An Evaluation through Simulation of Electrolarynx Control based on Statistical F_0 Prediction for Multiple Speakers

Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{ko-t, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

***Abstract*—An electrolarynx is a device that artificially generates excitation sounds to produce electrolaryngeal (EL) speech. Although proficient laryngectomees can produce intelligible EL speech by using this device, it sounds quite unnatural due to the mechanical excitation. To address this issue, we have proposed several EL speech enhancement methods using statistical voice conversion and showed that statistical prediction of excitation parameters, such as F_0 patterns, was essential to significantly improve naturalness of EL speech. Based on this result, we have also proposed a direct control method of F_0 patterns of excitation sounds generated from electrolarynx based on the statistical excitation prediction as an EL speech enhancement method applicable to face-to-face conversation. In our previous work, this direct control method was evaluated through its simulation using only a single laryngectomee's EL speech and it was demonstrated that this method enables to improve naturalness of EL speech while preserving its listenability. However, because quality of EL speech highly depends on proficiency of each laryngectomee, it is necessary to apply this method to other laryngectomees and evaluate its effectiveness. In addition, we need to evaluate this method from more various perspectives, such as not only naturalness and listenability but also intelligibility. In this paper, we apply the direct control method to multiple speakers consisting of two real laryngectomees and one non-laryngectomee and evaluate its performance through simulations in terms of naturalness, listenability, and intelligibility. The experimental results demonstrate that the proposed method yields significant improvements in naturalness of EL speech for multiple laryngectomees while keeping its listenability and intelligibility high enough.**

I. Introduction

Electrolaryngeal (EL) speech is produced by one of the major alternative speaking methods for laryngectomees. EL speech is produced using an electrolarynx, which is typically held against the neck to mechanically generate artificial excitation signals. The generated excitation signals are conducted into the speaker's oral cavity, and they are articulated to produce EL speech. EL speech is relatively intelligible but its naturalness is very low owing to the fundamental frequency (F_0) patterns of the mechanically generated excitation signals.

To address this issue of EL speech, there have been proposed several control methods of F_0 patterns of the excitation signals generated from an electrolarynx additionally using intentionally controllable signals, such as expiratory air pressure [1], up and down switch control by a finger [2], and forearm movements [3]. Although these methods can change the F_0 patterns, it is inherently difficult to intentionally control those signals to generate natural F_0 patterns varying corresponding to linguistic contents.

To generate more natural F_0 patterns, we have proposed a control method [4] based on statistical excitation prediction [5] [6] [7]. In this framework, F_0 patterns are predicted not using some additional signals as in the other control methods but using only the produced EL speech signals as shown in Fig. 1. Statistical voice conversion techniques [8] [9] have been successfully applied to this prediction processing. Relatively natural F_0 patterns are predicted using statistics extracted in advance from a parallel data consisting of utterance pairs of EL speech and natural speech. Our preliminary experimental results through a simulation have demonstrated that the proposed method yields significant improvements of naturalness while causing no degradation in listenability compared to the original EL speech. However, in our preliminary experiment, we evaluated the effectiveness of the proposed method using only one laryngectomee. Because quality of EL speech highly depends on proficiency of each laryngectomee and it also affects F_0 prediction accuracy, it is necessary to apply the proposed method to other laryngectomees and evaluate its effectiveness. Moreover, we need to evaluate it from various perspectives, such as not only naturalness and listenability but also intelligibility.

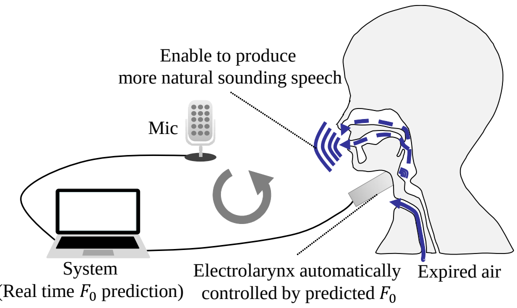

Fig. 1. A direct electrolarynx control methods using real time statistical F_0 prediction for laryngectomees.

In this paper, we apply the proposed method to multiple speakers consisting of two real laryngectomees and one non-laryngectomee and evaluate its performance in terms of naturalness, listenability, and intelligibility. A simulation experiment is conducted in the evaluation as done in our previous work [4]. As a result, it is shown that the proposed method yields significant improvements in naturalness of EL speech while preserving its listenability and intelligibility high enough for multiple speakers.

II. Electrolarynx control with statistical F_0 prediction

A. Statistical F_0 Prediction

This method is a part of statistical voice conversion techniques [8] [9] to predict F_0 patterns of normal speech from spectral parameters of EL speech. It consists of training and prediction processes.

In the training process, we train a Gaussian mixture model (GMM) to model the joint probability density [10] of the source and target features using the corresponding joint feature vector set generated by performing automatic frame alignment such as Dynamic Time Wrapping (DTW). As the source features, spectral segment features of EL speech are extracted from mel-cepstra at multiple frames around the current frame [11]. As the target features, smoothed continuous F_0 patterns [7] are extracted from natural speech.

In the prediction process, the smoothed continuous F_0 patterns of the target normal speech are predicted over all frames utterance by utterance from the spectral segment features of EL speech using the trained GMM based on maximum likelihood estimation of speech parameter trajectories considering global variance (GV) [9]. Finally, silence frames are automatically detected using waveform power of EL speech and unvoiced excitation signals are generated only at those frames. Note that a real time prediction process can be achieved by using a computationally efficient real-time voice conversion method [12] based on the low-delay conversion algorithm [13].

B. Direct Control of Electrolarynx

Our proposed system to directly control the excitation signals generated from an electrolarynx is shown in the left side of Fig. 2. This system consisting of prediction and articulation processes. In the prediction process, F_0 values are predicted frame by frame using the real-time voice conversion method from EL speech produced by a laryngectomee. In the articulation process, the laryngectomee articulate the excitation sounds generated from the electrolarynx based on the predicted F_0 values to produce the EL speech. Therefore, this system allows laryngectomees to directly produce enhanced EL speech with more natural F_0 patterns corresponding to linguistic contents.

In this system, the produced EL speech suffers from a misalignment between spectral information determined by articulation and the predicted F_0 patterns because the real-time statistical F_0 prediction causes a constant processing delay of 50 to 70 msec [12]. Namely, F_0 patterns are constantly delayed from the spectral information. Although we have found that this delay doesn't cause any adverse effects on naturalness and listenability of EL speech for only a single laryngectomee, it is necessary to more investigate this effect on EL speech for other laryngectomees.

Moreover, the statistical F_0 prediction is affected by acoustic mismatches between the training and prediction processes. In the training process, the traditional EL speech with usual excitation sounds is used as an input for the prediction. On the other hand, in the prediction process, the enhanced EL speech with more natural excitation sounds is used. Although only spectral features are used as an input feature for the prediction, they are also affected by F_0 values of the excitation signals in particular if a very simple spectral analysis method such as fast Fourier transform (FFT) is used.

To address this issue, we have proposed two approaches, a model-based approach and a feature extraction approach. The former approach uses EL speech samples resynthesized by widely changing F_0 values to train the GMM accepting EL speech with various F_0 values. The latter approach uses a spectral analysis method robust to the periodicity of the excitation signals. STRAIGHT analysis [14] is used in this paper. To significantly reduce computational cost of STRAIGHT analysis, the predicted F_0 value is directly used to skip an F_0 extraction process.

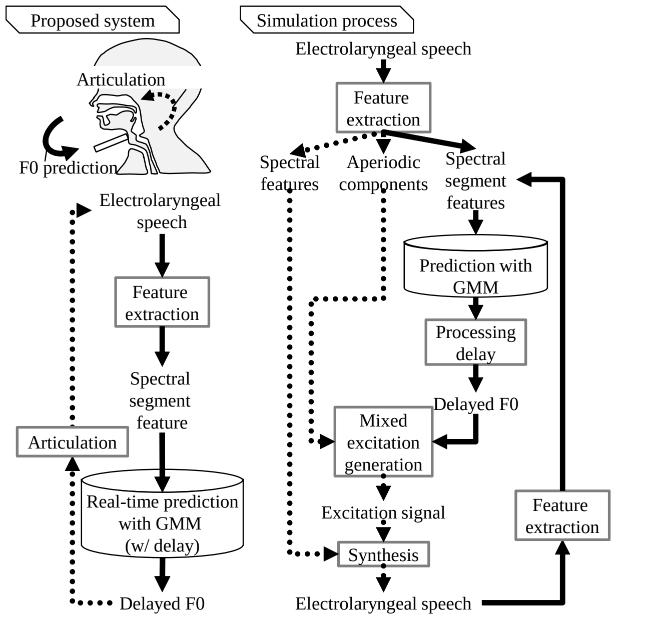

Fig. 2. The proposed system and its simulation implementation.

III. Evaluation method through simulation

As the first step for implementation of the proposed system, we investigate the performance of the proposed system a simulation experiment as in our previous work [4]. The simulated implementation of the proposed system is also shown in Fig. 2. EL speech signals produced by the excitation signals based on the predicted F_0 values are artificially generated using STRAIGHT analysis/synthesis method. In this paper, we use the batch-type conversion algorithm of which conversion accuracy is almost equivalent to that of the low-delay conversion algorithm.

At first, 1) we extract spectral envelope parameters and aperiodic components (AC) [15] from the original EL speech in advance by using STRAIGHT analysis. These features capture acoustic properties depending on articulation and the excitation signals leaked out from the electrolarynx, except for periodicity of the excitation signals. These are used to approximate the EL speech production process. Then, 2) spectral segment features are extracted from EL speech and F_0 patterns are predicted by the statistical F_0 prediction. 3) The predicted F_0 patterns are delayed to consider the delay time caused by real time prediction process. 4) Using the delayed F_0 patterns and the extracted AC, excitation signals are generated using the mixed excitation model [16]. 5) Finally, the enhanced EL speech is approximately synthesized by filtering the generated excitation signals with the extracted spectral envelope parameters. Note that this is a result of using the spectral segment features extracted from the original EL speech, and therefore it is not affected by the predicted F_0 patterns. To consider the impact of the predicted F_0 patterns on the spectral segment features, 6) the spectral segment features are extracted again from the synthesized EL speech and F_0 pattern prediction is also performed again using the extracted spectral segment features. Step 3) to step 6) are iteratively repeated until the predicted F_0 patterns converge. If they converge, the proposed system may be expected to work stably because the EL speech produced with the predicted F_0 patterns is consistent with that used in the spectral segment feature extraction.

IV. Experimental Evaluation

A. Experimental Conditions

We conducted an objective evaluation for evaluating prediction accuracy of F_0 patterns and three subjective evaluations on intelligibility, naturalness, and listenability. The source speech were EL speech of two laryngectomees and speech of another non-disabled speaker, and the target speech was normal speech of one non-disabled speaker. Sampling frequency was set to 16 kHz.

We employed FFT analysis or STRAIGHT analysis to extract the mel-cepstra of EL speech as the spectrum parameters. Note that F_0 values of EL speech in STRAIGHT analysis were constantly set to 100 Hz, which was almost equal to F_0 values of the excitation signals generated by the electrolarynxes. The frame shift length was set to 5 msec. We extract segment feature from the mel-cepstra at current $\pm$ 4 frames. F_0 values of normal speech were extracted with STRAIGHT F_0 analysis and continuous F_0 patterns were generated as the target feature using a low-pass filter with 10 Hz cut-off frequency. The mean F_0 value of normal speech was around 220 Hz.

We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training, and the remaining 10 utterance pairs were used for evaluation. The number of mixture components was set to 32. Speaker-dependent GMMs were trained. In the training data generation process described in Section II-B, F_0 values were shifted to 150, 200, and 250 Hz, and totally 160 EL speech samples were used to train the GMM. The processing delay time in the simulation experiment was set to 70 msec.

The EL speech generated by the following four systems were mainly evaluated:

- **EL**: Original EL speech
- **BASELINE**: Enhanced speech by our previously proposed hybrid system, where the enhanced speech was generated with vocoding process and presented from a loud speaker [7]. Therefore, the enhanced speech was not affected by the processing delay and the predicted F_0 values unlike the proposed direct control system.
- **MIX**: Enhanced speech through the simulation with the processing delay using the GMM trained with the training data generation process.
- **STRAIGHT**: Enhanced speech through the simulation with the processing delay using robust spectral analysis with STRAIGHT.

In the objective evaluation, the correlation coefficient between the predicted and natural F_0 patterns was calculated. To clarify the impact of the acoustic mismatches caused by the predicted F_0 values on the statistical F_0 prediction accuracy, we also evaluated two simulation systems without the training data generation process or the robust spectral analysis, ”NORMAL” and ”NORMAL+matched.” In ”NORMAL+matched”, to reduce the effect of the acoustic mismathces, the predicted F_0 patterns were shifted so that their average was equal to that of the training data (i.e., 100 Hz). This modification was not performed in ”NORMAL.”

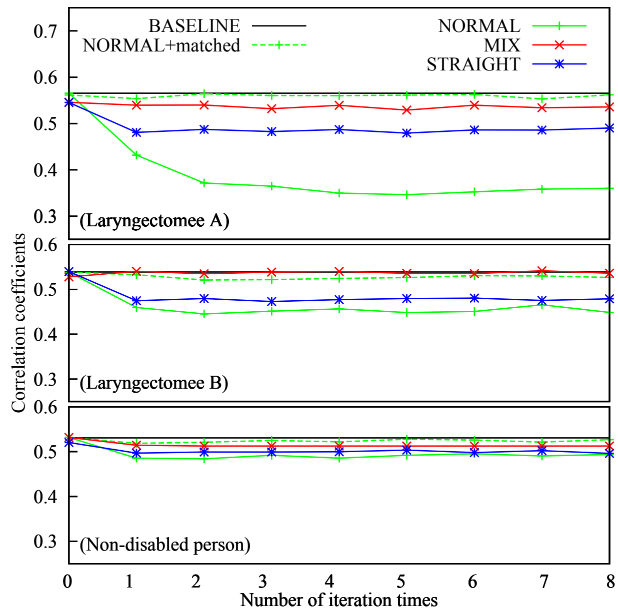

Fig. 3. Prediction accuracy for F_0 correlation coefficient of each person.

TABLE I
Result of dictation test on intelligibility.

	Word correct rate [%]	Word accuracy [%]	Number of replays
Laryngectomee A			
EL	97.88	98.15	1.39
BASELINE	96.72	96.45	1.79
MIX	98.43	97.73	1.21
STRAIGHT	98.29	98.15	1.40
Laryngectomee B			
EL	97.03	92.76	1.58
BASELINE	94.85	88.92	1.67
MIX	95.94	93.89	1.46
STRAIGHT	96.80	94.46	1.67
Non-disabled person			
EL	97.09	94.74	1.87
BASELINE	96.76	93.18	2.27
MIX	97.11	95.60	1.58
STRAIGHT	97.21	94.03	1.87

In the subjective evaluations, we conducted a dictation test on intelligibility and two opinion tests on naturalness and listenability. The opinion score was set to a 5-point scale (i.e., 1 (very poor) to 5 (excellent)). The number of listeners was 13 in the dictation test and 5 in the opinion tests.

B. Experimental Results

Figure 3 shows the result of the objective evaluation. We can see that correlation coefficients of all systems converge and the simulation process works reasonably well. If the acoustic mismatches are not caused by the predicted F_0, such as the system “NORMAL+matched”, the correlation coefficient is constant over the iterative process in the simulation. On the other hand, it can be observed from “NORMAL” that the correlation coefficient significantly degrades in the mismatched situations. This degradation is effectively alleviated by using the training data generation “MIX” or the robust spectral analysis “STRAIGHT.”

Table I shows the result of the dictation test on intelligibility. We found that all systems tend to have high scores and there is no large difference between each system of the non-disabled person. On the other hands, as for laryngectomees, although

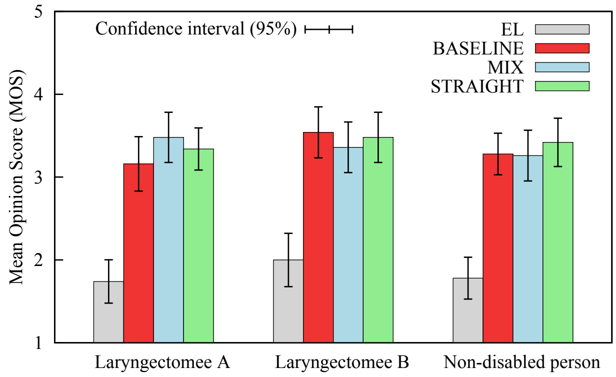

Fig. 4. Result of opinion test on naturalness.

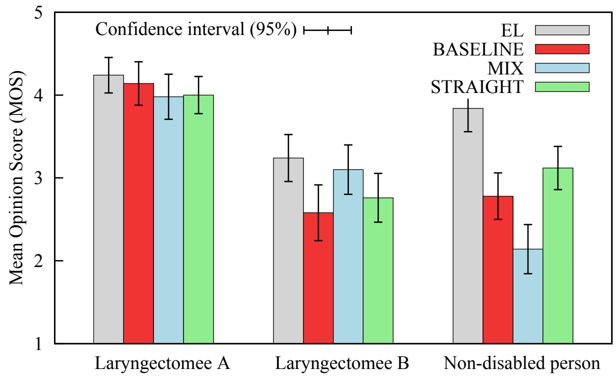

Fig. 5. Result of opinion test on listenability.

"BASELINE" tends to cause slight degrade, we found that simulated systems, such as "MIX" and "STRAIGHT", could reduce these adverse effects. Especially, as for the word accuracy, we could find these improvements clearly. Moreover, as for the system "MIX", the number of replays are less than the others systems.

Figure 4 shows the result of the opinion test on naturalness. The original EL speech is very unnatural, but its naturalness can be significantly improved by "BASELINE" as reported in [7]. The proposed systems "MIX" and "STRAIGHT" can also significantly improve the naturalness. Because no statistically significant difference can be observed between "BASELINE" and the proposed systems "MIX" and "STRAIGHT", it is revealed that misalignment of F_0 patterns does not cause any degradation in naturalness.

Figure 5 shows the result of the opinion test on listenability. We found that the results depend on each person. As for laryngectomy A, there is no difference between all systems. On the other hands, as for larymgectomee B and non disabled person, enhanced systems caused degradations. There is a possibility that is the influence caused by synthesized speech using VOCODER. Note that even if we perceive that it is hard to understand the linguistic contents, actually we understand those contents as shown in Table I.

V. Conclusions

In this paper, for multiple laryngectomees, we constructed electrolaryngeal (EL) speech enhancement systems that directly controls F_0 values of the excitation signals generated by an electrolarynx based on statistical excitation prediction. Moreover, we performed more detailed evaluations including a dictation test on intelligibility. As for evaluations, we conducted simulation experiments to evaluate the effectiveness of the proposed system, investigating whether or not the enhanced EL speech is significantly affected by the processing delay of F_0 prediction and acoustic mismatches caused by the dynamically predicted F_0 values, which are always observed in the proposed system. The experimental results have shown that they cause no significant differences in either naturalness or intelligibility and the proposed system can significantly improve naturalness of EL speech while preserving its high intelligibility for multiple laryngenctomees.

VI. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26280060.

References

[1] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima, "Design of a new electrolarynx having a pitch control function," in *Proc. 3rd IEEE International Workshop of Robot and Human Communication*, Jul 1994, pp. 198–203.

[2] Y. Kikuchi and H. Kasuya, "Development and evaluation of pitch adjustable electrolarynx," in *Proc. Speech Prosody 2004, International Conference.*, March 2004, pp. 761–764.

[3] K. Matsui, K. Kimura, Y. Nakatoh, and Y. O. Kato, "Development of electrolarynx with hands-free prosody control," in *Proc. SSW8*, August 2013, pp. 273–277.

[4] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Direct f0 control of an electrolarynx based on statistical excitation feature prediction and its evaluation through simulation," in *Proc. INTERSPEECH*, Sep 2014.

[5] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," in *Proc. Speech Communication*, vol. 54, no. 1, January 2012, pp. 134 – 146.

[6] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 172–183, January 2014.

[7] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation," *IEICE Transactions on Information and Systems*, June 2014.

[8] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, March 1998.

[9] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, November 2007.

[10] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, May 1998, pp. 285–288.

[11] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, November 2012.

[12] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." in *Proc. INTERSPEECH*, September 2012.

[13] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," in *Proc. INTERSPEECH*, September 2008, pp. 1076–1079.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," in *Proc. Speech Communication*, vol. 27, no. 3. Elsevier, April 1999, pp. 187–207.

[15] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. MAVEBA*, September 2001, pp. 13–15.

[16] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, September 2006, pp. 2266–2269.