Electrolaryngeal Speech Modification towards Singing Aid System for Laryngectomees

Kazuho Morikawa* and Tomoki Toda†
* Graduate School of Informatics, Nagoya University, Japan
E-mail: morikawa.kazuho@h.mbox.nagoya-u.ac.jp
† Information Technology Center, Nagoya University, Japan
E-mail: tomoki@icts.nagoya-u.ac.jp

***Abstract*—Towards the development of a singing aid system for laryngectomees, we propose a method for converting electrolaryngeal (EL) speech produced by using an electrolarynx into more naturally sounding singing voices. Singing by using the electrolarynx is less flexible because the pitch of EL speech is determined by the source excitation signal mechanically produced by the electrolarynx, and therefore, it is necessary to embed melodies of songs to be sung in advance to the electrolarynx. In addition, sound quality of singing voices produced by the electrolarynx is severely degraded by an adverse effect of its mechanical excitation sounds emitted outside as noise. To address these problems, the proposed conversion method uses 1) pitch control by playing a musical instrument and 2) noise suppression. In the pitch control, pitch patterns of music sounds played simultaneously in singing with the electrolaryx are modified so that they have specific characteristics usually observed in singing voices, and then, the modified pitch patterns are used as the target pitch patterns in the conversion from EL speech into singing voices. In the noise suppression, spectral subtraction is used to suppress the leaked excitation sounds. The experimental results demonstrate that 1) naturalness of singing voices is significantly improved by the noise suppression and 2) the pitch pattern modification is not necessarily effective in the conversion from EL speech into singing voices.**

I. Introduction

Laryngectomees are people whose larynxes have been removed in surgery for diseases, such as laryngeal cancer. They are unable to generate glottal excitation sounds as they have lost their vocal folds, and consequently, they suffer from serious vocal disorder, i.e., they are unable to speak and sing in a usual manner. Speaking is an important behavior in speech communication with others. Singing is also an important behavior to enjoy our life. Therefore, this vocal disorder causes significantly large degradation in quality of life (QoL) of laryngectomees.

To allow laryngectomees to speak again, several alternative speaking methods have been proposed. One of the typical methods is the use of a electrolarynx, which is a device to mechanically generate source excitation sounds. This method has several advantages compared to the other alternative speaking methods; e.g., electrolaryngeal (EL) speech produced by using the electrolarynx is quite intelligible; and it is easy to learn how to use the electrolarynx. On the other hand, this method also has disadvantages. One of the biggest disadvantages is that EL speech sounds mechanical and unnatural due to the use of mechanically generated source excitation sounds. F_0 patterns of the source excitation sounds generated by the electrolarynx are usually flat or are given by predetermined patterns because it is essentially difficult to generate natural F_0 patterns corresponding to linguistic contents in real time. The use of these F_0 patterns causes significant degradation in naturalness of EL speech. Moreover, the sound excitation signals produced by the electrolarynx are usually emitted outside as noise. Therefore, sound quality of EL speech is easily degraded by this adverse effect.

There have been proposed various attempts at addressing the issues of EL speech. To enhance EL speech, several approaches based on signal processing have been proposed, such as noise suppression based on comb filtering [1], smoothing of acoustic features [2], formant manipulation [3], and noise suppression based on auditory masking [4]. Recently statistical approaches based on statistical voice conversion techniques [5] have also been proposed to significantly improve naturalness of EL speech, such as statistical voice conversion from EL speech into natural voices [6], speaker identity control based on one-to-many eigenvoice conversion [7], F_0 control based on statistical feature prediction [8], and direct F0 control of the electrolarynx [8]. It is expected that naturalness of EL speech will be improved more by the further development of these speaking aid techniques.

On the other hand, there have been few attempts at developing singing aid techniques for laryngectomees. In singing, it is essential to control pitch of EL speech, i.e., F_0 pattern of EL speech. One existing approach is to set F_0 patterns corresponding to melodies of predetermined songs and embed them in advance to the electrolarynx as a function to sing these embedded songs. However, varieties of the embedded songs are limited. Moreover, flexibility in singing is very low as only the predetermined songs are allowed laryngectomees to sing. To achieve more flexible singing aid, it is necessary to develop a function capable of freely controlling pitch of EL speech as laryngectomees want. Furthermore, it is also important to develop a technique to produce naturally sounding singing voices rather than mechanically sounding ones.

To address these issues, we aim to develop a new singing aid system for laryngectomees based on pitch control by playing a musical instrument while singing with the electrolarynx. Towards the development of such a system, in this paper we propose a conversion method from EL speech into singing

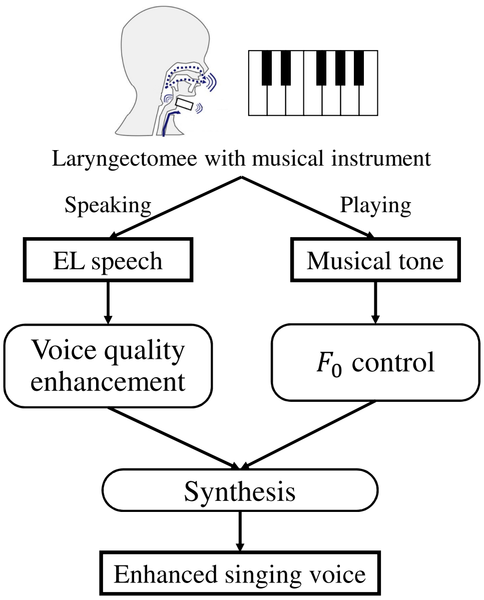

Fig. 1. Process of the proposed singing aid system using electrolarynx for laryngectomee.

voices based on 1) F_0 pattern control of EL speech using pitch of music sounds and 2) EL speech enhancement using noise suppression. In the F_0 pattern control, the F_0 pattern modification method for converting normal voices into singing voices [9], [10] is applied to EL speech conversion. In the EL speech enhancement, the noise suppression method based on spectral subtraction for EL speech [11] is employed to improve sound quality of singing voices generated from EL speech. The effectiveness of the proposed methods is investigated by conducting subjective evaluations.

II. Singing Aid System for Laryngectomees

Figure 1 shows an overview of the proposed singing aid system for laryngectomees. In the proposed system, a laryngectomee sings a song with the electrolarynx simultaneously playing its melody with a musical instrument. A speech analysis-synthesis technique is used to generate naturally sounding singing voices using phonetic information extracted from the EL speech and pitch information extracted from the musical sounds.

In this paper, we focus on two enhancement processes in the proposed system, 1) F_0 pattern control and 2) voice quality enhancement. In the F_0 pattern control, the F_0 patterns extracted from the musical sounds are modified so that they have specific characteristics usually observed in natural singing voices. In the voice qualit enhancement, spectral subtraction is performed to suppress the source excitation signals emitted outside from the electrolarynx.

III. Proposed Conversion Methods from EL Speech to Singing Voice

In this section, we describe the proposed methods for converting EL speech into singing voices based on the F_0 pattern control and the voice quality enhancement. In the F_0 pattern control, we apply the conventional technique for converting speaking voices into singing voices [9], [10] to the proposed system. In the voice quality enhancement, we apply the conventional EL speech enhancement technique based on the spectral subtraction [11] to the proposed system.

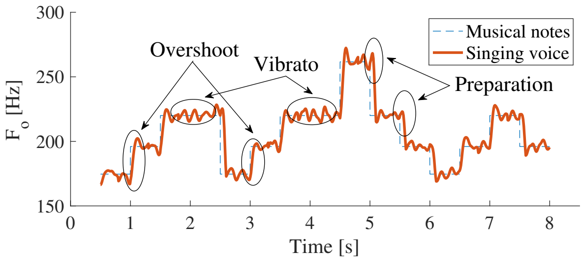

Fig. 2. F_0 patterns of musical notes and singing voice.

A. F_0 pattern control

It has been reported in [9] that specific characteristics observed in acoustic parameters of natural singing voices. In this paper, we focus on those related to F_0 patterns and implement a modification technique to add them to F_0 patterns extracted from the musical sounds.

In the conventional speech-to-singing method [10], the following four types of F_0 fluctuations are focused on as the typical characteristics of F_0 patterns of natural singing voices:

Overshoot: a deflection exceeding the target note after a note change.

Vibrato: a quasi-periodic frequency modulation (4-7 Hz).

Preparation: a deflection in the direction opposite to a note change observed just before the note change.

Fine fluctuation: an irregular frequency fluctuation higher than 10 Hz.

Figure 2 shows the F_0 patterns of musical notes and those of singing voices to demonstrate these four types of F_0 fluctuations.

In the proposed method, these four types of F_0 fluctuations are added to the F_0 patterns extracted from the music signals in the same manner as used in the speech-to-singing method [10]. The overshoot, vibrato, and preparation are added by applying the following infinite impulse response (IIR) filter to the extracted F_0 patterns:

$$H(s) = \frac{k}{s^2 + 2\zeta\omega s + \omega^2} \tag{1}$$

where ω is the natural frequency, ζ is the damping coefficient, k is the proportional gain of the system. Overshoot and preparation are expressed by damping model ($0 < |\zeta| < 1$), and vibrato is expressed by oscillation model ($|\zeta| = 0$). These parameters are detemined manually in this paper. On the other hand, the fine fluctuation is generated by using white noise. White noise signals are high-pass-filtered with cut off frequency of 10 Hz and its amplitude is normalized so that its maximum is 5 Hz.

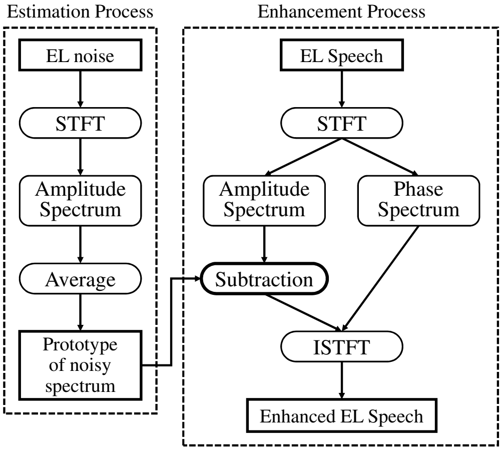

Fig. 3. Process of spectral subtraction for EL speech.

B. *Voice quality enhancement*

It has been reported in [11] that the spectral subtraction method [12] is very effective to reduce the adverse effect of the sound excitation signals emitted outside from the electrolarynx for enhancing EL speech sound quality. It is expected that the same enhancement effect will be observed in the conversion from EL speech into natural singing voices in the proposed method.

Figure 3 shows the noise suppression process based on the spectral subtraction for EL speech. To accurately estimate the averaged amplitude spectrum of noisy signals as a prototype of noisy spectrum, we record the leaked source excitation signals of the electrolarynx in advance as the EL noise with a close-talking microphone. In this recording, a speaker uses the electrolaryx in the same manner as used in speaking or singing while keeping his/her mouth closed in order to minimize the effect of the produced EL speech on the recorded signals. The prototype of noisy spectrum $|\hat{L}(\omega)|$ is calculated by averaging amplitude spectrum of the EL noise. In enhancement process, the amplitude spectrum of noise-suppressed EL speech $|\hat{S}(\omega, t)|$ is calculated as follows:

$$|\hat{S}(\omega,t)| = \begin{cases} |Y(\omega,t)| - 2|\hat{L}(\omega)|, & (|Y(\omega,t)| > 2|\hat{L}(\omega)|) \\ 0, & (\text{otherwise}) \end{cases} \quad (2)$$

where $|Y(\omega, t)|$ is the amplitude spectrum of EL speech. Finally, enhanced EL speech is obtained by inverse short-time Fourier transformation (ISTFT) using the amplitude spectrum of noise-suppressed EL speech and the original phase spectrum of the EL speech.

After noise suppression, spectral envelope and aperiodicity of EL speech are produced by analyzing it. However, there is a case where EL speech can't be analyzed into desired parameters due to incorrect F_0 estimation and failure of unvoice/voice (U/V) decision. In order to address this problem, F_0 of EL speech is given in advance under the assumption that it is almost constant, and on/off information of the electrolarynx is regarded as U/V information.

TABLE I
PARAMETER SETTINGS OF F_0 CONTROL

	Duration [ms]	ω	ζ	k
Overshoot	300	34.8	0.403	34.8
Preparation	100	34.8	0.603	34.8
Vibrato	-	34.8	0	0.0164

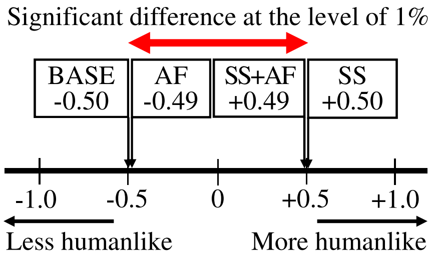

Fig. 4. Result of subjective evaluation to investigate the effectiveness of the proposed singing aid system using spectral subtraction.

IV. Experimental Evaluations

A. *Experimental conditions*

In order to investigate the effectiveness of the proposed method, we conducted a subjective evaluation on naturalness of singing voices. As an experiment through the simulation, we used music sounds generated with MIDI and recorded EL speech samples sung by one non-disabled male person using an electrolarynx while listening to the music sounds. Six phrases of Japanese songs were recorded. These recorded EL speech samples and the music sounds were used to generate singing voices in the proposed system.

In the evaluation, the following 4 methods were evaluated:

BASE: directly used the F_0 patterns of the music sounds.
AF: BASE + the proposed F_0 pattern modification.
SS: BASE + the proposed voice quality enhancement.
SS+AF: the proposed F_0 pattern modification and voice quality. enhancement

The preference test on naturalness of singing voices was performed with Scheffé's method. The number of listeners was 10. In the spectral subtraction, we used FFT spectrum with 1024 points. In the F_0 pattern modification, WORLD [13] was used as the analysis-synthesis method. The filter parameters for the F_0 pattern modification were set as shown in Table 1.

B. *Experimental results*

Figure 4 shows a result of the preference test. We can see that the naturalness of singing voices is significantly improved by using the voice quality enhancement based on the spectral subtraction. This result corresponds to that observed in the conventional EL speech enhancement processing [11]. On the other hand, the F_0 pattern modification is not necessary to improve the naturalness of singing voices. This result is different from that observed in the conventional speech-to-singing processing [10].

Figure 5 shows several examples of F_0 patterns and spectrograms of the EL speech, the converted voice with BASE, and the converted voice with SS+AF. We can see that the F_0 patterns of BASE have relatively large fluctuations even

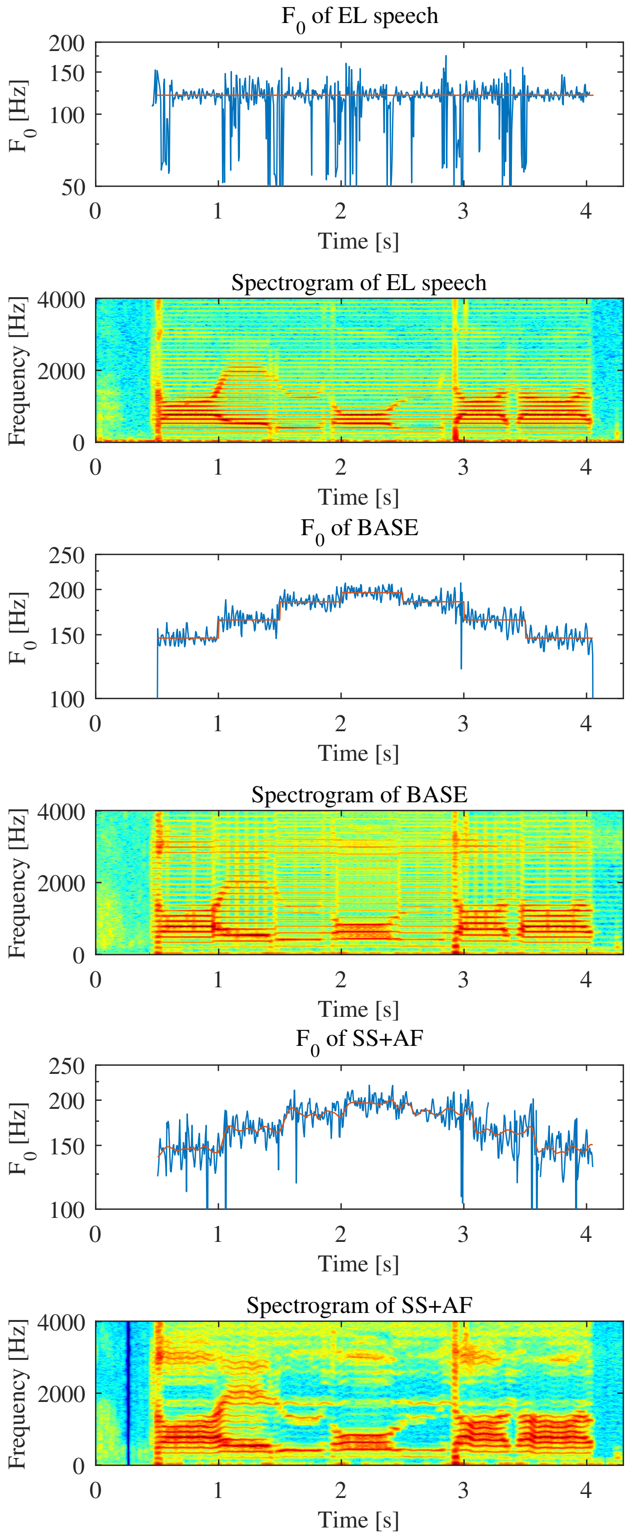

Fig. 5. Example of F_0 patterns and spectrograms of EL speech and synthesized singing voice. In the first, third and fifth graph, blue lines show results of F_0 analysis of them and brown lines show F_0 patterns given in synthesis of them.

though step-like F_0 patterns are used in generation. It is expected that these fluctuations come from the spectrogram extracted from the EL speech. Consequently, the effect of F_0 pattern modification on naturalness of singing voices is limited.

V. Conclusions

In this paper, towards the development of a singing aid system allowing laryngectomees to flexibly sing a song as they want, we have proposed the conversion methods from electrolaryngeal (EL) speech into more naturally sounding singing voices. As a result of a subjective evaluation on the naturalness of singing voices, it has been demonstrated that 1) the naturalness of singing voices is significantly improved by the voice quality enhancement based on the noise suppression for the EL speech and 2) the F_0 pattern modification doesn't always yield naturalness improvements in the proposed conversion. In future work, we plan to investigate statistical approaches to the F_0 pattern conversion and the spectral conversion, and also to implement a real-time conversion system.

VI. Acknowledgment

Part of this work was supported by JST, Sakigake JPMPJR1657.

References

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

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

[3] H. R. Sharifzadeh, I. V. McLoughlin, F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. Biomedical Engineering*, Vol. 57, No. 10, pp. 2448–2458, 2010.

[4] H. Liu, Q. Zhao, M. Wan, S. Wang, "Enhancement of electrolarynx speech based on auditory masking," *IEEE Trans. Biomedical Engineering*, Vol. 53, No. 5, pp. 865–874, 2006.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[6] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, Vol. 54, No. 1, pp. 134–146, 2012.

[7] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, K. Shikano, "Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 1, pp. 172–183, 2014.

[8] K. Tanaka, T. Toda, G. Neubig, S. Sakti, S. Nakamura, "Direct F0 Control of an Electrolarynx based on Statistical Excitation Feature Prediction and its Evaluation through Simulation," *INTERSPEECH*, pp.31–35, 2014.

[9] T. Saitou, M. Unoki, M. Akagi, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, Vol.46, N0.3 pp.405–417, 2005.

[10] T. Saitou, M. Goto, M. Unoki, M. Akagi, "Speech-to-Singing Synthesis System: Vocal Conversion from Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices," *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.215–218, 2007.

[11] K. Tanaka, T. Toda, G. Neubig, S. Sakti, S. Nakamura, "A Hybrid Approach to Electrolaryngeal Speach Enhancement Based on Noise Reduction and Sattistical Excitation Generation," *IEICE Transactions on Information and Systems*, Vol.97, No.6, pp.1429–1437, 2014.

[12] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust, Speech, Signal Processing*, Vol.27, No.2, pp.113–120, 1979.

[13] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, Vol. E99-D, No. 7, pp. 1877–1884, 2016.