PAPER

A Vibration Control Method of an Electrolarynx Based on Statistical F_0 Pattern Prediction

Kou TANAKA[†a)], ***Nonmember***, **Tomoki TODA**[††b)], ***and*** **Satoshi NAKAMURA**[†c)], ***Members***

SUMMARY This paper presents a novel speaking aid system to help laryngectomees produce more naturally sounding electrolaryngeal (EL) speech. An electrolarynx is an external device to generate excitation signals, instead of vibration of the vocal folds. Although the conventional EL speech is quite intelligible, its naturalness suffers from the unnatural fundamental frequency (F_0) patterns of the mechanically generated excitation signals. To improve the naturalness of EL speech, we have proposed EL speech enhancement methods using statistical F_0 pattern prediction. In these methods, the original EL speech recorded by a microphone is presented from a loudspeaker after performing the speech enhancement. These methods are effective for some situation, such as telecommunication, but it is not suitable for face-to-face conversation because not only the enhanced EL speech but also the original EL speech is presented to listeners. In this paper, to develop an EL speech enhancement also effective for face-to-face conversation, we propose a method for directly controlling F_0 patterns of the excitation signals to be generated from the electrolarynx using the statistical F_0 prediction. To get an "actual feel" of the proposed system, we also implement a prototype system. By using the prototype system, we find latency issues caused by a real-time processing. To address these latency issues, we furthermore propose segmental continuous F_0 pattern modeling and forthcoming F_0 pattern modeling. With evaluations through simulation, we demonstrate that our proposed system is capable of effectively addressing the issues of latency and those of electrolarynx in term of the naturalness.

key words: *laryngectomee, electrolarynx, voice restoration, speech enhancement, statistical F_0 pattern prediction*

1. Introduction

Speech is a common tool in human communication. Since speech is produced by the vocal apparatus, the produced sounds are physically constrained by the conditions of the human body. Unfortunately, there are many people with disabilities that prevent them from producing speech freely, leading to communication barriers and degrading their Quality of Life (QoL). A typical example is laryngectomees who have undergone an operation to remove them larynges including the vocal folds for several reasons such as injury and laryngeal cancer. Their ability to generate sound source excitation signals is severely impaired because they no longer have their vocal folds, although their vocal tracts remain.

An electrolarynx is an medical device for laryngectomees to produce electrolaryngeal (EL) speech by mechanically generating artificial excitation signals. The generated excitation signals are conducted into the speaker's oral cavity through the neck, and are articulated to produce EL speech as shown in Fig. 1. 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 techniques 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 switches 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 make it possible to control the F_0 patterns without conscious operation, some methods using other physical signals generated by articulation, such as neck surface electromyography (EMG) and intramuscular cricothyroid (CT) EMG, have been proposed [4]–[6]. Although the CT EMG has a strong correlation (higher than 0.9) with F_0 patterns, the CT muscles are accessible only through invasive needle electrodes. On the other hands, the surface EMG is easily measured, but the F_0 patterns predicted by using the surface EMG are still unnatural compared with those of normal speech, and what is worse, its quality strongly depends on the position of measuring instrument.

To improve naturalness of EL speech, we have proposed several EL speech enhanced methods based on statistical voice conversion techniques [7]–[9]. In these methods, acoustic features of EL speech are converted into

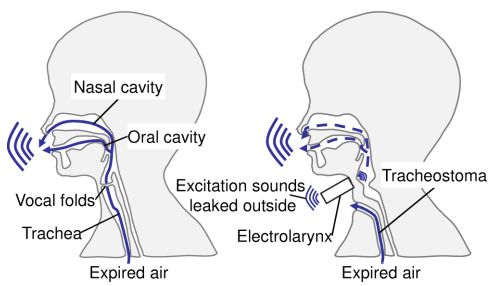

Fig. 1 Speech production mechanisms of non-disabled people (left figure) and total laryngectomees (right figure).

Manuscript received December 8, 2016.
Manuscript revised April 27, 2017.
Manuscript publicized May 23, 2017.

[†]The authors are with the Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630–0192 Japan.
[††]The author is with the Nagoya University, Nagoya-shi, 464–0814 Japan.
a) E-mail: ko-t@is.naist.jp
b) E-mail: tomoki@icts.nagoya-u.ac.jp
c) E-mail: s-nakamura@is.naist.jp

DOI: 10.1587/transinf.2016EDP7485

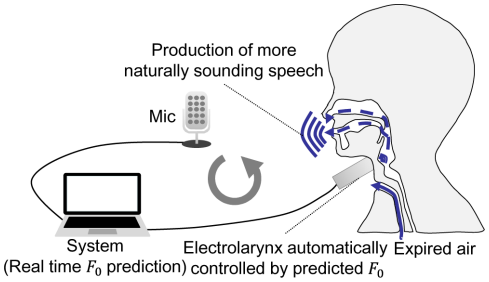

Fig. 2 A proposed system to control F_0 patterns of excitation signals of an electrolarynx using statistical F_0 pattern prediction for laryngectomees.

those of normal speech using Gaussian mixture models (GMMs) [7]–[9]. We have shown that F_0 pattern replacement from the mechanically generated ones into those predicted from the spectral sequence of the EL speech using the GMM significantly improves naturalness of EL speech while preserving its intelligibility [9]. On the other hand, the use of these enhancement methods needs to use a loudspeaker to present the enhanced EL speech. This requirement strongly restricts situations where these enhancement methods are available, e.g. telecommunication presenting only the enhanced speech to the listener. By contrast, in face-to-face conversation where the listener is close to the speaker, this requirement is essential drawback because not only enhanced EL speech but also original EL speech are presented to the listener at the same time.

In this paper, we propose an EL speech enhancement system (Fig. 2) effective for any situation, including face-to-face conversation. F_0 patterns of the excitation signals produced by the electrolarynx are directly controlled using real-time statistical F_0 pattern prediction. Namely, an F_0 value at a current frame is predicted in real-time from the EL speech produced by articulating the excitation signals with previously predicted F_0 values. This proposed system has the potential to allow laryngectomees to directly produce enhanced EL speech with more natural F_0 patterns than the original EL speech, and present only the enhanced EL speech to the listener. To get an "actual feel" of the proposed system, we also implement a prototype system. By using the prototype system, we find latency issues caused by a real-time processing. To address the latency issues, we furthermore propose segmental continuous F_0 pattern modeling and forthcoming F_0 pattern modeling. With evaluations through simulation, we demonstrate that our proposed system is capable of effectively addressing the issues of latency and those of electrolarynx in term of the naturalness.

2. Statistical F_0 Pattern Prediction

Our proposed enhancement system uses a statistical F_0 pattern prediction, which is a part of voice conversion techniques [10], [11], to predict F_0 patterns of normal speech from spectral features of EL speech. It consists of training and prediction processes as shown in Fig. 3. A joint probability density function [12] of F_0 patterns of normal speech and spectral features of EL speech is trained using a parallel data set consisting of utterance pairs of EL speech and normal speech. With properly trained parameters, the most likely F_0 patterns of normal speech given spectral features of EL speech can be found by maximum likelihood estimation of trajectory.

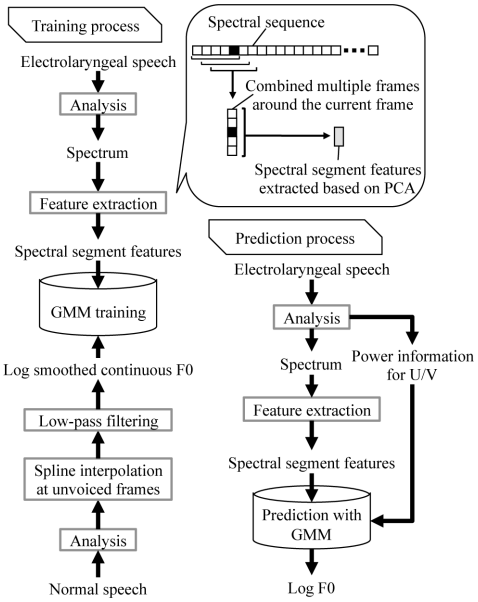

Fig. 3 The training and prediction process.

2.1 Feature Extraction

The spectral structure of some phonemes of EL speech is unstably because of the production mechanism of EL speech, such as totally voiced speech. To address this issue, we use the following segment feature $\boldsymbol{X}_t$ [13] extracted by applying principal component analysis (PCA) to the stacked vector consisting of the mel-cepstra of multiple frames around the current frame t as source feature:

$$\boldsymbol{X}_t = \boldsymbol{C}[\boldsymbol{x}_{t-i}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+i}^\top]^\top + \boldsymbol{d} \tag{1}$$

where $^\top$ is transposition, and $\boldsymbol{C}$ and $\boldsymbol{d}$ are a transformation matrix and a bias vector extracted by PCA, respectively.

As a target feature, we use $\boldsymbol{Y}_t = [y_t^\top, \Delta y_t^\top]^\top$ consisting of the static and dynamic features of smooth and continuous F_0 (CF_0) patterns of normal speech. To simplify characteristics of the parameter sequence to be modeled, CF_0 are obtained by removing rapid movements [14] with low-pass filtering after interpolating F_0 values at unvoiced frames. This modification is reasonable because 1) it is difficult to accurately model and reproduce these rapid movements with a GMM and 2) a constant value at the unvoiced frames, clearly different from F_0 values (e.g., 0), disturbs accurate modeling of F_0 trajectory.

2.2 Training Process

Let $\boldsymbol{\lambda}_G$ be the parameters of the following joint probability density function of source and target features defined as a

GMM:

$$p([\boldsymbol{X}_t^\top, \boldsymbol{Y}_y^\top]^\top|\boldsymbol{\lambda}_G) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right), \quad (2)$$

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \quad (3)$$

where α_m is a m-th mixture component weight, and $\mathcal{N}(\cdot; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ denotes a m-th Gaussian distribution with a mean vector $\boldsymbol{\mu}_m$ and a covariance matrix $\boldsymbol{\Sigma}_m$. The mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ consists of a mean vector $\boldsymbol{\mu}_m^{(X)}$ of source features and a mean vector $\boldsymbol{\mu}_m^{(Y)}$ of target features. The covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ consists of source and target covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ and cross-covariance matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$. The total number of mixture components is M. The corresponding joint feature vectors can be obtained by performing automatic frame alignment with Dynamic Time Warping (DTW). To align F_0 patterns of normal speech and spectral parameters of EL speech, we use alignments which are obtained by using spectral parameters of EL speech and normal speech in the same manner as in [8], [13].

2.3 Batch-Type Prediction Process

With properly trained parameters, the most likely F_0 pattern $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is predicted from given source feature sequence $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_T^\top]^\top$ as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}_G) \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (4)$$

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}_G) = \sum_{\boldsymbol{m}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}_G) P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}_G) \quad (5)$$

$$\simeq P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G) P(\hat{\boldsymbol{m}}|\boldsymbol{X}, \boldsymbol{\lambda}_G), \quad (6)$$

where $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_T^\top]^\top$ denotes the joint static and dynamic feature sequence, $\boldsymbol{W}$ is a transform matrix to extend the static feature sequence into the static and dynamic feature sequence [15]. To avoid the complicated formula $\sum_{\boldsymbol{m}}$ in Eq. (5), we adopt the suboptimum mixture component sequence $\hat{\boldsymbol{m}} = \{\hat{m}_1, \ldots, \hat{m}_T\}$,

$$\hat{m}_t = \underset{m}{\operatorname{argmax}}\, P(m|\boldsymbol{X}_t, \boldsymbol{\lambda}_G), \quad (7)$$

$$P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}_G) = \mathcal{N}(\boldsymbol{Y}; \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y|X)}, \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)}, \boldsymbol{D}_{\hat{m}_t}^{(Y|X)}), \quad (8)$$

$$\boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)} = \boldsymbol{\mu}_{\hat{m}_t}^{(Y)} + \boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)} \boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_{\hat{m}_t}^{(X)}\right), \quad (9)$$

$$\boldsymbol{D}_{\hat{m}_t}^{(Y|X)} = \boldsymbol{\Sigma}_{\hat{m}_t}^{(YY)} - \boldsymbol{\Sigma}_{\hat{m}_t}^{(YX)} \boldsymbol{\Sigma}_{\hat{m}_t}^{(XX)^{-1}} \boldsymbol{\Sigma}_{\hat{m}_t}^{(XY)}, \quad (10)$$

where $\boldsymbol{E}_{\hat{m}_t,t}^{(Y|X)}$ 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 $\hat{m}_t$.

Finally, the maximum-likelihood estimation of F_0 patterns $\hat{\boldsymbol{y}}$ is analytically determined as follows:

$$\hat{\boldsymbol{y}} = (\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{W})^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y|X)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y|X)}. \quad (11)$$

Note that after predicting CF_0 patterns over all frames, only silence frames are automatically detected by using waveform power [9].

2.4 Real-Time Prediction Process

The real-time prediction process is achieved by using a computationally efficient real-time voice conversion method [16] based on a low-delay conversion algorithm [17]. 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. Treating the segment vectors as a latent variable, the following linear dynamical system can be designed:

$$\boldsymbol{y}^{(t)} = \boldsymbol{J}\boldsymbol{y}^{(t-1)} + [\boldsymbol{0}_{1\times L}, \mu_{\hat{m}_t,t}^{(y|X)} + n_{\hat{m}_t}^{(y|X)}]^\top, \quad (12)$$

$$\mu_{\hat{m}_t,t}^{(\Delta y|X)} = \boldsymbol{w}\boldsymbol{y}^{(t)} + n_{\hat{m}_t}^{(\Delta y|X)}, \quad (13)$$

where the state transition matrix $\boldsymbol{J}$ just shifts the previous segment vector $\boldsymbol{y}^{(t-1)}$, and the transformation matrix $\boldsymbol{w}$ 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 only diagonal components of the conditional 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 maximum-likelihood 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.

3. Control Strategy of an Electrolarynx Based on Statistical F_0 Pattern Prediction

Our proposed enhancement system (shown in Fig. 2) 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 prediction algorithm mentioned in Sect. 2.4. In the articulation process, to produce the EL speech, the laryngectomee articulates the excitation signals of the electrolarynx reflecting predicted F_0 values. Therefore, this system allows laryngectomees to directly produce enhanced EL speech with more naturally sounding F_0 patterns corresponding to linguistic contents because the source spectral features of EL speech capture the linguistic contents.

3.1 Implementation of Prototype System

A prototype one of our proposed enhancement system was

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

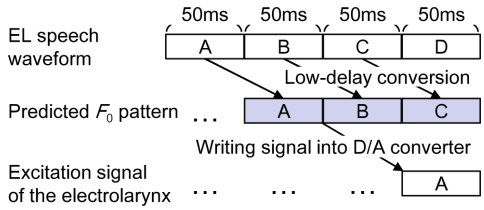

Fig. 4 The latency caused by each process of our prototype system.

developed using a microphone, a laptop, and a digital/analog (D/A) converter shown in Table 1. As shown in Fig. 2, EL speech produced from a mouth of a laryngectomee is detected by 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 prediction algorithm. The predicted F_0 values are linearly converted to voltage values to control the F_0 values of the excitation signals. Then, through the D/A converter connected from the laptop to the electrolarynx, an electric signal corresponding to the determined voltage values is generated. Finally, the electrolarynx generates the excitation signals reflecting the predicted F_0 values according to the input electric signal generated from the D/A converter.

As mentioned in the previous section, the F_0 patterns are constantly delayed owing to the latency of the real-time prediction process. Moreover, additional latency is caused in our prototype system because of the use of D/A converter. Figure 4 shows the latency caused by each process of our prototype system. For the real-time prediction process, 50 msec latency is caused in our conventional implementation [16]. For the D/A part to convey the digital signals, it takes around 50 msec. Consequently, the whole D/A part causes 100 msec latency because the digital signal to be written needs to be determined before starting writing. In total, 150 msec latency is caused in the prototype system†.

3.2 A Simulation Experiment

To flexibly investigate the performance of our proposed control method, we also design a simulation method of EL speech production process using the controlled electrolarynx. The simulated process is shown in the right side of Fig. 5. EL speech signals produced by articulating the excitation signals based on the predicted F_0 values are artificially generated using the STRAIGHT [19] analysis/synthesis method.

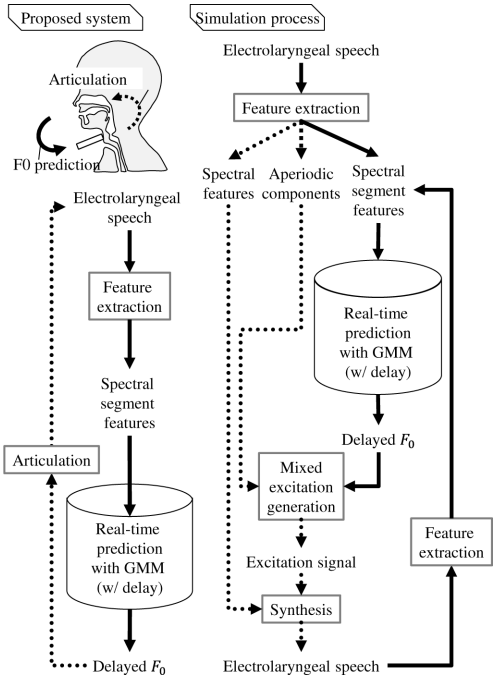

Fig. 5 The proposed system and its simulation implementation.

At first, 1) we extract spectral envelope parameters and aperiodic components [20] from the original EL speech in advance by using STRAIGHT analysis. These features to approximate the EL speech production process capture acoustic properties determined by articulation and the excitation signals leaking out as noise from the electrolarynx, except for the periodicity of the excitation signals. Then, 2) spectral segment features are extracted from EL speech, and F_0 patterns of normal speech are predicted from them based on the real-time F_0 pattern prediction. 3) The predicted F_0 patterns are just delayed to consider the delay time caused by the whole process of our prototype system, such as mentioned D/A part. 4) Using the delayed F_0 patterns and the extracted aperiodic components, excitation signals are generated based on the mixed excitation model [21] to replace actual excitation signals of the electrolarynx. 5) Finally, the enhanced EL speech is approximately synthesized by filtering the generated excitation signals with the extracted spectral envelope parameters reflecting the articulation. Note that in our prototype system, the F_0 values of the enhanced EL speech suffer from those of previous time step because the F_0 values are predicted from EL speech reflecting the F_0 values of previous time step. However, the above-mentioned processing without an iterative update, Step 3) to 5), results in the F_0 prediction using the spectral segment features extracted from the original EL speech. To reflect the impact of the predicted F_0 values of previous time step, 6) the spectral segment features are extracted again from the synthesized EL speech and F_0 pattern prediction is also performed

†Note that the 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 voice conversion 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.

again using the extracted spectral segment features. Step 3) to 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.

4. Addressing Latency Issues

Through the use of the prototype system, we confirmed that it yields significant improvements in the naturalness of EL speech while preserving its high intelligibility. However, we also found that the naturalness of enhanced EL speech tends to be lower than that yielded by the batch-type prediction.

As mentioned in Sect. 2.4, the latency to predict F_0 patterns is inherent in our proposed enhancement system. It has been reported in a spectral conversion task [17] 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 latency for the F_0 prediction accuracy. It is possible that longer delay will be required because F_0 is a suprasegmental feature, which has a strong correlation over a wider range compared to segmental features, such as spectral features. Moreover, in our prototype system mentioned in Sect. 3.1, the additional latency is caused by using D/A converter to convey predicted F_0 values to the electrolarynx. This latency on our proposed system leads to asynchronous problem between articulation and F_0 patterns of excitation signals generated by the electrolarynx. To address these issues, we also propose the use of segmented continuous F_0 patterns as trained target features and forthcoming F_0 prediction for reducing the latency caused by the real-time prediction process while preserving F_0 prediction accuracy at the level of the batch-type prediction process.

4.1 Segmented Continuous F_0 Patterns

In the previous CF_0 modeling method, the prediction process given in Eq. (4) 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 predetermined threshold. An example of the segmented CF_0 patterns is shown in Fig. 6. 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.

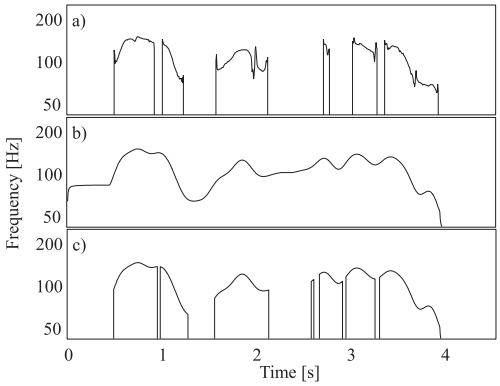

Fig. 6 a) F_0 patterns extracted from normal speech, b) smooth and continuous F_0 patterns interpolated at unvoiced frames, and c) segmented CF_0 patterns of (b) extracted by using the power of the waveform.

4.2 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 function $P([\boldsymbol{X}_t^\top, \boldsymbol{Y}_{t+F}^\top]^\top|\boldsymbol{\lambda}_G)$ of the source features at time frame t, $\boldsymbol{X}_t$ and the target features at time 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 trade-off 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.

5. Experimental Evaluation

5.1 Experimental Conditions

We conducted 5 objective evaluations to examine the performance of the proposed methods and 1 subjective evaluation to examine the naturalness 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 segmented 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 for the validity of the proposed simulation experiment to simulate our proposed enhancement system. The fourth evaluation is conducted to investigate the nega-

tive impacts caused by latency on the proposed system and to examine the effectiveness of the proposed segmented CF_0 pattern modeling. The last objective 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 [22]. 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 features. 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 ± 4 frames. On the other hand, F_0 values of normal speech were extracted with STRAIGHT F_0 analysis [19] 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. To obtain the time alignment path between source and target speakers, the numbers of mixture components of the GMM trained for spectral parameters conversion were set to 64, and the mel-cepstral distortion without power information was 5.09 dB.

5.2 Best Number of Mixture Components

To choose the best setting from a variety number of mixture components for later evaluations, 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. As shown in Fig. 7, the best number of mixture components is 32 for F_0, 16 for CF_0, and 16 for segmented CF_0. We found that reducing the variability of F_0 patterns such as rapid movements, we achieved to train F_0 patterns with a smaller number of mixture components. Moreover, as reported in [9], we also confirmed that CF_0 brings better performance compared with the original F_0 because continuous sequence makes it possible to consider inter-frame correlation over an utterance. The proposed segmented CF_0 preserves such an improvement relatively well while minimizing degradation of the prediction accuracy.

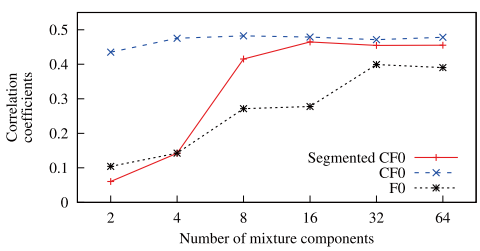

Fig. 7 Prediction accuracy of batch-type prediction.

5.3 Comparison of Batch-Type Prediction and Real-Time Prediction

As mentioned in Sect. 4.1, 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.

The result is shown in Fig. 8. As for the F_0 pattern, even if setting the delay time to 85 msec (corresponding to $L = 10$), a quite high correlation coefficient 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 similar to the case of F_0 pattern 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. The CF_0 pattern modeling achieves high prediction accuracy while requiring the large size of memory corresponding to the required delay time in the prediction process. As we expected, the segmented CF_0 modeling converges faster compared with the CF_0 pattern modeling because the number of frames considering inter-frame correlation is limited. The segmented CF_0 modeling reduces the required size of memory and the computational costs to predict F_0 patterns while degrading the prediction accuracy compared with CF_0 pattern modeling.

5.4 Comparison of Prototype System and Simulated System

The F_0 patterns predicted by the prototype system strongly correlate to those with the simulated system, 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. This result allows us to replace the evaluations of our prototype system into those of

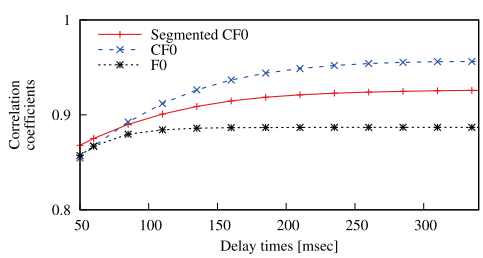

Fig. 8 Comparison of batch-type prediction and real-time prediction for each F_0 pattern.

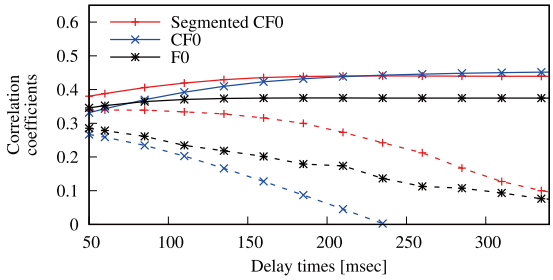

Fig. 9 Prediction accuracy of real-time prediction for each F_0 pattern. Solid lines are results w/ delay time correction at the time of evaluation, and dash lines are results w/o delay time correction.

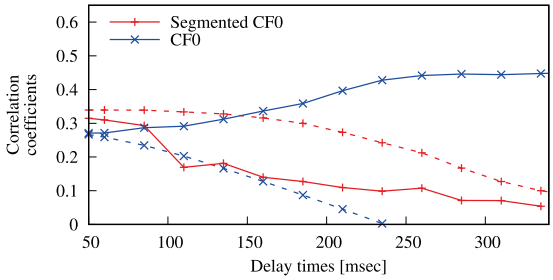

Fig. 10 Comparison of basic modeling (dash lines) and forthcoming modeling (solid lines).

the simulated system.

5.5 Negative Impacts Caused by Latency

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 the prediction accuracy, we also evaluate predicted F_0 patterns with delay time correction at time of evaluation. As shown in solid lines in Fig. 9, 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 the calculation of the correlation coefficient.

The result is shown in Fig. 9. As for the solid lines, we confirmed a similar tendency to the results in Sect. 5.3. As for the F_0 pattern, we found that 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 Fig. 7. As for the CF_0 pattern, the converged prediction accuracy is significantly higher than that in the F_0 pattern, as also observed in Fig. 7, and its convergence is very slow. To achieve sufficient prediction accuracy, the delay time needs to be set to around 250 msec. 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. As for the dash lines, the delay time is set to longer, the prediction accuracy gets lower. However, the segmented CF_0 pattern makes it possible to alleviate the negative impact of latency compared with the other baseline F_0 modeling.

5.6 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 proposed segmented CF_0 pattern, and its effectiveness was examined. The result is shown in Fig. 10. 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 for CF_0 pattern modeling. 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. However, as for the segmented CF_0 patterns, even if we apply the proposed forthcoming F_0 prediction, its prediction accuracy is not improved. This result shows that restricting the unit considering inter-frame correlation makes it difficult to predict F_0 values at frames far away from the current one than those at closer frames.

5.7 Naturalness of Predicted F_0 patterns

Through the simulation experiment, we evaluated the naturalness of F_0 patterns predicted by using our proposed F_0 modeling. The term "naturalness" is used to indicate a score that was measured by asking the listener to subjectively evaluate whether the evaluated speech is similar to natural human speech or not. In the opinion tests, 5 listeners evaluated each speech quality using a 5-scaled opinion score (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Excellent). The number of listeners was 5 and each listener evaluates 10 sentences per one system. Hence, each system is evaluated with 50 sentences. Comparison methods are following 4 systems:

EL Original EL speech

Batch Enhanced EL speech with CF_0 patterns predicted by batch-type prediction algorithm. This is a baseline system.

RT Enhanced EL speech with the real-time prediction algorithm for segmented CF_0 pattern modeling (delay time: 85 msec). This is a simulated system of our proposed system.

Forthcoming Enhanced EL speech with forthcoming F_0 prediction on real-time prediction algorithm for CF_0 patterns modeling (delay time: 265 msec). This is also a simulated system.

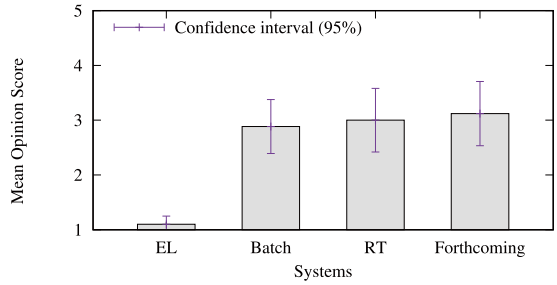

Fig. 11 Naturalness of Predicted F_0 patterns.

The result is shown in Fig. 11. As reported in [9], we confirmed that **Batch** is significantly improved compared with **EL** by predicting F_0 patterns based on statistical F_0 patterns. For our proposed methods **RT** and **Forthcoming**, we achieved that two proposed systems caused no degradation compared with **Batch**. These results show that our proposed methods successfully overcome the latency issues mentioned in Sect. 4.

6. Conclusion

In this paper, we have proposed a new electrolarynx capable of automatically controlling F_0 patterns of its excitation signals based on statistical F_0 pattern prediction. Moreover, we have also proposed two methods to address the latency issues caused by the whole process of our proposed enhancement system: segmented continuous F_0 patterns modeling and forthcoming F_0 modeling. In additionally, we have also designed the simulation experiment of our proposed enhancement system to alleviate several construction costs, such as recording to evaluate new proposal. Through implementing a prototype system and its simulation, we have demonstrated that our proposed system is capable of effectively addressing the issues of electrolarynx.

Acknowledgements

This work was supported in part of JSPS KAKENHI Grant Numbers: 15J10727 and 17H01763, and JST PRESTO Grant Number JPMJPR1657, and the authors would like to thank Mr. Sugai of Densei Communication Inc., Japan, for advice 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," Proc. 3rd IEEE International Workshop of Robot and Human Communication, pp.198–203, July 1994.
[2] Y. Kikuchi and H. Kasuya, "Development and evaluation of pitch adjustable electrolarynx," Proc. Speech Prosody 2004, International Conference., pp.761–764, March 2004.
[3] K. Matsui, K. Kimura, Y. Nakatoh, and Y.O. Kato, "Development of electrolarynx with hands-free prosody control," Proc. SSW8, pp.273–277, Aug. 2013.
[4] B. Roubeau, C. Chevrie-Muller, and J.L.S. Guily, "Electromyographic activity of strap and cricothyroid muscles in pitch change," Proc. Acta Otolaryngologica, vol.117, no.3, pp.459–464, May 1997.
[5] T. Shipp, E.T. Doherty, and P. Morrissey, "Predicting vocal frequency from selected physiologic measures," Proc. J. Acoust. Soc. Am, vol.66, pp.678–684, 1979.
[6] K. Nakamura, M. Janke, M. Wand, and T. Schultz, "Estimation of fundamental frequency from surface electromyographic data: EMG-to-F_0," Proc. ICASSP, pp.573–576, May 2011.
[7] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," Speech Commun., vol.54, no.1, pp.134 –146, Jan. 2012.
[8] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "A laryngeal speech enhancement based on one-to-many eigenvoice conversion," IEEE/ACM Trans. Audio Speech & Language Process., vol.22, no.1, pp.172–183, Jan. 2014.
[9] 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 Trans. Inf. & Syst., vol.E97-D, no.6, pp.1429–1437, June 2014.
[10] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp.131–142, March 1998.
[11] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio Speech & Language Process., vol.15, no.8, pp.2222–2235, Nov. 2007.
[12] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," 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," IEEE Trans. Audio Speech & Language Process., 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] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315–1318, June 2000.
[16] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion." Proc. INTERSPEECH, pp.94–97, Sept. 2012.
[17] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," Proc. INTERSPEECH, pp.1076–1079, Sept. 2008.
[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," Proc. INTERSPEECH, pp.3072–3076, Aug. 2013.
[19] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," Proc. Speech Commun., vol.27, no.3-4, pp.187–207, April 1999.
[20] 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," Proc. MAVEBA, pp.13–15, Sept. 2001.
[21] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," Proc. INTERSPEECH, pp.2266–2269, Sept. 2006.
[22] M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech database," ATR Technical Report, TR-I-0166, Sept. 1990.

Kou Tanaka received his B.E. degree from the Department of Mathematics and Informatics, Faculty of Human Development, Kobe University, Hyogo, Japan in 2012 and his M.E. degree from the Graduate School of Information Science, NAIST, Nara, Japan, in 2014, respectively. He is currently in the doctoral course at NAIST. He is a student member of ISCA, and ASJ.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005–011) and an Associate Professor (2011–015) at NAIST. From 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was a director of ATR Spoken Language Communication Research Laboratories in 2000–008, and a vice president of ATR in 2007–008. He was a director general of Keihanna Research Laboratories, National Institute of Information and Communications Technology, Japan in 2009–010. He is currently a professor and a director of Augmented Human Communication laboratory, Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of spoken dialog system, speech-to-speech translation. He is one of the leaders of speech-to-speech translation research projects including C-STAR, IWSLT and A-STAR. He headed the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received LREC Antonio Zampoli Award, the Commendation for Science and Technology by the Ministry of Science and Technology in Japan. He is an elected board member of ISCA, International Speech Communication Association, and an elected member of IEEE SPS, speech and language TC.