10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Generative Modeling of F_0 Contours Leveraged by Phrase Structure and Its Application to Statistical Focus Control

Yuma Shirahata, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Japan
{shirahata, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we propose a statistical generative model of fundamental frequency (F_0) contours that incorporates a phrase structure of Japanese ("*bunsetsu*"), and apply this model to control of the focus point in a sentence. Fujisaki model is a mathematical model that formulates F_0 contours as the superposition of phrase and accent components, considering the control mechanism of vocal fold vibration. In the Fujisaki model, model parameters are closely related to linguistic information. Thus, flexible and interpretable conversion of F_0 contours corresponding to linguistic information is achieved by changing the model parameters. Recently, a method of treating the Fujisaki model as a stochastic model has been proposed. In this method, the model parameters are inferred from observed F_0 contours by a maximum likelihood manner. However, since there are no constraints of linguistic information in inference, unnatural parameters are occasionally estimated. In the proposed method, occurrence of phrase commands is linked to the boundaries of bunsetsu, and then the Fujisaki model parameters and phrase structure correspond to each other. It enables simultaneous modeling of two different F_0 contours in every bunsetsu unit. The proposed modeling can be applied to pairs of neutral and focused utterances, and it enables bunsetsu-by-bunsetsu focus control . Experimental results show that the proposed method achieved reasonable control of focus in 74% accuracy rate compared with natural speech. Though there is room for improvement in naturalness, the proposed scheme achieves interpretable conversion of prosody.

Index Terms: speech F_0 contours, Fujisaki model, hidden Markov model, EM algorithm

1. Introduction

The fundamental frequency (F_0) of speech conveys various kinds of information such as linguistic, non-linguistic and paralinguistic information. These kinds of information are expressed not in a frame, but in a relatively long time structure, such as a word or a clause. In statistical parametric speech synthesis, F_0 contours are often modeled by using hidden Marcov models (HMM) or deep neural networks (DNN) [1, 2, 3], which process F_0 contours in a frame-by-frame manner. While these frameworks have an advantage that they can use the raw F_0 value as a training data directly, they inherently handle few information of time structure, and have difficulty in capturing long time structures. Speech synthesis model utilizing recurrent neural networks (RNN) or end-to-end synthesis models implicitly incorporate time structures. However, these models does not explicitly tie linguistic information to model parameters. Thus, there is no guarantee that they exactly represent physical phenomena such as vocal fold vibration [4, 5].

On the other hand, Fujisaki model is one of the solutions to these issues, since it formulates the process of generating F_0 contours of human speech and directly models both time structure and the relationship between linguistic information and parameters of the model [6]. In this model, F_0 contours in logarithmic scale are described as the superposition of phrase and accent components. These components correspond to the pitch variations of phrase units and those of accent units, respectively. Since the Fujisaki model incorporates the process of human speech with an explicit formula, it has an advantage to generating natural F_0 contours. Another advantage of handling F_0 contours in the Fujisaki model framework is that we can obtain clear relationship between generated F_0 contours and their background linguistic information. Hence, the framework enables flexible and interpretable control of prosodic features. Namely, for example, we can impose a focus on a specific word, by increasing magnitude of the corresponding accent command [7].

Recently, a stochastic model of speech F_0 contours has been proposed by translating the Fujisaki model into a probabilistic generative model [8]. This model expresses the process to generate a sequence of phrase and accent commands as a stochastic process represented by HMM. In this model, the Fujisaki model parameters are automatically estimated from observed F_0 contours in a maximum likelihood manner. In addition, as an extension of this stochastic model, a method of converting an F_0 contour of a speaker to that of another speaker has been studied [9]. Though these stochastic models are powerful framework, they adopt naive HMM topologies. Thus, the advantage of the Fujisaki model that it explicitly incorporates the relationship between linguistic information and model parameters, is not sufficiently utilized. Imposing some linguistic restriction to the HMM topologies makes the framework much more effective.

In this paper, we propose a stochastic model incorporating bunsetsu, which is a phrase structure of Japanese. In this model, a bunsetsu is utilized as a minimal unit in which the HMM topologies are controlled, and the occurrence of phrase commands is linked to the boundary of bunsetsu. This model enables simultaneous modeling of two different F_0 contours in every bunsetsu unit and can be applied to the modeling of neutral and focused utterances. This paper also investigates the method of converting an F_0 contour of a neutral utterance to that of an utterance in which a target bunsetsu is focused, by applying the simultaneous modeling.

The rest of this paper is organized as follows: Section 2 briefly introduces the original Fujisaki model and its expansion to statistical modeling. Section 3 describes the proposed stochastic model incorporating bunsetsu. Section 4 presents the experimental evaluations. Section 5 concludes this paper.

10.21437/SSW.2019-41

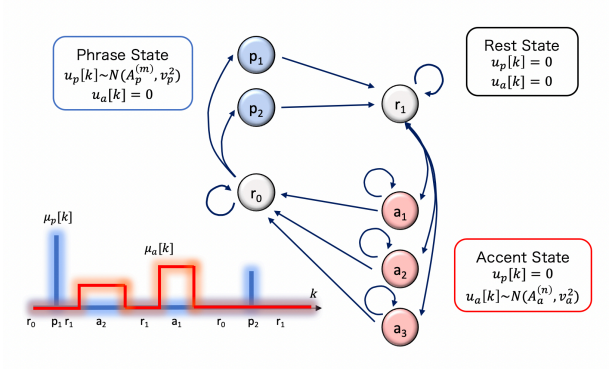

Figure 1: *The state transition topology of Fujisaki model commands.* p, a, r *denotes phrase states, accent states, rest states, respectively.*

2. Generative Model of Speech F_0 Contours

2.1. Fujisaki Model

The original Fujisaki model formulates an F_0 contour in logarithmic scale, $y(t)$, where t is time, as the superposition of a phrase component $x_{\mathrm{p}}(t)$, an accent component $x_{\mathrm{a}}(t)$ and a base component x_{b}:

$$y(t) = x_{\mathrm{p}}(t) + x_{\mathrm{a}}(t) + x_{\mathrm{b}}. \tag{1}$$

The phrase component $x_{\mathrm{p}}(t)$ represents long-scale pitch variations over the duration of prosodic units, and the accent component $x_{\mathrm{a}}(t)$ represents relatively short-scale pitch variations in accent units. x_{b} is a constant value which represents the lower bound of the speaker's $\log F_0$. The phrase component is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command $u_{\mathrm{p}}(t)$, while the accent component is generated by another second-order, critically-damped linear filter in response to a stepwise accent command $u_{\mathrm{a}}(t)$:

$$x_{\mathrm{p}}(t) = G_{\mathrm{p}}(t) * u_{\mathrm{p}}(t), \tag{2}$$
$$x_{\mathrm{a}}(t) = G_{\mathrm{a}}(t) * u_{\mathrm{a}}(t). \tag{3}$$

In the equation, $*$ denotes convolution over time, and $G_{\mathrm{p}}(t)$, $G_{\mathrm{a}}(t)$ are described as

$$G_{\mathrm{p}}(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{4}$$

$$G_{\mathrm{a}}(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{5}$$

where α, β are natural angular frequencies of the two second-order systems. It is generally said that $\alpha = 3.0$ rad/s and $\beta = 20.0$ rad/s can be used as default values [8].

2.2. Stochastic Model of F_0 Contours

A method of modeling the generative process of speech F_0 contours based on a discrete-time version of Fujisaki model has been proposed [8]. Multiple constraints are required in the original Fujisaki model; 1) phrase and accent commands should be impulse sequences and rectangular pulses, respectively. 2) Both of the commands are non-negative, and 3) they are not allowed to overlap with each other. In order to meet these requirements,

Table 1: *The parameters to define the HMM.*

Output sequence:	$\{\boldsymbol{u}[k]\}_{k=1}^{K}$
State sequence:	$\{s_k\}_{k=1}^{K}$
Output distribution:	$P(\boldsymbol{u}[k] \mid s_k) = \mathcal{N}(\boldsymbol{u}[k]; \boldsymbol{\mu}[k], \boldsymbol{\Sigma}[k])$
	$\boldsymbol{\mu}[k] = \begin{cases} (0,0)^\top & (s_k \in \mathrm{r}_0, \mathrm{r}_1) \\ (A_{\mathrm{p}}^{(m)}, 0)^\top & (s_k \in \mathrm{p}_m) \\ (0, A_{\mathrm{a}}^{(n)})^\top & (s_k \in \mathrm{a}_n) \end{cases}$
Transition probability:	$\phi_{i',i} = P(s_k = i \mid s_{k-1} = i')$

the $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ pair, i.e., $\boldsymbol{u}[k] = (u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^\top$ was treated as a model parameter, using a hidden Markov model (HMM). $\boldsymbol{u}[k]$ was modeled as the output of the HMM illustrated in Figure 1. The output distribution of each HMM state is a Gaussian distribution

$$\boldsymbol{u}[k] \sim \mathcal{N}(\boldsymbol{u}[k]; \boldsymbol{\mu}[k], \boldsymbol{\Sigma}[k]), \tag{6}$$

where s_k denotes the HMM state in discrete-time k and $\boldsymbol{\mu}[k] = (\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k])^\top$, $\boldsymbol{\Sigma}[k] = \mathrm{diag}(v_{\mathrm{p}}^2, v_{\mathrm{a}}^2)$ are the mean vector and the covariance matrix of $\boldsymbol{u}[k]$, respectively: Thus, the HMM to generate the command sequence is defined by the parameters shown in Table 1. The output sequence of the above HMM $u_{\mathrm{p}}[k], u_{\mathrm{a}}[k]$ is then convoluted with different second-order filters $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$, respectively, to generate the phrase and accent component $x_{\mathrm{p}}[k], x_{\mathrm{a}}[k]$:

$$x_{\mathrm{p}}[k] = u_{\mathrm{p}}[k] * G_{\mathrm{p}}[k], \tag{7}$$
$$x_{\mathrm{a}}[k] = u_{\mathrm{a}}[k] * G_{\mathrm{a}}[k], \tag{8}$$

where $*$ denotes convolution over k and $G_{\mathrm{p}}[k], G_{\mathrm{a}}[k]$ are the discrete-time representation of (4) and (5), respectively. The logarithmic F_0 contour is then described as

$$x[k] = x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + x_{\mathrm{b}}, \tag{9}$$

where x_{b} denotes the lower bound of the $\log F_0$ contour. Generally, in a stochastic process, it is regarded that observations include a uncertainty. In the case of the generation process of F_0, observed F_0 contours reflects this uncertainty. In order to incorporate this uncertainty of observed F_0 contours, an observed F_0 contour $y[k]$ is modeled as the superposition of the above $x[k]$ and a noise component $x_{\mathrm{n}}[k] \sim \mathcal{N}(0, v_{\mathrm{n}}[k]^2)$:

$$y[k] = x[k] + x_{\mathrm{n}}[k]. \tag{10}$$

By marginalizing $x_{\mathrm{n}}[k]$ out, the probability density function of $\boldsymbol{y} = \{y[k]\}_{k=1}^{K}$, given $\boldsymbol{u} = \{u[k]\}_{k=1}^{K}$, can be written as follows:

$$P(\boldsymbol{y} \mid \boldsymbol{u}) = \prod_{k=1}^{K} \mathcal{N}(y[k]; x[k], v_{\mathrm{n}}^2[k]), \tag{11}$$

where $x[k]$ is derived from (7)-(9).

2.3. Statistical Vocabulary Model of Fujisaki Model Commands

In the stochastic model of F_0 contours introduced in Section 2.2, the generated patterns of the Fujisaki model commands, i.e., phrase and accent commands, are restricted by the state transition topology of HMM. Since the restriction was limited

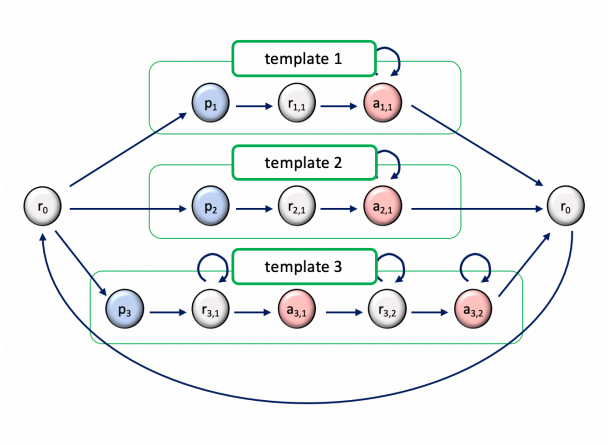

Figure 2: *The state transition topology in statistical vocabulary model. Transition is restricted by templates.* p, a, r *denotes phrase states, accent states, rest states, respectively.*

only to the condition of the original Fujisaki model described in Section 2.2, however, the model sometimes generates parameters that are invalid in terms of linguistic structure. Statistical vocabulary model has been proposed to improve this naive restriction by incorporating linguistic structure to the HMM topologies [10]. In this model, the state transition topology of HMM was changed as illustrated in Figure 2. This model handles the sequence of the Fujisaki model commands as connections of "template", a left-to-right HMM that consists of one phrase state and one or two following accent states. Compared to the original topology, longer-scale unit of pronunciation is directly modeled, which is linguistically more reasonable. Phrase structures are extracted from multiple sentences in the vocabulary model, while model parameters are estimated independently sentence by sentence in the original model.

2.4. Simultaneous Generative Model of Multiple F_0 Contours

In statistical mapping such as voice conversion utilizing Gaussian mixture models (GMM), conversion from input features to output ones is achieved by constructing a joint generative model of these two features. Similarly to this approach, a generative model that simultaneously generates two different F_0 contours and its application to the conversion of F_0 contours has been proposed [9]. In this generative model, a pair of parallel F_0 contours $(\boldsymbol{y}^{(A)}, \boldsymbol{y}^{(B)})$ from different two speakers or speaking styles are treated simultaneously in the same topology of HMM as that illustrated in Figure 1, while the output distribution is modeled by joint Gaussian. Namely, the probability density function of two speakers' phrase and accent command $\boldsymbol{u}^{(A)}[k] = (\boldsymbol{u}_{\mathrm{p}}^{(A)}[k], \boldsymbol{u}_{\mathrm{a}}^{(A)}[k])^\top$, $\boldsymbol{u}^{(B)}[k] = (\boldsymbol{u}_{\mathrm{p}}^{(B)}[k], \boldsymbol{u}_{\mathrm{a}}^{(B)}[k])^\top$ are formulated as

$$P\left(\boldsymbol{u}^{(A)}[k] \,\middle|\, s_k\right) = \mathcal{N}\left(\boldsymbol{u}^{(A)}[k]; \boldsymbol{\mu}[k]^{(A)}, \boldsymbol{\Sigma}[k]^{(A)}\right), \quad (12)$$

$$P\left(\boldsymbol{u}^{(B)}[k] \,\middle|\, s_k\right) = \mathcal{N}\left(\boldsymbol{u}^{(B)}[k]; \boldsymbol{\mu}[k]^{(B)}, \boldsymbol{\Sigma}[k]^{(B)}\right). \quad (13)$$

Note that s_k is shared between speaker A and B, while the mean vector $\boldsymbol{\mu}[k]$ and covariance matrix $\boldsymbol{\Sigma}[k]$ are treated separately. Using this model, conversion of an F_0 contour $\boldsymbol{y}^{(A)}$ to that of a different speaker or a speaking style $\boldsymbol{y}^{(B)}$ can be achieved by maximizing the conditional likelihood described in [9].

3. Bunsetsu-Incorporated Simultaneous Generative Model of Multiple F_0 Contours

3.1. Bunsetsu-Incorporated Statistical Vocabulary Model

This section describes a stochastic generative model of F_0 contours that explicitly incorporates a phrase structure of Japanese, which is called bunsetsu. In an ideal Fujisaki model, model parameters and linguistic information completely correspond to each other. However, once the parameters are estimated from observed F_0 contours, this correspondence is not always satisfied. Although the stochastic model introduced in Section 2.2 is a powerful scheme to estimate the model parameters, it still has difficulty in meeting this correspondence well. Ideally speaking, generative models for F_0 contours should balance the flexibility of the stochastic model and the transparency of linguistic information in the forward process of the Fujisaki model.

The statistical vocabulary model in Section 2.3 tries to incorporate linguistic structures in the stochastic framework. In that model, a sequence of the HMM states called template is introduced to the topologies of HMM as a representation of linguistic structures. Although this model tries to utilize linguistic information, its ability to link model parameters to linguistic information is still limited because there is no explicit restriction from the linguistic structures.

The proposed model aims to achieve the correspondence between phrase commands and phrase structures, by explicitly utilizing bunsetsu information. Bunsetsu information becomes seeds of the templates of HMM topologies. In this model, the template and bunsetsu are in one-to-one correspondence and phrase commands are allocated only to the beginning of each template (each bunsetsu). For simplification, phrase state is always allocated in every beginning of bunsetsu. To represent a bunsetsu that has no phrase command, some templates that have a phrase command of 0 magnitude are prepared.

3.2. Simultaneous Generative Model of Multiple F_0 Contours for Focus Control

This section proposes a generative model of a pair of F_0 contours that can be utilized for focus control. Focus control is important to convey information appropriately, especially when a part of utterance should be stressed, and it can be achieved by controlling F_0 contours [7]. In order to achieve a conversion of a F_0 contour, such as neutral to focused, a joint generative model is necessary, as described in Section 2.4. In the simultaneous generative model of multiple F_0 contours, a pair of parallel F_0 contours were modeled simultaneously and it enabled conversion of F_0 contours into another speaker or speaking styles.

By integrating this conversion method with the proposed model introduced in Section 3.1, control of F_0 contours in every bunsetsu unit can be achieved, and focus control can be regarded as an expansion of this bunsetsu-by-bunsetsu control. The proposed model targets F_0 conversion from an utterance without any focus (neutral utterance) to an utterance with a focus in one of the bunsetsu (focused utterance).

Similarly to Section 2.4, we treat a pair of commands $\boldsymbol{u}[k] = (\boldsymbol{u}^{\mathrm{N}}[k], \boldsymbol{u}^{\mathrm{F}}[k])$ as a model parameter, where $\boldsymbol{u}^{\mathrm{N}}[k] = (\boldsymbol{u}_{\mathrm{p}}^{\mathrm{N}}[k], \boldsymbol{u}_{\mathrm{a}}^{\mathrm{N}}[k])^\top$ are the phrase and accent commands of a neutral utterance, and $\boldsymbol{u}^{\mathrm{F}}[k] = (\boldsymbol{u}_{\mathrm{p}}^{\mathrm{F}}[k], \boldsymbol{u}_{\mathrm{a}}^{\mathrm{F}}[k])^\top$ are those of a focused utterance. $\boldsymbol{u}^{\mathrm{N}}[k]$ and $\boldsymbol{u}^{\mathrm{F}}[k]$ are assumed to be normally

Table 2: *The parameters to define the HMM of the proposed method.*

Output sequence:	$\{\boldsymbol{u}[k]\}_{k=1}^{K}$
State sequence:	$\{s_k\}_{k=1}^{K}$
Output distribution:	$P\left(\boldsymbol{u}^i[k] \mid s_k\right) = \mathcal{N}\left(\boldsymbol{u}^i[k]; \boldsymbol{\mu}[k]^i, \boldsymbol{\Sigma}[k]^i\right)$
	$\boldsymbol{\mu}[k]^i = \begin{cases} (0,0)^\top & (s_k \in \mathrm{r}_0, \mathrm{r}_1) \\ (A_{\mathrm{p}}^{i(m_{\mathrm{p}})}, 0)^\top & (s_k \in \mathrm{p}_{m_{\mathrm{p}}}) \\ (0, A_{\mathrm{a}}^{i(m_{\mathrm{a}})})^\top & (s_k \in \mathrm{a}_{m_{\mathrm{a}}}) \end{cases}$
Mean sequence:	$\{\boldsymbol{\mu}[k] = \boldsymbol{\mu}[k]\}_{k=1}^{K}$
Transition probability:	$\phi_{i',i} = P\left(s_k = i \mid s_{k-1} = i'\right)$

distributed:

$$\boldsymbol{u}^i[k] \sim \mathcal{N}\left(\boldsymbol{u}^i[k]; \boldsymbol{\mu}[k]^i, \boldsymbol{\Sigma}[k]^i\right), \tag{14}$$

where $i \in \{\mathrm{N}, \mathrm{F}\}$ and $\boldsymbol{\mu}[k]^i, \boldsymbol{\Sigma}[k]^i$ denote the mean vector and the covariance matrix of $\boldsymbol{u}^i[k]$. The mean vector of the HMM output is $\boldsymbol{\mu}[k] = (\boldsymbol{\mu}^{\mathrm{N}}[k], \boldsymbol{\mu}^{\mathrm{F}}[k])$, where $\boldsymbol{\mu}^i[k] = (\mu_{\mathrm{p}}^i[k], \mu_{\mathrm{a}}^i[k])^\top$.

For focus control, it is convenient to directly treat the difference of commands' magnitude of neutral utterances and focused utterances [7]. Therefore, we regard the difference $\boldsymbol{\mu}^{\mathrm{d}}[k] = \boldsymbol{\mu}^{\mathrm{F}}[k] - \boldsymbol{\mu}^{\mathrm{N}}[k]$ as an estimated model parameter instead of $\boldsymbol{\mu}^{\mathrm{F}}[k]$. In order to model the different properties in bunsetsu with/without focus (d0, d1), we separately treat these properties, and describe $\boldsymbol{\mu}^{\mathrm{F}}[k]$ as

$$\boldsymbol{\mu}^{\mathrm{F}}[k] = \begin{cases} \boldsymbol{\mu}^{\mathrm{N}}[k] + \boldsymbol{\mu}^{\mathrm{d0}}[k] & (k \in \textit{focus}) \\ \boldsymbol{\mu}^{\mathrm{N}}[k] + \boldsymbol{\mu}^{\mathrm{d1}}[k] & (k \notin \textit{focus}) \end{cases}, \tag{15}$$

where *focus* denotes the group of time that corresponds to focused bunsetsu in phone alignment. Consequently, the proposed HMM can be defined by the parameters shown in Table 2.

3.3. Process of Parameter Optimization

Here, we derive an algorithm for training the model parameters $\boldsymbol{u}$ and $\boldsymbol{\theta} = \{A_j^{\mathrm{N}(m_j)}, A_j^{\mathrm{d0}(m_j)}, A_j^{\mathrm{d1}(m_j)}\}_{m_j=1}^{M_j}$, where $j \in \{\mathrm{p}, \mathrm{a}\}$ and M_j denotes the number of state j, by locally maximizing $P\left(\boldsymbol{u}, \boldsymbol{\theta} \mid \boldsymbol{y}\right)$, utilizing Expectation-Maximization (EM) algorithm. By introducing $\boldsymbol{s}$ as a latent variable, $P\left(\boldsymbol{u}, \boldsymbol{\theta} \mid \boldsymbol{y}\right)$ can be written as

$$\begin{aligned} P\left(\boldsymbol{u}, \boldsymbol{\theta} \mid \boldsymbol{y}\right) &= \sum_{\boldsymbol{s}} P\left(\boldsymbol{u}, \boldsymbol{s}, \boldsymbol{\theta} \mid \boldsymbol{y}\right) \\ &\overset{c}{=} \sum_{\boldsymbol{s}} P\left(\boldsymbol{y} \mid \boldsymbol{u}\right) P\left(\boldsymbol{u} \mid \boldsymbol{s}, \boldsymbol{\theta}\right) P(\boldsymbol{s}), \end{aligned} \tag{16}$$

where $\overset{c}{=}$ denotes equal except constant values. (16) can be maximized by introducing the following Q-function:

$$\begin{aligned} &Q(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{u}', \boldsymbol{\theta}') = \sum_{\boldsymbol{s}} P\left(\boldsymbol{s} \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) \log P\left(\boldsymbol{u}, \boldsymbol{s}, \boldsymbol{\theta} \mid \boldsymbol{y}\right) \\ &\overset{c}{=} \log P\left(\boldsymbol{y} \mid \boldsymbol{u}\right) + \sum_{\boldsymbol{s}} P\left(\boldsymbol{s} \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) \log P\left(\boldsymbol{u} \mid \boldsymbol{s}, \boldsymbol{\theta}\right) P(\boldsymbol{s}) \\ &= \log P\left(\boldsymbol{y}^{\mathrm{N}} \;\middle|\; \boldsymbol{u}^{\mathrm{N}}\right) + \log P\left(\boldsymbol{y}^{\mathrm{F}} \;\middle|\; \boldsymbol{u}^{\mathrm{F}}\right) \\ &+ \sum_{\boldsymbol{s}} P\left(\boldsymbol{s} \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) \left(\log P\left(\boldsymbol{u}^{\mathrm{N}} \;\middle|\; \boldsymbol{s}, \boldsymbol{\theta}\right) P\left(\boldsymbol{u}^{\mathrm{F}} \;\middle|\; \boldsymbol{s}, \boldsymbol{\theta}\right) P(\boldsymbol{s})\right). \end{aligned} \tag{17}$$

The $\boldsymbol{u}$ and $\boldsymbol{\theta}$ that locally maximize $P\left(\boldsymbol{u}, \boldsymbol{\theta} \mid \boldsymbol{y}\right)$ are obtained, by calculating $P\left(\boldsymbol{s} \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right)$ (E-step) and increasing $Q(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{u}', \boldsymbol{\theta}')$ as to $\boldsymbol{u}, \boldsymbol{\theta}$ (M-step) alternately. In the E-step, the occupation count $P\left(s_k = t \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right)$ are calculated using Viterbi approximation. Viterbi path is calculated only using the neutral utterance for two reasons. First, considering the fact that $\boldsymbol{s}$ is estimated only by the neutral utterance when an F_0 contour is converted, it is more natural to obtain $\boldsymbol{s}$ from the neutral utterance when training. Second, this causes $\boldsymbol{u}^{\mathrm{N}}$ to take a closer value to $\boldsymbol{\mu}^{\mathrm{N}}$ than using $\boldsymbol{u}^{\mathrm{F}}$ in addition, and stabilizes the value of $\boldsymbol{\theta}$ and $\boldsymbol{u}$. Viterbi approximation is used instead of Forward-Backward algorithm, because it is convenient to meet the constraint that phrase state is always allocated in every beginning of bunsetsu. In the M-step, we can derive the lower bound function of $Q(\boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{u}', \boldsymbol{\theta}')$ by using the Jensen's inequality. Then, by differentiating the lower bound function by $u_j^i[l]$, the estimated values are obtained as

$$u_j^i[l] = \frac{\sum_t \frac{P\left(s_l = t \middle| \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) \mu_{j,t}^i[l]}{v_{j,t}^2} + \sum_{k=l}^{K} \frac{y^i[k] G_j[k-l]}{v_{\mathrm{n}}^i[k]^2}}{\sum_t \frac{P(s_l = t \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}')}{v_{j,t}^2} + \sum_{k=l}^{K} \frac{G_j[k-l]^2}{v_{\mathrm{n}}^i[k]^2 \lambda_{j,k,l}^j}}. \tag{18}$$

Next, $\boldsymbol{\theta} = \{A_j^{\mathrm{N}(m_j)}, A_j^{\mathrm{d0}(m_j)}, A_j^{\mathrm{d1}(m_j)}\}_{m_j=1}^{M_j}$ is updated by

$$A_j^{\mathrm{N}(m_j)} = \frac{\sum_{k=1}^{K} \sum_{t \in j_{m_j}} \frac{P\left(s_k = t \middle| \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) u_j^{\mathrm{N}}[k]}{v_{j,t}^2}}{\sum_{k=1}^{K} \sum_{t \in j_{m_j}} \frac{P(s_k = t \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}')}{v_{j,t}^2}}, \tag{19}$$

$$A_j^{\mathrm{d0}(m_j)} = \frac{\sum_{k \in \textit{focus}} \sum_{t \in j_{m_j}} \frac{P\left(s_k = t \middle| \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) (u_j^{\mathrm{F}}[k] - \mu_{j,t}^{\mathrm{N}})}{v_{j,t}^2}}{\sum_{k \in \textit{focus}} \sum_{t \in j_{m_j}} \frac{P(s_k = t \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}')}{v_{j,t}^2}}, \tag{20}$$

$$A_j^{\mathrm{d1}(m_j)} = \frac{\sum_{k \notin \textit{focus}} \sum_{t \in j_{m_j}} \frac{P\left(s_k = t \middle| \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}'\right) (u_j^{\mathrm{F}}[k] - \mu_{j,t}^{\mathrm{N}})}{v_{j,t}^2}}{\sum_{k \notin \textit{focus}} \sum_{t \in j_{m_j}} \frac{P(s_k = t \mid \boldsymbol{y}, \boldsymbol{u}', \boldsymbol{\theta}')}{v_{j,t}^2}}. \tag{21}$$

3.4. Process of Training Statistical Vocabulary Model

The HMM topologies of statistical vocabulary model are complicated as shown in Figure 2. Hence, it is essential to stabilize the training. In the proposed model, a coarse-to-fine strategy was adopted. Namely, this model achieves stable training process by gradually increasing the number of templates. The detailed process is as follows:

1. Initially, two templates (A, B) were prepared. Both of them consist of one phrase state, one rest state and one accent state. In A, A_{p} was fixed at 0.
2. Template C was added based on template B. C has one phrase and one rest state which were copied from B, and two accent states both of which were copied from B.
3. All existing templates were duplicated and two identical A_{p} were replaced by $A_{\mathrm{p}} - \delta_{\mathrm{p}}$ and $A_{\mathrm{p}} + \delta_{\mathrm{p}}$, respectively.
4. All existing templates were duplicated and two identical A_{a} were replaced by $A_{\mathrm{a}} - \delta_{\mathrm{a}}$ and $A_{\mathrm{a}} + \delta_{\mathrm{a}}$, respectively.
5. 3. and 4. were executed alternately.

HMM was retrained step by step in the above process.

3.5. Process of Focus Control

After training the HMM through the process proposed in Section 3.3 and Section 3.4, the conversion of F_0 contours are executed as follows:

$$\begin{aligned}\hat{\boldsymbol{y}}^{\mathrm{F}} =& \arg\max_{\boldsymbol{y}^{\mathrm{F}}} P\left(\boldsymbol{y}^{\mathrm{F}} \,\middle|\, \boldsymbol{y}^{\mathrm{N}}, \boldsymbol{\theta}\right) \\ =& \arg\max_{\boldsymbol{y}^{\mathrm{F}}} \int \sum_{\boldsymbol{s}} P\left(\boldsymbol{y}^{\mathrm{F}} \,\middle|\, \boldsymbol{u}^{\mathrm{F}}, \boldsymbol{\theta}\right) P\left(\boldsymbol{u}^{\mathrm{F}} \,\middle|\, \boldsymbol{s}, \boldsymbol{\theta}\right) \\ & P\left(\boldsymbol{s} \,\middle|\, \boldsymbol{u}^{\mathrm{N}}, \boldsymbol{\theta}\right) P\left(\boldsymbol{u}^{\mathrm{N}} \,\middle|\, \boldsymbol{y}^{\mathrm{N}}, \boldsymbol{\theta}\right) d\boldsymbol{u}^{\mathrm{N}} d\boldsymbol{u}^{\mathrm{F}}. \end{aligned} \tag{22}$$

For simplification, the above equation can be approximated by replacing the integral of $\boldsymbol{u}$ and the summation of $\boldsymbol{s}$ with the maximum values:

$$\begin{aligned}\hat{\boldsymbol{y}}^{\mathrm{F}} =& \arg\max_{\boldsymbol{y}^{\mathrm{F}}} P\left(\boldsymbol{y}^{\mathrm{F}} \,\middle|\, \hat{\boldsymbol{u}}^{\mathrm{F}}, \boldsymbol{\theta}\right) P\left(\hat{\boldsymbol{u}}^{\mathrm{F}} \,\middle|\, \hat{\boldsymbol{s}}, \boldsymbol{\theta}\right) \\ & P\left(\hat{\boldsymbol{s}} \,\middle|\, \hat{\boldsymbol{u}}^{\mathrm{N}}, \boldsymbol{\theta}\right) P\left(\hat{\boldsymbol{u}}^{\mathrm{N}} \,\middle|\, \boldsymbol{y}^{\mathrm{N}}, \boldsymbol{\theta}\right), \end{aligned} \tag{23}$$

s.t.

$$\begin{aligned}\hat{\boldsymbol{u}}^{\mathrm{N}} &= \arg\max_{\boldsymbol{u}^{\mathrm{N}}} P\left(\boldsymbol{u}^{\mathrm{N}} \,\middle|\, \boldsymbol{y}^{\mathrm{N}}, \boldsymbol{\theta}\right), \\ \hat{\boldsymbol{s}} &= \arg\max_{\boldsymbol{s}} P\left(\boldsymbol{s} \,\middle|\, \hat{\boldsymbol{u}}^{\mathrm{N}}, \boldsymbol{\theta}\right), \\ \hat{\boldsymbol{u}}^{\mathrm{F}} &= \arg\max_{\boldsymbol{u}^{\mathrm{F}}} P\left(\boldsymbol{u}^{\mathrm{F}} \,\middle|\, \hat{\boldsymbol{s}}, \boldsymbol{\theta}\right). \end{aligned}$$

(23) is equivalent to the maximization about $\boldsymbol{u}^{\mathrm{N}}, \boldsymbol{s}, \boldsymbol{u}^{\mathrm{F}}, \boldsymbol{y}^{\mathrm{F}}$ in order. The best $\hat{\boldsymbol{u}}^{\mathrm{N}}$ is calculated by (18), running the EM algorithm with $\boldsymbol{\theta}$ fixed. Then, $\hat{\boldsymbol{s}}$ is calculated using $\hat{\boldsymbol{u}}^{\mathrm{N}}$ and Viterbi algorithm. $\hat{\boldsymbol{u}}^{\mathrm{F}}$ is the mean sequence from $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{y}}^{\mathrm{F}}$ can be obtained by $\hat{\boldsymbol{u}}^{\mathrm{F}}$ and (7)-(9).

4. Experiments of Focus Control

4.1. Experimental Conditions

In order to evaluate the performance of the proposed method, an experiment of focus control was carried out. As materials, utterances from a female narrator were recorded. From 503 phonetically balanced sentences in ATR Japanese speech database [11], 50 sentences in subset A (a01-50) was selected. Since each sentence includes multiple bunsetsu, utterances in neutral style and with a focus on a specific bunsetsu were uttered by the narrator. Finally, 171 utterances (50 utterances are in neutral style, and 121 utterances include a focus) were utilized. A parallel data between a neutral utterance and a focused utterance were derived by dynamic time warping (DTW) of Mel-cepstra extracted by SPTK [1]. The parallel utterances were divided into 10 groups, A (a01-a05), B (a06-a10), $\cdots$, J (a46-a50). Four groups (A, B, C, and D) were selected for evaluation. To generate the test utterances, one group was selected and the remaining groups were used for training the HMM models for the selected group. Note that each selected group was not included in the groups for training. F_0 contours were extracted using the method in [12]. The initial values of $\boldsymbol{u}$ were set at the values obtained by Narusawa's method [13], shifting the position of phrase command to the nearest start time of bunsetsu, which is obtained by forced alignment with phoneme sequence. The constant parameters were fixed at $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $v_{\mathrm{p}}^2 = 0.3^2, v_{\mathrm{a}}^2 = 0.1^2, v_{\mathrm{n}}^2[k] = 10^{15}$ for unvoiced regions and

[1] http://sp-tk.sourceforge.net

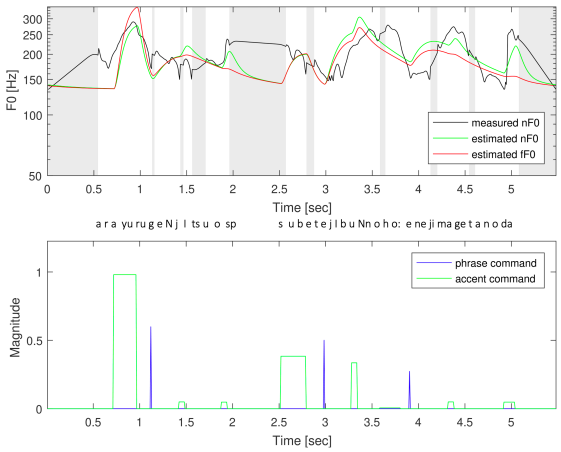

Figure 3: *The above figure shows the result of F_0 conversion using a Japanese sentence, which is "*arayuru genjitsuwo subete jibunno hoe nejimagetanoda*" in Japanese. The natural neutral F_0 contour "measured nF_0" is approximated by "estimated nF_0" and then converted to "estimated fF_0" in which the first bunsetsu "*arayuru*" is focused. The gray scale range represents unvoiced regions. The below figure shows the phrase and accent commands of "estimated fF_0".*

$v_{\mathrm{n}}^2[k] = 0.03^2$ for voiced regions. x_{b} was set at the minimum $\log F_0$ of voiced regions of neutral utterances. $\delta_{\mathrm{p}}, \delta_{\mathrm{a}}$ were set at 0.1 times $v_{\mathrm{p}}, v_{\mathrm{a}}$, respectively. The number of the iterations in each step in the process in Section 3.4 was set at 5.

We converted F_0 contours of neutral utterances to those of focused utterances by the constructed HMM. When we conduct the conversion, EM algorithm was run 3 times, fixing the values of $\boldsymbol{\theta}$. The utterances were generated by WORLD vocoder [14] with the converted F_0 contours and the spectral envelopes extracted from the input neutral utterances.

We also carried out an evaluation of the naturalness and the focused position of each voice, to subjectively evaluate the synthesized voices. The voices for evaluation consisted of natural or synthesized and neutral or focused utterances. Subjects were 9 native Japanese speakers. Each subject listened to the voices in random order and evaluated each voice. The naturalness of voices was evaluated by mean opinion score (MOS) of 5 point scale. For evaluation of the performance of focus control, the subjects were asked to point out which the focused bunsetsu is in target utterances. A choice that there is no focus in the utterance is permitted. The accuracy rate was calculated as the comprehensibility of the focused position.

4.2. Experimental Results

Figure 3 shows an example of the result of the converted F_0 contour. From this figure, it can be said that estimated focused F_0 is higher than estimated neutral F_0 in the focused bunsetsu, "*arayuru*" in Japanese, while it is slightly lower in other bunsetsu. These characteristics were observed in most of the converted F_0 contours. This fact demonstrates that the proposed model can achieve focus control in every bunsetsu unit.

Table 3 shows the results of the naturalness in MOS. In the table, the naturalness of synthesized speech is inferior to that of natural speech. It can also be said that focused utterances are less natural than neutral utterances in synthesized speech. This may be caused by the method of conversion discussed in Section

Table 3: *Naturalness of natural speech and synthesized speech with 95% confidence intervals.*

Utterance Type	Natural Speech	Synthesized Speech
Neutral	4.407 ± 0.082	3.117 ± 0.191
Focus	4.127 ± 0.093	2.756 ± 0.074

Table 4: *Accuracy rate of focused position for focused utterance. p_c, p_n and p_i represent the rate of selecting correct focused position, no focused position, incorrect focused position, respectively.*

Pattern	Natural Speech	Synthesized Speech
p_c	0.59	0.44
p_n	0.18	0.22
p_i	0.23	0.34

3.5. In Section 3.5, $\hat{\boldsymbol{u}}^{\mathrm{N}}$ is obtained regarding to $P(\boldsymbol{u}^{\mathrm{N}}|\boldsymbol{y}^{\mathrm{N}})$, while $\hat{\boldsymbol{u}}^{F}$ is calculated as the mean vector from $\hat{\boldsymbol{s}}$, which is the summation of $\boldsymbol{\mu}^{\mathrm{N}}$ and $\boldsymbol{\mu}^{\mathrm{d}}$. Since there is one more process to obtain $\hat{\boldsymbol{u}}^{\mathrm{F}}$, it gets more unstable than $\hat{\boldsymbol{u}}^{\mathrm{N}}$. It indicates that there is room for improvement in the naturalness of synthesized speech.

Table 4 shows the accuracy rate of focused position for focused utterances. In the table, p_c, p_n and p_i represent the rate of selecting correct focused position, no focused position, incorrect focused position, respectively. Although it is expected that p_c occupies almost 100 % in natural speech, it occupies only 59% of the result. This result shows that it is not an easy task for listeners to identify the focused bunsetsu in natural speech communication. In the synthesized speech, 74% $(= 0.44/0.59)$ in accuracy rate was achieved compared with natural speech. It indicates that the proposed scheme achieved reasonable conversion of prosody. From intuition, the differences of command amplitudes in focused bunsetsu were expected to be positive, while those in non-focused bunsetsu to be slightly negative values. Actually the trained model included some parameters that had the opposite trends. This shows the diversity of the representation of the focus information, and it should be investigated in further works.

5. Conclusions

This paper has proposed a stochastic generative model of F_0 contours incorporating bunsetsu, which is a phrase structure of Japanese. The proposed model combines the powerful stochastic generative model of F_0 contours with information of linguistic structure, by restricting the occurrence position of phrase commands to the every beginning of bunsetsu. This model makes it possible to handle two different F_0 contours simultaneously, conditioned by bunsetsu unit. In this paper, the model was applied to pairs of neutral and focused utterances, and enabled focus control in every bunsetsu unit. Experimental evaluations demonstrated that the proposed model can achieve reasonable focus control of F_0 contours in 74% accuracy rate of focus position compared with natural speech.

On the other hand, there are further works that should be improved in the proposed model. For example, the simultaneous generative model assumes that parallel data has the same command in the same time and only the magnitude differs. In real speech, the accent of word occasionally changes when a focus is placed on a bunsetsu. Thus, a model that incorporates the information of accent type and accent connection should be investigated. Also, integration of the proposed scheme with the frame-wise modeling such as DNN or HMM is another further work to be investigated.

6. Acknowledgements

This research and development work was supported by the Ministry of Internal Affairs and Communications.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *proceedings of ICASSP*. IEEE, 2013, pp. 7962–7966.

[4] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," 2017.

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[7] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of japanese based on the generation process model," *Proceedings of ICASSP*, 2009.

[8] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden Markov convolutive mixture model for pitch contour analysis of speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] T. Ishihara, Y. Kota, and K. Hirokazu, "Prosody conversion based on joint generative model of F0 contours," in *Proceedings of ASJ Spring Meeting*, no. 3-6-19, Mar. 2014, pp. 369–372.

[10] T. Ishihara, K. Hirokazu, Y. Kota, S. Daisuke, and S. Sagayama, "Statistical vocabulary model underlying command sequences of Fujisaki model for speech F0 contour analysis," in *Proceedings of ASJ Spring Meeting*, no. 1-7-9, Mar. 2013.

[11] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357 – 363, 1990.

[12] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and f0 contour generating process model," *IEICE Technical Report*, vol. 110, pp. 29–34, 2010.

[13] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *proceedings of ICASSP*, vol. 1. IEEE, 2002, pp. I–509.

[14] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.