO-SPEECH)

To enhance the three types of alaryngeal speech, namely, ES speech, EL speech, and silent EL speech, statistical VC approaches for converting alaryngeal speech into normal speech have been proposed. In AL-to-Speech based on VC, the converted speech sounds similar to normal speech because the converted speech parameters are basically determined according to the statistics extracted from the normal speech in a probabilistic manner. Furthermore, by applying one-to-many EVC to AL-to-speech, AL-to-Speech allows users to flexibly control the speech quality of the converted speech. In this section, we describe AL-to-Speech based on VC and one-to-many EVC for the enhancement of the three types of alaryngeal speech.

A. Feature Extraction in AL-to-Speech

In AL-to-Speech, three types of acoustic feature of normal speech, namely, spectrum, aperiodic components, and F_0, are separately estimated from the acoustic features of each type of alaryngeal speech. Then, the estimated acoustic features are used in vocoding to generate converted speech. To estimate the acoustic features of normal speech by AL-to-Speech, we need to decide which acoustic features of each type of alaryngeal speech are used as the input feature. However, we have little choice because most of the acoustic features of alaryngeal speech are less informative as mentioned in Section II. Table I shows characteristics of acoustic features of each types of alaryngeal speech.

Because the spectrum of ES speech is the only informative acoustic feature of ES speech even if it unstably varies, we use the spectrum of ES speech as an input feature to estimate the spectrum and aperiodic components of normal speech, which smoothly vary according to the phoneme. On the other hand, for F_0 estimation, we assume that an acoustic cue of the pitch of ES speech could be included in the spectrum with power as mentioned in Section II. On the basis of this assumption, we also use the spectrum of ES speech as an input feature to estimate F_0 of normal speech. This is similar to the conventional estimation methods of F_0 from mel-frequency cepstral coefficients [23], [24] but this is an estimation process of F_0 from the spectrum of unvoiced speech rather than voiced speech, and therefore, this estimation process is much more difficult compared with the conventional ones. Furthermore, although one reasonable conversion process is to estimate F_0 corresponding to the pitch of ES speech, it is not straightforward to prepare target F_0 values in training process as such an F_0 is difficult to extract from ES speech. To address this issue, we record normal speech uttered by non laryngectomee so that its pitch sounds similar to that of ES speech and use F_0 extracted from the recorded normal speech as the target in training. Namely, F_0 corresponding to the pitch of ES speech is approximated with F_0 of normal speech uttered by a different speaker.

Although the spectra of EL speech and silent EL speech change according to the phoneme, it is significantly different from those of normal speech. F_0 values of EL speech and silent EL speech are mechanically decided independent of utterance content. Moreover, because an electrolarynx is driven during an utterance, the aperiodic components of EL speech and silent EL speech depending on only the mechanical excitation signals are not informative as input feature. Therefore, in the acoustic features of EL speech and silent EL speech, only the spectrum is informative, and then, we use the spectra of EL speech and silent EL speech as input feature to estimate the spectrum, aperiodic components, and F_0 of normal speech uttered by non laryngectomee.

In AL-to-Speech for these three types of alaryngeal speech, the spectrum of each alaryngeal speech is used as input feature to estimate each acoustic feature of each type of alaryngeal speech, respectively. However, directly using the spectrum of alaryngeal speech causes the degradation of the converted speech because the spectrum structures of some phonemes of alaryngeal speech are often collapsed owing to difficulties of producing them. To address these issues, we use a spectral segment feature extracted from multiple frames [25] as follows:

$$\boldsymbol{X}'_t = \boldsymbol{C}\boldsymbol{X}_t + \boldsymbol{d}, \tag{10}$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top]^\top$ is a joint vector generated by concatenating a spectral parameter vector $\boldsymbol{x}_t$ at the current frame and those at $\pm i$ preceding and succeeding frames. Because this joint vector includes significantly redundant information, dimensionality reduction with principal component analysis (PCA) is performed for the joint vector $\boldsymbol{X}_t$ in order to extract a spectral segment feature $\boldsymbol{X}'_t$ at frame t, where $\boldsymbol{C}$ and $\boldsymbol{d}$

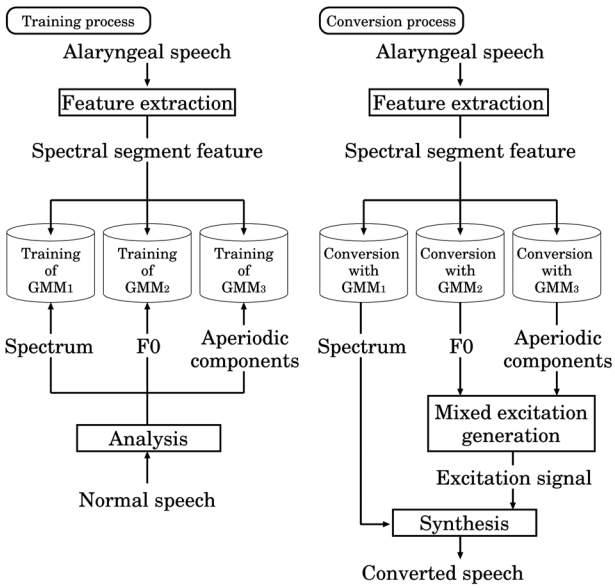


Fig. 3. Training process and conversion process of AL-to-Speech.

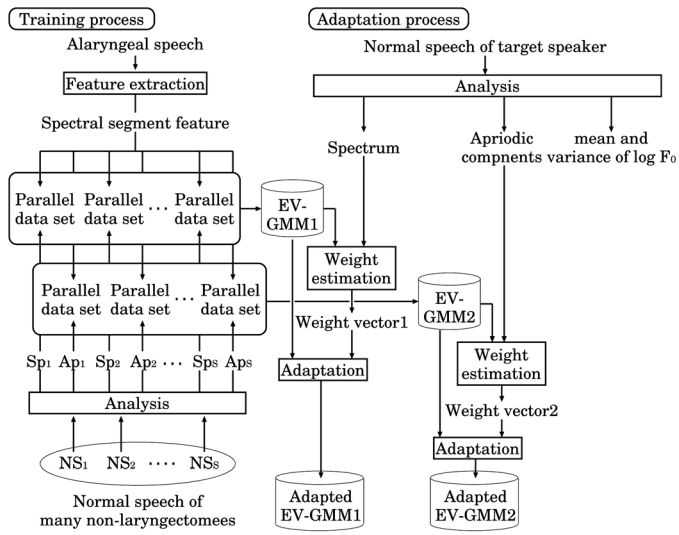


Fig. 4. Training process and adaptation process in AL-to-Speech based on one-to-many EVC. "Sp_s" and "Ap_s" show spectral features and aperiodic components of normal speech of the sth speaker NS_s, respectively.

are a transformation matrix and a bias vector extracted by PCA, respectively.

B. AL-to-Speech Based on VC

AL-to-Speech based on VC consists of a training process and a conversion process shown in Fig. 3.

Training Process: In the training process, we record utterance pairs of alaryngeal speech and normal speech uttered by a non laryngectomee. To convert the spectral segment feature of alaryngeal speech into three speech parameters of the target normal speech: namely, (1) spectral features, (2) log-scaled F_0, and (3) aperiodic components, we independently train three GMMs modeling joint probability densities of the spectral segment feature of alaryngeal speech and individual target speech parameters using the corresponding joint feature vector sets. For the F_0 feature, a constant value clearly different from F_0 values (e.g., a value much less than the minimum F_0 value) is used to represent unvoiced frames [26]. In F_0 estimation for ES speech, F_0 extracted from normal speech recorded so as to be similar to the pitch of ES speech is used as the target.

Conversion Process: In the conversion process, spectral segment features are extracted from alaryngeal speech. Then, individual converted speech parameters are independently estimated from the extracted spectral segment features using each of the trained GMMs. In the F_0 estimation, unvoiced/voiced decision is also performed using a manually setting threshold to detect the constant value to represent unvoiced frames. After estimating the converted spectral features, the converted log-scaled F_0, and the converted aperiodic components, the excitation signal is generated using STRAIGHT mixed excitation on the basis of the converted F_0 values and the converted aperiodic components [15]. Finally, the converted speech is synthesized by filtering the generated excitation signal with the converted spectral features.

C. AL-to-Speech Based on EVC

In AL-to-Speech based on the basic VC, it is difficult to recover speaker individuality of alaryngeal speech owing to the predefined voice quality of converted speech, although the sound quality of converted speech is improved. To flexibly control the converted voice quality, we further apply one-to-many EVC to AL-to-Speech. The EVC technique allows users to manually control the voice quality of converted speech. Furthermore, conversion models can be adapted using a small number of utterances of target speech in a text-independent manner. Therefore, if recorded normal speech data of the laryngectomee before undergoing the total laryngectomy still exist, AL-to-Speech based on EVC can estimate the converted speech that sounds similar to the laryngectomee's original voice. In this section, we describe the AL-to-Speech system on the basis of the assumption that a few utterances of the laryngectomee's original speech have been kept. AL-to-Speech based on EVC consists of training, adaptation, and conversion processes. Fig. 4 shows the training and adaptation process in AL-to-Speech based on one-to-many EVC.

Training Process: In the training process, we independently train two one-to-many EV-GMMs: namely, a one-to-many EV-GMM for estimating the converted spectral feature and a one-to-many EV-GMM for estimating the converted aperiodic components, for each alaryngeal speech. To train these two EV-GMMs, we use multiple parallel data sets consisting of alaryngeal speech data uttered by the laryngectomee and prestored normal speech data uttered by many non laryngectomees. The EV-GMM for the spectral estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored spectral features of the non laryngectomees. On the other hand, the EV-GMM for the aperiodic component estimation is trained using multiple joint feature vector sets consisting of the spectral segment features of the laryngectomee and the prestored aperiodic components of the non laryngectomees. In this paper, we perform

the training method on the basis of Speaker Adaptive Training (SAT) [27] for the EV-GMM [28].

The EV-GMM is adapted to a new target speaker by adjusting the weight vector so that the marginal likelihood for given target speech features is maximized as shown in (9). This adaptation process is effective if speaker-dependent characteristics are well captured by short-term features, such as spectrum and aperiodic components. On the other hand, it is essentially difficult to control speaker-dependent characteristics captured by long-term features, such as F_0 patterns. Therefore, instead of the EV-GMM, a well-trained speaker-dependent GMM is used to estimate the F_0 patterns from the spectral segment sequence of alaryngeal speech. In AL-to-Speech for ES speech, to develop the GMM for estimating the F_0 patterns corresponding to the perceived pitch information of ES speech, we use F_0 values extracted from normal speech uttered by a non laryngectomee as an imitating prosody of ES speech in the training as the output features. This process is the same as in the AL-to-Speech based on VC described in Section IV-B. To develop a GMM for F_0 estimation in EL speech and silent EL speech, speaker-dependent GMMs are separately trained for all target speakers. Then, the GMM achieving the highest F_0 estimation accuracy is manually selected.

Adaptation and Conversion Processes: Assuming that a few speech samples uttered by laryngectomees before undergoing total laryngectomy are available as adaptation data, the EV-GMM is flexibly adapted to the target voice quality by automatically determining the weight vector in a text-independent manner [12]. The weight vectors of the EV-GMMs for the spectral and aperiodic estimations are independently estimated using the spectral features and aperiodic components extracted from the given target speech samples. The converted spectral feature vectors and aperiodic components are independently estimated using the adapted EV-GMMs. In the F_0 estimation, the global speaker-dependent characteristics of F_0 patterns are simply controlled. A log-scaled F_0 sequence is first estimated with the selected speaker-dependent GMM, and then further converted so that its mean μ_x and standard deviation σ_x are equal to those of the adaptation speech data, μ_y and σ_y, as follows:

$$\log y_t = \frac{\sigma_y}{\sigma_x}(\log x_t - \mu_x) + \mu_y, \tag{11}$$

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 AL-to-Speech based on VC/EVC methods, we conducted experimental evaluations using several criteria. Then, we explicitly indicate the advantage of each alaryngeal speech when applying AL-to-Speech.

A. Experimental Conditions

We recorded 50 phonetically balanced sentences of ES speech uttered by one Japanese male laryngectomee, those of EL speech and silent EL speech uttered by another Japanese male laryngectomee, and those of normal speech uttered by each of 40 Japanese non laryngectomees. The speech data of 30 non laryngectomees were used for training and those of the other 10 non laryngectomees were used as the target data for evaluation. From the 50 recorded sentences of each speaker, 40 were used as the training or adaptation data and the remaining 10 were used as the test data. The sampling frequency was set to 16 kHz.

The 0th through 24th mel-cepstral coefficients were used as spectral parameters. STRAIGHT analysis [17], which is F_0 adaptive analysis to extract accurate spectral envelope by effectively removing the effect of F_0 periodicity on spectrum, was employed for normal speech. On the other hand, mel-cepstrum analysis [29] was employed for alaryngeal speech since F_0 of alaryngeal speech is not informative.

As the source excitation features of normal speech, we used log-scaled F_0 values and aperiodic components 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 frame shift was 5 ms. To extract the spectral segment feature of ES speech, current and ± 8 frames were used for spectral and aperiodic component estimations and current and ± 16 frames were used for F_0 estimation. For EL speech and silent EL speech, current and ± 8 frames were used for each parameter estimation. These numbers of frames per segment were experimentally optimized [13].

The EV-GMMs for spectral and aperiodic component estimations were trained for each type of alaryngeal speech. The numbers of eigenvectors and mixture components were set to 29 and 64 in every EV-GMM, respectively. The EV-GMMs were adapted to the target speakers using 1, 2, 4, 8, 16, or 32 utterances of their normal speech data. For AL-to-Speech based on VC, the GMMs for spectral and aperiodic estimation were trained using a parallel dataset for each type of alaryngeal speech and normal speech of each target speaker. The number of training utterance pairs was set to 1, 2, 4, 8, 16 or 32. The number of mixture components was optimized manually depending on the training data size so that the best conversion accuracy in the evaluation data was obtained. Individual speaker-dependent GMMs for F_0 estimation were trained for all the 40 non laryngectomees. The GMM yielding the most natural F_0 pattern was then selected by listening to the converted speech. The same F_0 estimation process was performed for the EVC-based AL-to-Speech and VC-based AL-to-Speech. Full covariance matrices were used in every GMM/EV-GMM.

B. Objective Evaluations

We evaluated the effectiveness of AL-to-Speech for each type of alaryngeal speech with estimation accuracy of each acoustic feature, i.e., spectrum, aperiodic components, and F_0.

Estimation Accuracy of Spectrum and Aperiodic Components: Figs. 5 and 6 show mel-cepstral distortion and root mean square error (RMSE) on aperiodic components as a function of the number of adaptation utterances used in EVC or of utterance pairs used in VC, respectively. EVC shows a significantly smaller mel-cepstral distortion and RMSE than VC in each type of alaryngeal speech enhancement when the amount of the target normal speech data is small. Even if only one arbitrary utterance of the target normal speech is available in EVC, its conversion performance is almost equivalent to or

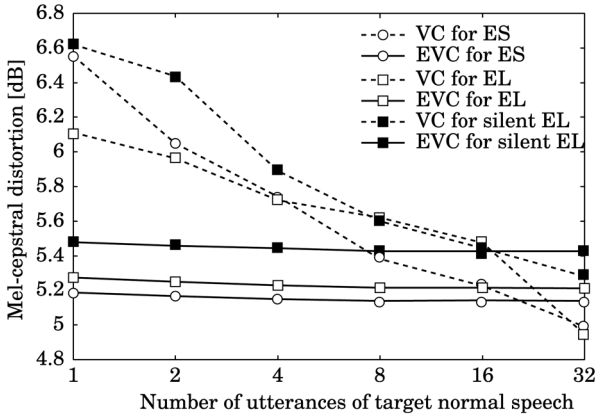


Fig. 5. Mel-cepstral distortion as a function of the number of utterances of target normal speech (i.e., utterance pairs in VC or adaptation utterances in EVC).

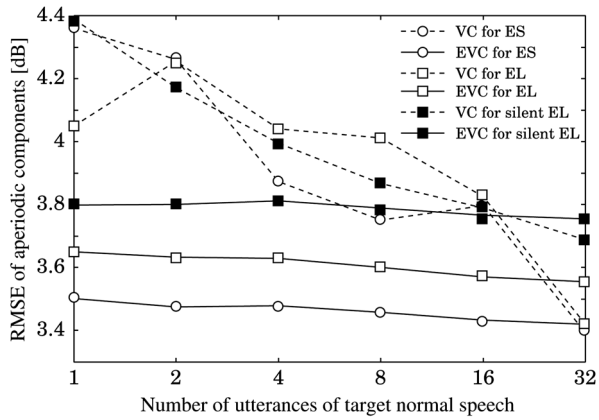


Fig. 6. RMSE on aperiodic components as a function of the number of utterances of target normal speech (i.e., utterance pairs in VC or adaptation utterances in EVC).

better than that of VC using 16 parallel utterance pairs. It is also observed that ES speech yields the best conversion accuracy, and silent EL speech yields the worst among the three types of alaryngeal speech.

F_0 Estimation Accuracy: We also evaluated the F_0 estimation accuracy in AL-to-Speech for each type of alaryngeal speech using F_0 correlation coefficient and Unvoiced/Voiced (U/V) error between converted speech and target normal speech. To demonstrate the F_0 estimation accuracy for various speakers in AL-to-Speech, the results calculated using individual speaker-dependent GMMs for the 40 non laryngectomees are shown in Table II. For ES speech, the results for another non laryngectomee who uttered normal speech so that its pitch sounded similar to that of ES speech are also shown as "ES pitch." ES speech yields the best estimation accuracy among the three types of alaryngeal speech. Additionally, the estimation accuracy is significantly improved using the GMM developed with the normal speech, the F_0 patterns of which correspond well to the pitch patterns of ES speech.

The final results for the 10 target non laryngectomees from the test data are shown in Table III. The GMM for "ES pitch" was used in ES speech enhancement, and manually selected speaker-dependent GMMs were used in the EL/silent EL speech enhancement. Namely, the speaker used in the model training is different from the target speakers. It is observed that, for EL

TABLE II
F_0 ESTIMATION ACCURACIES FOR VARIOUS TARGET SPEAKERS USING CORRESPONDING TARGET-SPEAKER-DEPENDENT GMMS

	Correlation	U/V error [%]
ES	0.58	12.39 ($V \to U : 6.59, U \to V : 5.80$)
EL	0.40	13.20 ($V \to U : 4.92, U \to V : 8.28$)
Silent EL	0.42	14.02 ($V \to U : 6.89, U \to V : 7.13$)
ES pitch	0.68	8.36 ($V \to U : 4.30, U \to V : 4.05$)

TABLE III
F_0 ESTIMATION ACCURACIES FOR ACTUAL TARGET SPEAKERS IN EVALUATION USING WELL-TRAINED SPEAKER-DEPENDENT GMMS

	Correlation	U/V error [%]
ES pitch	0.62	13.88 ($V \to U : 10.70, U \to V : 3.18$)
EL	0.51	12.05 ($V \to U : 7.13, U \to V : 4.92$)
Silent EL	0.45	13.78 ($V \to U : 8.92, U \to V : 4.86$)

speech and silent EL speech, the estimation accuracy of the selected GMMs is higher than that of various speaker-dependent GMMs shown in Table II, even though a speaker different from the target speakers is used in the training. To generate a natural F_0 pattern in AL-to-Speech, it is useful to select an optimum speaker for training rather than to directly use the same speaker as the actual target speaker since the F_0 estimation accuracy largely varies among different speakers. It is also observed that ES speech enhancement yields better F_0 correlation than the others.

C. Subjective Evaluations

We evaluated the effectiveness of AL-to-Speech for each types of alaryngeal speech with speech quality, listenability, intelligibility, and speaker individuality. In this paper, the term "listenability" is used to indicate a score that was measured by asking the listener to subjectively evaluate how easy it was to understand the utterance. On the other hand, the term "intelligibility" is used to indicate a score that was calculated by asking the listener to write down the content of the utterance, and measuring the accuracy of transcription.[2]

Opinion Test on Speech Quality and Listenability: We conducted opinion tests on speech quality and listenability. In the opinion test of speech quality and listenability, 8 listeners evaluated 9 types of speech including original alaryngeal speech and converted speech with AL-to-Speech based on VC/EVC in ES speech, EL speech, and silent EL speech. The VC-based AL-to-Speech used 32 utterance pairs for GMM training. On the other hand, only one utterance was used as adaptation data for the EVC-based AL-to-Speech. The GV was considered in the conversion process. For the EVC-based AL-to-Speech, the mean vector of the GV probability density was set to the GV extracted from the adaptation utterance for each target non laryngectomee and the covariance matrix was fixed to that calculated using the GVs extracted from all utterances of the non laryngectomees used in the training of the EV-GMM. The opinion score in each test was set to a 5-point scale. Individual speech samples were normalized so that loudness of each sample was almost the same as each other. Note that signal-to-noise ratio was kept in each sample before and after the normalization. We

[2]Note that the term "intelligibility" in [9] is used in the sense of "listenability."

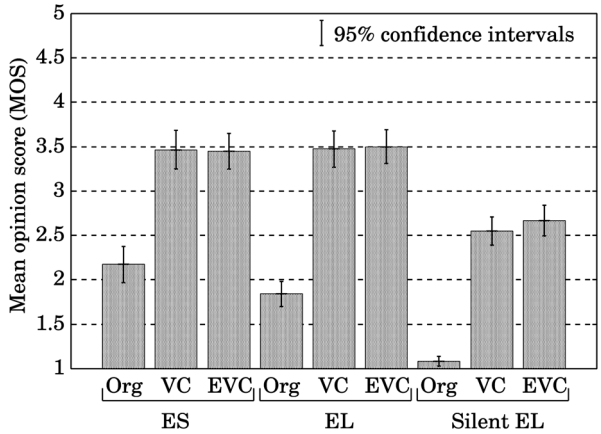


Fig. 7. Result of opinion test of speech quality. "Org", "VC", and "EVC" indicate original alaryngeal speech, converted speech by AL-to-Speech based on VC trained with 32 utterance pairs, and converted speech by AL-to-Speech based on EVC adapted with one utterance of target speech, respectively.

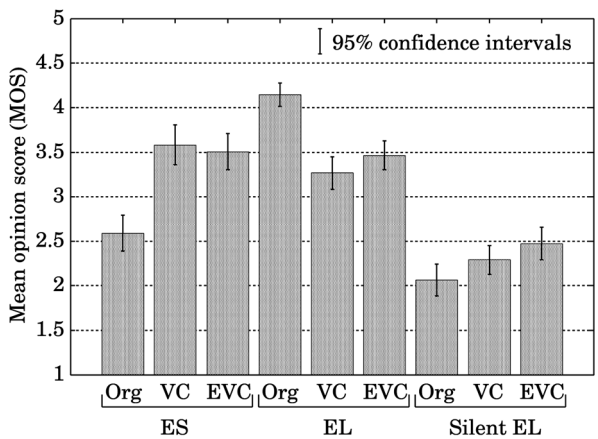


Fig. 8. Result of opinion test of listenability.

asked the listeners to evaluate speech samples with a wide range of the score from 1 to 5; i.e., a higher score was better speech quality or listenability. In each test, the individual listeners listened to several speech samples before starting the test to make their own score range as stable as possible. Each listener evaluated 135 speech samples in the test.

Figs. 7 and 8 show the results of the opinion tests of speech quality and listenability, respectively. All the AL-to-Speech methods yield significant improvements in speech quality compared with that of the original alaryngeal speech. The speech quality of the enhanced silent EL speech is lower than those of the enhanced ES speech and enhanced EL speech but it is significantly higher than that of each type of original alaryngeal speech. The listenability of ES speech and silent EL speech are also improved by AL-to-Speech. On the other hand, the listenability of EL speech slightly degrades from that of the original EL speech by AL-to-Speech, as observed in our previous work [3]. The speech quality and listenability enhanced by the EVC-based AL-to-Speech are almost equivalent to those enhanced by the VC-based AL-to-Speech. Note that the EVC-based method requires only one arbitrary utterance of the target normal speech, whereas the VC-based method requires 32 utterance pairs of alaryngeal speech and the target normal speech.

TABLE IV
RESULT OF DICTATION TEST ON INTELLIGIBILITY

	Word correct [%]	Word accuracy [%]	Number of replays
ES	87.76	84.30	2.23
EL	92.89	90.93	2.70
Silent EL	66.42	64.71	2.70
ES-EVC	79.90	76.96	2.93
EL-EVC	89.22	87.50	1.90
Silent EL-EVC	84.80	82.84	2.57

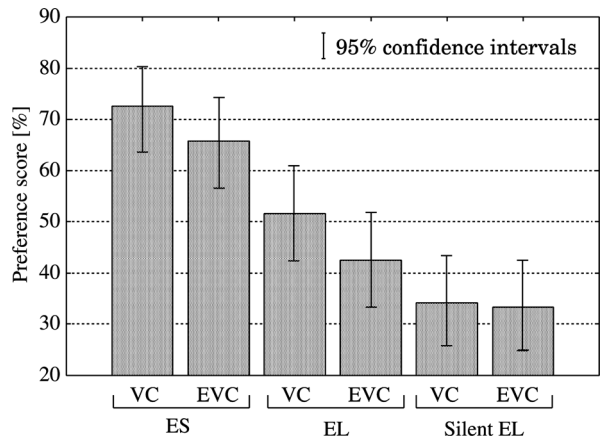


Fig. 9. Result of preference test of speaker individuality.

Dictation Test on Intelligibility: We conducted a manual dictation test. In this test, 6 listeners evaluated 6 types of speech including original alaryngeal speech and converted speech by AL-to-Speech based on EVC. Only one utterance was used as adaptation data in the adaptation process of EV-GMMs. We allowed listeners to replay the same stimulus as much as they want.

Table IV shows word correct, word accuracy, and the average number of replays by listeners. Before conversion, EL speech is the best, ES speech is the next, and silent EL speech is the worst in intelligibility. We found that articulation of the laryngectomee who uttered EL speech and silent EL speech is clearer than that of the other laryngectomee who uttered ES speech. Because quality of silent EL speech is significantly degraded by the small-powered excitation and body-conductive recording, silent EL speech is less intelligible than ES speech even if it is more clearly articulated than ES speech.

Intelligibility of silent EL speech is significantly improved by AL-to-Speech based on EVC. On the other hand, intelligibility of ES speech and EL speech are degraded by AL-to-Speech based on EVC. Consequently, after conversion, EL speech is still the best, silent EL speech is the next, and ES speech is the worst. In AL-to-Speech, conversion errors are inevitable and they tend to cause degradation in intelligibility. Moreover, if articulation of input speech is unclear and unstable, larger degradation in intelligibility tends to be caused by conversion as observed in ES speech.

On the other hand, intelligibility of silent EL speech is significantly improved by conversion. As articulation of silent EL speech is relatively clear and stable, the degradation of intelligibility caused by conversion tends to be smaller. Moreover, AL-to-Speech well addresses a problem of very low quality

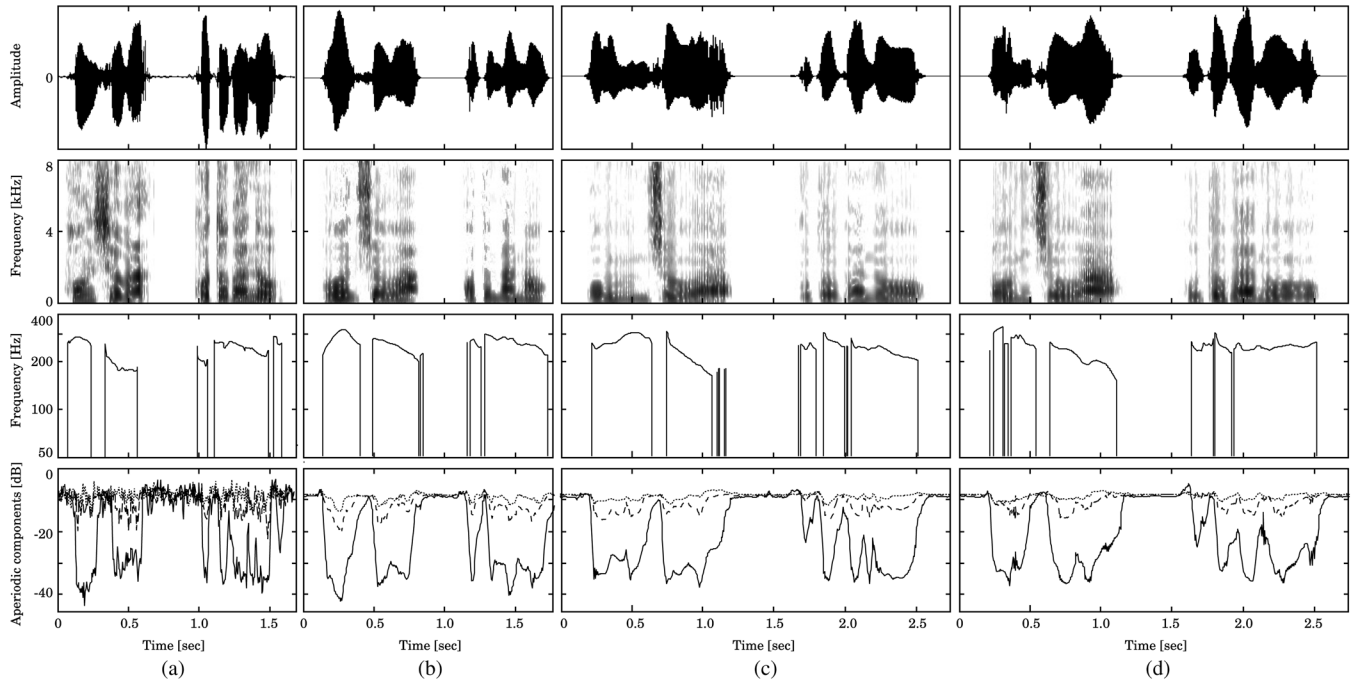


Fig. 10. Example of acoustic features, i.e., waveforms, spectrograms, F_0 contours, and aperiodic components of (a) normal speech and three types of converted speech by AL-to-Speech based on EVC from (b) ES speech, (c) EL speech, and (d) silent EL speech in the same sentence fragment /h o N sy o h a k o t o b a n o/. In aperiodic components, the solid line, coarse broken line, and fine broken line represent low band, middle band, and high band of aperiodic components, respectively.

causing degradation in intelligibility of silent EL speech. Consequently, intelligibility of silent EL speech is better than ES speech after conversion. These results suggest that our proposed method is basically capable of improving intelligibility of alaryngeal speech if its intelligibility is degraded by some factors, such as low speech quality, except for less articulated sounds.

Preference Test on Speaker Individuality: We also conducted ABX test as a preference test to evaluate speaker individuality. In the preference test, 6 listeners evaluated 6 types of speech consisting of converted speech by the VC/EVC-based AL-to-Speech in EL speech, ES speech, and silent EL speech. In this test, listeners heard target speech sample and two speech samples from among 6 types of converted speech, and then, they chose speech sample that has more similar speaker individuality as target speech sample.

Each listener evaluated 60 pairs in the test.

The training data used in VC and the adaptation data used in EVC were the same as those used in the opinion tests.

Fig. 9 shows the result of the preference test.

Every type of the converted speech was compared against all the others and the preference score of each was calculated as the ratio of the number of samples selected as having better speaker individuality to the total number of samples presented to listeners.

We can observe the same tendency as that in Fig. 6. Enhanced ES speech yields the best speaker individuality and enhanced silent EL speech yields the worst among the three types of alaryngeal speech. Even if using only one arbitrary utterance of the target speaker is used in the EVC-based method, its performance is close to that of the VC-based method using 32 parallel utterances of the target speaker. This result shows that the EVC-based method is capable of effectively adjusting speaker individuality of the converted speech.

D. Example of the Converted Speech by AL-to-Speech Based on EVC

Fig. 10 shows an example of the acoustic features of target normal speech and converted speech from the three types of alaryngeal speech by AL-to-Speech based on EVC. These samples were converted from each alaryngeal speech shown in Fig. 2. We can see that the acoustic features of each converted speech come closer to those of normal speech than to those of each type of alaryngeal speech. In the spectrogram, the spectral structure of converted speech from ES speech became clearer and stably varies compared with those of ES speech. Moreover, the spectral structure at high frequency that could not be observed in EL speech and silent EL speech is observed in converted speech from EL speech and silent EL speech. F_0 of converted speech from each type of alaryngeal speech can capture the coarse structure of those of normal speech. Furthermore, although over smoothing occurs, the aperiodic components of converted speech are similar to those of normal speech. Therefore, the AL-to-Speech based on EVC is significantly effective for the enhancement of alaryngeal speech.

VI. Conclusions

In this paper, we presented AL-to-Speech as enhancement methods based on VC and one-to-many EVC for three types of alaryngeal speech, namely, esophageal speech (ES speech), electrolaryngeal speech (EL speech), and body-conducted silent electrolaryngeal speech (silent EL speech). These methods convert a spectral segment feature into spectrum,

aperiodic components, and F_0 of normal speech independently using different GMMs or EV-GMMs. The experimental results suggested that (1) the proposed methods significantly improve the speech quality of each type of alaryngeal speech, (2) the proposed methods also improve the listenability of ES speech and silent EL speech, (3) the proposed methods also significantly improve the intelligibility of silent EL speech, (4) AL-to-Speech based on eigenvoice conversion (EVC) is capable of effectively adjusting the voice quality of enhanced speech to the target voice quality using only one arbitrary utterance of the target voice.

Although the individual types of alaryngeal speech has their own merits and demerits in terms of speech quality, listenability, intelligibility, speaker individuality, difficulties of learning how to produce, and so on. Some of the demerits are effectively addressed by AL-to-Speech. The results presented in this paper could be helpful for laryngectomees to decide which type of alaryngeal speech they use. We plan to develop and evaluate the AL-to-Speech systems for other alaryngeal speech and further implement a real-time conversion process for AL-to-Speech.

Acknowledgment

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, H. Saruwatari, and K. Shikano, "Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion," *IEICE Trans. Inf. Syst.*, vol. E93-D, no. 7, pp. 1909–1917, Jul. 2010.

[2] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) recognition," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 1, pp. 1–8, Jan. 2006.

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "The use of air-pressure sensor in electrolaryngeal speech enhancement based on statistical voice conversion," in *Proc. Interspeech*, Sep. 2010, pp. 1628–1631.

[4] A. Hisada and H. Sawada, "Real-time clarification of esophageal speech using a comb filter," in *Proc. ICDVRAT*, Sep. 2002, pp. 39–46.

[5] K. Matsui, N. Hara, N. Kobayashi, and H. Hirose, "Enhancement of esophageal speech using formant synthesis," in *Proc. ICASSP*, May 1999, pp. 1831–1834.

[6] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 10, pp. 2448–2458, Oct. 2010.

[7] H. Liu, Q. Zhao, M. Wan, and S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 5, pp. 865–874, May 2006.

[8] G. Aguilar-Torres, M. Nakano-Miyatake, and H. Perez-Meana, "Enhancement and restoration of alaryngeal speech signals," in *Proc. 16th IEEE Int. Conf. Electron., Commun. Comput. (CONIELECOMP '06)*, Feb. 2006, p. 30.

[9] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *Proc. ICASSP*, May 2011, pp. 5136–5139.

[10] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[11] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[12] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, Apr. 2007, pp. 1249–1252.

[13] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models," *IEICE Trans. Inf. Syst.*, vol. E93-D, no. 9, pp. 2472–2482, Sep. 2010.

[14] 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," in *Proc. EUROSPEECH*, Sep. 1999, pp. 2781–2784.

[15] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Interspeech '06–ICSLP*, pp. 2266–2269, Sep. 2006.

[16] 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," in *Proc. MAVEBA*, Sep. 2001.

[17] 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 Commun.*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[18] T. Hirahara, M. Otani, S. Shimizu, T. Toda, and K. Nakamura, "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Commun.*, vol. 52, no. 4, pp. 301–313, Apr. 2010.

[19] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

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

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

[22] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[23] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 24–33, Dec. 2007.

[24] J. Darch, B. Milner, and S. Vaseghi, "Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures," *J. Acoust. Soc. Amer.*, vol. 124, no. 6, pp. 3989–4000, Dec. 2008.

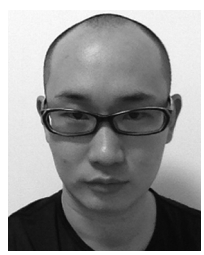
[25] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum with a Gaussian mixture model," *Speech Commun.*, vol. 50, no. 3, pp. 215–227, Mar. 2008.

[26] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 9, pp. 2505–2517, Sep. 2012.

[27] T. Anastasakos, J. McDonough, S. R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, vol. 2, pp. 1137–1140.

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

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

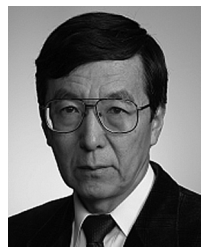


Hironori Doi graduated from the Department of Information and Image Science, Faculty of Engineering, Chiba University, Japan in 2008. He received his M.E. and D.E. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2009 and 2012, respectively. He was a Research Fellow of JSPS in NAIST from 2011 to 2012. He currently works for DWANGO Co., Ltd., Japan. He mainly studies singing voice conversion and speaking-aid systems.

Tomoki Toda earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nitech, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visiting Researcher at the NICT, Kyoto, Japan, since May 2006. From April 2003 to March 2006, he was a Visiting Researcher at the ATR SLC Research Laboratories, Kyoto, Japan. He was also a Visiting Researcher at the LTI, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the CUED, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech and language processing. He received 11 awards including the 2009 Young Author Best Paper Award from the IEEE SPS and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal). He was a member of the SLTC of the IEEE SPS from 2007 to 2009.

Keigo Nakamura graduated from the Course of Multimedia Studies, Faculty of Education and Human Sciences, Yokohama National University, Japan in 2005. He received his M.E. and D.E. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2007 and 2010, respectively. He was a Research Fellow of JSPS in NAIST from 2009 to 2011. He was also a Visiting Researcher at the Karlsruhe Institute of Technology, Karlsruhe, Germany, from April 2010 to February 2011. He currently works for Rakuten, Inc., Japan. He mainly studies speaking-aid systems for laryngectomees.

Hiroshi Saruwatari (M'00) was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E., and Ph.D., degrees from Nagoya University in 1991, 1993, and 2000, respectively. He joined Intelligent System Laboratory, SECOM Co., Ltd., Tokyo, Japan, in 1993, where he engaged in the research on the ultrasonic array system for the acoustic imaging. He is currently an Associate Professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include noise reduction, array signal processing, blind source separation, and sound field reproduction. He received paper awards from IEICE in 2001 and 2006, from Telecommunications Advancement Foundation in 2004 and 2009, and from IEEE-IROS2005 in 2006. He won the first prize in IEEE MLSP2007 Data Analysis Competition for BSS. Prof. Saruwatari is a member of the IEICE, Japan VR Society, and the Acoustical Society of Japan.



Kiyohiro Shikano received the B.S., M.S., and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972, and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. From 1972, he had been working at NTT Laboratories, where he had been engaged in speech recognition research. During 1990-1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986-1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech recognition and speech synthesis research. During 1984-1986, he was a visiting scientist in Carnegie Mellon University. He received the IEICE (Institute of Electronics, Information and Communication Engineers of Japan) Yonezawa Prize in 1975, IEEE Signal Processing Society 1990 Senior Award in 1991, the Technical Development Award from ASJ (Acoustical Society of Japan) in 1994, IPSJ (Information Processing Society of Japan) Yamashita SIG Research Award in 2000, Paper Award from the Virtual Reality Society of Japan in 2001, IEICE Paper Award in 2005 and 2006, and IEICE Inose Best Paper Award in 2005. He is a fellow member of IEEE, IEICE, and IPSJ. He is a member of ASJ, Japan VR Society, and International Speech Communication Association.