INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Physically Constrained Statistical F_0 Prediction for Electrolaryngeal Speech Enhancement

Kou Tanaka[1]*, Hirokazu Kameoka*[2]*, Tomoki Toda*[3]*, and Satoshi Nakamura*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[3]Information Technology Center, Nagoya University, Japan

ko-t@is.naist.jp, kameoka.hirokazu@lab.ntt.co.jp, tomoki@icts.nagoya-u.ac.jp, s-nakamura@is.naist.jp

Abstract

Electrolaryngeal (EL) speech produced by a laryngectomee using an electrolarynx to mechanically generate artificial excitation sounds severely suffers from unnatural fundamental frequency (F_0) patterns caused by monotonic excitation sounds. To address this issue, we have previously proposed EL speech enhancement systems using statistical F_0 pattern prediction methods based on a Gaussian Mixture Model (GMM), making it possible to predict the underlying F_0 pattern of EL speech from its spectral feature sequence. Our previous work revealed that the naturalness of the predicted F_0 pattern can be improved by incorporating a physically based generative model of F_0 patterns into the GMM-based statistical F_0 prediction system within a Product-of-Expert framework. However, one drawback of this method is that it requires an iterative procedure to obtain a predicted F_0 pattern, making it difficult to realize a real-time system. In this paper, we propose yet another approach to physically based statistical F_0 pattern prediction by using a HMM-GMM framework. This approach is noteworthy in that it allows to generate an F_0 pattern that is both statistically likely and physically natural without iterative procedures. Experimental results demonstrated that the proposed method was capable of generating F_0 patterns more similar to those in normal speech than the conventional GMM-based method.

Index Terms: electrolaryngeal speech, statistical F_0 prediction, generative model, speech enhancement

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 human body. Unfortunately, there are many people with disabilities that prevent them from producing speech freely, leading to communication barriers and degrading Quality of Life (QoL). Laryngectomees are people with disabilities and have undergone an operation to remove the larynx including the vocal folds for several reasons such as injury and laryngeal cancer. Their ability to generate excitation sounds is severely impaired, because they no longer have their vocal folds although their vocal tracts remain. One alternative means of producing speech sounds is the use of electrolaryngeal (EL) speech, which allows them to use the excitation sounds mechanically generated from an electrolarynx. EL speech is reasonably intelligible, but somewhat unnatural particularly due to the monotonic excitation sounds.

To address this issue, we have previously proposed methods that make it possible to convert acoustic features of EL speech to those of normal-sounding speech by predicting the fundamental frequency (F_0) pattern from the spectrum sequence of the EL speech based on Gaussian Mixture Models (GMMs) [1, 2, 3]. With a similar aim, whisper-to-speech conversion [4] and whisper-to-audible speech conversion [5] have been proposed. These methods have successfully shown to improve the naturalness of EL speech [1, 2] while preserving its intelligibility [3]. However, the predicted F_0 patterns are still unnatural compared with that in normal speech. One possible reason is that the predicted F_0 patterns were not necessarily guaranteed to satisfy the physical constraints of the actual control mechanism of the thyroid cartilage, even though they were optimal in a statistical sense.

As for the generative process of F_0 patterns, one of the authors have proposed a statistical model [7, 8, 9] formulated as a stochastic counterpart of the Fujisaki model [10], a well-founded mathematical model representing the control mechanism of vocal fold vibration. The Fujisaki model [10] assumes that a F_0 pattern on a logarithmic scale is the superposition of a phrase component, an accent component and a constant value. The phrase and accent components are assumed to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. The model reported in [7, 8, 9] has made it possible to estimate the underlying parameters of the Fujisaki model that best explain the given F_0 pattern, by using powerful statistical inference techniques. We previously proposed unifying this model and the GMM-based F_0 pattern prediction model within a Product-of-Experts (PoE) framework [6]. Although this framework showed improvement in the prediction accuracy, it turned out that it was not suitable to real-time processing because of the iterative procedure required in the prediction process.

In this paper, we propose yet another approach that makes it possible to use the generative model of F_0 patterns for statistical F_0 pattern prediction using an HMM-GMM. Specifically, we model the joint probability distribution of a sequence of the phrase/accent commands and a spectral feature sequence. As we shall see in the following, the state sequence of our HMM is associated with a sequence of the phrase/accent commands of the Fujisaki model. Here, we assume that this HMM emits a spectral feature vector at each state. By doing so, our HMM can be used as a model to predict a phrase/accent command sequence from an input spectral feature sequence. With properly trained parameters, the most likely phrase/accent command sequence given an input spectral feature sequence can be found by state decoding. Experimental results in term of prediction accuracy revealed that the proposed method was capable of predicting F_0 patterns more similar to those in normal speech than

http://dx.doi.org/10.21437/Interspeech.2017-688

the conventional GMM-based method.

2. Statistical F_0 Pattern Prediction Based on GMM

We briefly review our statistical F_0 prediction method [1, 2, 3], based on statistical voice conversion techniques [11, 12]. The aim of this method is to predict F_0 patterns from the spectral features of EL speech. As with voice conversion methods, it consists of training and prediction processes.

Let $\boldsymbol{\lambda}_G$ be the parameters of the joint probability density $p((\boldsymbol{x}[t]^\top, \boldsymbol{o}[t]^\top)^\top|\boldsymbol{\lambda}_G)$ defined as a GMM, where $^\top$ is transposition, and $\boldsymbol{x}[t]$ and $\boldsymbol{o}[t]$ are a source feature and a target feature at time frame t, respectively. The corresponding joint feature vectors can be obtained by performing automatic frame alignment with Dynamic Time Warping. As a source feature, the spectral segment feature $\boldsymbol{x}[t]$ of EL speech is obtained by applying principal component analysis (PCA) to the stacked vector consisting of the mel-cepstra of multiple frames around the current frame t [13]. The target feature $\boldsymbol{o}[t] = (y[t], \Delta y[t])^\top$ consists 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. At training time, the parameters of the GMM are trained using the parallel data of source and target features.

At test time, 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:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\ p(\boldsymbol{o}|\boldsymbol{x}, \boldsymbol{\lambda}_G) \quad \text{subject to} \quad \boldsymbol{o} = \boldsymbol{W}\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{o} = (\boldsymbol{o}[1]^\top, \ldots, \boldsymbol{o}[T]^\top)^\top$ is the joint static and dynamic feature vector sequence, and $\boldsymbol{W}$ is a constant matrix that transforms the static feature vector sequence $\boldsymbol{y}$ to $\boldsymbol{o}$.

3. Generative Model of F_0 Patterns

The generative model of F_0 patterns [7, 8, 9] is a stochastic counterpart of a discrete-time version of the Fujisaki model [10]. The Fujisaki model (shown in Fig. 1) assumes that F_0 patterns $y[t]$ on a logarithmic scale is given as the sum of three components:

$$y[t] = y_{\mathrm{p}}[t] + y_{\mathrm{a}}[t] + \mu_{\mathrm{b}}, \tag{2}$$

where $y_{\mathrm{p}}[t]$ and $y_{\mathrm{a}}[t]$ are a phrase component and an accent component at time frame t, and μ_{b} is a constant value, respectively. The phrase and accent components are assumed to be the outputs of different second-order critically damped filters, $g_{\mathrm{p}}[t]$ and $g_{\mathrm{a}}[t]$, excited with Dirac deltas $u_{\mathrm{p}}[t]$ (phrase commands) and rectangular pulses $u_{\mathrm{a}}[t]$ (accent commands), respectively:

$$y_{\mathrm{p}}[t] = g_{\mathrm{p}}[t] * u_{\mathrm{p}}[t], \tag{3}$$
$$y_{\mathrm{a}}[t] = g_{\mathrm{a}}[t] * u_{\mathrm{a}}[t], \tag{4}$$

where $*$ is convolution over time.

The key idea of the generative model proposed in [7, 8, 9] is that a sequence of phrase/accent command pair is modeled as an output sequence of the following path-restricted hidden Markov model (HMM) with Gaussian emission densities (shown in Fig. 2) so that estimating the state transition of the HMM directly amounts to estimating the Fujisaki-model parameters. This HMM $\boldsymbol{\lambda}_F$ consists of M+N+2 states, r_0, r_1, p_0,. . ., p_{M-1}, a_0,. . ., a_{N-1}. M and N are the numbers of possible values that

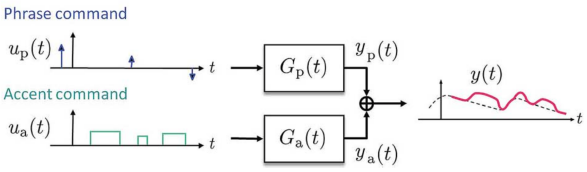

Figure 1: *Original Fujisaki model [10].*

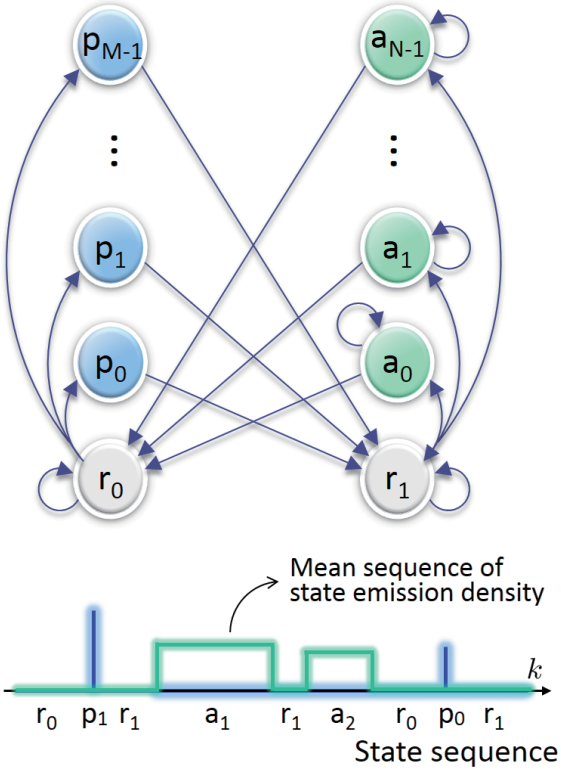

Figure 2: *Command function modeling with HMM.*

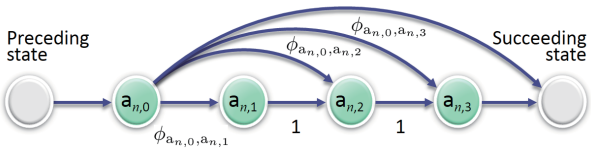

Figure 3: *State splitting.*

the magnitude of each phrase/accent command can take. It can thus be understood as the resolution of magnitude "quantization". At states r_0 and r_1, the means $\mu_{\mathrm{p}}[t] = \mathbb{E}[u_{\mathrm{p}}[t]]$ and $\mu_{\mathrm{a}}[t] = \mathbb{E}[u_{\mathrm{a}}[t]]$ are both restricted to zero. At states p_0,. . ., p_{M-1}, $\mu_{\mathrm{p}}[t]$ can take non-zero values, whereas $\mu_{\mathrm{a}}[t]$ is still restricted to zero. At states a_0,. . ., a_{N-1}, $\mu_{\mathrm{a}}[t]$ can take non-zero values, whereas $\mu_{\mathrm{p}}[t]$ is forced to be zero. The path constraint shown in Fig. 2 restricts $\mu_{\mathrm{p}}[t]$ to consisting of isolated deltas and $\mu_{\mathrm{a}}[t]$ to consisting of rectangular pulses. Furthermore, we split each state (except for p_m) into a certain number of substates (as in Fig. 3) so that we can directly assign probabilities to the durations of self-transition.

Output sequence: $\boldsymbol{u}[t] = (u_{\rm p}[t], u_{\rm a}[t])^\top$ $(t = 1, \ldots, T)$,
Set of states: $\boldsymbol{S} = \{{\rm r}_0, {\rm p}_0, \ldots, {\rm p}_{M-1}, {\rm r}_1, {\rm a}_0, \ldots, {\rm a}_{N-1}\}$,
State sequence: $\boldsymbol{s} = \{s[t] \in \boldsymbol{S} | t = 1, \ldots, T\}$,
Emission densities: $p(\boldsymbol{u}[t]|s[t] = i) = \mathcal{N}(\boldsymbol{u}[t]; \boldsymbol{c}_i, \boldsymbol{\Sigma}_i)$,

$$\boldsymbol{c}_i = \begin{cases} (0,0)^\top & (i \in {\rm r}_0, {\rm r}_1) \\ (\mu_{\rm p}^{(m)}, 0)^\top & (i \in {\rm p}_m) \\ (0, \mu_{\rm a}^{(n)})^\top & (i \in {\rm a}_n) \end{cases}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} v_{{\rm p},i}^2 & 0 \\ 0 & v_{{\rm a},i}^2 \end{bmatrix},$$

State transition probabilities: $\phi_{\hat{\imath},i} = p(s[t] = i | s[t-1] = \hat{\imath})$.

By using the conditional density $p(\boldsymbol{y}|\boldsymbol{u})$ reflecting the constraints Eq. (2,3,4) and the HMM likelihood $p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$, we can describe the probability density of $\boldsymbol{y}$ as:

$$p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\lambda}_F) = \int p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{s}, \boldsymbol{\lambda}_F) \, {\rm d}\boldsymbol{u} \tag{5}$$

$$= \mathcal{N}(\boldsymbol{y}; \boldsymbol{m}, \boldsymbol{\Gamma}), \tag{6}$$

where $\boldsymbol{m}$ is given as the F_0 pattern of the Fujisaki model given a phrase/accent command sequence $\boldsymbol{s}$ [7, 8, 9].

4. Proposed F_0 Pattern Prediction Based on HMM-GMM

4.1. Joint Modeling with HMM-GMM

Here, we propose designing the joint probability distribution of a sequence of the phrase/accent commands of target speech and a spectral feature sequence of source speech. The state sequence of the HMM introduced in Sec. 3 is associated with a sequence of the phrase/accent commands of the Fujisaki model. Thus, if we assume that this HMM $\boldsymbol{\lambda}_h$ emits a spectral feature vector at each state according to a GMM emission density, the joint probability of output and state sequences of this HMM-GMM can be used as a model to predict a phrase/accent command sequence from an input spectral feature sequence. With properly trained parameters, the most likely phrase/accent command sequence given an input spectral feature sequence can be found by state decoding.

We can write the joint *p.d.f.* of the spectral feature sequence $\boldsymbol{x}$ of EL speech and the state sequence $\boldsymbol{s}$ of the phrase/accent commands of normal speech as:

$$p(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{\lambda}_h) = p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{\lambda}_h) p(\boldsymbol{s}|\boldsymbol{\lambda}_h), \tag{7}$$

where

$$p(\boldsymbol{s}|\boldsymbol{\lambda}_h) = \prod_t p(s[t]|s[t-1], \boldsymbol{\lambda}_h), \tag{8}$$

$$p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{\lambda}_h) = \prod_t p(\boldsymbol{x}[t]|\boldsymbol{s}[t], \boldsymbol{\lambda}_h). \tag{9}$$

We assume the state emission density to be a GMM:

$$p(\boldsymbol{x}[t]|s[t] = i) = \sum_m \alpha_{i,m} \mathcal{N}(\boldsymbol{x}[t]; \boldsymbol{\mu}_{i,m}, \boldsymbol{\Sigma}_{i,m}). \tag{10}$$

4.2. Training Process

As with the method in Sec. 2, we use the parallel data of source and target speech obtained using DTW for training. First, we train the state transition probabilities of the path-restricted HMM so that the HMM is ensured to produce only the phrase/accent command sequences that are likely to occur in normal speech. By setting the transition probability of a particular pair of states at exactly 0, we can impose constraints on state transitions. This is in particular convenient since phrase/accent command sequences have several requirements to be met [10].

We then use the method in [7, 8, 9] to extract the phrase/accent commands from target speech and determine the state sequence $\boldsymbol{s}$. With fixed $\boldsymbol{s}$, we can train the state emission density of each state by using the spectral feature vectors of source EL speech at all the frames assigned to that state. To avoid overfitting, we consider reducing the parameters of the HMM by parameter tying. Specifically, we partition the substates in each state into three segments so that all the substates belonging to the same segment have exactly the same emission densities. As for the spectral feature, we use the spectral segment feature used in the method described in Sec. 2. Namely, the spectral segment feature $\boldsymbol{x}[t]$ is obtained by applying PCA to the stacked vector consisting of the mel-cepstra of multiple frames around the current frame t.

4.3. Prediction Process

Given the spectral feature sequence $\boldsymbol{x}$ of EL speech, we can predict the most likely state sequence $\hat{\boldsymbol{s}}$ by maximizing the following conditional probability:

$$\hat{\boldsymbol{s}} = \operatorname*{argmax}_{\boldsymbol{s}} p(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{\lambda}_h) \tag{11}$$

$$= \operatorname*{argmax}_{s} p(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{\lambda}_h). \tag{12}$$

Given $\hat{\boldsymbol{s}}$, we can obtain the F_0 pattern $\hat{\boldsymbol{y}}$ by maximizing Eq. (5) with respect to $\boldsymbol{y}$. Since $p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\lambda}_F)$ is given as a multivariate normal distribution [9], $\hat{\boldsymbol{y}}$ is simply given by its mean. In contrast with the conventional GMM-based approach, the present framework always ensures the predicted F_0 pattern to follow the Fujisaki model. This restriction is perhaps too strong, but may be advantageous in that it can effectively avoid generating unnatural patterns.

5. Experimental Evaluation

5.1. Experimental Conditions

We objectively evaluated the performance of the proposed method with the following F_0 correlation coefficients $f_{\rm corr}$ between the predicted F_0 patterns $\hat{\boldsymbol{y}}$ and target F_0 patterns $\boldsymbol{y}$ at voiced frames:

$$f_{\rm corr} = \frac{\sigma_{\hat{\boldsymbol{y}}\boldsymbol{y}}^2}{\sigma_{\hat{\boldsymbol{y}}} \sigma_{\boldsymbol{y}}}, \tag{13}$$

where $\sigma_{\hat{\boldsymbol{y}}\boldsymbol{y}}^2$ is a covariance of the predicted and the target F_0 patterns, and $\sigma_{\hat{\boldsymbol{y}}}$ and $\sigma_{\boldsymbol{y}}$ are standard deviations of the predicted and the aliged reference F_0 patterns, respectively. We chose the GMM-based method [1, 2, 3] as a baseline method for comparison. We also tested a simple postfiltering approach that consists of performing the GMM-based F_0 prediction followed by fitting the Fujisaki model to the predicted F_0 pattern using the method of [7, 8, 9].

For the training, we recorded EL speech uttered by one normal person (male) and normal speech uttered by one different normal person (male). Each speaker uttered about 503 sentences in the ATR phonetically balanced sentence set [14]. We used 450 sentences as training data, and remaining 53 sentences as test data. The sampling frequency was set to 16 kHz. The frame shift was set to 5 ms. The frame length of feature extraction was set to 25 ms. The dimension of mel-cepstra including waveform power information was set to 25. The concatenated

frames were ± 4 frames around current frame when we perform the PCA. Thus, we compressed a $25 * (4 + 1 + 4) = 225$ dimensions feature into a 50 dimensions feature. The extraction of mel-cepstra of EL speech uses FFT analysis. The extraction of mel-cepstra of normal speech uses STRAIGHT analysis [15]. The extraction of F_0 of normal speech is performed by TEMPO [15]. The level of amplitude of phrase command and accent command were set to 3, respectively. The number of mixture components are 16 and 4 for the GMM-based method and the proposed method, respectively. In this paper, to address the smaller amount of spectral features on phrase commands compared with that on accent commands, we modified the Dirac deltas of phrase command to rectangular pulses of 5 frames which amplitude is one-fifth amplitude of corresponding Dirac delta.

The methods selected for comparison were:

- *gmm*: the GMM-based maximum likelihood trajectory estimation of F_0 patterns by using the F_0 feature of normal speech as the target feature.
- *gmm+post*: As the post processing, fit the Fujisaki model to the above predicted F_0 patterns,
- *proposed*: Predict F_0 patterns with the proposed HMM-GMM based method.

5.2. Experimental Results

Table 1 shows that using the postfiltering slightly improved the prediction accuracy obtained with the GMM-based method. This is because the predicted F_0 patterns are modified so that the physical constraint expressed by the Fujisaki model is satisfied (see Fig. 4). We confirmed that the proposed method yielded even higher accuracy than both of the baseline methods. This implies the advantage of the proposed method, which allows for the direct prediction of the phrase/accent commands of the Fujisaki model. These show that modeled natural duration and amplitude of F_0 command help us to generate F_0 patterns from spectral features and the use of fixed transition probability is effective.

5.3. Discussion

Our proposed method allows us to directly predict the phrase /accent commands from the spectral features. This allows us to modify the F_0 patterns within the physical constraint underlying the generative process of F_0 patterns. It should be noted that the present model can also be developed into our previously proposed methods, such as [3] and [6], for further improvements.

While the GMM-based method allows us to predict raw F_0 values, the phrase commands and accent commands are quantized into 3 levels in the proposed method, described in Sec. 3 and 5. The accuracy of F_0 pattern prediction suffers from this negative impact of quantization.

To further improve the prediction accuracy, we also plan to incorporate the model employed in [16]. In normal speech, many phrases or sentences share the same intonation pattern.

Table 1: *F_0 correlation coefficients*

gmm	0.40
gmm+post	0.41
proposed	0.50

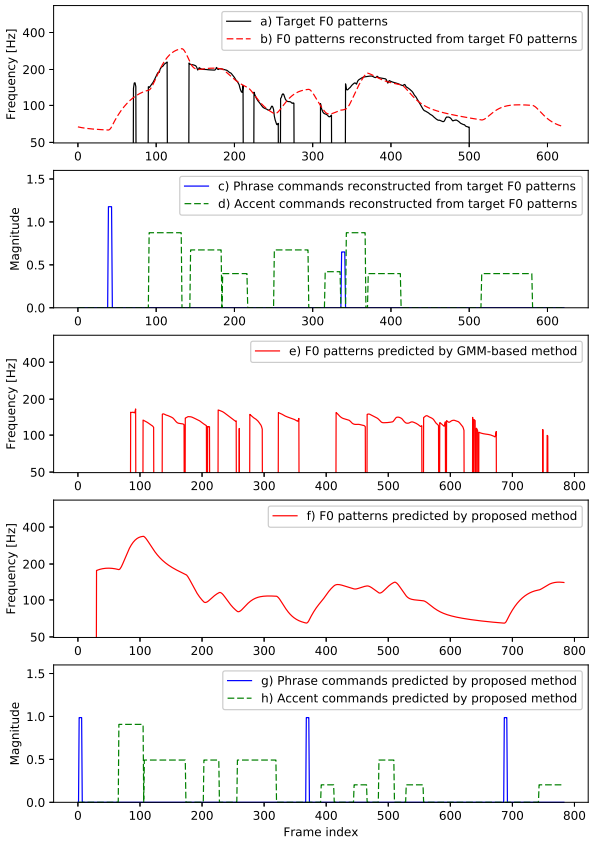

Figure 4: *Sample of F0 patterns and phrase and accent commands.*

This is because an intonation pattern is usually determined by the grammatical structure of an uttered sentence or the accent type associated with each phrase. Thus, it would be natural to hypothesize that all phrase and accent command sequences are drawn from a vocabulary consisting of relatively small number of intonation pattern templates. By using a dictionary consisting of a finite number of left-to-right HMM templates, we can assume that a sequence of the phrase and accent command pairs is generated according to a concatenation of those templates. This can be modeled by an HMM topology employed in [16].

6. Conclusions

In this paper, to improve F_0 prediction performance in electrolaryngeal speech enhancement, we proposed a HMM-GMM based F_0 pattern prediction that combined two conventional methods, a statistical F_0 pattern prediction method and a statistical F_0 pattern modeling method based on its generative process. Experimental results revealed that the proposed method outperformed our conventional method in terms of prediction accuracy.

7. Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 26280060, and by JST, PRESTO Grant Number JPMJPR1657.

8. References

[1] 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, pp. 134– 146, January 2012.

[2] 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 Transactionson,* vol. 22, no. 1, pp. 172–183, January 2014.

[3] 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,* vol. E97-D, no. 6, pp. 1429–1437, June 2014.

[4] J. Li, I. V. McLoughlin, L. Dai, and Z. Ling, "Whisper-to-speech conversion using restricted Boltzmann machine arrays," *Electronics Letters,* vol. 50, no. 24, pp. 1781–1782, November 2014.

[5] M. Janke, M. Wand, T. Heistermann and T. Schultz, "Fundamental frequency generation for whisper-to-audible speech conversion," *Proc. ICASSP,* pp. 2598–2602, May 2014.

[6] K. Tanaka, H. Kameoka, T. Toda, and S. Nakamura, "Statistical $\boldsymbol{F_0}$ prediction for electrolaryngeal speech enhancement considering generative process of $\boldsymbol{F_0}$ contours within product of experts framework," *Proc. ICASSP,* pp. 5665–5669, March 2016.

[7] H. Kameoka, J. Le Roux, Y. Ohishi, "A statistical model of speech F_0 contours," *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition,* pp. 43–48, September 2010.

[8] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation," *Proc. Speech Prosody,* pp. 175–178, May 2012.

[9] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *Audio, Speech, and Language Processing, IEEE/ACM Transactionson,* vol. 23, no. 6, pp. 1042–1053, June 2015.

[10] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal Fold Physiology: Voice Production, Mechanisms and Functions,* pp. 347–355, 1998.

[11] 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.

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

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

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

[15] 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," in *Proc. Speech Communication,* Vol. 27, No. 3-4, pp. 187–207, April 1999.

[16] T. Ishihara, H. Kameoka, K. Yoshizato, D. Saito, S. Sagayama, "Probabilistic speech F_0 contour model incorporating statistical vocabulary model of phrase-accent command sequence," in *Proc. Interspeech,* pp. 1017–1021, August 2013.