Real-Time Vibration Control of An Electrolarynx based on Statistical F_0 Contour Prediction

Kou Tanaka[1], Tomoki Toda[2], Graham Neubig[1], and Satoshi Nakamura[1]
[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, JAPAN
Email: {ko-t, neubig, s-nakamura}@is.naist.jp
[2]Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, JAPAN
Email: tomoki@icts.nagoya-u.ac.jp

***Abstract*—An electrolarynx is a speaking aid device to artificially generate excitation sounds to help laryngectomees produce electrolaryngeal (EL) speech. Although EL speech is quite intelligible, its naturalness significantly suffers from the unnatural fundamental frequency (F_0) patterns of the mechanical excitation sounds. To make it possible to produce more naturally sounding EL speech, we have proposed a method to automatically control F_0 patterns of the excitation sounds generated from the electrolarynx based on the statistical F_0 prediction, which predicts F_0 patterns from the produced EL speech in real-time. In our previous work, we have developed a prototype system by implementing the proposed real-time prediction method in an actual, physical electrolarynx, and through the use of the prototype system, we have found that improvements of the naturalness of EL speech yielded by the prototype system tend to be lower than that yielded by the batch-type prediction. In this paper, we examine negative impacts caused by latency of the real-time prediction on the F_0 prediction accuracy, and to alleviate them, we also propose two methods, 1) modeling of segmented continuous F_0 (CF_0) patterns and 2) prediction of forthcoming F_0 values. The experimental results demonstrate that 1) the conventional real-time prediction method needs a large delay to predict CF_0 patterns and 2) the proposed methods have positive impacts on the real-time prediction.**

I. Introduction

Production of electrolaryngeal (EL) speech is 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 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.

To address this issue of EL speech, several methods have been proposed to control 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 switched controlled by a finger [2], and forearm movements [3]. Although these methods can change the F_0 patterns, it is inherently difficult to control these signals to generate natural F_0 patterns corresponding to linguistic content of the speech.

To generate more natural F_0 patterns, we have proposed a method to control F_0 values [4] based on statistical F_0 prediction [5] [6] [7]. In this framework, F_0 patterns are predicted not according to signals consciously provided by the speaker 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 task. 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. Our preliminary experimental results through a simulation 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 [4][10].

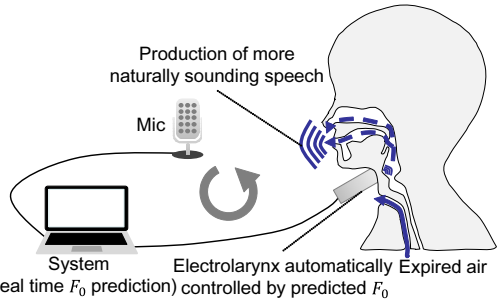

Fig. 1. Proposed system to directly control an electrolarynx using real-time statistical F_0 prediction for laryngectomees.

In our previous work [11], we have developed a prototype system by implementing our proposed F_0 control method of an electrolarynx based on statistical F_0 prediction. We have confirmed that the naturalness of EL speech is significantly improved by the use of our prototype system. On the other hand, we have also found that improvements of the naturalness of EL speech yielded by the prototype system tend to be lower than that yielded by our conventional system based on batch-type prediction. It is possible that the real-time statistical F_0 prediction requires a large delay time to ensure sufficient F_0 prediction accuracy. Moreover, because the prototype system also needs additional processing delay, its large delay time significantly degrades naturalness of the enhanced EL speech in the form of causing larger mismatch between articulation behavior and the predicted F_0 patterns.

In this paper, we examine the effect of the delay time on the real-time statistical F_0 prediction. Moreover, to address the negative impacts caused by latency of the real-time prediction on the F_0 prediction accuracy, we propose two methods: 1) modeling of segmented continuous F_0 (CF_0) patterns to shorten the required delay time in real-time statistical F_0 prediction and 2) prediction of forthcoming F_0 values to cancel the impact of the processing delay of the prototype system. Through the experimental evaluation, we demonstrate that 1) the delay time required to predict CF_0 patterns in the conventional prediction method can be significantly reduced by using the segmented CF_0 modeling strategy, and 2) the negative impacts of the processing delay can be effectively

alleviated by predicting the forthcoming F_0 values.

II. Direct Control of Electrolarynx based on Statistical F_0 Prediction [4]

A. Strategy of Direct Control of Electrolarynx

Our proposed system (shown in Fig. 1) to directly control F_0 patterns of the excitation signals generated from an electrolarynx consists of prediction and articulation processes. In the prediction process, the F_0 value is predicted from EL speech produced by a laryngectomee frame by frame using the real-time voice conversion algorithm. In the articulation process, the laryngectomee produces the EL speech by articulating the excitation sounds generated from the electrolarynx based on the predicted F_0 values. Therefore, this system allows laryngectomees to directly produce enhanced EL speech with more natural F_0 patterns corresponding to linguistic contents.

B. Statistical F_0 Prediction Method

We briefly review our statistical F_0 prediction method [5] [6] [7], which exploits the idea of statistical voice conversion techniques [8] [9]. The aim of this method is to predict F_0 contours from the spectral parameters of EL speech. We describe training, batch-type prediction, real-time prediction processes and problems of delay arising in real-time processing.

In the training process, a parameter set $\boldsymbol{\lambda}$ of the joint probability density function $P([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top|\boldsymbol{\lambda})$ modeled by a Gaussian mixture model (GMM) is trained [12] using a parallel data set consisting of utterance pairs of the EL speech and target normal speech, where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ denote source features and target features at time frame t, respectively, and $\top$ denotes transposition. The corresponding joint feature vectors can be obtained by performing automatic frame alignment between the EL speech and the target normal speech with dynamic time warping. As the source features $\boldsymbol{X}_t$, the spectral segment features of EL speech are extracted on a frame-by-frame basis from the mel-cepstra at multiple frames around the current frame t [13]. The target features $\boldsymbol{Y}_t = [y_t, \Delta y_t]^\top$ consist of the static and delta (time derivative) components of the log-scaled F_0 value y_t, extracted on a frame-by-frame basis from the target normal speech. Note that to improve prediction accuracy, we use continuous F_0 (CF_0) patterns which interpolate unvoiced frames of F_0 patterns by using spline interpolation, and remove micro-prosody [14].

In the batch-type prediction process, given the spectral segment sequence $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ of EL speech, the most likely F_0 sequence $\boldsymbol{y} = [y_1, \ldots, y_T]^\top$ can be obtained as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \tag{1}$$

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) &= \sum_{\boldsymbol{m}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) \\ &\simeq P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}) P(\hat{\boldsymbol{m}}|\boldsymbol{X}, \boldsymbol{\lambda}), \end{aligned} \tag{2}$$

$$\hat{\boldsymbol{m}} = \underset{m}{\operatorname{argmax}}\, P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}), \tag{3}$$

$$\begin{aligned} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}) &= \mathcal{N}(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y|X)}, \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)}) \\ &= \textstyle\prod_{t=1}^{T} \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)}, \boldsymbol{D}_{\hat{m}_t}^{(Y|X)}), \end{aligned} \tag{4}$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$ denotes the joint static and dynamic feature vector sequence, $\boldsymbol{W}$ is a linear transform to append dynamic features into the static feature sequence, $\boldsymbol{m} = \{m_1, \ldots, m_T\}$ indicates a sequence of mixture indices, $\boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)}$ is the conditional mean vector at frame t, which is given by the mixture-dependent linear transformation of the source feature vector $\boldsymbol{X}_t$, and $\boldsymbol{D}_{\hat{m}_t}^{(Y|X)}$ is the conditional covariance matrix depending of the mixture component $\hat{m}_t$. The ML estimate of F_0 sequence $\hat{\boldsymbol{y}}$ is analytically determined as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y|X)}. \tag{5}$$

The real-time prediction process is achieved by using a computationally efficient real-time voice conversion method [15] based on a low-delay conversion algorithm [16]. To approximate the batch-type prediction process with the frame-wise prediction process, we divide the F_0 sequence $\boldsymbol{y}$ into overlapped $(L+1)$-dimensional segment vectors $\boldsymbol{y}^{(t)} = [y_{t-L}, \ldots, y_t]^\top$ at individual frames. With treating the segment vectors as a latent variable, the following linear dynamical system can be designed [17]:

$$\boldsymbol{y}^{(t)} = \boldsymbol{J}\boldsymbol{y}^{(t-1)} + [\boldsymbol{0}_{1\times L}, \mu_{\hat{m}_t,t}^{(y|X)} + n_{\hat{m}_t}^{(y|X)}]^\top, \tag{6}$$

$$\mu_{\hat{m}_t,t}^{(\Delta y|X)} = \boldsymbol{w}\boldsymbol{y}^{(t)} + n_{\hat{m}_t}^{(\Delta y|X)}, \tag{7}$$

where the state transition matrix $\boldsymbol{J}$ just shifts the previous segment vector $\boldsymbol{y}^{(t-1)}$, and the linear transform $\boldsymbol{w}$ to calculate the dynamic features at frame t from the segment vector. The observation $\mu_{\hat{m}_t,t}^{(\Delta y|X)}$, a parameter $\mu_{\hat{m}_t,t}^{(y|X)}$, process noise $n_{\hat{m}_t}^{(y|X)}$, and observation noise $n_{\hat{m}_t}^{(\Delta y|X)}$ are described with the conditional mean vector $\boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)}$ and covariance matrix $\boldsymbol{D}_{\hat{m}_t}^{(Y|X)}$ at frame t. The segment vector is recursively updated frame by frame with Kalman filtering, and its first component y_{t-L} is used as the ML estimate $\hat{y}_{t-L}$. Therefore, the F_0 value at frame t is determined by considering all past frames, a current frame, and next L frames.

C. Effect of Delay Time on Real-time F_0 Prediction Accuracy

In the previous work [16], it has been reported in a spectral conversion task that the delay time depending on the segment feature length L in the real-time prediction process requires around 50 to 70 msec to maintain the conversion accuracy of the batch-type prediction process. On the other hand, no previous work has examined the effect of the delay time on the F_0 prediction accuracy. It is possible that longer delay will be required because F_0 is a suprasegmental feature, which has strong correlation over a wider range compared to segmental features, such as spectral parameters. This effect is expected to be stronger in CF_0 patterns.

III. Development of Prototype System

A. Implementation of Real-Time F_0 Control

A prototype system based on our proposed system was developed using a laptop and a digital/analog (D/A) converter shown in Table I. As shown in Fig. 1, EL speech produced from mouth of a laryngectomee is detected with a usual 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

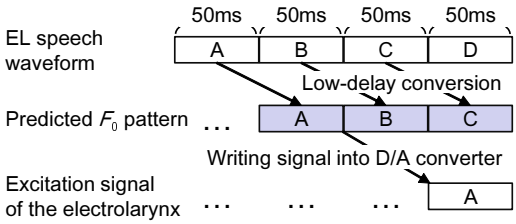

Fig. 2. Latency caused by each process

electrolarynx generates the excitation signal with the predicted F_0 values according to the input electric signal generated from the D/A converter.

As described in the previous section, the F_0 patterns are constantly delayed owing to latency of the real-time statistical F_0 prediction process. Moreover, additional latency is caused in the prototype system by the D/A converter. Figure 2 shows latency caused by each step. In our implementation, 50 msec latency is caused by the real-time statistical F_0 prediction. For the D/A part, it takes around 50 msec to write the digital signal on the D/A converter, where the digital signal to be written needs to be determined before starting writing. Consequently, the D/A part causes 100 msec latency. In total, 150 msec latency is caused in the prototype system[1].

B. Negative Impacts Caused by Latency

The F_0 patterns predicted by the prototype system strongly correlate to those by the simulated system [10], with a correlation coefficient higher than 0.9. This high correlation demonstrates that the proposed implementation is effective and the simulated system is able to effectively approximate the results of the prototype system. Through the use of the prototype system, we confirmed that it yields significant improvements in naturalness of EL speech while preserving its high intelligibility as expected in the previous evaluation [10]. However, we also found that the naturalness of enhanced EL speech tends to be lower than that yielded by the batch-type prediction. We assume that this degradation is caused by 1) the shorter delay time compared to the required delay time on the real-time prediction of CF_0 patterns, and 2) additional processing delay of the prototype system. It is necessary to compare the F_0 patterns predicted by the real-time prediction with those predicted by the batch-type prediction.

TABLE I
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

[1]Note that latency in the D/A part could be addressed by the development of a special device for the electrolarynx. Moreover, we have successfully implemented statistical VC processing on a digital signal processor (DSP) [18]. It is thus expected that all processors could be embedded into the electrolarynx and total latency will be decreased to the 50 msec caused by the real-time statistical F_0 prediction.

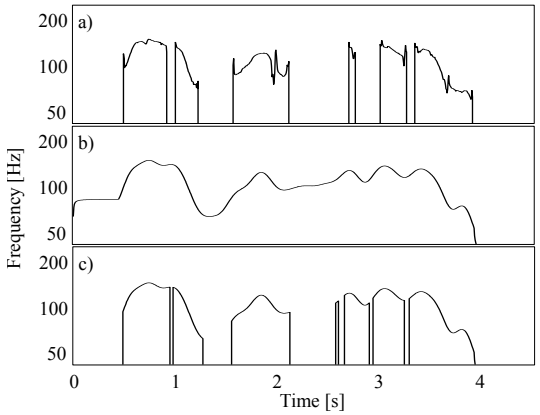

Fig. 3. a) F_0 patterns extracted from normal speech, b) smoothed continuous F_0 patterns interpolated at unvoiced frames, and c) segmented CF_0 patterns of (b) extracted by using the power of waveform.

IV. Proposed Methods for Addressing Latency Issues

On our prototype system, the produced EL speech suffers from a misalignment between articulation and the constantly delayed F_0 patterns. To address this issue, we propose two methods for reducing the delay time caused by the real-time process while preserving F_0 prediction accuracy at the level of the batch-type prediction process.

A. Segmented Continuous F_0 Patterns

In the previous CF_0 modeling method, the prediction process given in Eq. (1) is performed utterance by utterance. Because inter-frame correlation over an utterance is considered in this process, a long delay is required in real-time prediction to achieve sufficient prediction accuracy.

To reduce the delay time, we propose a segmented CF_0 pattern modeling method to make the range of which we consider inter-frame correlation shorter than an utterance. Shorter segments are first extracted from each utterance, and then, CF_0 patterns of individual segments (i.e., segmented CF_0 patterns) are modeled and predicted separately. In this paper, we determine the individual segments by extracting time frames of which the waveform power is over a pre-determined threshold. An example of the segmented CF_0 patterns is shown in Fig. 3. Note that the segmented CF_0 patterns are still different from the original F_0 pattern, which is segmented by unvoiced frames, in that 1) the segmented CF_0 patterns can also include unvoiced frames, and thus they tend to be longer than segments observed in the original F_0 patterns, and 2) each segmented CF_0 pattern varied more smoothly than the original F_0 patterns.

B. Forthcoming F_0 Prediction

In order to cancel the misalignment between articulation and the constantly delayed F_0 patterns predicted in the real-time process, we investigate the possibility of predicting forthcoming F_0 values. We train the GMM for modeling the joint probability density $P([\boldsymbol{X}_t^\top, \boldsymbol{Y}_{t+F}^\top]^\top|\boldsymbol{\lambda})$ of the source features at frame t, $\boldsymbol{X}_t$ and the target features at frame $t+F$, $\boldsymbol{Y}_{t+F}$. The trained GMM is used to predict the F_0 value at F frames ahead. For example, if the latency of the prototype system is set to 200 msec, we train the GMM to predict the F_0 values at 200 msec ahead. Consequently, there is no mismatch between articulation and the predicted F_0 patterns.

It is expected that there is a tradeoff between the prediction accuracy and the setting of F; i.e., larger F accepts a longer delay time in the real-time prediction process, which makes the real-time prediction accuracy close to the batch-type prediction accuracy; on the other hand, it is obviously more difficult to predict F_0 values at frames far away from the current one than those at closer frames.

V. Experimental Evaluation

A. Experimental Conditions

We conducted four types of objective evaluation to examine the performance of the developed prototype system and the effectiveness of the proposed methods. The first evaluation is a comparison of the prediction accuracy among three types of F_0 pattern modeling, F_0, CF_0, and the proposed segmental CF_0 (Seg CF_0), in the batch-type prediction process. The second evaluation is a comparison of the accuracy of batch-type F_0 prediction and real-time F_0 prediction. The third evaluation is conducted to examine the effectiveness of the proposed segmental CF_0 modeling method in the real-time prediction process. The last evaluation is conducted to examine the effectiveness of the proposed forthcoming F_0 prediction method. The source speech was EL speech uttered by a 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 [19]. 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. Sampling frequency was set to 16 kHz.

We employed FFT analysis with a 25 msec hanning window to extract the mel-cepstra of EL speech as the spectral parameters. The frame shift length was set to 5 msec. As the source features, the spectral segment features were extracted from the mel-cepstra at the current $\pm$ 4 frames. On the other hand, F_0 values of normal speech were extracted with STRAIGHT F_0 analysis [20] and CF_0 patterns were generated as the target feature using a low-pass filter with 10 Hz cut-off frequency. Moreover, the target F_0 patterns were shifted so that their mean value was equal to 100 Hz to predict F_0 patterns suitable for the source male speaker.

B. Comparison of F_0 Modeling Methods in Batch-type F_0 Prediction

We evaluated the prediction accuracy of each F_0 pattern modeling method in the batch-type process using the correlation coefficient between the predicted F_0 pattern and the target F_0 pattern. The number of mixture components of the GMM was optimized separately in each modeling method.

Table II shows the results as well as the number of mixture components optimized for each modeling method. CF_0 yields a significant improvement in the prediction accuracy compared to the original F_0 as reported in [7]. The proposed segmented CF_0 modeling method (Seg CF_0) preserves such an improvement relatively well while minimizing degradation of the prediction accuracy.

C. Comparison of the Accuracy of Batch-type F_0 Prediction and Real-time F_0 Prediction

As mentioned in Section IV-A, it is possible in the real-time prediction that the larger delay time is required in the CF_0 pattern than in the F_0 pattern to achieve the prediction accuracy comparable to that of the batch-type prediction. To examine this possibility, we calculated a correlation coefficient between the F_0 pattern predicted by the real-time prediction with various settings of the delay time and that by the batch-type prediction.

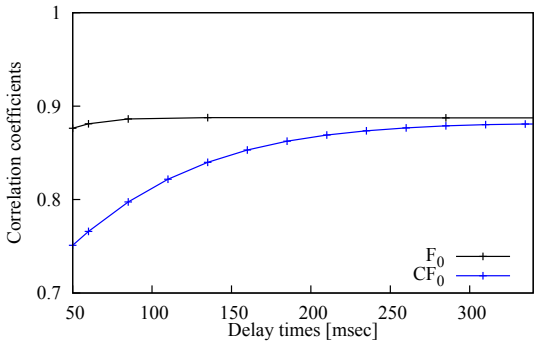

Fig. 4. Comparison of the accuracy of batch-type F_0 prediction and real-time F_0 prediction.

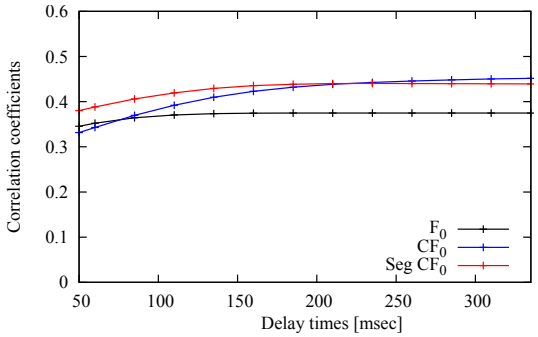

Fig. 5. Relationship of the delay time and each F_0 pattern on real-time F_0 prediction accuracy (w/ delay time correction at the time of evaluation).

The result of a comparison between the CF_0 and F_0 patterns is shown in Fig. 4. As for the F_0 pattern, even if setting the delay time to 60 msec (corresponding to $L = 5$), a quite high correlation coefficient (around 0.89) is achieved. On the other hand, as for the CF_0 pattern, the predicted patterns are quite different from those by the batch-type process, showing that the correlation coefficient is less than 0.8 when setting the delay time to less than 85 msec. Moreover, its accuracy convergence is much slower compared to that observed in the F_0 pattern. Consequently, in the CF_0 pattern, the delay time needs to be set to around 250 msec to achieve the prediction accuracy comparable to that of the batch-type prediction.

D. Evaluation of the Proposed Segmental CF_0 Modeling

We evaluated the real-time prediction accuracy of each F_0 modeling method using the correlation coefficient between the predicted F_0 pattern and the target F_0 pattern. To evaluate only

TABLE II
Accuracy of Batch-type Prediction

	F_0	CF_0	Seg CF_0
The number of mixture components	32	16	16
F_0 correlation coefficients	0.40	0.48	0.46

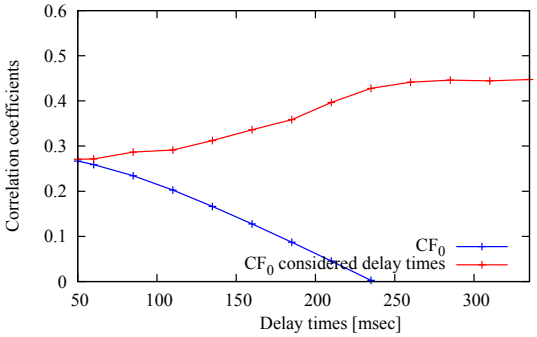

Fig. 6. Comparison of basic CF_0 patterns modeling and forthcoming CF_0 patterns modeling (w/o delay time correction at the time of evaluation).

the prediction accuracy, the effect of the misalignment between the predicted and the target F_0 patterns, which is observed on the prototype system, was removed in this evaluation by shifting the predicted F_0 patterns according to the delay time settings in calculation of the correlation coefficient.

The results are shown in Fig. 5. As for the F_0 pattern, although the prediction accuracy quickly converges at around 60 msec of the delay time, the resulting correlation coefficient is lower than 0.4 because the prediction accuracy of the batch-type prediction is also low, as shown in Table II. As for the CF_0 pattern, the converged prediction accuracy is significantly higher than that in the F_0 pattern, as also observed in Table II, and its convergence is very slow. To achieve sufficient prediction accuracy, the delay time needs to be set to around 250 msec, as also observed in Fig. 4. On the other hand, the use of the proposed segmented CF_0 patterns makes the convergence faster than that of the CF_0 patterns while preserving its prediction accuracy. Consequently, the prediction accuracy comparable to the batch-type prediction can be achieved by setting the delay time to around 150 msec.

E. Evaluation of the Proposed Forthcoming F_0 Prediction

We evaluated the real-time prediction accuracy also considering the effect of the misalignment between articulation and the delayed F_0 patterns predicted in the real-time process, which was observed in a practical situation, using the correlation coefficient between the predicted F_0 pattern without any correction of the delay time and the target F_0 pattern.

The proposed forthcoming F_0 prediction method was applied to the CF_0 pattern and its effectiveness was examined. The result is shown in Fig. 6. If not using the proposed forthcoming F_0 prediction, the delay time is set to longer, the prediction accuracy gets lower. This result shows that the adverse effect of the misalignment on the actual prediction accuracy is significantly large. This issue is well addressed by using the proposed forthcoming F_0 prediction. Consequently, by setting the delay time to around 250 msec, the real-time prediction with the proposed forthcoming F_0 prediction method makes it possible to achieve prediction accuracy comparable to that of the batch-type prediction.

VI. Conclusion

In this paper, we have examined negative impacts caused by latency of the real-time prediction on F_0 prediction accuracy, and to alleviate them, we have also proposed two methods, 1) modeling of segmented continuous F_0 (CF_0) patterns and 2) prediction of forthcoming F_0 values. The experimental results have demonstrated that 1) the conventional real-time prediction needs a large delay to predict CF_0 patterns and 2) the proposed methods have positive impacts on the real-time prediction.

Acknowledgment

This work was supported in part of JSPS KAKENHI Grant Numbers: 15J10727 and 26280060, and the authors would like to thank Mr. Sugai of Densei Communication Inc., Japan, for advise to control an electrolarynx.

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*, pp. 198–203, Jul. 1994.

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

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

[4] 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.

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

[6] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "A laryngeal 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, Jan. 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*, Jun. 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, Mar. 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, Nov. 2007.

[10] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "An inter-speaker evaluation through simulation of electrolarynx control based on statistical F_0 prediction," *Proc. APSIPA ASC,* Dec. 2014.

[11] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "An enhanced electrolarynx with automatic fundamental frequency control based on statistical prediction," *Proc. ASSETS,* pp. 435–436, Sep. 2015.

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

[13] 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, Nov. 2012.

[14] K. J. Kohler, "Macro and micro F_0 in the synthesis of intonation," *Papers in Laboratory Phonology I,* pp. 115–138, 1990.

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

[16] 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*, pp. 1076–1079, Sep. 2008.

[17] T. Toda, "Augmented speech production based on real-time statistical voice conversion," *Proc. GlobalSIP,* pp. 755–759, Dec. 2014.

[18] T. Moriguchi, T. Toda, M. Sano, H. Sato, G. Neubig, S. Sakti, and S. Nakamura, "A Digital Signal Processor Implementation of Silent/Electrolaryngeal Speech Enhancement based on Real-Time Statistical Voice Conversion," in *Proc. INTERSPEECH*, pp. 3072–3076, Aug. 2013.

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

[20] 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 *Speech Communication*, vol. 27, no. 3. Elsevier, pp. 187–207, Apr. 1999.