An Enhanced Electrolarynx with Automatic Fundamental Frequency Control based on Statistical Prediction

Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti and Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, Japan
{ko-t, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

ABSTRACT

An electrolarynx is a type of speaking aid device which is able to mechanically generate excitation sounds to help laryngectomees produce electrolaryngeal (EL) speech. Although EL speech is quite intelligible, its naturalness suffers from monotonous fundamental frequency patterns of the mechanical excitation sounds. To make it possible to generate more natural excitation sounds, we have proposed a method to automatically control the fundamental frequency of the sounds generated by the electrolarynx based on a statistical prediction model, which predicts the fundamental frequency patterns from the produced EL speech in real-time. In this paper, we develop a prototype system by implementing the proposed control method in an actual, physical electrolarynx and evaluate its performance.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues - Assistive technologies for persons with disabilities

Keywords

electrolaryngeal speech, automatic fundamental frequency control

1. INTRODUCTION

Electrolaryngeal (EL) speech is one of the major speaking methods used by laryngectomees who are people who have had their larynx usually due to laryngeal cancer. 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 are articulated to produce EL speech. EL speech is relatively intelligible but its naturalness is very low owing to unnatural fundamental frequency (F_0) patterns of the mechanically generated excitation signals. Consequently, quality of life of laryngectomees is significantly degraded.

ASSETS'15, October 26–28, 2015, Lisbon, Portugal.
ACM 978-1-4503-3400-6/15/10.
DOI: http://dx.doi.org/10.1145/2700648.2811340.

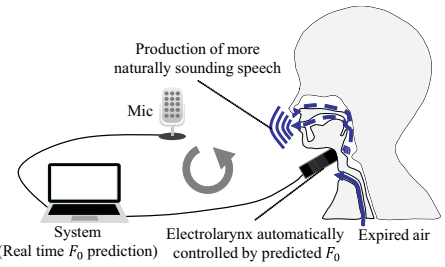

Figure 1: Proposed system to directly control electrolarynx using real-time statistical F_0 prediction.

To generate more natural F_0 patterns, we have proposed a method to control F_0 based on the statistical F_0 prediction [1]. In our proposed system based on this method, F_0 patterns are predicted from the produced EL speech signals as shown in Fig. 1. Relatively natural F_0 patterns can be predicted using statistics extracted in advance from parallel data consisting of utterance pairs of EL speech and natural speech. Therefore, this system allows laryngectomees to directly produce more natural EL speech in the same manner as in the traditional speaking method using the conventional electrolarynx. Our preliminary experimental results through simulation [1] have demonstrated that the proposed method yields significant improvements in naturalness while causing no degradation in listenability and intelligibility compared to the original EL speech.

In this paper, we develop a prototype system by implementing our proposed F_0 control method in an actual, physical electrolarynx and evaluate its performance. The experimental results demonstrate that the prototype system generates more natural excitation sounds, as in the simulation.

2. METHODOLOGY

Direct Control of Excitation Signals of Electrolarynx based on Statistical F_0 Prediction: Our proposed system allows a laryngectomee to produce EL speech with predicted F_0 patterns using two processes: prediction and articulation [1]. In the prediction process, the F_0 value is predicted frame by frame using the real-time voice conversion algorithm [2] from EL speech produced by the laryngectomee. This process causes a constant processing delay of 50 ms to make it possible to predict relatively natural F_0 patterns varying corresponding to linguistic contents. In the articulation process, the laryngectomee produces the EL speech by articulating the excitation sounds generated

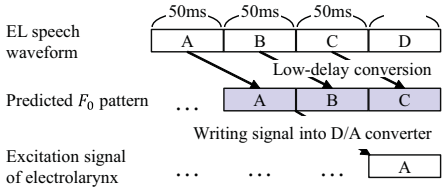

Figure 2: Latency caused by each process

from the electrolarynx based on the predicted F_0 values. These two processes are simultaneously and continuously performed. Consequently, EL speech with the predicted F_0 patterns can be directly produced by the laryngectomee but it always suffers from misalignment between the articulated sounds and F_0 patterns caused by the processing delay of 50 ms. The results of our previous evaluation through simulation have demonstrated that the impact of this misalignment on perception is small [1].

Development of Prototype System: A prototype system based on our proposed technique has been developed using a laptop and a digital/analog (D/A) converter shown in Table 1. As shown in Fig. 1, EL speech produced from mouth of a laryngectomee is detected with a standard close-talk microphone. The EL speech signal is recorded on a laptop and F_0 patterns of normal speech are predicted on the fly by using the real-time statistical F_0 prediction. The predicted F_0 values are linearly converted to voltage values to control the F_0 values of the excitation signal generated by an electrolarynx. Then, an electric signal corresponding to the determined voltage values is generated with the D/A converter connected from the laptop to the electrolarynx. The electrolarynx changes the F_0 values of the excitation signal according to the input electric signal generated from the D/A converter. As shown in Fig. 2, additional latency is caused by the D/A converter in the prototype system. It takes around 50 ms to write the digital signal on the D/A converter. Moreover, the digital signal to be written needs to be determined before starting writing. Consequently, the D/A part always causes 100 ms latency. In total, 150 ms latency is caused in the prototype system. Note that this latency in the D/A part may be addressed by the development of a special device for the electrolarynx.

3. EVALUATION AND RESULTS

We conducted an objective evaluation for evaluating prediction accuracy of F_0 patterns generated by the developed prototype compared to that confirmed in the simulation process [1]. The source speech was EL speech uttered by one non-disabled male speaker, and the target speech was normal speech uttered by a professional female speaker. Each speaker uttered about 50 sentences in the ATR phonetically balanced sentence set [3]. We conducted a 5-fold cross validation test in which 40 utterance pairs were used for training of a statistical F_0 prediction model, and the remaining 10 utterance pairs were used for evaluation. Sampling frequency was set to 16 kHz.

Table 1: Electronic devices on the prototype system

Electrolarynx	Yourtone
Microphone	Crown CM-311A
CPU of the laptop	Intel(R) Core(TM) i5-4200U
D/A converter	AIO-160802AY-USB

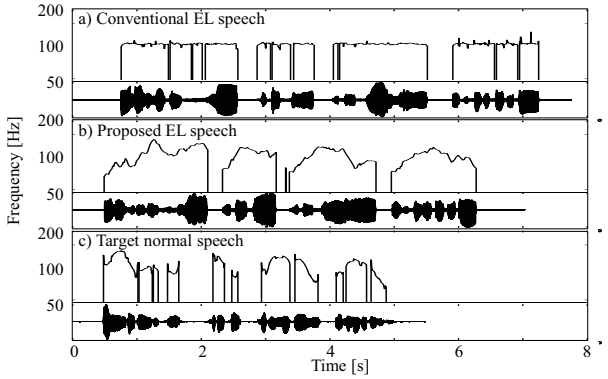

Figure 3: Example of waveforms and F_0 patterns of EL speech signals generated by conventional and proposed systems and those of target natural speech.

Experimental results: The F_0 correlation coefficient between the prototype system and the simulation process is 0.91. This result shows that F_0 patterns predicted by the prototype system strongly correlate to those by the simulation process, which have already been confirmed to be effective for improving naturalness of EL speech [1]. An example of EL speech signals and their F_0 patterns are shown in Fig. 3. We can see that the prototype system makes it possible to produce EL speech with more naturally varying F_0 patterns compared to the conventional EL speech.

4. CONCLUSIONS

In this paper, we have developed a prototype system by implementing our proposed F_0 control method of an electrolarynx based on the statistical F_0 prediction technique and evaluate its performance. The experimental results have demonstrated that the prototype system enables a speaker to produce more naturally sounding electrolaryngeal speech.

5. ACKNOWLEDGMENTS

This work was supported in part of JSPS KAKENHI Grant Numbers: 26280060 and the authors would like to thank Mr. Y. Sugai of Densei Communication Inc., Japan, for advise to control an electrolarynx.

6. REFERENCES

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Direct F_0 control of an electrolarynx based on statistical excitation feature prediction and its evaluation through simulation," in *Proc. INTERSPEECH*, Sep 2014.

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

[3] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database," ATR Technical Report, TR-I-0166, Sep 1990.