Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 54 (2012) 134–146

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech

Keigo Nakamura *, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan

Received 23 February 2011; received in revised form 2 July 2011; accepted 14 July 2011
Available online 26 July 2011

Abstract

An electrolarynx (EL) is a medical device that generates sound source signals to provide laryngectomees with a voice. In this article we focus on two problems of speech produced with an EL (EL speech). One problem is that EL speech is extremely unnatural and the other is that sound source signals with high energy are generated by an EL, and therefore, the signals often annoy surrounding people. To address these two problems, in this article we propose three speaking-aid systems that enhance three different types of EL speech signals: EL speech, EL speech using an air-pressure sensor (EL-air speech), and silent EL speech. The air-pressure sensor enables a laryngectomee to manipulate the F_0 contours of EL speech using exhaled air that flows from the tracheostoma. Silent EL speech is produced with a new sound source unit that generates signals with extremely low energy. Our speaking-aid systems address the poor quality of EL speech using voice conversion (VC), which transforms acoustic features so that it appears as if the speech is uttered by another person. Our systems estimate spectral parameters, F_0 and aperiodic components independently. The result of experimental evaluations demonstrates that the use of an air-pressure sensor dramatically improves F_0 estimation accuracy. Moreover, it is revealed that the converted speech signals are preferred to source EL speech.

Keywords: Electrolaryngeal speech; Voice conversion; Speaking-aid system; Speech enhancement; Airpressure sensor; Silence excitation; Non-audible murmur; Laryngectomee

1. Introduction

More than 12000 people in the United States are estimated to receive the diagnosis of laryngeal cancer in 2008 (Jemal et al., 2008). There are some treatments to cure laryngeal cancer such as chemotherapy, radiation therapy, partial laryngectomy (Forastiere et al., 2003; Laccourreye et al., 1996) or total laryngectomy. Total laryngectomy, which removed the whole larynx including the vocal folds, was the standard treatment for locally advanced disease until the early 1990s (Forastiere et al., 2003). It means that there are many total laryngectomees that have been undergone the total laryngectomy.[1] For example, it is said that there are 12000 laryngectomees in Japan (Ifukube, 2003). After laryngectomy, laryngectomees cannot speak using the vibration of their vocal folds as shown in Fig. 1. By undergoing a laryngectomy, the quality of life of laryngectomees suffers because of the loss of their original voices (Carr et al., 2000). Therefore, voice rehabilitation for laryngectomees is an important research topic.

There are three major types of alaryngeal speech called esophageal speech, tracheo-esophageal (T-E) shunt speech, and artificial laryngeal speech such as electrolaryngeal speech (EL speech) (Singer and Blom, 1980; Hashiba

* Corresponding author. Tel.: +81 743 72 5287; fax: +81 743 72 5289.
E-mail address: kei_go@nifty.com (K. Nakamura).

[1] The words of laryngectomee and laryngectomy respectively denote total laryngectomee and total laryngectomy in this article.

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.07.007

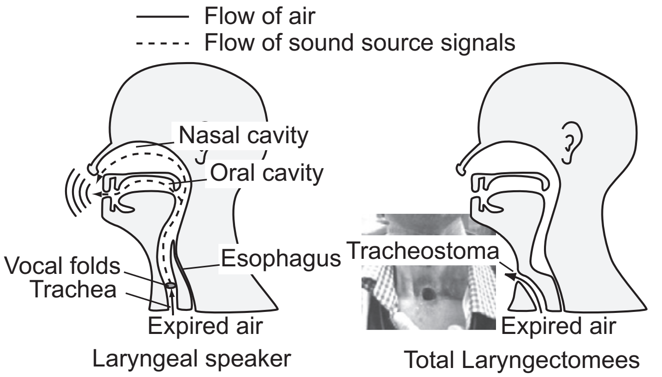

Fig. 1. Anatomical image of laryngal speakers and total laryngectomees.

et al., 2001; Hayes, 1951; Williams and Watson, 1985; Hočevar-Boltežar and Žargi, 2001). The main difference among these alaryngeal speech signals is how the speaker obtains an alternative sound source vibration to vocal fold vibration. There are some types of artificial larynxes such as an electrolarynx(EL) or a pneumatic artificial larynx (also known as a whistle-type artificial larynx). An EL is one of the most commonly used alaryngeal speech (Williams and Watson, 1985; Carr et al., 2000), and therefore, we focus on EL speech.

When a speaker produces EL speech, an EL is pressed against the speaker's lower jaw, and the signals are transmitted to the speaker's oral cavity through the speaker's skin. Finally, the speaker produces EL speech by articulating with the sound source signals generated by the EL. One of the advantages of using an EL is that speakers easily learn how to produce EL speech. Another advantage is that EL speech is possible for people with less physical fitness such as elderly laryngectomees. Two problems with EL speech that are focused on in this article are (1) the unnaturalness of EL speech and (2) the radiation of noisy sound source signals. One of the main defects of EL speech is its unchangeable F_0 pattern. Because of this unchangeable F_0 pattern, EL speech sounds obviously artificial and unnatural even when speakers are proficient in using an EL. Another defect of EL speech is the high energy of vibrations generated by the EL itself. The sound source signals are often radiated from the location of attachment of the EL, and the radiated noise often annoys surrounding people; especially in quiet environments such as inside a library. Moreover, the speaker might also be concerned that he or she will annoy people nearby because of the radiated sound source signals. Therefore, the enhancement of EL speech quality is an important issue in ensuring the smooth interpersonal speech communication of laryngectomees.

Many researchers and developers have attempted to address problems of EL speech (Uemi et al., 1994; secom.co., 2011; Murakami et al., 2004; Saikachi et al., 2009; Liu et al., 2006; Takahashi et al., 2001; Goldstein et al., 2004). Uemi *et al.* developed an EL with an air-pressure sensor (Uemi et al., 1994), which enables laryngectomees to control the F_0 pattern using the exhaled air that flows from the tracheostoma that is a hole located on the speaker's neck for breathing. Murakami *et al.* enhanced EL speech by employing a feature transformation technique (Murakami et al., 2004). In their approach, numerous conversion rules are developed from training data and then applied to test data. Saikachi *et al.* proposed a method to control F_0 of EL speech based on root-mean-square amplitude (Saikachi et al., 2009). Their concept of obtaining reasonable F_0 contours to make EL speech naturally sound is same as that of this article. Liu *et al.* proposed a method to reduce the radiation noises based on spectral subtraction considering auditory masking (Liu et al., 2006). Novel artificial larynxes were also proposed. For example, Takahashi *et al.* developed an intra-oral electrolarynx for people who could not acquire common alaryngeal speech or for early post-surgery speech rehabilitation (Takahashi et al., 2001). Goldstein *et al.* also developed hands-free EL triggered by neck muscle electromyographic activity (Goldstein et al., 2004). Development of hands-free EL is another very important mission for improving the quality of life of laryngectomees. However, developing new devices is out of the scope of this article. Despite these studies for enhancing EL speech, some problems still remain. In the study of Murakami et al. (2004), input utterances cannot be converted when no conversion rules are found. A method proposed by Saikachi et al. (2009) requires pre-laryngectomy speech. Moreover, not only amplitude but also spectral information might be needed to estimate reasonable F_0 contours. The approach of Liu et al. (2006) might remove radiation noises; however, the problem of radiation noises is not the only problem in EL speech. There are huge gaps between EL speech and normal speech, and these gaps have not been filled by these previous studies.

We proposed a speaking-aid system that enhanced EL speech using a statistical voice conversion (VC) technique with Gaussian mixture models (GMMs) to address the unnaturalness of EL speech and the problem of radiation noises (Nakamura et al., 2007). Our speaking-aid system consists of four parts that (1) generate sound source signals, (2) record the produced EL speech, (3) convert the recorded EL speech, and (4) present the converted speech. We also introduced a new sound source device that generates signals with extremely low energy. Since the sound source signals of EL speech is given from outside, the voice quality of EL speech is mainly defined by the given source signals. Existing ELs have to generate sound source signals with high energy to make the volume of EL speech close to that of normal speech, although the energy of the radiated noises is also getting higher when the volume of sound source signals is raised. Our aid system proposed in our previous work, on the other hand, includes a VC procedure, and therefore, using an existing EL is not essential. We have had an idea to suppress the energy of radiated noises by using a new sound source unit that generates signals with extremely low energy. The produced EL speech with the silent signals (silent EL speech) is also very low energy. It is difficult to make an absolute definition what silent EL speech is. In this article, we regard EL speech produced with the new sound source unit generating extremely low energy as silent EL speech. Although a usual air-conductive microphone is difficult to detect the silent EL speech owing to its sensitivity to external noises, a special body-conductive microphone is capable of detecting it with high quality. As a result of experimental evaluations, the effectiveness of this system was demonstrated by using silent EL speech produced by a laryngeal speaker (Nakamura et al., 2007). However, two important problems have still remained; (1) the system has been evaluated using silent EL speech produced by not a laryngectomee but a laryngeal speaker and (2) input speech data have been converted into not normal speech but whispered voice that does not have F_0 information.

In this article we address two problems of EL speech; one is the unnaturalness of EL speech, and the other one is noisy sound source signals radiated from the location of the EL attachment. For addressing the problem of unnaturalness, we propose a speaking-aid system that converts EL speech into normal speech using a VC technique. This system presents converted EL speech that includes F_0 information estimated from the EL speech. Moreover, this system removes noisy sound source signals using the VC technique. For more improving the naturalness, we propose another speaking-aid system that accepts EL-air speech. In this system, VC effectively uses F_0 information of EL-air speech. We also propose the other speaking-aid system that accepts silent EL speech. This system addresses the problem of noisy sound source signals; however, the speech quality of silent EL speech is still same as that of EL speech. Therefore, VC should be applied to silent EL speech as well. The shared part between these three aid systems and the aid system described in our previous article (Nakamura et al., 2007) is the four components constructing each system. On the other hand, the following points are updated from our previous work; (1) two more aid systems are proposed to accept EL speech and EL-air speech, (2) source speech is converted into not whispered voice but normal speech, and (3) three aid systems are all experimentally evaluated in the same experimental conditions, in which source speech signals are produced by not a laryngeal speaker but a laryngectomee. Moreover, there are three features of our aid systems compared to other aid systems proposed in previous studies described above; (1) each system can convert an arbitrary sentence, (2) each system uses statistics of normal speech to estimate F_0 contours, and (3) each system can address two problems focused on in this article simultaneously.

This article is organized as follows. In Section 2, the key techniques used in this article are explained. Our proposed systems are described in Section 3 and experimentally evaluated in Section 4 and discussed in Section 5. Finally, we conclude this article in Section 6.

2. Key techniques

2.1. VC based on maximum likelihood estimation (Kain and Macon, 1998; Toda et al., 2007)

Not only the primary role of speech to convey linguistic information but also secondary information conveyed by speech such as speaker individuality play an important role in interpersonal speech communication. VC generally modifies speech signals of a given source speaker while maintaining the linguistic information so that it appears as if another speaker (the target speaker) utters the speech. VC is useful for many applications such as voice responses, text reader systems, and so forth. It is often convenient to specify the desired modification of source acoustic characteristics with reference to an existing target acoustic characteristic.

A statistical VC method using a GMM-based maximum likelihood estimation is used in applicative studies such as Toda and Shikano (2005). This VC technique consists of training and conversion parts. In the training part, a parallel data set, which consists of sentence pairs of the source and the target speakers, is used to train a GMM. Namely, these two speakers need to utter the same sentence set to develop training data. In each sentence pair, time alignment is automatically performed by a dynamic time warping (DTW) procedure to build joint vectors of the source and the target features. Next, the joint probability density consisting of the source and the target data is modeled by a GMM (Kain and Macon, 1998). In the conversion part, the conditional probability density function of the target data given the source data is used to generate the converted target data (Toda et al., 2007). After the GMM training, any sentences can be converted with the trained GMM, even if it is not included in the training data. A main problem in maximum likelihood estimation is that the estimated parameters tend to be oversmoothed. Oversmoothing of acoustic parameters tends to significantly reduce their variations, and it causes noticeable degradation in naturalness of synthetic speech quality (Toda et al., 2007). To address the oversmoothing problem, Toda *et al.* proposed the consideration of the global variance (GV) of acoustic features over a time sequence (Toda et al., 2007).

2.1.1. Training procedure to describe joint probability density function (Kain and Macon, 1998)

Let $\boldsymbol{x}_t = [x_t(1), \ldots, x_t(d_x)]^\top$ and $\boldsymbol{y}_t = [y_t(1), \ldots, y_t(d_y)]^\top$ be static source and target feature vectors at frame t where d_x and d_y denote the dimensions of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively, and $\top$ denotes transposition. The joint probability density of these vectors $\boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top$ is described by a GMM as follows:

$$P(\boldsymbol{z}_t|\lambda^{(z)}) = \sum_{m=1}^{M} \omega_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right), \quad (1)$$

where m is the index of mixture components, M is the number of mixture components, ω_m is the weight of the mth mixture component, and $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ represents a Gaussian distribution including a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. $\lambda^{(z)}$ is a model parameter set including weights, mean vectors, and covariance matrices. The mth mean vector and the covariance matrix are written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ represent the mean vectors of the mth mixture component for the source and target features. $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ represent the covariance matrices of the mth mixture component for the source and target features, and $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ represent the cross-covariance matrices of the mth mixture component for the source and target features, respectively. The model parameters are estimated by the expectation-maximization (EM) algorithm (Dempster et al., 1977).

The probability density of the GV of the target static feature vectors over a time sequence is modeled by a Gaussian distribution as

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

where the GV $\boldsymbol{v}(y) = [v(1),\ldots,v(d),\ldots,v(D)]^\top$ is calculated as follows:

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

The parameter set $\lambda^{(v)}$ consists of the mean vector $\boldsymbol{\mu}^{(v)}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}^{(vv)}$. The GV is calculated utterance by utterance in this article.

2.1.2. Conversion procedure considering dynamic features (Toda et al., 2007)

Let $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ be the source and target joint feature vectors of the static and dynamic feature vectors for frame t, respectively. Then, time sequences of the source and target feature vectors are written as $\boldsymbol{X} = [\boldsymbol{X}_1^\top,\ldots,\boldsymbol{X}_t^\top,\ldots,\boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top,\ldots,\boldsymbol{Y}_t^\top,\ldots,\boldsymbol{Y}_T^\top]^\top$, respectively. The converted static feature sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top,\ldots,\hat{\boldsymbol{y}}_t^\top,\ldots,\hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing the product of the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the probability density of the GV as follows:

$$\hat{\boldsymbol{y}} = \arg\max P(\boldsymbol{Y}|\boldsymbol{X},\lambda^{(Z)})^{\omega} P(\boldsymbol{v}(y)|\lambda^{(v)}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{W}$ extends the static feature vector to a joint feature vector consisting of the static and dynamic feature vectors. The constant ω denotes the weight factor used to control the balance between these two likelihoods. In this article we set ω as the ratio between the numbers of dimensions of $\boldsymbol{v}(y)$ and $\boldsymbol{Y}$, namely, $\frac{1}{2T}$.

To efficiently perform the conversion process, we employ the approximation method proposed in (Toda et al., 2007). Originally, all mixture component sequences should be considered to calculate $P(\boldsymbol{Y}|\ \boldsymbol{X},\lambda^{(Z)})$ such as

$$P(\boldsymbol{Y}|\boldsymbol{X},\lambda^{(Z)}) = \sum_{all\,\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y},\lambda^{(Z)})P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\lambda^{(Z)}), \tag{6}$$

where $\boldsymbol{m} = \{m_1,\ldots,m_t,\ldots,m_T\}$ is a mixture component sequence. Eq. (6) is efficiently approximated by suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X},\lambda^{(Z)}) \approx P(\hat{\boldsymbol{m}}|\boldsymbol{X},\boldsymbol{Y},\lambda^{(Z)})P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\lambda^{(Z)}), \tag{7}$$

$$\hat{\boldsymbol{m}} = \arg\max P(\boldsymbol{m}|\boldsymbol{X},\lambda^{(Z)}). \tag{8}$$

Using this suboptimum mixture component sequence, the target static feature sequence $\hat{\boldsymbol{y}}$ is estimated as

$$\hat{\boldsymbol{y}} = \arg\max P(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X},\hat{\boldsymbol{m}},\lambda^{(Z)})^{\omega} P(\boldsymbol{v}(y)|\lambda^{(v)}). \tag{9}$$

2.2. Air-pressure sensor for an EL (Uemi et al., 1994)

The air-pressure sensor shown in Fig. 2 enables laryngectomees to manipulate the number of vibrations of an EL using exhaled air from the tracheostoma. This air-pressure sensor is connected to an existing EL with the name 'yourtone' (Hashiba et al., 2001). EL-air speech is produced by (1) drawing air into the lungs, (2) covering the tracheostoma with the air-pressure sensor, (3) expelling this air to drive the vibrator, and (4) articulating the sound source signals while holding the EL and the air-pressure sensor with both hands. Note that the location of attachment of the EL and the method of articulation are the same as those in the production of EL speech. Because this air-pressure sensor is a closed-type that does not allow the passage of air when the air-pressure sensor covers the tracheostoma, the speaker needs to repeat the above process after every pause in speech. The circuit to convert the air pressure to the number of vibrations is built into the main body of the EL. A threshold that defines the lower limit of the air pressure required to turn on the vibrations is set in this EL. If this threshold is too low, EL generates

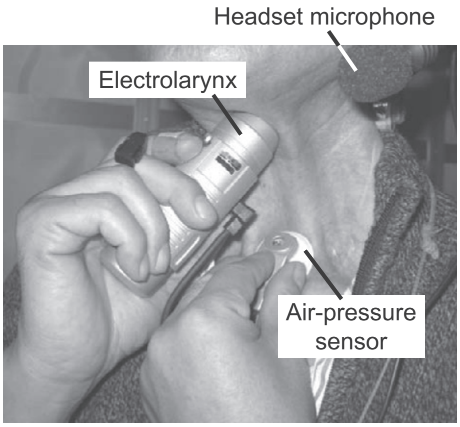

Fig. 2. Photograph of man using EL with air-pressure sensor.

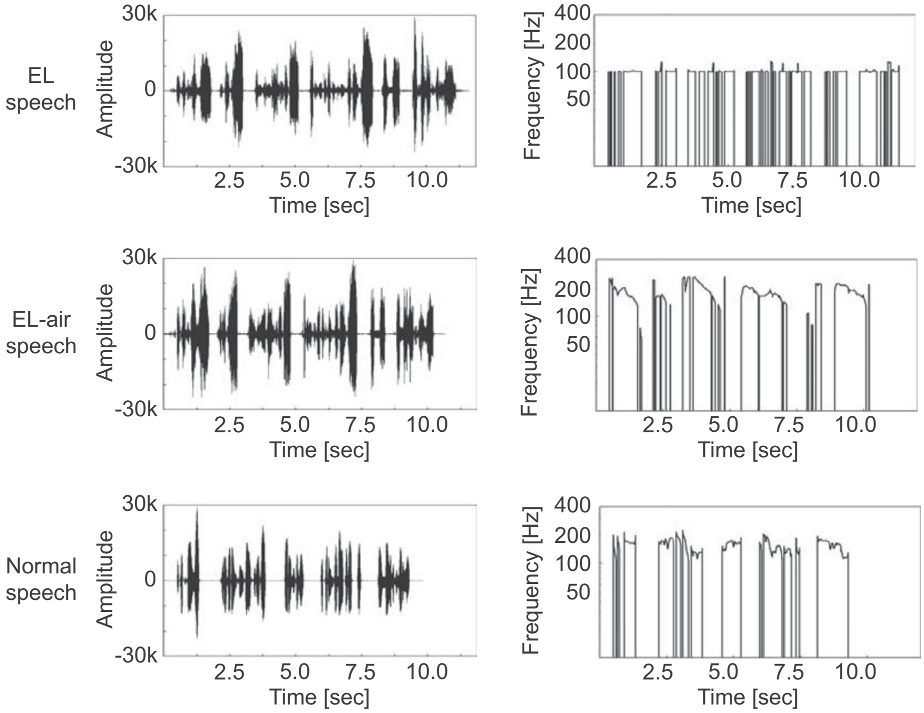

Fig. 3. Examples of waveforms and F_0 contours of EL speech, EL-air speech, and normal speech.

vibrations all the time. On the other hand, if it is too high, the speaker needs very strong air pressure to generate the vibrations. Therefore, a suitable threshold should be adjusted for each speaker in advance.

Fig. 3 shows examples of waveforms and F_0 contours of EL speech, EL-air speech, and normal speech. The EL and EL-air speech signals are produced by the same laryngectomee, and the normal speech is produced by a laryngeal speaker. As shown in the figure, rich F_0 contours are observed in EL-air speech, resulting in it sounding more natural than EL speech. However, EL-air speech is still less natural than normal speech. Moreover, it is difficult for a speaker to intentionally switch between unvoiced and voiced (U/V) sounds when producing EL-air speech, since an EL always vibrates when the speaker produces EL-air speech.

2.3. Non-audible murmur (NAM) microphone (Nakajima et al., 2006)

Nakajima *et al.* defined a special speech that consists of articulated respiratory sound without vocal-fold vibration transmitted through the soft tissues of the head as non-audible murmur (NAM) (Nakajima et al., 2006). Since speech signals of NAM have extremely low energy, it is too difficult to capture those signals with normal air-conductive microphones such as a head-set microphone. Therefore, it is required to capture speech signals of NAM using a special microphone called a NAM microphone (Nakajima et al., 2005). NAM is expected to be used as a novel technique in silent speech interfaces, and applicative studies on NAM and NAM microphones have been conducted (Toda and Shikano, 2005; Nakajima et al., 2006; Nakagiri et al., 2006; Miyamoto et al., 2009; Toda et al., 2009).

A NAM microphone is attached to the body on the sternocleidomastoid behind the neck as shown in Fig. 4. The

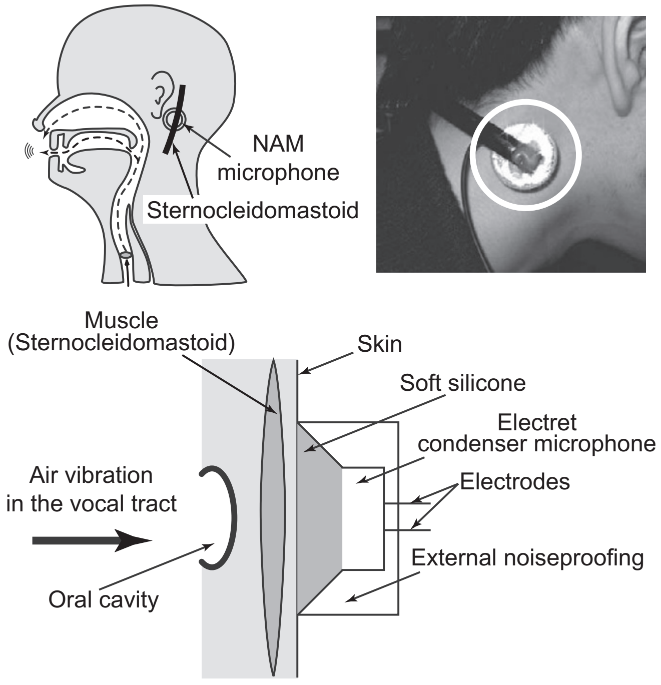

Fig. 4. Location of attachment and structure of NAM microphone.

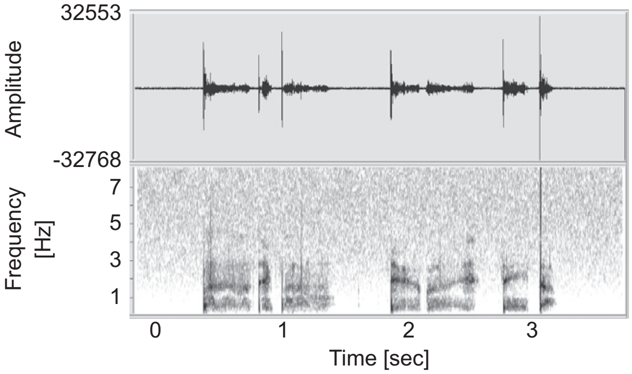

Fig. 5. Example of NAM signals.

waveform in Fig. 5 shows that the dynamic ranges of vowels and consonants are extremely different and that plosive consonants such as /k/ can be impulsively observed. The dynamic range of NAM microphones is larger than that of normal air-conductive microphones; this is one advantage of NAM microphones, since they can detect not only speech signals of NAM but also those of normal speech. In the spectrogram in Fig. 5, spectral components above 4 kHz almost disappear. It is said that this phenomenon is caused by mainly two reasons; one reason is that the radiation features of the oral cavity are lost and the other one is that NAM signals are low-pass filtered when they pass through the muscle (Nakajima et al., 2006). On the other hand, the first and second formants, which are important in speech processing, are clearly observed.

2.4. Sound source unit generating signals with extremely low energy (Hosoi and Sakaguchi, 2003)

Hosoi and Sakaguchi proposed a novel unit as a new sound source unit for silent speech communication (Hosoi and Sakaguchi, 2003). This unit, which is available from Toshin Co. Ltd. (http, xxxx) with the model number BR-41, generates signals with extremely low energy, which cannot be captured by people around the speaker. A photograph of this unit and the vibrator without the cover are shown in Fig. 6. This unit was originally developed as a bone-conductive receiver, and therefore, it can generate arbitrary signals.

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

3.1. Speaking-aid system for EL speech

3.1.1. Framework of the system

To address the problem of the unnaturalness of EL speech, in this article we propose a speaking-aid system for enhancing EL speech. An overview of this system is shown in Fig. 7. This system consists of four parts that (1) generate sound source signals, (2) record the EL speech, (3) convert the EL speech, and (4) present the converted normal speech. The location where the EL is attached is the same as that where a laryngectomee usually sets the EL.

When a laryngectomee speaks with this system, listeners would hear not only converted speech but also the source EL speech, making the use of our proposed system limited. In other words, it is not suitable for laryngectomees to use this system in face-to-face conversations. This proposed system, however, is expected to be effective in some situations where listeners do not have to hear the source EL speech such as when used in telecommunication systems. In telecommunication systems, listeners must understand what a laryngectomee says from only the speech signals, and therefore, converted speech should dominate the source EL speech to allow the conversation to smoothly flow. When our aid system is used in telecommunication systems, it is enough to transfer only converted speech so that listeners do not have to hear the source EL speech. Regarding telecommunications, for example, conventional telephones encode normal speech signals. In other words, current EL speech is unsuitable for use with telecommunications, making it difficult for laryngectomees to use telephones. However, our proposed system enables laryngectomees to conduct smooth communication on the telephone.

3.1.2. VC for EL speech

In the VC from EL speech to normal speech (EL-to-Speech), three acoustic features of the spectral parameters,

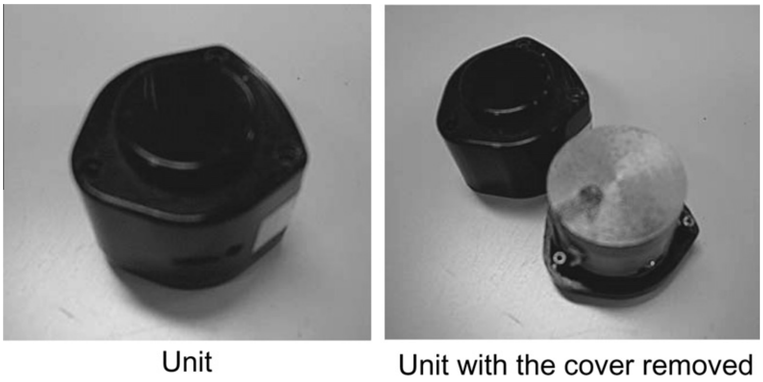

Fig. 6. Photographs of sound source unit generating signals with extremely low energy.

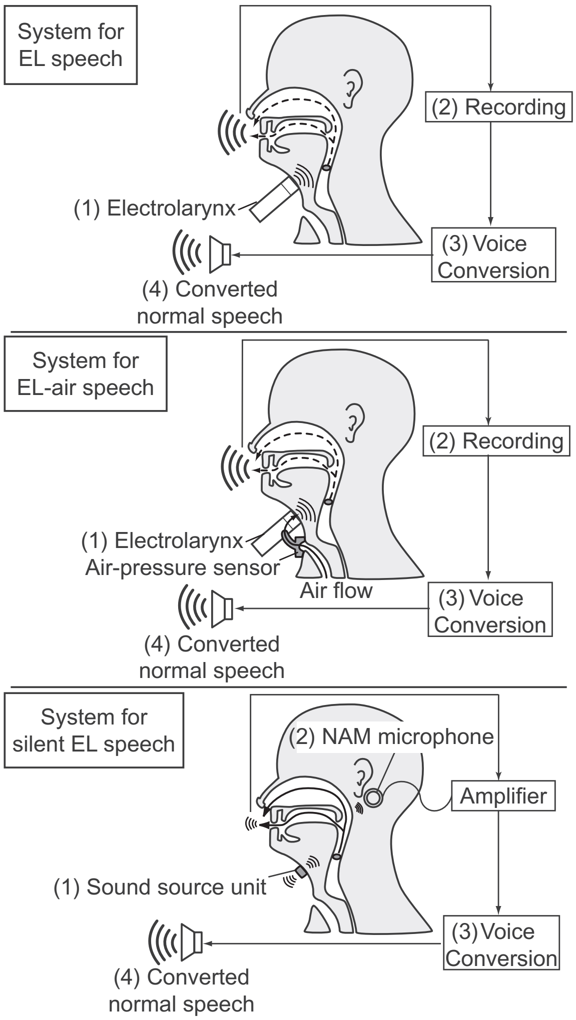

Fig. 7. Overview of proposed speaking-aid system for three types of EL speech. Top accepts EL speech, middle does EL-air speech, and bottom does silent EL speech.

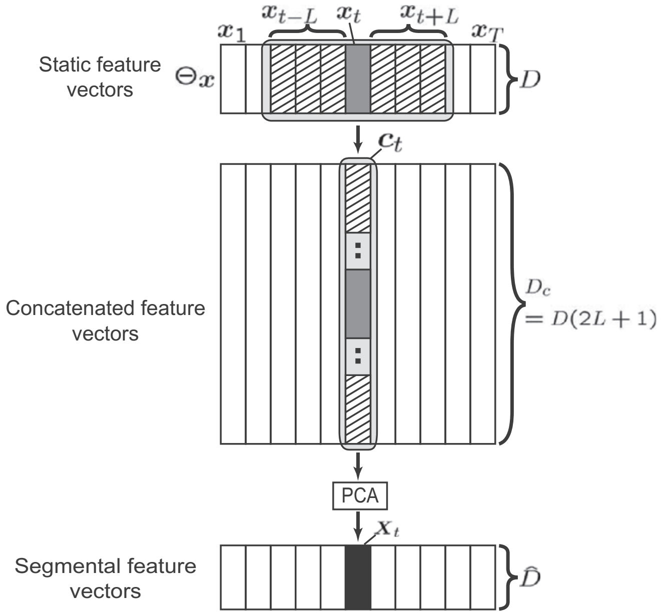

Fig. 8. Flow chart of the construction of segmental feature vectors from static feature vectors.

Table 1
Source and target acoustic features for EL-to-Speech.

Source	Target
Spectrum	Spectrum F_0 Aperiodic components

F_0 , and aperiodic components (Kawahara et al., 2001) are independently estimated from only source spectral parameters. Aperiodic components, which are used to construct mixed excitation signals (Ohtani et al., 2006), indicate the strength of noises in each frequency band.

For the source data, we set a segmental feature vector in this article that includes information over multiple frames. This idea originates from another VC framework used to convert NAM to normal speech (Toda and Shikano, 2005). Fig. 8 shows a flow chart of the construction of segmental feature vectors from static spectral parameter vectors. Let $\Theta_{\boldsymbol{x}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T\}$ be a set of source static feature vectors, where $\boldsymbol{x}_t = [x_t(1), \ldots, x_t(d), \ldots, x_t(D_x)]$ is a D_x-dimensional feature vector. Let $\boldsymbol{c}_t = [\boldsymbol{x}_{t-L}^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_{t+L}^\top]^\top$ be a $D_c = D(2L+1)$-dimensional concatenated feature vector over the current $\pm L(L \geqslant 1)$ frames. Then, the $\widehat{D}$-dimensional segmental feature vector $\boldsymbol{X}_t$ at frame t is extracted from $\boldsymbol{c}_t$ by principal component analysis (PCA). For the target data, joint feature vectors consisting of static and delta features are constructed. Delta features are calculated from the previous and succeeding frames.

Table 1 shows the source and target acoustic features of EL-to-Speech. The training data of the GMM to estimate target spectral parameters are joint vectors consisting of the segmental feature vectors of source spectra and joint feature vectors of target spectra. In the estimation of F_0 and aperiodic components, individual GMMs are trained using joint feature vectors of the segmental feature vectors and those of target *Log*-scaled F_0 or aperiodic components.

In the conversion procedure, source segmental feature vectors are constructed by the same approach as that in the training part. Acoustic parameters are estimated by the method described in Section 2.

3.2. Speaking-aid system for EL-air speech

3.2.1. Framework of the system

In this article we propose another speaking-aid system that accepts EL-air speech. An overview of this system is shown in Fig. 7. This system is expected to estimate more natural F_0 contours by using the air-pressure sensor. The four components consisting of this system, the usage, and the situations in which this system can be used are the same as those for the proposed system for EL speech described in Section 3.1.1.

3.2.2. VC for EL-air speech

As shown in Fig. 3, the F_0 contours of EL-air speech vary more than those of EL speech. However, it is difficult for users to intentionally switch between U/V sounds when producing EL-air speech, and therefore, many U/V errors are observed in the F_0 contours of EL-air speech. In other words, the F_0 contours of EL-air speech are not sufficiently accurate to present natural speech, and therefore, it is necessary to convert not only the spectral parameters but also the F_0 features of EL-air speech.

Table 2 shows the relationship between the source and target acoustic features in the VC from EL-air speech to normal speech (EL-air-to-Speech). For the estimation of target spectral parameters and aperiodic components, the segmental feature vectors created from spectral parameters of EL-air speech are used for the source data, which are constructed by the same manner shown in Fig. 8.

For the estimation of F_0 , two methods are considered for constructing source segmental features: (1) concatenate the static feature vectors of spectra and F_0 , and then construct segmental feature vectors, or (2) construct segmental feature vectors of spectra and F_0 separately, and then concatenate these vectors (see Fig. 9). The first method includes a risk that the F_0 information might not be presented when the segmental feature vectors are constructed. Therefore, in this article we employ the second method. *Log*-scaled F_0 values are extracted from EL-air speech and normal speech to be used for the static F_0 features. For the target F_0 features, joint feature vectors of static and delta F_0 values are constructed.

Table 2
Source and target acoustic features for EL-air-to-Speech.

Source	Target
Spectrum	Spectrum
Spectrum and F_0	F_0
Spectrum	Aperiodic components

3.2.3. Data recording of training data for EL-air-to-Speech

It is essential in VC to use source and target features that correlate with each other. To obtain these data, a laryngectomee trained how to control F_0 using the air-pressure sensor for one month. The laryngectomee further trained to control F_0 for another three weeks so that the pitch of EL-air speech sounds similar to that of the target normal speech. EL-air speech was recorded after this training.

However, we noticed that it was too difficult for the laryngectomee to mimic the target pitch pattern by controlling F_0 with exhaled air. Moreover, the F_0 patterns of the recorded EL-air speech were significantly different from those of the target speech. Therefore, we additionally recorded target normal speech for the recorded EL-air speech. In this recording, a target speaker uttered target speech while mimicking the pitch patterns of the recorded EL-air speech as naturally as possible. Note that the F_0 contours of the recorded EL-air speech are still different from those of the re-recorded target speech. For example, the F_0 contours of the recorded EL-air speech vary discontinuously, although those of the target normal speech smoothly vary. These differences are expected to be removed by VC in our speaking-aid system.

3.3. Speaking-aid system for silent EL speech

3.3.1. Framework of the system

The speaking-aid system for silent EL speech is shown in Fig. 7. In our previous work, input silent EL speech was only converted to a whispered voice (Nakamura et al., 2007). However, in this article we convert silent EL speech

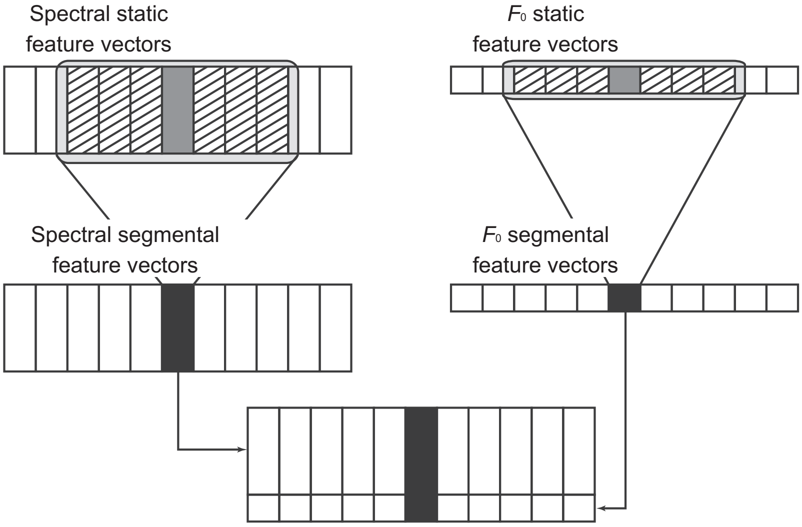

Fig. 9. Flow chart of the construction of segmental feature vectors from spectral and F_0 feature vectors.

to normal speech. This system addresses not only the unnaturalness of EL speech but also the problem of noisy sound source signals. The four components of this system are the same as those of the other speaking-aid systems described above. To address the problem of noisy sound source signals, this system employs the new sound source unit described in Section 2.4. Produced silent EL speech is captured by the NAM microphone described in Section 2.3 as well as NAM speech signals. In this article, silent EL speech is recorded in a sound proof room that is extremely quiet space. Note that silent EL speech has energy to provide auditory feedback to the speaker; however, the energy is extremely low. Therefore, silent EL speech is hardly heard by listeners, and it is easily masked by external noises. As a result of the masking phenomenon, it is expected that this system can be used in daily conversations as well as with telecommunication systems.

3.3.2. VC for silent EL speech

VC from silent EL to normal speech (silent-EL-to-Speech) is performed by the method used for EL-to-Speech described in Section 3.1.2. Three acoustic features of spectral parameters, F_0 , and aperiodic components are estimated from only the spectral parameters of silent EL speech. Segmental feature vectors are constructed in the manner shown in Fig. 8 and are set as the source features. Joint features consisting of static and delta features are set as the target features.

4. Experimental evaluations

4.1. Experimental conditions

The source speaker was a laryngectomee (Japanese male), who was proficient in speaking with an EL. The target speaker was a laryngeal speaker (also a Japanese male). Both speakers recorded 50 phoneme-balanced sentences, which served as training data, and 30 utterances from newspaper articles, which served as test data. The source speaker recorded three types of alaryngeal speech: EL speech, silent EL speech using a pulse train with a frequency of 100 Hz, and EL-air speech. The source speaker used an EL named ‘yourtone’ (Hashiba et al., 2001) to produce EL speech and EL-air speech. The target speaker recorded normal speech. EL-air speech and normal speech were recorded in the manner described in Section 3.2.3. The recorded target speech data were shared among the settings of VC for each source speech. All speech signals were recorded in a sound proof room. Silent EL speech was recorded with a NAM microphone. Other speech signals, which were EL speech, EL-air speech, and the target normal speech, were recorded with a head-set microphone.

The number of mixture components of the GMMs used to estimate spectral parameters and aperiodic components was set to 32 and the number of mixture components of the GMM used to estimate F_0 was set to 16, 32, 64, and 128, respectively. The 0 through 24 mel-cepstral coefficients, which were extracted by mel-cepstral analysis (Fukada et al., 1992), were used as the source spectral parameters, where the 0 coefficient captured power information. The concatenating frame length for the source segmental feature vectors was set to 8. After the concatenation of frames, 50- and 2-dimensional components were extracted frame by frame to construct spectral and F_0 segmental feature vectors, respectively. F_0 contours of the source speech signals were automatically extracted using a robust algorithm for pitch tracking (Talkin, 1995). Acoustic features of the target speech were extracted by STRAIGHT analysis (Kawahara et al., 1999).

Following mel-cepstral distortion was used to measure the spectral conversion accuracy.

$$\mathrm{Mel-cd[dB]} = \frac{1}{T}\sum_{t=1}^{T}\frac{10\sqrt{2\sum_{d=1}^{D}\{(tar_t[d]-conv_t[d])^2\}}}{\ln 10}, \tag{10}$$

where $tar_d[d]$ and $conv_t[d]$ were dth coefficients of the target and converted mel-cepstrum at the frame t, respectively. T was the total number of frames. The accuracy of F_0 was evaluated by U/V decision error rates and the correlation coefficient between frames for which the estimated and target frames both correspond to be voiced as shown in Fig. 10. Target F_0 contours of the test data including U/V information were given as references for calculating U/V decision error rates. The correlation coefficients were calculated utterance by utterance in this article.

Six laryngeal speakers subjectively and independently evaluated stimuli in terms of (1) naturalness, (2) intelligibility, and (3) preference in this order, which were all rated using a five-point-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Our aid systems were expected to be used in our daily life. Therefore, we evaluated the systems using not isolated words but continuous sentences. For conducting writing test to evaluate the intelligibility, we needed to prepare lots of sentences to prevent subjects from learning the contents. On the other hand, it was difficult for the source speaker to produce many

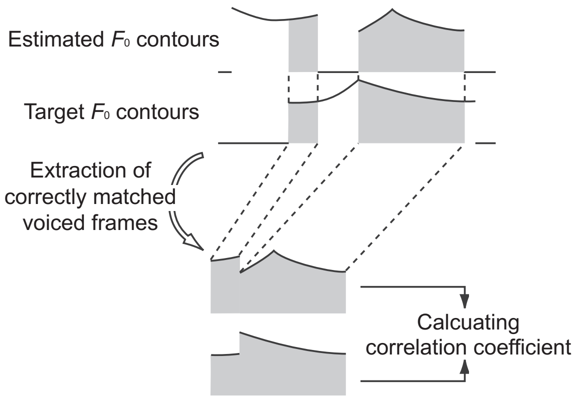
Estimated F_0 contours
Target F_0 contours
Extraction of correctly matched voiced frames
Calcuating correlation coefficient

Fig. 10. Calculation of correlation coefficients between voiced frames of estimated and target F_0 contours.

sentences for each source speech. Therefore, we conducted an opinion test on intelligibility. Five different sentences out of 30 test sentences were randomly selected for each speaker. Note that the subjects conducted perceptual evaluations for three times using different test data set. Seven types of stimuli were evaluated: analysis-synthesized target normal speech, three types of recorded source speech signals (EL speech, silent EL speech, and EL-air speech), and three types of converted speech signals from each source speech. All stimuli were randomly given using a headphone in a sound proof room. Converted and the target speech waveforms were synthesized using a mel log spectrum approximation filter (Imai et al., 1983). When synthesizing the converted speech waveforms, the GV parameters of only the converted spectra were taken into account. F_0 contours of the test data were estimated using 64 mixture components.

4.2. Experimental results

4.2.1. Objective results

Table 3 shows the results of mel-cepstral distortion. As shown in the table, VC strongly enhances the spectra of source speech signals. The result of spectral conversion from EL speech or EL-air speech is better than that from silent EL speech. The result of EL-air speech conversion is slightly better than that of EL speech conversion in terms of the average and the standard deviation; however, the performances of VC for EL speech and EL-air speech are very close each other.

Table 3
Mel-cepstral distortions without power information.

Source speech	Source – Target ± S.D. [dB]	Converted – Target ± S.D. [dB]
Silent EL speech	11.42 ± 0.36	4.55 ± 0.24
EL speech	8.96 ± 0.31	4.25 ± 0.26
EL-air speech	9.51 ± 0.30	4.12 ± 0.20

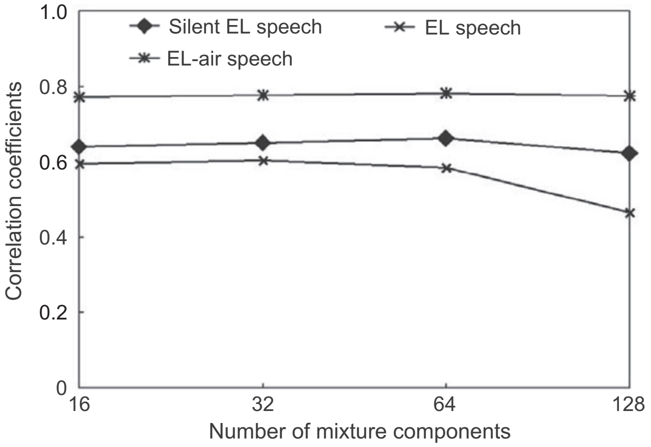

Fig. 11. Correlation of F_0 contours between target and converted F_0 contours.

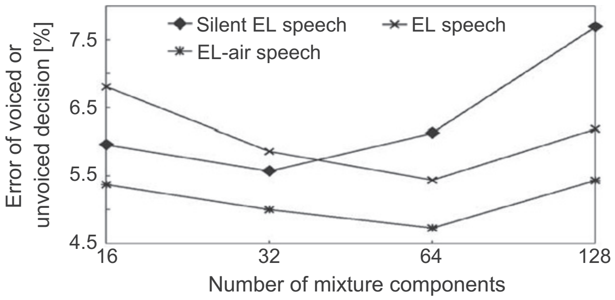

Fig. 12. U/V errors for converted F_0 contours.

Fig. 11 shows the correlation coefficients between voiced values of target and converted F_0 contours. From this figure, the results of using silent EL speech and EL speech are almost the same. Overtraining, however, occurred in the result of EL speech with 128 mixture components. On the other hand, the result using EL-air speech is better than that using the other two speech signals, and moreover, this result remains almost the same for different numbers of mixture components.

Fig. 12 shows U/V decision error rates for converted F_0 contours. According this figure, EL-air speech produces fewer U/V decision errors than the other two speech signals, even though overtraining is observed when GMM has 128 mixture components.

Fig. 13 shows examples of F_0 contours of source EL-air speech, converted speech, and the corresponding target speech. As shown in this figure, VC is highly effective for making the F_0 contours of the EL-air speech smoothly and continuously vary while suitably switching between U/V decisions.

4.2.2. Subjective results

Fig. 14 shows the mean opinion score (MOS) for each test.

The result shows that our systems dramatically improve the naturalness of each source speech. Namely, the results of the converted speech have the significance to the corresponding source speech considering 5% significance level. However, the results of the three converted speech does not have the significance each other considering 5% significance level, since the naturalness of the converted silent EL speech, the converted EL speech, and the converted EL-air speech varies between 2.31 and 3.09, 2.94 and 3.53, and 3.08 and 3.92, respectively. This result shows that our speaking-aid systems are highly effective for improving the naturalness of three types of EL speech signals.

The intelligibility of the converted silent EL speech is scored to vary between 1.77 and 2.49, and that of the source silent EL speech is scored to vary between 1.13 and 1.47. Therefore, the converted silent EL speech has the significance to source EL speech on intelligibility.

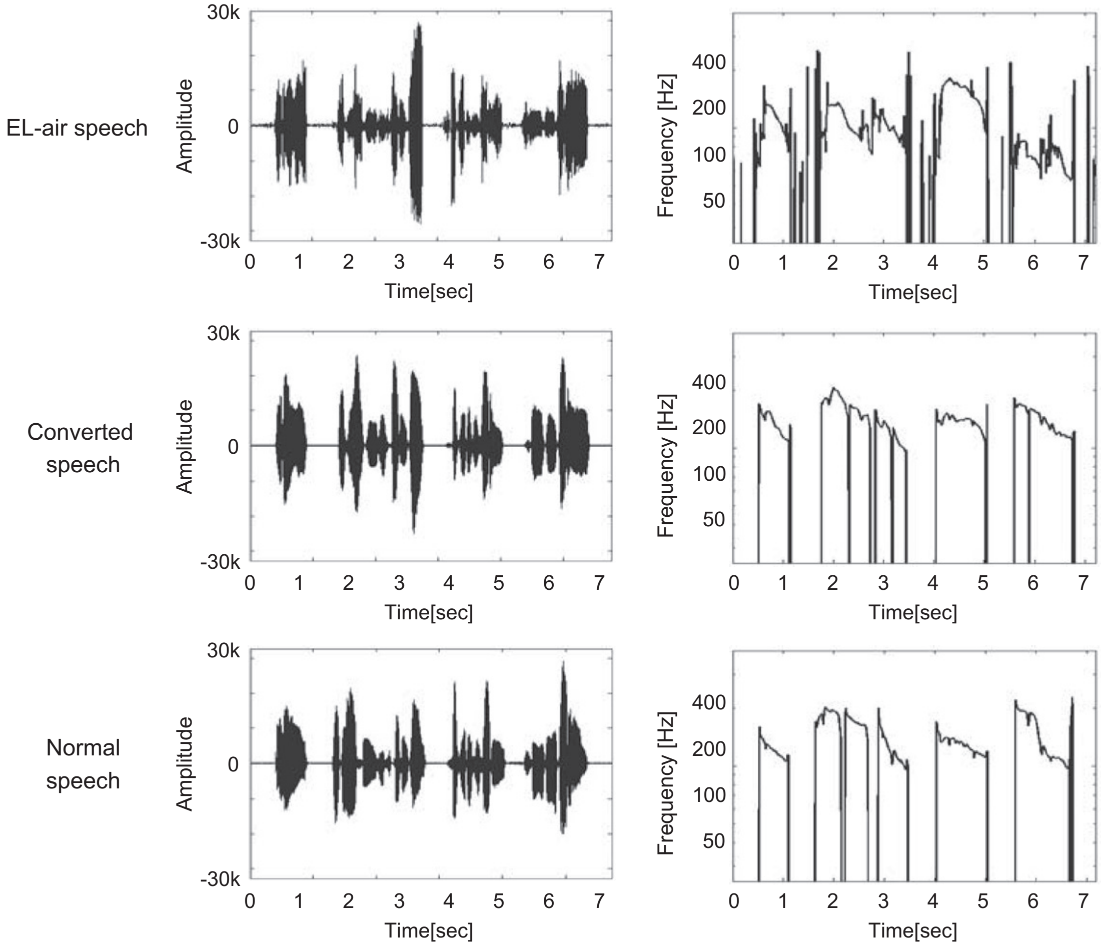

Fig. 13. Examples of F_0 contours in EL-air-to-Speech.

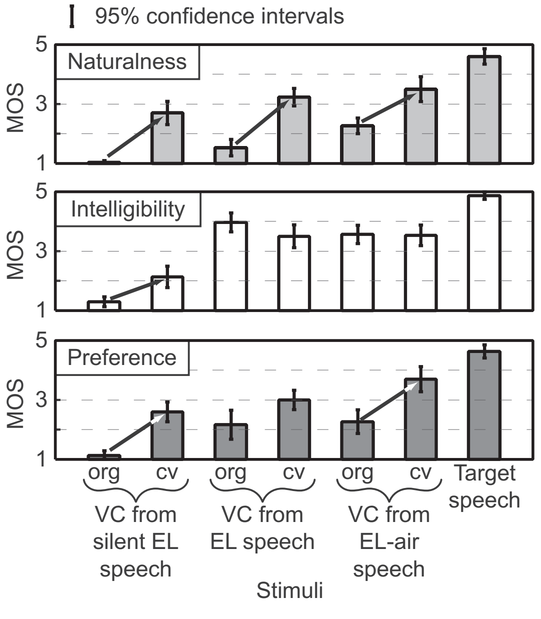

Fig. 14. Result of subjective evaluation by six laryngeal speakers. ‘org’ and ‘cv’ denote original and converted speech, respectively. Arrows in figure shows that improvement was observed considering 5% significance level.

However, other two VC experiments have no significance between before and after VC procedure. Although the mean value of the intelligiblity of the converted EL speech is slightly degraded than that of the source EL speech, no significant differences are observed when we consider the 95% confidence intervals. The intelligibility of the converted EL-air speech is similar to that of the source EL-air speech, although each result is better than the result of the converted silent EL speech.

The preference for each converted speech scored higher than that for the corresponding source speech. The preference of the EL speech conversion also has the significance considering 5% significance level. Namely, the preference of the source and the converted EL speech vary between 1.68 and 2.66 and 2.67 and 3.3, respectively. It is interesting that the preference for the source EL-air speech is similar to that for source EL speech, although EL-air speech includes more F_0 information than EL speech. Although the mean value of the preference of converted EL-air speech is better than that of the converted EL speech, there is no statistical significance considering 5% significance level since the preference of the converted EL speech varies between 2.67 and 3.33 and that of the EL-air speech varies between 3.28 and 4.12.

5. Discussion

Differences of mel-cepstral distortions after VC are not large, on the other hand, certain tendency is appeared. Since the target speech is shared among each VC experiment, the result of the mel-cepstral distortion is caused by the difference of the type of the source speech, namely, the amount of information source speech has. Since the sound source signals of EL speech and EL-air speech are larger than those of silent EL speech, the result of VC from EL speech or EL-air speech is better than that from silent EL speech. Moreover, EL speech and EL-air speech signals were captured by an air-conductive microphone, whereas silent EL speech signals are captured by a body-conductive microphone. Therefore, EL and EL-air speech contain much more information than silent EL speech. On the other hand, the difference between EL speech and EL-air speech is only the existence of the air-pressure sensor, which mainly affects not spectral information but F_0 information. Therefore, it is reasonable to obtain similar results of VC between EL speech and EL-air speech.

A reason why the mean value of the intelligibility of the source EL speech is higher than that of the source EL-air speech is a difference of training periods to produce EL speech or EL-air speech. The source speaker has used an EL in his daily life for more than 10 years, and therefore, the speaker well knows how to produce intelligible EL speech. The source speaker trained how to produce EL-air speech and control its F_0 using his exhaled air for around one month and three weeks. After the speech recording, the source speaker gave us a comment that he might be able to produce more natural and intelligible EL-air speech if he could get more training periods. As a result, it is reasonable that the intelligibility of the source EL speech is higher than that of the source EL-air speech, since the duration of training of the source speaker to produce EL-air speech is shorter than that to produce EL speech.

VC surely reduced radiation noises. However, the current VC procedure might cause another distortion on intelligibility, which is our future work. As a result, there are no significant differences in the intelligibility before and after converting EL speech and EL-air speech.

The subjective result of preference has a similar tendency to that of naturalness than that of intelligibility. This is because the improvement of the naturalness affected the subjects much more than the difference of intelligibility. Moreover, the result of preference of the converted EL-air speech is better than that of the converted EL speech. Considering the objective result of spectral and F_0 estimation accuracy, the result of preference score comes from the improvement of F_0 estimation accuracy. The gap between the converted EL-air speech and target speech in the preference result is the combination of a defect of intelligibility due to the spectral estimation and that of naturalness due to F_0 estimation. For more improvement of F_0 estimation, it might be effective to use more training data.

From these results of spectral and F_0 estimation, we conclude that VC dramatically improves the spectral performance of source speech signals, and the use of the air-pressure sensor effectively improves the converted speech quality.

6. Conclusion

In this article we proposed three speaking-aid systems for EL speech, EL-air speech, and silent EL speech with the aim of addressing two problems EL speech: its unnaturalness and noisy sound source signals. We introduced a statistical VC technique for addressing the unnaturalness of EL speech and employed a new sound source unit, which generated signals with extremely low energy, to address the problem of noisy sound source signals. We also introduced an air-pressure sensor that enables laryngectomees to manipulate the F_0 contours of an EL using their exhaled air. This device was expected to enable the estimation of more natural F_0 contours.

As the result of experimental evaluations, it was shown that VC greatly enhanced the spectral quality. The use of the air-pressure sensor dramatically improved the correlation coefficients of F_0 contours using only voiced frames from 0.58 to 0.78 and also suppressed the proportion of U/V decision errors from 5.44% to 4.73%. Moreover, subjective evaluations indicated that our speaking-aid systems dramatically improved the naturalness of EL speech and converted speech is preferred to the corresponding source EL speech.

Acknowledgments

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

References

Carr, M.M., Schmidbauer, J.A., Majaess, L., Smith, R.L., 2000. Communication after laryngectomy: An assesment of quality of life. Arch. Otolaryngol. Head and Neck Surg. 122 (1), 39–43.

Carr, M.M., Schmidbauer, J.A., Majaess, L., Smith, R.L., 2000. Communication after laryngectomy: an assessment of quality of life. Otolaryngol. Head and Neck Surg. 122 (1), 39–43.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodological) 39 (1), 1–38.

Forastiere, A.A., Goepfert, H., Maor, M., Pajak, T.F., Weber, R., Morrison, W., Glisson, B., Trotti, A., Ridge, J.A., Chao, C., Peters, G., Lee, D., Leaf, A., Ensley, J., Cooper, J., 2003. Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. The New England J. Med. 349 (22), 2091–2098.

Fukada, T., Tokuda, K., Kobayashi, T., Imai, S. 1992. An adaptive algorithm for mel-cepstral analysis of speech. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, pp. 137–140.

Goldstein, E.A., Heaton, J.T., Kobler, J.B., Stanley, G.B., Hillman, R.E., 2004. Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. IEEE Trans. Biomed. Eng. 51 (2).

Hashiba, M., Uemi, N., Oikawa, M., Yamaguchi, Y., Sugai, Y., Ifukube, T., 2001. Industrialization of the electrolarynx with a pitch control function and its evaluation. IEICE Trans. Inform. Syst. J94-D-II (6), 1240–1247, in Japanese.

Hayes, M., 1951. Rehabilitation of the laryngectomee. A Cancer J. Clinic. 1, 147–152.

Hočevar-Boltežar, I., Žargi, M., 2001. Communication after laryngectomy. Radiat. Oncol. 35 (4), 249–254.

Hosoi, Y., Sakaguchi, T. 2003. Silent Voice Input System Without Exhalation –Theory and Applications, Technical Report of IEICE, SP2003-105, pp. 13–16.

http://www.toshin-ha.co.jp/ (confirmed on 23.02.2011, in Japanese).

Ifukube, T., 2003. Sound-based Assistive Technology for the Disabled. Corona publishing Co. Ltd., in Japanese.

Imai, S., Sumita, K., Furuichi, C., 1983. Mel log spectrum approximation (MLSA) filter for speech synthesis. Electron. Commun. Jpn. 66 (2), 10–18.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., Thun, M.J., 2008. Cancer statistics, 2008. A Cancer J. Clinic. 58, 71–96.

Kain, A., Macon, M.W. 1998. Spectral voice conversion for text-to-speech synthesis. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 285–288.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. 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. 27 (3–4), 187–207.

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. 2nd Models Anal. Vocal Emissions for Biomed. Appl. (MAVEBA).

Laccourreye, O., Weinstein, G., Naudo, P., Cauchois, R., Laccourreye, H., Brasnu, D., 1996. Supracricoid partial laryngectomy after failed laryngeal radiation therapy. The Laryngoscope 106 (4), 495–498.

Liu, H., Zhao, Q., Wan, M., Wang, S., 2006. Enhancement of electrolarynx speech based on auditoy masking. IEEE Trans. Biomed. Eng. 53 (5), 865–874.

Miyamoto, D., Nakamura, K., Toda, T., Saruwatari, H., Shikano, K. 2009. Acoustic compensation methods for body transmitted speech conversion. In: IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP2009), pp. 3901–3904.

Murakami, K., Araki, K., Hiroshige, M., Tochinai, K., 2004. A method for speech transform from electrolaryngeal speech to normal speech. IEICE Trans. Inform. Syst. J87-D-I (11), 1030–1040, in Japanese.

Nakagiri, M., Toda, T., Kashioka, H., Shikano, K., 2006. Improving body transmitted unvoiced speech with statistical voice conversion. Proc. Interspeech 2006, 2270–2273.

Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2005. Remodeling of the sensor for non-audible murmur (NAM). Proc. Interspeech 2005, 293–296.

Nakajima, Y., Kashioka, H., Campbell, N., Shikano, K., 2006. Non-audible murmur (NAM) Recognition. IEICE Trans. Inform. Syst. E89-D (1), 1–8.

Nakamura, K., Toda, T., Saruwatari, H., Shikano, K., 2007. A speech communication aid system for total laryngectomees using voice conversion of body transmitted artificial speech. IEICE Trans. Inform. Syst. J90-D (3), 780–787, in Japanese.

Ohtani, Y., Toda, T., Saruwatari, H., Shikano, K., 2006. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. Proc. Interspeech 2006, 2266–2269.

Saikachi, Y., Stevens, K.N., Hillman, R.E., 2009. Development and perceptual evaluation of amplitude-based *F*0 control in electrolarynx speech. J. Speech, Lang. Hearing Res. 52, 1360–1369.

http://www.secom.co.jp/personal/medical/myvoice.html (confirmed on 23.02.2011, in Japanese).

Singer, M.I., Blom, E.D., 1980. An endoscopic technique for restoration of voice after laryngectomy. The Anal. Otol. Rhinol. Laryngol. 89, 529–533.

Takahashi, H., Nakano, M., Okusa, T., Hatamura, Y., Kikuchi, Y., Kaga, K., 2001. A voice-generation system using an intramouth vibrator. J. Artif. Organs 4 (4), 288–294.

Talkin, D., 1995. A robust algorithm for pitch tracking. In: Speech Coding and Synthesis. Elsevier Science, pp. 495–518.

Toda, T., Shikano, K., 2005. NAM-to-Speech conversion with Gaussian mixture models. Proc. Interspeech 2005, 1957–1960.

Toda, T., Black, A.W., Tokuda, K., 2007. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio, Speech Lang. Process. 15 (8), 2222–2235.

Toda, T., Nakamura, K., Nagai, T., Kaino, T., Nakajima, Y., Shikano, K., 2009. Technologies for processing body conductive speech detected with non-audible murmur microphone. Proc. Interspeech 2009, 632–635.

Uemi, N., Ifukube, T., Takahashi, M., Matsushima, J. 1994. Design of a new electrolarynx having a pitch control function. In: Proc. 3rd IEEE Internat. Workshop of Robot and Human Communication, pp. 198–203.

Williams, S.E., Watson, J.B., 1985. Differences in speaking proficiencies in three laryngectomee groups. Arch. Otolaryngol. Head and Neck Surg. 111, 216–219.