Generative modeling of speech F_0 contours

Hirokazu Kameoka[1,2], *Kota Yoshizato*[1], *Tatsuma Ishihara*[1],
Yasunori Ohishi[2], *Kunio Kashino*[2], *Shigeki Sagayama*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation

{kameoka,yoshizato,ishihara,sagayama}@hil.t.u-tokyo.ac.jp,
{kameoka.hirokazu,ohishi.yasunori,kashino.kunio}@lab.ntt.co.jp

Abstract

This paper introduces our ongoing work on generative modeling of speech fundamental frequency (F_0) contours for estimating prosodic features from raw speech data. The present F_0 contour model is formulated by translating the Fujisaki model, a well-founded mathematical model representing the control mechanism of vocal fold vibration, into a probabilistic model described as a discrete-time stochastic process. The motivation behind this formulation is two fold. One is to derive a general parameter estimation framework for the Fujisaki model, allowing for the introduction of powerful statistical methods. The other is to construct an automatically trainable version of the Fujisaki model so that in future it can be used to develop a statistical speaking style conversion system or incorporated into existing text-to-speech synthesis systems to improve the naturalness and intelligibility of computer-generated speech. We also briefly introduce a generative model of F_0 contours of singing voice developed under the same spirit.

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

1. Introduction

Prosody aids the listener in interpreting an utterance by grouping words into larger information units and drawing attention to specific words. It also plays an important role in conveying various types of non-linguistic information such as the identity, intention, attitude and mood of the speaker. Since the voice fundamental frequency (F_0) contour is an important acoustic correlate of many prosodic constructs, modeling and analyzing F_0 contours can be potentially useful for many speech applications such as speech synthesis, speaker identification, speech conversion and dialogue systems, in which prosodic information plays a significant role.

An F_0 contour consists of long term pitch variations over the duration of prosodic units and short term pitch variations in accented syllables. The former usually contribute in phrasing while the latter contribute in accentuation during an utterance. These two types of pitch variations can be interpreted as the manifestations of two independent movements by the thyroid cartilage. The Fujisaki model [1] is a well-founded mathematical model that describes an F_0 contour as the sum of these two contributions. This model is known to approximate actual F_0 contours of speech surprisingly well when the model parameters are chosen appropriately, and its validity has been shown for many, typologically diverse languages. For this reason, and thanks to the intuitive association of the model parameters with the mechanical factors in the control mechanism of phonation, the Fujisaki model has been widely used with notable success to design F_0 contours for synthesizing natural speech. Since a prosodic feature in speech is predominantly characterized by the levels and timings of the phrase and accent components, one important challenge is to solve an inverse problem of estimating the Fujisaki-model parameters automatically from a raw F_0 contour.

However, this problem has been a difficult task. Several techniques have already been developed ([2–4], to name just a few), but so far with limited success due to the difficulty in searching for optimal parameters under the constraints imposed in the Fujisaki model. To overcome this difficulty, we have been concerned with translating the Fujisaki model into a probabilistic model such that one can make the best use of powerful methods in statistical estimation theory (such as expectation-maximization algorithm and Viterbi algorithm) for the parameter estimation. The other important motivation for this formulation is to construct an automatically trainable version of the Fujisaki model so that in future it can be used to develop a statistical speaking style conversion system or incorporated into existing text-to-speech synthesis systems to improve the naturalness and intelligibility of computer-generated speech. This paper introduces our ongoing work along with some new ideas on generative modeling of speech fundamental frequency (F_0) contours based on a probabilistic reformulation of the Fujisaki model [5–8]. We also briefly introduce our recent work on generative modeling of F_0 contours of singing voice [9] developed under the same spirit.

2. Original Fujisaki Model

The Fujisaki model [1], shown in Fig. 1, assumes that an F_0 contour on a logarithmic scale, $y(t)$, where t is time, is the superposition of three components: a phrase component $x_{\rm p}(t)$, an accent component $x_{\rm a}(t)$, and a base component $x_{\rm b}$:

$$y(t) = x_{\rm p}(t) + x_{\rm a}(t) + x_b. \quad (1)$$

The phrase component $x_{\rm p}(t)$ consists of the major-scale pitch variations over the duration of the prosodic units, and the accent component $x_{\rm a}(t)$ consists of the smaller-scale pitch variations in accented syllables. These two components are modeled as the outputs of second-order critically damped filters, one being excited with a command function $u_{\rm p}(t)$ consisting of Dirac deltas (phrase commands), and the other with $u_{\rm a}(t)$ consisting

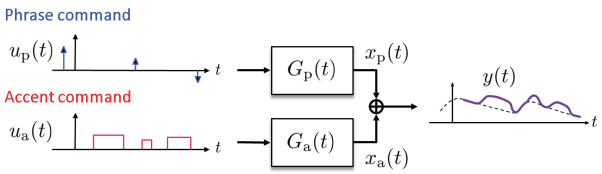

Figure 1: *Original Fujisaki model [1].*

of rectangular pulses (accent commands):

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

$$G_\mathrm{p}(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{3}$$

$$x_\mathrm{a}(t) = G_\mathrm{a}(t) * u_\mathrm{a}(t), \tag{4}$$

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

where $*$ denotes convolution over time. The baseline component x_b is a constant value related to the lower bound of the speaker's F_0, below which no regular vocal fold vibration can be maintained. α and β are natural angular frequencies of the two second-order systems, which are known to be almost constant within an utterance as well as across utterances for a particular speaker. It has been shown that $\alpha = 3$ rad/s and $\beta = 20$ rad/s can be used as default values.

3. Discretized Fujisaki model

In this section, we apply a backward difference s-to-z transform to the phrase and accent control mechanisms described as continuous-time linear systems in order to obtain a discrete-time version of the Fujisaki model [5]. The reason for the discretization will be made apparent later. The transfer function (Laplace transform) of the impulse response of the phrase control mechanism is given in the s-domain as

$$\mathcal{G}_\mathrm{p}(s) = \mathcal{L}\big[G_\mathrm{p}(t)\big] = \frac{\alpha^2}{(s+\alpha)^2}. \tag{6}$$

The backward difference transform approximates the time differential operator s by the backward difference operator in the z-domain such that $s \simeq (1 - z^{-1})/t_0$, where t_0 is the sampling period of the discrete-time representation. By undertaking this transform, the transfer function of the inverse system $\mathcal{G}_\mathrm{p}^{-1}(s)$ can be written in the z-domain as

$$\mathcal{G}_\mathrm{p}^{-1}(z) = a_2 z^{-2} + a_1 z^{-1} + a_0, \tag{7}$$

with $a_2 = (\psi - 1)^2$, $a_1 = -2\psi(\psi - 1)$, and $a_0 = \psi^2$, where $\psi = 1 + 1/(\alpha t_0)$. Let us use $u_\mathrm{p}[k]$ and $x_\mathrm{p}[k]$ to denote the discrete-time version of the phrase command function and phrase component, respectively, where k is the discrete-time index. Then, $x_\mathrm{p}[k]$ can thus be regarded as the output of a constrained all-pole system whose characteristics are governed by a single parameter ψ (or α), such that

$$u_\mathrm{p}[k] = a_0 x_\mathrm{p}[k] + a_1 x_\mathrm{p}[k-1] + a_2 x_\mathrm{p}[k-2]. \tag{8}$$

In the same way, the relationship between the accent command function $u_\mathrm{a}[k]$ and the accent component $x_\mathrm{a}[k]$ is described as

$$u_\mathrm{a}[k] = b_0 x_\mathrm{a}[k] + b_1 x_\mathrm{a}[k-1] + b_2 x_\mathrm{a}[k-2], \tag{9}$$

with $b_2 = (\varphi - 1)^2$, $b_1 = -2\varphi(\varphi - 1)$, and $b_0 = \varphi^2$, where $\varphi = 1 + 1/(\beta t_0)$. Altogether, the discrete-time version of the Fujisaki model can be expressed as the superposition of the three components: $x_\mathrm{p}[k] + x_\mathrm{a}[k] + x_\mathrm{b}$.

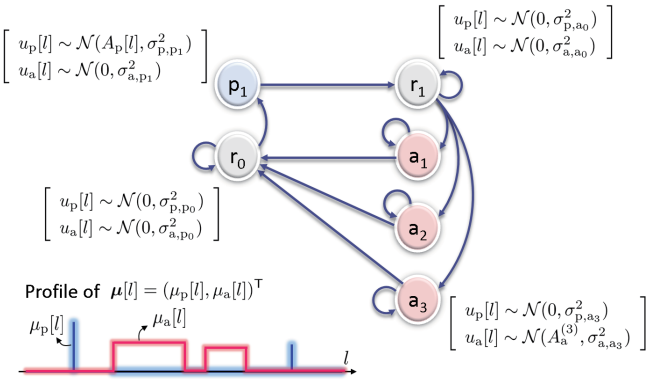

Figure 2: *Command function modeling with HMM[5]. In state* r_0*,* $\mu_\mathrm{p}[k]$ *and* $\mu_\mathrm{a}[k]$ *are both constrained to be zero. In state* p_1*,* $\mu_\mathrm{p}[k]$ *can take a non-zero value,* $A_\mathrm{p}[k]$*, whereas* $\mu_\mathrm{a}[k]$ *is still restricted to zero. In state* p_1*, no self-transitions are allowed. In state* r_1*,* $\mu_\mathrm{p}[k]$ *and* $\mu_\mathrm{a}[k]$ *become zero again. This path constraint restricts* $\mu_\mathrm{p}[k]$ *to consisting of isolated deltas. State* a_0 *leads to states* $\mathrm{a}_1, \ldots, \mathrm{a}_N$*, in each of which* $\mu_\mathrm{a}[k]$ *can take a different non-zero value* $A^{(n)}_\mathrm{a}$*, whereas* $\mu_\mathrm{p}[k]$ *is forced to be zero. Direct state transitions from state* a_n *to state* $\mathrm{a}_{n'}$ *without passing through state* r_1 *are not allowed. This constraint restricts* $\mu_\mathrm{a}[k]$ *to consisting of rectangular pulses.*

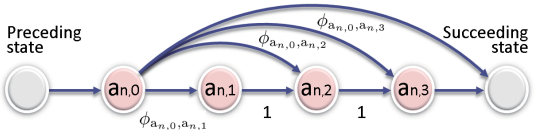

Figure 3: *Duration-explicit representation of the hidden states [6]. The splitting of state* a_n *into substates* $\mathrm{a}_{n,0}$*,* $\mathrm{a}_{n,1}$*,* $\mathrm{a}_{n,2}$*, and* $\mathrm{a}_{n,3}$ *allows us to parametrize the duration of each hidden state. For example,* $\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,1}}$ *corresponds to the probability of staying at state* a_n *with 4 consecutive times.*

4. Generative model of speech F_0 contours

Here, we model the generative process of a speech F_0 contour based on the discrete-time version of the Fujisaki model.

We first describe the process for generating the phrase and accent command functions, $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$. In the original Fujisaki model, it is required that the phrase commands must consist of Dirac deltas and the accent commands must consist of rectangular pulses. In addition, they are not allowed to overlap each other. To incorporate these requirements, we proposed in [5] to model the $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$ pair, i.e., $\boldsymbol{o}[k] = (u_\mathrm{p}[k], u_\mathrm{a}[k])^\mathsf{T}$, using a hidden Markov model (HMM) with the specific topology illustrated in Fig. 2. The output distribution of each state is a Gaussian distribution

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]\right), \tag{10}$$

$$\boldsymbol{\nu}[k] = \begin{bmatrix} \mu_\mathrm{p}[k] \\ \mu_\mathrm{a}[k] \end{bmatrix}, \quad \boldsymbol{\Upsilon}[k] = \begin{bmatrix} \upsilon^2_\mathrm{p}[k] & 0 \\ 0 & \upsilon^2_\mathrm{a}[k] \end{bmatrix}. \tag{11}$$

where the mean vector $\boldsymbol{\nu}[k]$ and covariance matrix $\boldsymbol{\Upsilon}[k]$ evolve in time as a result of the state transition. To parameterize the durations of the self transitions, we proposed in [6] to split each state into a certain number of substates such that they all have exactly the same emission densities. Fig. 3 shows an example of the splitting of state a_n. The number of substates is set at a sufficiently large value and the transition probability from substate $\mathrm{a}_{n,m}$ to substate $\mathrm{a}_{n,m+1}$ is set at 1 for $m \neq 0$. This state splitting allows us to flexibly control the durations for which the process stays in state a_n through the settings of the transition probability. The transition probability from substate $\mathrm{a}_{n,0}$ to substate

$\mathrm{a}_{n,m}$ $(m \geq 1)$ corresponds to the probability of the present HMM generating a rectangular pulse that has a particular duration. In the same way, we split states r_0 and r_1 to parameterize the probability of the spacing between phrase and accent commands. Henceforth, we use the notation $\mathrm{r}_0 = \{\mathrm{r}_{0,0}, \mathrm{r}_{0,1}, \ldots\}$, $\mathrm{r}_1 = \{\mathrm{r}_{1,0}, \mathrm{r}_{1,1}, \ldots\}$, and $\mathrm{a}_n = \{\mathrm{a}_{n,0}, \mathrm{a}_{n,1}, \ldots\}$. The HMM is defined as follows:

$$
\begin{aligned}
\text{Output sequence:}\quad & \{\boldsymbol{o}[k]\}_{k=1}^{K} \\
\text{Set of states:}\quad & \mathcal{S} = \{\mathrm{r}_0, \mathrm{p}_1, \mathrm{r}_1, \mathrm{a}_1, \ldots, \mathrm{a}_N\} \\
\text{State sequence:}\quad & \{s_k\}_{k=1}^{K} \\
\text{Output distribution:}\quad & P(\boldsymbol{o}[k]|s_k) = \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]) \\
& \boldsymbol{\nu}[k] = \begin{cases} (0,0)^{\mathsf{T}} & (s_k \in \mathrm{r}_0, \mathrm{r}_1) \\ (A_{\mathrm{p}}[k], 0)^{\mathsf{T}} & (s_k = \mathrm{p}_1) \\ (0, A_{\mathrm{a}}^{(n)})^{\mathsf{T}} & (s_k \in \mathrm{a}_n) \end{cases} \\
& \boldsymbol{\Upsilon}[k] = \begin{bmatrix} \sigma_{\mathrm{p},s_k}^2 & 0 \\ 0 & \sigma_{\mathrm{a},s_k}^2 \end{bmatrix} \\
\text{Transition probability:}\quad & \phi_{i',i} = \log P(s_k = i | s_{k-1} = i')
\end{aligned}
$$

In [8], we further proposed designing the transition network of the HMM under the hypothesis that phrase and accent command sequences are governed by a vocabulary model. Refer to [8] for more details.

Given the state sequence $\boldsymbol{s} = \{s_k\}_{k=1}^{K}$, the above HMM generates the $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ pair. From Eq. (8) and Eq. (9), $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ are then fed through different all-pole systems to generate the phrase and accent components, $x_{\mathrm{p}}[k]$ and $x_{\mathrm{a}}[k]$. It should be noted that in non-tonal languages such as standard Japanese, the phrase and accent commands must be non-negative. The treatment of the non-negativity constraint on $x_{\mathrm{p}}[k]$ and $x_{\mathrm{a}}[k]$ is discussed in [6, 7].

For real speech F_0 contours, observed F_0s should not always be considered reliable. For example, F_0 estimates obtained with a pitch extractor in unvoiced regions would be totally unreliable. When performing parameter inference, we would want to trust only reliable observations and neglect unreliable ones. To incorporate the degree of uncertainty of F_0 observations, we consider modeling an observed F_0 contour $y[k]$ as a superposition of the "ideal" F_0 contour, i.e., $x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + x_{\mathrm{b}}$, and a noise component $x_{\mathrm{n}}[k] \sim \mathcal{N}\left(0, v_{\mathrm{n}}^2[k]\right)$, where $v_{\mathrm{n}}^2[k]$ represents the degree of uncertainty of the F_0 observation at time k, which is assumed to be given. The entire F_0 contour is thus given by

$$
y[k] = x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + x_{\mathrm{b}} + x_{\mathrm{n}}[k], \tag{12}
$$

where x_{b} denotes the baseline component.

Now, let us define

$$
\begin{aligned}
&\boldsymbol{u}_{\mathrm{p}} = (u_{\mathrm{p}}[1], \ldots, u_{\mathrm{p}}[K])^{\mathsf{T}}, \quad \boldsymbol{u}_{\mathrm{a}} = (u_{\mathrm{a}}[1], \ldots, u_{\mathrm{a}}[K])^{\mathsf{T}}, \\
&\boldsymbol{\mu}_{\mathrm{p}} = (\mu_{\mathrm{p}}[1], \ldots, \mu_{\mathrm{p}}[K])^{\mathsf{T}}, \quad \boldsymbol{\mu}_{\mathrm{a}} = (\mu_{\mathrm{a}}[1], \ldots, \mu_{\mathrm{a}}[K])^{\mathsf{T}}, \\
&\boldsymbol{x}_{\mathrm{p}} = (x_{\mathrm{p}}[1], \ldots, x_{\mathrm{p}}[K])^{\mathsf{T}}, \quad \boldsymbol{x}_{\mathrm{a}} = (x_{\mathrm{a}}[1], \ldots, x_{\mathrm{a}}[K])^{\mathsf{T}}, \\
&\boldsymbol{x}_{\mathrm{n}} = (x_{\mathrm{n}}[1], \ldots, x_{\mathrm{n}}[K])^{\mathsf{T}}, \quad \boldsymbol{y} = (y[1], \cdots, y[K])^{\mathsf{T}}.
\end{aligned}
$$

Then, we can write $\boldsymbol{u}_{\mathrm{p}}$ and $\boldsymbol{u}_{\mathrm{a}}$ as

$$
\boldsymbol{u}_{\mathrm{p}} = \boldsymbol{A}\boldsymbol{x}_{\mathrm{p}}, \tag{13}
$$

$$
\boldsymbol{u}_{\mathrm{a}} = \boldsymbol{B}\boldsymbol{x}_{\mathrm{a}}, \tag{14}
$$

where

$$
\boldsymbol{A} = \begin{bmatrix} a_0 & & O \\ a_1 & a_0 & \\ a_2 & a_1 & a_0 \\ & \ddots & \ddots & \ddots \\ O & & a_2 & a_1 & a_0 \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} b_0 & & O \\ b_1 & b_0 & \\ b_2 & b_1 & a_0 \\ & \ddots & \ddots & \ddots \\ O & & b_2 & b_1 & b_0 \end{bmatrix}. \tag{15}
$$

$\boldsymbol{s}$ and $\boldsymbol{\theta} = \{\{A_{\mathrm{p}}[k]\}_{k=1}^{K}, \{A_{\mathrm{a}}^{(n)}\}_{n=1}^{N}\}$ are the free parameters to be estimated. Obviously, estimating $\boldsymbol{s}$ and $\boldsymbol{\theta}$ corresponds to estimating the command sequences, i.e., the Fujisaki model parameters. To sum up, the likelihood function of the Fujisaki model parameters $\boldsymbol{s}$ and $\boldsymbol{\theta}$ given $\boldsymbol{y}$ is given as

$$
\begin{aligned}
P(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\theta}) &= \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}, \\
\boldsymbol{\mu} &= \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}} + x_{\mathrm{b}}\mathbf{1}, \\
\boldsymbol{\Sigma} &= \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}\left(\boldsymbol{A}^{\mathsf{T}}\right)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}\left(\boldsymbol{B}^{\mathsf{T}}\right)^{-1} + \boldsymbol{\Sigma}_{\mathrm{n}},
\end{aligned} \tag{16}
$$

where

$$
\boldsymbol{\Sigma}_{\mathrm{p}} = \mathrm{diag}(v_{\mathrm{p}}^2[1], \ldots, v_{\mathrm{p}}^2[K]), \tag{17}
$$

$$
\boldsymbol{\Sigma}_{\mathrm{a}} = \mathrm{diag}(v_{\mathrm{a}}^2[1], \ldots, v_{\mathrm{a}}^2[K]), \tag{18}
$$

$$
\boldsymbol{\Sigma}_{\mathrm{n}} = \mathrm{diag}(v_{\mathrm{n}}^2[1], \ldots, v_{\mathrm{n}}^2[K]). \tag{19}
$$

$P(\boldsymbol{s})$ is given by the product of the state transition probabilities: $P(\boldsymbol{s}) = \phi_{s_1} \prod_{k=2}^{K} \phi_{s_k, s_{k-1}}$.

Readers are referred to [5–8] for detailed derivations of the parameter inference algorithms: [5, 6] describe an iterative algorithm for maximizing the posterior density $P(\boldsymbol{s}, \boldsymbol{\theta}|\boldsymbol{y}) \propto P(\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{\theta})$ with respect to $\boldsymbol{s}$ and $\boldsymbol{\theta}$. The deviced algorithm is based on an expectation-maximization (EM) algorithm consisting in performing a Viterbi algorithm in the M-step. [7] describes a parameter estimation algorithm under non-negativity constraints on the phrase and accent components.

5. Singing voice F_0 contour modeling

So far, we have focused on the modeling of F_0 contours in normal speech. Here, we briefly introduce our recent work on the generative modeling of singing voice F_0 contours.

The F_0 contours of a singing voice consist of two types of dynamic components. One is called a note component, which is influenced by the physical constraints of the vocal folds, such as overshoot, preparation and fine fluctuations [12, 13]. The other is called an expression component, which corresponds to a mixture of singer's musical expressive intentions, such as vibrato and portamento [14]. Most previous papers have reported that these dynamic components strongly affect singing-voice perception, and that the former relates to the naturalness and individuality of a singing voice while the latter relates to singing styles and skills [15]. Hence, automatic decomposition of a raw F_0 contour into these components can be potentially beneficial for many applications such as singer identification, singing skill evaluation, singing voice synthesis and singer style conversion.

Motivated by the above, we have been concerned with modifying the aforementioned F_0 contour model so as to adapt to singing voice F_0 contours. The proposed model assumes that the process of generating singing voice F_0 contours is structurally similar to the Fujisaki model (Fig. 4). The note and expression components are assumed to be the outputs of second-order linear systems driven by the note and expression commands, corresponding to the musical note sequence and the musical expressive intentions, respectively. The note component contains the note transition and overshoot, whereas the expression component contains vibrato and portamento. The F_0 contour is then modeled as a superposition of these two components. With the same strategy described in Sec. 4, we can translate this model into a probabilistic model and derive a parameter inference algorithm. For further details, refer to [9].

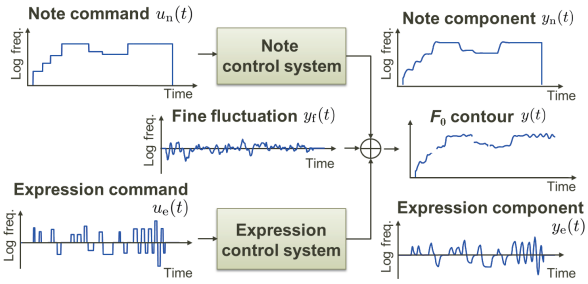

Figure 4: *The proposed model of a singing voice F_0 contour [9].*

6. Experimental evaluation

We quantitatively evaluated the parameter estimation accuracy of the present algorithm using real speech data, excerpted from the ATR Japanese speech database B-set [10]. This database consists of 503 phonetically balanced sentences. We selected speech samples of one male speaker. We used Fujisaki model parameters that had been manually annotated by an expert in the field of speech prosody as the ground truth data, where the baseline component was set at $\log 60$ Hz. F_0 contours were extracted using a method we had previously developed [11], from which the Fujisaki model parameters were estimated using the present algorithm. The number of substates in the HMM and the transition probability $\phi_{i',i}$ were determined according to the manually annotated data of the first 200 sentences. The parameter estimation algorithm was then tested on the remaining 303 sentences. We evaluated the accuracy of the parameter estimation based on the following two criteria: $\log F_0$ RMSE (root mean squared error) and detection rates. Our aim was to confirm whether the present model and algorithm can achieve high model reconstruction accuracy while keeping the meaningfulness of the model parameters. $\log F_0$ RMSE was used to evaluate the reconstruction accuracy, which measures the root mean squared error between an observed F_0 contour and the estimated F_0 contour. The detection rate was calculated by performing matching between the estimated and ground truth command sequences on a command-by-command basis by using a dynamic programming algorithm, where the time difference between the estimated and ground truth commands shorter than 0.3 seconds was considered "matched" and the local distance was set at zero (otherwise the local distance was set at 1). Let N_{E}, N_{A} be the total numbers of commands in the estimated and ground truth command sequences, N_{M} be the number of the matched commands between the two sequences, N_{Esum}, N_{Asum}, and N_{Msum} be the sum of N_{E}, N_{A}, N_{M} for all 303 sentences. We defined the insertion error rate E_{I} as $(N_{\mathrm{Esum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, the deletion error rate E_{D} as $(N_{\mathrm{Asum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, and the detection rate D as $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$.

We chose Narusawa's method [4] as a baseline method. The present method obtained a detection rate of 69.5% while the baseline method obtained 68.8%. This result confirms that our method was comparable to a state-of-the-art Fujisaki model parameter extractor in terms of the detection rate. As for the $\log F_0$ RMSE, on the other hand, the present method obtained 0.0611, while the baseline method obtained 0.1719. This result confirms that our method was superior to the conventional method in terms of the goodness-of-fit property. Fig. 5 shows an example of the estimated phrase and accent commands obtained from a raw F_0 contour.

To evaluate the pure behavior of the present parameter estimation algorithm, we also conducted a command estimation experiment using a synthetic F_0 contours. The synthetic F_0 contours were artificially created using the original Fujisaki model with the abovementioned, manually annotated command sequences. All other experimental conditions were the same as above. The present method obtained a detection rate of 83.4% while Narusawa's method only obtained 72.6%. With this experiment, the present method was shown to be significantly superior to the conventional method in terms of the detection rate of the command sequences.

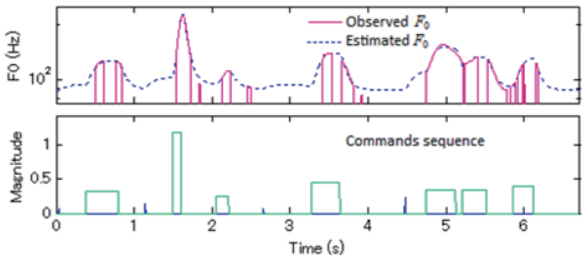

Figure 5: *An example of estimated phrase and accent commands (in blue and green, respectively), along with an observed F_0 contour in solid red line and the optimized F_0 contour model in dotted line [7].*

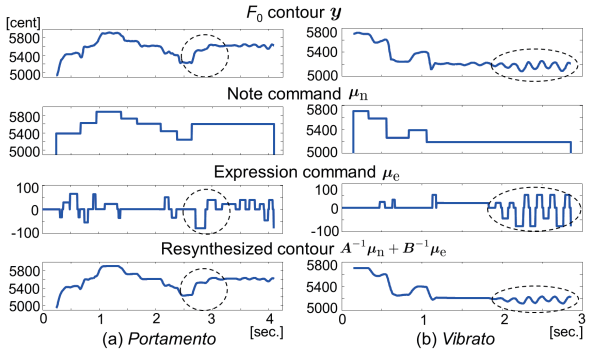

Figure 6: *Estimation of note and expression commands [9].*

Fig. 6 shows some examples of the decompositions of singing voice F_0 contours into note and expression components by the method desribed in Sec. 5. As can be seen from these examples, pitch variations related to portamento and vibrato were successfully separated from an observed F_0 contour.

7. Conclusion

This paper introduced our ongoing work on generative modeling of F_0 contours in speech and singing voice based on the Fujisaki model. One important contribution of our work is that the Fujisaki model has successfully been translated into an automatically trainable model. We believe that this will open the door to applying the present model to many speech applications such as speech synthesis, speaker identification, speech conversion, and dialogue systems, in which prosodic information plays a significant role. Future work includes incorporating the present model into the HMM-based speech synthesis system (HTS) [16] in such a way that the Fujisaki-model parameters can be learned from a speech corpus in a unified manner.

8. Acknowledgements

We thank Prof. Keikichi Hirose (The University of Tokyo) for kindly providing us with the manually annotated data associated with the ATR speech samples, Dr. Daisuke Saito (The University of Tokyo) and Dr. Daichi Mochihashi (The Institute of Statistical Mathematics) for fruitful discussions.

9. References

[1] H. Fujisaki, *In Vocal Physiology: Voice Production, Mechanisms and Functions*, Raven Press, 1988.

[2] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *J. Acoust. Soc. Jpn (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[3] H. Mixdorf, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *Proc. ICASSP*, 2000, vol. 3, pp. 1281–1284.

[4] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. ICASSP*, 2002, pp. 509–512.

[5] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F_0 contours," in *Proc. SAPA*, 2010, pp. 43–48.

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

[7] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden Markov convolutive mixture model for pitch contour analysis of speech," in *Proc. Interspeech 2012*, 2012.

[8] T. Ishihara, H. Kameoka, K. Yoshizato, D. Saito, and S. Sagayama, "Probabilistic speech F_0 contour model incorporating statistical vocabulary model of phrase-accent command sequence," submitted to Interspeech 2013.

[9] Y. Ohishi, H. Kameoka, D. Mochihashi, K. Kashino, "A stochastic model of singing voice F_0 contours for characterizing expressive dynamic components," in *Proc. Interspeech 2012*, 2012.

[10] 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. 27, pp. 187–207, 1999.

[11] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and F_0 contour generating process model," in *Tech. Rep. IEICE*, 2010, in Japanese.

[12] J. Sundberg, *The science of the singing*, Northern Illinois University Press, 1987.

[13] G. de Krom and G. Bloothooft, "Timing and accuracy of fundamental frequency changes in singing," in Proc. ICPhS95.

[14] C. E. Seashore, "A musical ornament, the vibrato," in *Psychology of Music*, 1938, pp. 33–52.

[15] T. Saitou et al., "Acoustic and perceptual effects of vocal training in amateur male singing," in Proc. EUROSPEECH 2009.

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP, 2000, pp. 1315–1318.