Electrolaryngeal Speech Enhancement Based on Statistical Voice Conversion

Keigo Nakamura1, Tomoki Toda1, Hiroshi Saruwatari1, and Kiyohiro Shikano1

1Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{kei-naka, tomoki, sawatari, shikano}@is.naist.jp

Abstract

This paper proposes a speaking-aid system for laryngectomees using GMM-based voice conversion that converts electrolaryngeal speech (EL speech) to normal speech. Because valid F0 information cannot be obtained from the EL speech, we have so far converted the EL speech to whispering. This paper conducts the EL speech conversion to normal speech using F0 counters estimated from the spectral information of the EL speech. In this paper, we experimentally evaluate these two types of output speech of our speaking-aid system from several points of view. The experimental results demonstrate that the converted normal speech is preferred to the converted whisper.

Index Terms: Electrolarynx, Laryngectomee, Voice conversion, Speaking-aid

1. Introduction

An electrolarynx is a medical device to enable laryngectomees who have undergone laryngectomy to rehabilitate their voice by giving external sound excitations through skin on the lower jaw. Some problems are included in the speech using an electrolarynx (EL speech), although laryngectomees can get their voices back. One of the most dominant problems is the naturalness of the EL speech. Because a basic electrolarynx drives vibrators with the same cycle, the EL speech is heard with a monotone pitch. The non-human-like sound of the voice gives not only the speaker but also people around the speaker uncomfortable feelings in their verbal communications.

Some electrolarynxes have been developed in Japan to provide users a more natural voice [1, 2]. One electrolarynx named ‘youtone’ tries to achieve the natural voice by using an optional air-pressure sensor [1]. Users can control the intonation with their breath expelled from the tracheostoma which is a hole connecting to their lungs. Both hands are, however, occupied to hold the electrolarynx and the air-pressure sensor, and therefore, the electrolarynx with the air-pressure sensor limits the situations in which users can speak with it. Another electrolarynx named ‘myvoice’ has a pre-defined F0 pattern inside. Users can speak with certain intonation; however, the generated speech is still artificial because the F0 counters are fixed. Another approach comes from the viewpoint of software processing. Murakami et al. have proposed a method of converting the EL speech to normal speech using a conversion dictionary [3]. This approach works in some situations; on the other hand, some utterances cannot be accepted.

We have proposed a speaking-aid system for laryngectomees in which input EL speech is converted to whispering, which is unvoiced hoarse speech using the GMM-based voice conversion technique [4]. This system is supposed to be used in telecommunication. The voice conversion from speech with invalid F0 information to a natural whisper has been proposed as an application of silent speech [5]. Because valid F0 information of the basic EL speech would not be obtained, the EL speech has so far been converted to whispering. This aid system has been experimentally evaluated to demonstrate that the naturalness of the converted whisper is scored higher than that of the original EL speech. Although it seems to work, it would be difficult for users to conduct all verbal communications with whisper.

This paper proposes another speaking-aid system for laryngectomees in which input EL speech is converted to normal speech using the GMM-based voice conversion technique. The main issue of the voice conversion from the EL speech to the normal speech is the estimation of natural F0 counters, which cannot be obtained from the basic EL speech. We use spectral information to estimate both the spectra and the F0 counters. This paper experimentally evaluates the converted whispering and normal speech in order to demonstrate that these are preferable to conventional EL speech and worthy of being set as the output in our aid systems.

This paper is organized as follows. In Section 2, the proposed EL speech enhancement method based on voice conversion is described. In Section 3, converted normal speech is experimentally evaluated. Finally, this paper is concluded in Section 4.

2. Voice Conversion for EL speech enhancement

Our aid system employs statistical voice conversion based on the maximum likelihood criterion using Gaussian mixture models (GMMs) to describe relationships between input and output acoustic features [6]. All procedures of the training and the conversion are automatically conducted without using linguistic information.

2.1. Training process

Let \( x_t = [x_t(1), \ldots, x_t(d_x)]^\top \) and \( y_t = [y_t(1), \ldots, y_t(d_y)]^\top \) be a static input and output feature vector at frame \( t \), respectively, where \( d_x \) and \( d_y \) denote the dimensions of \( x_t \) and \( y_t \), respectively, and \( \top \) denotes transposition. In order to compensate for lost acoustic characteristics at some phonemes such as /h/ due to the laryngectomy, a segment feature vector \( X_t \) is constructed by following processes [5] to be used as an input speech parameter. First, several frames \( (t \pm L) \) are concatenated as \( c_t = [x_{t-L}, \ldots, x_t, \ldots, x_{t+L}]^\top \). And then, a low dimensional feature vector \( X_t \) is extracted from \( c_t \) with PCA procedure as follows: \( X_t = A_c c_t + b_c \), where \( A_c \) and \( b_c \) are the PCA matrix and the bias vector of \( c_t \), respectively. A joint feature vector of output static and dynamic features \( Y_t = [y_t, \Delta y_t]^\top \) is constructed as an output speech parameter.

Training data consisting of input and output speech param-
eters are automatically time-aligned by a dynamic time warping procedure in advance to construct joint parameter vectors \( [X_i^T, Y_i^T] \), \( i=1, \ldots, T \) where \( T \) denotes the number of frames. A GMM models the joint probability density of the input and the output parameter vectors \( \{X_i, Y_i\}_{i=1}^M \) as follows:

\[
P(X_i, Y_i | \lambda) = \sum_{m=1}^M w_m \mathcal{N}( [X_i^T, Y_i^T]^\top ; \mu_m^{(X,Y)}, \Sigma_m^{(X,Y)} )
\]

\[
\mu_m^{(X,Y)} = \begin{bmatrix} \mu_m^{(X)} \\ \mu_m^{(Y)} \end{bmatrix}, \quad \Sigma_m^{(X,Y)} = \begin{bmatrix} \Sigma_m^{(X,X)} & \Sigma_m^{(X,Y)} \\ \Sigma_m^{(Y,X)} & \Sigma_m^{(Y,Y)} \end{bmatrix}
\]

where \( \mathcal{N}(\cdot; \mu, \Sigma) \) denotes the Gaussian distribution with a mean vector \( \mu \) and a covariance matrix \( \Sigma \). \( m \) denotes the mixture component index, and \( M \) denotes the total number of the mixture components. A parameter set of the GMM is denoted by \( \lambda \), which consists of weights \( w_m \), mean vectors \( \mu_m \), and full covariance matrices \( \Sigma_m \) for individual mixture components.

2.2. Basic conversion process

Let \( X = [X_1^T, \ldots, X_T^T]^\top \) and \( Y = [Y_1^T, \ldots, Y_T^T]^\top \) be a time sequence of the input and the output parameter vectors, respectively. The converted feature sequence \( \hat{y} = [\hat{y}_1, \ldots, \hat{y}_T]^\top \) is determined to maximize the likelihood of the conditional probability density of \( Y \) given \( X \) as follows:

\[
\hat{y} = \arg \max_y P(Y | X, \lambda) \quad \text{subject to} \quad Y = W y
\]

where \( W \) is a matrix to extend the static feature sequence to the parameter vector sequence consisting of the static and the dynamic features [8].

The converted speech quality can be more enhanced by considering the global variance (GV) factor [6].

2.3. Voice conversion from EL speech into whispering

Only the GMM to estimate spectral information is trained in the training process. In the conversion process, a time sequence of the input segment vectors \( X \) is constructed frame by frame, and then, the converted whisper is synthesized with the converted spectra and noise excitation.

2.4. Conversion into normal speech

Two GMMs to estimate spectral and \( F_0 \) information are trained. A segment vector \( X_t \), constructed on input spectral features is used to train both one GMM for spectral estimation and the other GMM for \( F_0 \) estimation. In the training procedure of the GMM for \( F_0 \) estimation, a joint \( F_0 \) vector \( Y_t^{(F_0)} = [y_t^{(F_0)}, \Delta y_t^{(F_0)}]^\top \) is used as an output \( F_0 \) parameter vector, where \( y_t^{(F_0)} \) and \( \Delta y_t^{(F_0)} \) denote a log-scaled static and a dynamic \( F_0 \) value at frame \( t \), respectively. In order to take time-alignment between input spectral vectors and output \( F_0 \) vectors, this paper first trains the GMM for spectral estimation. Next, the input spectral features are converted to the output data, and then, a time warping function between the converted and output spectral features is determined by dynamic time warping procedure. The resulting time warping function is used to take alignment between input spectral features and output \( F_0 \) features. Finally, a GMM for \( F_0 \) estimation is trained using joint vectors of \( X_t \) and \( Y_t \).

In the conversion process, first, a time sequence of the input segment vectors \( X \) is constructed frame by frame. Next, \( X \) is independently converted into the spectral and the \( F_0 \) sequence. Finally, the converted normal speech is obtained by filtering excitations, which is designed from the converted \( F_0 \) sequence, with the converted spectral sequence.

3. Experimental Evaluations

3.1. Experimental conditions

The input speaker was a Japanese male laryngectomee in his 50s, who is proficient in speaking with an electrolarynx. The output speaker was a Japanese male non-laryngectomee. Both speakers recorded 49 phoneme-balanced sentences. Out of those, 42 sentences were used as the training data for the GMMs, and the remaining 7 sentences were used as the test data.

The number of a GMM component to estimate spectral parameters was set to 32, and that of another GMM component to estimate \( F_0 \) parameters was set to 8. The 1st through 24th mel-cepstral coefficients, which are extracted by mel-cepstral analysis, were used as the spectral parameters in which the 0th coefficient captures power information. The spectral segment vector of the EL speech was constructed by the following procedures. First, the current, previous and succeeding eight frames were concatenated into one vector, and then, the dimension of the vector was compressed by PCA. As the result, a 50-dimensional vector in each frame was constructed. This segmental vector was used as the input data for both spectral and \( F_0 \) estimation. Acoustic features of the target speech were extracted by STRAIGHT analysis [9].

In the objective evaluation, the mel-cepstral distortion measured the spectral conversion accuracy. The \( F_0 \) accuracies were evaluated by the coefficients of correlation between target and converted \( F_0 \) counters, and voiced or unvoiced judgment errors.

In the subjective evaluation, five non-laryngectomees and the speaker himself evaluated the speech quality in view of 1) intelligibility, 2) naturalness, and 3) preference, which were all evaluated by five-scored opinion scale (1: Bad - 5: Excellent). Non-laryngectomees evaluated preference level from the perspective of the listeners. On the other hand, the laryngectomee evaluated stimuli from the perspective of the user. Four kinds of stimuli were given: 1) original EL speech, 2) converted whispering, 3) converted normal speech excited by only noises, and 4) converted normal speech excited by converted \( F_0 \) counters. 3) was expected to evaluate only spectral features by ignoring excitation features. The difference between 2) and 3) is only the output data of the voice conversion. The voice quality could be further enhanced by considering global variance (GV) parameters [6]. Giving stimuli to the subjects, GV parameters of the spectral features were introduced; on the other hand, those of the excitation features weren’t considered because the GV parameters would encourage the unnaturalness caused by the unsatisfied \( F_0 \) counters.

3.2. Experimental results

3.2.1. Objective results

Figure 1 shows the mel-cepstral distortion in each phoneme. The averaged mel-cepstral distortion between the converted and the target spectrum is 5.4 [dB] with powers and 4.6 [dB] without powers respectively. The mel-cepstral distortion between whispering and converted whispering is 5.5 [dB] with powers
Table 1: Decision errors of voiced or unvoiced frames. For example, “U→V” shows the rate of estimating an unvoiced frame as a voiced frame.

<table>
<thead>
<tr>
<th>U/V Decision</th>
<th>Rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>93.1 (U→U: 49.66, V→V: 43.44)</td>
</tr>
<tr>
<td>Error</td>
<td>6.9 (U→V: 5.13, V→U: 1.77)</td>
</tr>
</tbody>
</table>

and 4.7 [dB] without powers respectively; however, note that the approaches of the feature extraction are different.

The coefficient of correlation of counted and target 0\(_{\text{counters}}\) is 0.38. Table 1 shows misjudgment of voiced or unvoiced frames. Although the coefficient of correlation is not quite high, the judgment errors about voiced or unvoiced frames are not terrible.

3.2.2. Subjective results

Figure 2 shows the subjective results of non-laryngectomees and the laryngectomee. Similar tendencies are seen in the result of both non-laryngectomees and the laryngectomee.

A Intelligibility

The intelligibility of all converted speech is degraded from that of the original EL speech. This is the problem of the current GMM-based voice conversion approach, and we will address this problem in our future work.

Comparing among the converted speech, the intelligibilities of the converted normal speech are almost the same; on the other hand, those scores are degraded from the score of the converted whisper. From this result, we think that unnatural 0\(_{\text{counters}}\) counters are not affecting to the intelligibility so much. Decision of voiced or unvoiced frames, however, greatly affects the intelligibility although the decision errors of voiced or unvoiced frames are not so high, as shown in Table 1.

B Naturalness

Naturalness of all kinds of converted speech is highly scored compared with that of the original EL speech. This result is one of the most valuable in this paper compared to another voice conversion framework [10] in which unvoiced speech is converted to normal speech by the same conversion approach as this paper. We think that this is because the original EL speech is voiced speech with monotone pitch. We often recognize the pitch when we hear a whispered voice, even though its unvoiced speech doesn’t have 0\(_{\text{counters}}\). In other words, we might imagine the pitch in our mind when we hear some voices that don’t have 0\(_{\text{counters}}\) information inside. On the other hand, we cannot recognize natural pitch when we hear monotone EL speech because this speech has definite monotone 0\(_{\text{counters}}\) counters that play roles to suppress our imagination. In [10], the insufficient estimated 0\(_{\text{counters}}\) counters make the converted speech more unnatural than the original unvoiced speech. On the other hand, in our voice conversion from EL speech to normal speech, the naturalness of the estimated 0\(_{\text{counters}}\) counters rises above that of the original EL speech even though the converted 0\(_{\text{counters}}\) counters are still of unsatisfactory quality to achieve the target ones.

The converted normal speech with noise excitations is the lowest scored. This speech sounds extremely hoarse, and in other words, this speech sounds like speech uttered by another type of speech-disabled person.

The laryngectomee scored the converted speech as having high naturalness. From the interview after the evaluation, he explained that this result is derived because the naturalness is determined by the naturalness of the human voice. The laryngectomee also explained that he was able to utter extremely well in some test data so that the noises of the external sound signals did not flow from the lower jaw; however, the speech had only monotone pitch resulting in low naturalness. On the other hand, it concerns him that the speaker individuality of the converted speech is different from that of the laryngectomee. In spite of these ambivalences, the naturalness of the EL speech is the lowest scored because of its mechanical monotone pitch.

C Preference

This result can be seen as the total voice qualities. The converted normal speech with noise excitation isn’t preferred because of its unnaturalness. This speech wouldn’t be used as the output of our aid system. Both non-laryngectomees and the laryngectomee prefer the converted whisper or normal speech with converted 0\(_{\text{counters}}\) to original EL speech even though the estimated 0\(_{\text{counters}}\) accuracy is not satisfied as Fig. 3 shows. This result indicates that the voice conversion of EL speech to normal speech is effective, and these two kinds of speech would be set as the output in our aid system. Note that all listeners haven’t always scored with the similar tendencies. Only one non-laryngectomee who doesn’t prefer the synthesized speech has evaluated in a contrary way. Further improvements of the EL speech enhancement are necessary.

4. Conclusions

This paper proposed a speaking-aid system for laryngectomees that converts arbitrary EL speech utterances to normal ones.
using GMM-based voice conversion to investigate the output speech of our aid systems using the voice conversion. As a result of subjective evaluation, the naturalness of the converted normal speech is much higher than that of the original EL speech because the original EL speech has only monotone pitch. Moreover, both non-laryngectomees and the laryngectomene who is the speaker preferred converted speech to the original EL speech. From these results, voice conversion from EL speech to normal speech is worth being conducted. Our aid system is desirable to set normal speech and whispering as its output speech so that users can select either one they want.

We will try EL speech enhancement for different users and improve the estimation accuracy of $F_0$ counters. We will also employ the Eigen voice techniques [11] to remove concern about different speaker intelligibility.

5. Acknowledgements

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

6. References


Figure 3: An example of waveform, spectrogram, and $F_0$ counters of EL speech, converted normal speech, and normal speech, for a sentence fragment “i g sh w: k a N b a k r i ny w: y o: k u o sh u z a i sh i t a “, where a colon denotes the long vowel.