Generative Modeling of Voice Fundamental Frequency Contours

Hirokazu Kameoka, Kota Yoshizato, Tatsuma Ishihara, Kento Kadowaki, Yasunori Ohishi, and Kunio Kashino

***Abstract*—This paper introduces a generative model of voice fundamental frequency (F_0) contours that allows us to extract 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. There are two motivations behind this formulation. One is to derive a general parameter estimation framework for the Fujisaki model that allows the introduction of powerful statistical methods. The other is to construct an automatically trainable version of the Fujisaki model that we can incorporate into statistical-model-based text-to-speech synthesizers in such a way that the Fujisaki-model parameters can be learned from a speech corpus in a unified manner. It could also be useful for other speech applications such as emotion recognition, speaker identification, speech conversion and dialogue systems, in which prosodic information plays a significant role. We quantitatively evaluated the performance of the proposed Fujisaki model parameter extractor using real speech data. Experimental results revealed that our method was superior to a state-of-the-art Fujisaki model parameter extractor.**

***Index Terms*—Expectation-maximization algorithm, Fujisaki model, prosody, voice fundamental frequency contour.**

I. Introduction

PROSODY assists the listener to interpret 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 is 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. It is also important to note that F_0 contours indicate intonation in pitch accent languages.

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 to phrasing while the latter contribute to 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 [5], [6] is a well-founded mathematical model that describes an F_0 contour as the sum of these two contributions. This model approximates actual F_0 contours of speech fairly well when the model parameters are appropriately chosen, and its validity has been demonstrated for many typologically diverse languages [6]–[13]. Since prosodic features in speech are predominantly characterized by the levels and timings of the phrase and accent components, one important challenge is to solve the inverse problem of estimating the Fujisaki-model parameters automatically from a raw F_0 contour.

However, this problem has proved difficult to solve. Several techniques have already been developed [6], [14]–[18], but so far with limited success due to the difficulty of finding optimal parameters under the constraints imposed in the Fujisaki model. While the Fujisaki model describes a deterministic process for generating voice F_0 contours, this paper proposes formulating a stochastic counterpart of the Fujisaki model. As will be shown in the subsequent sections, this makes it possible to use powerful statistical inference techniques for estimating the underlying parameters of the Fujisaki model. Another important motivation for this formulation is to construct an automatically trainable version of the Fujisaki model that we can smoothly incorporate into text-to-speech synthesis systems or speech conversion systems so as to guarantee the naturalness of computer-generated speech.

The rest of this paper is organized as follows. Section II briefly reviews the original Fujisaki model. Section III describes a discrete-time version of the Fujisaki model, on which basis Section IV formulates a generative model of voice F_0 contours. Section V presents two iterative algorithms, which locally maximize the posterior density of the Fujisaki model parameters given an observed F_0 contour. Section VI presents experimental evaluations of the present method in terms of its ability as a Fujisaki model parameter extractor.

Manuscript received October 22, 2013; revised January 09, 2015; accepted March 13, 2015. Date of publication April 01, 2015; date of current version April 15, 2015. This work was supported by the JSPS KAKENHI under Grants 26730100 and 26280060. Earlier versions of this work were presented at SAPA 2010 [1], Speech Prosody 2012 [2], Interspeech 2012 [3], and Interspeech 2013 [4] as workshop/conference papers. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nobutaka Ono.

H. Kameoka is with the Graduate School of Information Science and Technology, The University of Tokyo, 113-8656, Japan, and also with NTT Communication Science Laboratories, NTT Corporation, 243-0198, Japan (e-mail: kameoka@hil.t.u-tokyo.ac.jp; kameoka.hirokazu@lab.ntt.co.jp).

K. Yoshizato, T. Ishihara, and K. Kadowaki are with the Graduate School of Information Science and Technology, The University of Tokyo, 113-8656, Japan (e-mail: yoshizato@hil.t.u-tokyo.ac.jp; ishihara@hil.t.u-tokyo.ac.jp; kadowaki@hil.t.u-tokyo.ac.jp).

Y. Ohishi and K. Kashino are with the NTT Communication Science Laboratories, NTT Corporation, 243-0198, Japan (e-mail: ohishi.yasunori@lab.ntt.co.jp; kashino.kunio@lab.ntt.co.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2418576

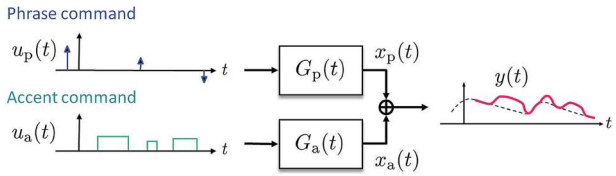


Fig. 1. Original Fujisaki model [5].

II. Original Fujisaki Model

The Fujisaki model [5], 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 value $\mu_{\rm b}$:

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

The phrase and accent components are considered to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. The phrase component $x_{\rm p}(t)$ consists of the large-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 of rectangular pulses (accent commands):

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

where $*$ denotes convolution over time and

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

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

$\mu_{\rm b}$ 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 [6], [15].

III. 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. The reason for the discretization will be made apparent later. The transfer function (the Laplace transform of the impulse response) of the phrase control mechanism is given in the s-domain as

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

The backward difference transform approximates the time differential operator s by the backward difference operator in the z-domain such that

$$s \simeq \frac{1 - z^{-1}}{t_0}, \tag{7}$$

where t_0 is the sampling period of the discrete-time representation. By undertaking this transform, the transfer function of the inverse system $\mathcal{G}_{\rm p}^{-1}(s)$ can be written in the z-domain as

$$\mathcal{G}_{\rm p}^{-1}(z) = f_{\rm p2} z^{-2} + f_{\rm p1} z^{-1} + f_{\rm p0}, \tag{8}$$

where

$$f_{\rm p2} = (\psi - 1)^2, \tag{9}$$
$$f_{\rm p1} = -2\psi(\psi - 1), \tag{10}$$
$$f_{\rm p0} = \psi^2, \tag{11}$$
$$\psi = 1 + 1/(\alpha t_0). \tag{12}$$

Let us use $u_{\rm p}[k]$ and $x_{\rm p}[k]$ to denote the discrete-time version of the phrase command function and phrase component, respectively, where k is the discrete-time index. $x_{\rm p}[k]$ can thus be regarded as the output of a constrained all-pole system whose characteristics are governed by a single parameter ψ (or α)

$$u_{\rm p}[k] = f_{\rm p0} x_{\rm p}[k] + f_{\rm p1} x_{\rm p}[k-1] + f_{\rm p2} x_{\rm p}[k-2]. \tag{13}$$

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

$$u_{\rm a}[k] = f_{\rm a0} x_{\rm a}[k] + f_{\rm a1} x_{\rm a}[k-1] + f_{\rm a2} x_{\rm a}[k-2], \tag{14}$$

where

$$f_{\rm a2} = (\varphi - 1)^2, \tag{15}$$
$$f_{\rm a1} = -2\varphi(\varphi - 1), \tag{16}$$
$$f_{\rm a0} = \varphi^2, \tag{17}$$
$$\varphi = 1 + 1/(\beta t_0). \tag{18}$$

Altogether, the discrete-time version of the Fujisaki model can be expressed as the superposition of the three components: $x_{\rm p}[k] + x_{\rm a}[k] + \mu_{\rm b}$.

IV. Generative Model of Voice F_0 Contours

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

A. Modeling Phrase and Accent Command Pair

We first describe the process for generating the phrase and accent command functions, $u_{\rm p}[k]$ and $u_{\rm a}[k]$. In the original Fujisaki model, they must satisfy the following requirements:

1) Phrase commands are a set of impulses and accent commands are a set of step-wise functions.
2) A phrase command occurs at the start of an utterance or after the offset of an accent command in the preceding phrase, and is followed by the onset of the next accent command. This means that a phrase command will not occur while an accent command is being activated.
3) The onset of an accent command is followed by its offset. This means that neighboring accent commands will not overlap each other.

According to assumption 2, $u_{\rm p}[k]$ and $u_{\rm a}[k]$ are reciprocally constrained and so they should not simply be modeled separately. One challenge as regards the estimation of the Fujisaki model parameters has been how to deal with the optimization problem under these constraints. As a convenient way of incorporating these requirements into the command functions,

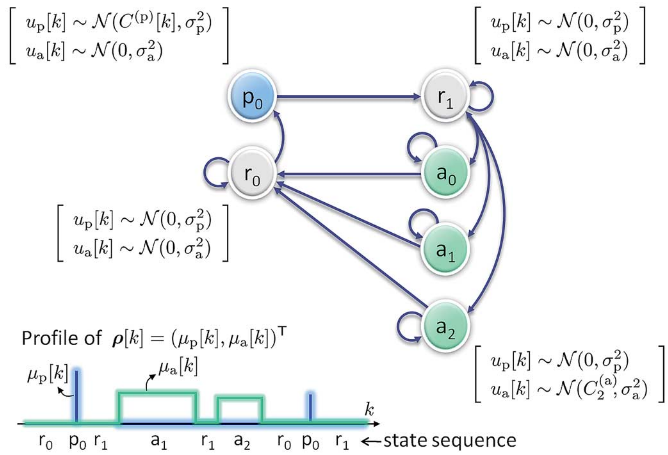


Fig. 2. Command function modeling with HMM.

we propose modeling the $u_{\rm p}[k]$ and $u_{\rm a}[k]$ pair using a hidden Markov model (HMM).

We denote the $u_{\rm p}[k]$ and $u_{\rm a}[k]$ pair by $\boldsymbol{o}[k]$ and assume that it is normally distributed:

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{\rho}[k], \boldsymbol{\Upsilon}\right), \tag{19}$$

where

$$\boldsymbol{o}[k] = \begin{bmatrix} u_{\rm p}[k] \\ u_{\rm a}[k] \end{bmatrix}, \ \boldsymbol{\rho}[k] = \begin{bmatrix} \mu_{\rm p}[k] \\ \mu_{\rm a}[k] \end{bmatrix}, \ \boldsymbol{\Upsilon} = \begin{bmatrix} \sigma_{\rm p}^2 & 0 \\ 0 & \sigma_{\rm a}^2 \end{bmatrix}.$$

Eq. (19) can be viewed as an HMM in which the output distribution of each state is a Gaussian distribution and the mean vector $\boldsymbol{\rho}[k]$ evolves over time as a result of the state transition. The mean vector $\boldsymbol{\rho}[k]$ consists of the means of the phrase and accent command functions, $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$. The use of an HMM allows us to incorporate assumptions 1–3 into $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ by constraining the path of the state transitions as illustrated in Fig. 2.

The present HMM consists of $N+3$ distinct states, $\rm r_0$, $\rm p_0$, $\rm r_1$ and $\rm a_0, \ldots, a_{N-1}$. N is the number of possible values that the magnitude of each accent command can take. It can thus be understood as the resolution of magnitude "quantization": the larger it becomes, the more finely the model is able to express the accent command function. In state $\rm r_0$, $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ are both restricted to zero. In state $\rm p_0$, $\mu_{\rm p}[k]$ can take a non-zero value, $C^{(p)}[k]$, whereas $\mu_{\rm a}[k]$ is still restricted to zero. In state $\rm p_0$, no self-transitions are allowed. In state $\rm r_1$, $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ become zero again. This path constraint restricts $\mu_{\rm p}[k]$ to consisting of isolated deltas. State $\rm r_1$ leads to states $\rm a_0, \ldots, a_{N-1}$, in each of which $\mu_{\rm a}[k]$ can take a different non-zero value $C_n^{(a)}$, whereas $\mu_{\rm p}[k]$ is forced to be zero. Direct state transitions from state $\rm a_n$ to state $\rm a'_n$ $(n \neq n')$ without passing through state $\rm r_1$ are not allowed. This constraint restricts $\mu_{\rm a}[k]$ to consisting of rectangular pulses. It should also be noted that this HMM ensures that no more than one command will be active at each point in time. The use of the HMM described above for modeling the command functions has been our primary reason for translating the Fujisaki model into its discrete-time counterpart.

The state segments correspond to the timings and durations of phrase and accent commands. If the statistical distributions of the state durations can be trained a priori, they can be useful for estimating the timings of phrase and accent commands. While an ordinary HMM assumes the state durations to be geometrically distributed, it would be more convenient if we were allowed to assume arbitrary distributions. To allow arbitrary distributions, we propose splitting 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 $\rm a_n$. 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 assume arbitrary distributions over the durations for which the process stays in state $\rm a_n$ 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 $\rm r_0$ and $\rm r_1$ to parameterize the probability of the spacing between phrase and accent commands. Note that this is equivalent to the explicit-duration HMM proposed by Ferguson [19]. Alternatively, the use of a hidden semi-Markov model [20], [21] would also be appropriate for the same purpose. 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 $\mathrm{a}_n = \{\mathrm{a}_{n,0}, \mathrm{a}_{n,1}, \ldots\}$. Let $\phi_{i',i}$ be the logarithm of the transition probability from state i' and i. To sum up, the present HMM is defined as follows:

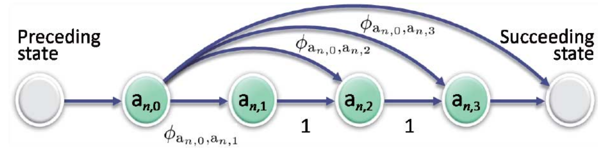


Fig. 3. A duration-explicit representation of the hidden states. Splitting state a_n into substates $\mathrm{a}_{n,0}$, $\mathrm{a}_{n,1}$, $\mathrm{a}_{n,2}$, and $\mathrm{a}_{n,3}$ allows us to parametrize the duration of each hidden state. For example, $\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,1}}$ corresponds to the probability of staying at state a_n with 4 consecutive times.

$$\begin{aligned}
\text{Output sequence: } & \{\boldsymbol{o}[k]\}_{k=0}^{K-1} \\
\text{Set of states: } & \mathcal{S} = \{\mathrm{r}_0, \mathrm{p}_0, \mathrm{r}_1, \mathrm{a}_0, \ldots, \mathrm{a}_{N-1}\} \\
\text{State sequence: } & \{s_k\}_{k=0}^{K-1} \\
\text{Output distribution: } & P(\boldsymbol{o}[k] | s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Upsilon}) \\
& \boldsymbol{c}_i[k] = \begin{cases} (0,0)^\top & (i \in \mathrm{r}_0) \\ (C^{(\rm p)}[k], 0)^\top & (i = \mathrm{p}_1) \\ (0,0)^\top & (i \in \mathrm{r}_1) \\ (0, C_n^{(\rm a)})^\top & (i \in \mathrm{a}_n) \end{cases} \\
& \boldsymbol{\Upsilon} = \begin{bmatrix} \sigma_{\rm p}^2 & 0 \\ 0 & \sigma_{\rm a}^2 \end{bmatrix} \\
\text{Transition probability: } & \phi_{i',i} = \log P(s_k = i | s_{k-1} = i')
\end{aligned}$$

The free parameters to be estimated in our command function model consist of the state sequence, $\{s_k\}_{k=0}^{K-1}$, and the mean and variance of the output distribution of each state, $\{C^{(p)}[k]\}_{k=0}^{K-1}$, $\{C_n^{(a)}\}_{n=0}^{N-1}$, $\{\sigma_{\rm p}^2, \sigma_{\rm a}^2\}$. Hereafter, we use $\boldsymbol{s}$ to denote $\{s_k\}_{k=0}^{K-1}$ and θ to denote the rest of the parameters:

$$\begin{aligned}
\boldsymbol{s} &:= \{s_k\}_{k=1}^{K}, \\
\theta &:= \{\{C^{(p)}[k]\}_{k=0}^{K-1}, \{C_n^{(a)}\}_{n=0}^{N-1}, \sigma_{\rm p}^2, \sigma_{\rm a}^2\}.
\end{aligned}$$

The generating process for the phrase and accent components is summarized as follows: First, the state sequence $\{s_k\}_{k=0}^{K-1}$ is generated according to a Markov chain. Given the state sequence $\{s_k\}_{k=0}^{K-1}$, the mean sequences $\{\mu_{\rm p}[k]\}_{k=0}^{K-1}$ and $\{\mu_{\rm a}[k]\}_{k=0}^{K-1}$ are determined by

$$\begin{bmatrix}\mu_{\rm p}[k]\\ \mu_{\rm a}[k]\end{bmatrix} = \boldsymbol{\rho}[k] = \boldsymbol{c}_{s_k}[k]. \tag{20}$$

Given $\boldsymbol{\rho}[k]$ and $\boldsymbol{\Upsilon}$ the present HMM generates the $u_{\rm p}[k]$ and $u_{\rm a}[k]$ pair according to Eq. (19). From Eq. (13) and Eq. (14), $u_{\rm p}[k]$ and $u_{\rm a}[k]$ are then fed through different all-pole systems to generate the phrase and accent components, $x_{\rm p}[k]$ and $x_{\rm a}[k]$.

B. Uncertainty of F_0 Observations

In the following, we assume that F_0 contours and voiced/unvoiced (V/UV) segments are obtained in advance by using a pitch extractor and a V/UV detector. For real speech, F_0 values 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 should 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., $x_{\rm p}[k]+x_{\rm a}[k]+\mu_{\rm b}$, and a normally distributed noise component

$$x_{\rm n}[k] \sim \mathcal{N}\left(0, \sigma_{\rm n}^2[k]\right), \tag{21}$$

where $\sigma_{\rm n}^2[k]$ represents the degree of uncertainty of the F_0 observation at time k, which is assumed to be given. For example, one simple way would be to set $\sigma_{\rm n}^2[k]$ at a small value near 0 for voiced regions and a sufficiently large value for unvoiced regions. By denoting

$$x_{\rm b}[k] = \mu_b + x_{\rm n}[k], \tag{22}$$

the entire F_0 contour is given by

$$y[k] = x_{\rm p}[k] + x_{\rm a}[k] + x_{\rm b}[k]. \tag{23}$$

C. Likelihood Function and Prior Probabilities

In this subsection, we derive the probability density function of an observed F_0 contour, $y[0], \ldots, y[K-1]$, based on the probabilistic modeling of the command functions and the reliability modeling presented in the previous subsections. From Eq. (19),

$$u_{\rm p}[k]|\theta, s_k \sim \mathcal{N}(\mu_{\rm p}[k], \sigma_{\rm p}^2), \tag{24}$$
$$u_{\rm a}[k]|\theta, s_k \sim \mathcal{N}(\mu_{\rm a}[k], \sigma_{\rm a}^2). \tag{25}$$

If we let $\boldsymbol{u}_{\rm p}$ and $\boldsymbol{u}_{\rm a}$ be

$$\boldsymbol{u}_{\rm p} = (u_{\rm p}[0], \ldots, u_{\rm p}[K-1])^\top, \tag{26}$$
$$\boldsymbol{u}_{\rm a} = (u_{\rm a}[0], \ldots, u_{\rm a}[K-1])^\top, \tag{27}$$

we can write Eqs. (24) and (25) in vector notation:

$$\boldsymbol{u}_{\rm p}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\rm p}, \boldsymbol{\Sigma}_{\rm p}), \tag{28}$$
$$\boldsymbol{u}_{\rm a}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\rm a}, \boldsymbol{\Sigma}_{\rm a}), \tag{29}$$

where

$$\boldsymbol{\mu}_{\rm p} = (\mu_{\rm p}[0], \ldots, \mu_{\rm p}[K-1])^\top, \tag{30}$$
$$\boldsymbol{\mu}_{\rm a} = (\mu_{\rm a}[0], \ldots, \mu_{\rm a}[K-1])^\top, \tag{31}$$
$$\boldsymbol{\Sigma}_{\rm p} = \sigma_{\rm p}^2 \boldsymbol{I}, \tag{32}$$
$$\boldsymbol{\Sigma}_{\rm a} = \sigma_{\rm a}^2 \boldsymbol{I}. \tag{33}$$

By using the linear equation given by Eqs. (13) and (14), the vectors consisting of the phrase and accent components

$$\boldsymbol{x}_{\rm p} = (x_{\rm p}[0], \ldots, x_{\rm p}[K-1])^\top, \tag{34}$$
$$\boldsymbol{x}_{\rm a} = (x_{\rm a}[0], \ldots, x_{\rm a}[K-1])^\top, \tag{35}$$

can be written in terms of $\boldsymbol{u}_{\rm p}$ and $\boldsymbol{u}_{\rm a}$,

$$\boldsymbol{u}_{\rm p} = \boldsymbol{F}_{\rm p}\boldsymbol{x}_{\rm p}, \tag{36}$$
$$\boldsymbol{u}_{\rm a} = \boldsymbol{F}_{\rm a}\boldsymbol{x}_{\rm a}, \tag{37}$$

where

$$\boldsymbol{F}_{\rm p} := \begin{bmatrix} f_{\rm p0} & & & & O \\ f_{\rm p1} & f_{\rm p0} & & & \\ f_{\rm p2} & f_{\rm p1} & f_{\rm p0} & & \\ & \ddots & \ddots & \ddots & \\ O & & f_{\rm p2} & f_{\rm p1} & f_{\rm p0} \end{bmatrix}, \tag{38}$$

$$\boldsymbol{F}_{\rm a} := \begin{bmatrix} f_{\rm a0} & & & & O \\ f_{\rm a1} & f_{\rm a0} & & & \\ f_{\rm a2} & f_{\rm a1} & f_{\rm a0} & & \\ & \ddots & \ddots & \ddots & \\ O & & f_{\rm a2} & f_{\rm a1} & f_{\rm a0} \end{bmatrix}. \tag{39}$$

Hence, it follows from Eqs. (28), (29), (36) and (37) that

$$\boldsymbol{x}_{\rm p}|\theta, \boldsymbol{s}, \psi \sim \mathcal{N}(\boldsymbol{F}_{\rm p}^{-1}\boldsymbol{\mu}_{\rm p}, \boldsymbol{F}_{\rm p}^{-1}\boldsymbol{\Sigma}_{\rm p}(\boldsymbol{F}_{\rm p}^{-1})^\top), \tag{40}$$
$$\boldsymbol{x}_{\rm a}|\theta, \boldsymbol{s}, \varphi \sim \mathcal{N}(\boldsymbol{F}_{\rm a}^{-1}\boldsymbol{\mu}_{\rm a}, \boldsymbol{F}_{\rm a}^{-1}\boldsymbol{\Sigma}_{\rm a}(\boldsymbol{F}_{\rm a}^{-1})^\top). \tag{41}$$

We refer to $x_{\rm b}[k]$ as the base component and let $\boldsymbol{x}_{\rm b}$ be

$$\boldsymbol{x}_{\rm b} = (x_{\rm b}[0], \ldots, x_{\rm b}[K-1])^\top. \tag{42}$$

It follows from Eqs. (21) and (22) that $\boldsymbol{x}_{\rm b}$ is normally distributed

$$\boldsymbol{x}_{\rm b}|\mu_{\rm b} \sim \mathcal{N}(\mu_{\rm b}\boldsymbol{1}, \boldsymbol{\Sigma}_{\rm b}), \tag{43}$$

where

$$\boldsymbol{1} = (1, \ldots, 1)^\top, \tag{44}$$
$$\boldsymbol{\Sigma}_{\rm b} = \mathrm{diag}(\sigma_{\rm n}^2[0], \ldots, \sigma_{\rm n}^2[K-1]). \tag{45}$$

Let us define a vector consisting of observed $F_0 s$ as

$$\boldsymbol{y} = (y[0], \ldots, y[K-1])^\top. \tag{46}$$

Hence,

$$\boldsymbol{y} = \boldsymbol{x}_{\rm p} + \boldsymbol{x}_{\rm a} + \boldsymbol{x}_{\rm b}. \tag{47}$$

We can conclude from Eqs. (47), (40), (41) and (43) that $\boldsymbol{y}$ is normally distributed such that

$$\boldsymbol{y}|\Theta \sim \mathcal{N}(\boldsymbol{F}_{\rm p}^{-1}\boldsymbol{\mu}_{\rm p} + \boldsymbol{F}_{\rm a}^{-1}\boldsymbol{\mu}_{\rm a} + \mu_{\rm b}\boldsymbol{1}, \\ \boldsymbol{F}_{\rm p}^{-1}\boldsymbol{\Sigma}_{\rm p}(\boldsymbol{F}_{\rm p}^{-1})^\top + \boldsymbol{F}_{\rm a}^{-1}\boldsymbol{\Sigma}_{\rm a}(F_{\rm a}^{-1})^\top + \boldsymbol{\Sigma}_{\rm b}), \tag{48}$$

where $\Theta := \{\theta, \boldsymbol{s}, \psi, \varphi, \mu_{\mathrm{b}}\}$. Overall, the likelihood function of the Fujisaki model parameters Θ given $\boldsymbol{y}$ can be written as

$$\begin{aligned} P(\boldsymbol{y}|\Theta) &= \frac{1}{\sqrt{(2\pi)^K|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}, \\ \boldsymbol{\mu} &= \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\boldsymbol{1}, \\ \boldsymbol{\Sigma} &= \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}(\boldsymbol{F}_{\mathrm{p}}^{\top})^{-1} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}(\boldsymbol{F}_{\mathrm{a}}^{\top})^{-1} + \boldsymbol{\Sigma}_{\mathrm{b}}. \end{aligned} \tag{49}$$

As for the prior probability of Θ, we assume that the phrase control parameter ψ, accent control parameter φ and state sequence $\{s[k]\}_{k=1}^K$ are independent of each other. Recall that we assumed in IV-A that $\{s[k]\}_{k=1}^K$ is a first-order Markov chain. We further assume that all other parameters are uniformly distributed. Thus,

$$P(\Theta) \propto P(\psi)P(\varphi)P(\boldsymbol{s}), \tag{50}$$

$$P(\boldsymbol{s}) = P(s_0)\prod_{k=1}^{K-1} P(s_k|s_{k-1}). \tag{51}$$

V. Parameter Optimization Algorithm

A. Expectation-Maximization (EM) approach

Here we describe an iterative algorithm, which locally maximizes the posterior density of Θ given $\boldsymbol{y}$, $P(\Theta|\boldsymbol{y}) \propto P(\boldsymbol{y}|\Theta)P(\Theta)$. By regarding the set consisting of the phrase, accent and base components, $\boldsymbol{x} := (\boldsymbol{x}_{\mathrm{p}}^\top, \boldsymbol{x}_{\mathrm{a}}^\top, \boldsymbol{x}_{\mathrm{b}}^\top)^\top$, as the complete data, this problem can be viewed as an incomplete data problem, which can be dealt with using the Expectation-Maximization (EM) algorithm [22], [23].

The log-likelihood of Θ given the complete data is given as

$$\begin{aligned} \log P(\boldsymbol{x}|\Theta) &\overset{\Theta}{=} \frac{1}{2}\log|\boldsymbol{\Lambda}^{-1}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^\top\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{m}), \\ \boldsymbol{x} &:= \begin{bmatrix}\boldsymbol{x}_{\mathrm{p}}\\ \boldsymbol{x}_{\mathrm{a}}\\ \boldsymbol{x}_{\mathrm{b}}\end{bmatrix}, \boldsymbol{m} := \begin{bmatrix}\boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}}\\ \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}}\\ \mu_{\mathrm{b}}\boldsymbol{1}\end{bmatrix}, \\ \boldsymbol{\Lambda}^{-1} &:= \begin{bmatrix}\boldsymbol{F}_{\mathrm{p}}^\top\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{F}_{\mathrm{a}}^\top\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\end{bmatrix}, \end{aligned} \tag{52}$$

where $\overset{\xi}{=}$ denotes equality up to a term independent of ξ. Taking the conditional expectation of Eq. (52) with respect to $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\Theta = \Theta'$, and then adding $\log P(\Theta)$ to both sides, we obtain an auxiliary function

$$\begin{aligned} Q(\Theta,\Theta') \overset{\Theta}{=} \frac{1}{2}\big[&\log|\boldsymbol{\Lambda}^{-1}| - \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta'\right]) \\ &+2\boldsymbol{m}^\top\boldsymbol{\Lambda}^{-1}\mathbb{E}\left[\boldsymbol{x}|\boldsymbol{y};\Theta'\right] - \boldsymbol{m}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\big] + \log P(\Theta). \end{aligned} \tag{53}$$

Because the relationship between the incomplete data and the complete data can be written as $\boldsymbol{y} = H\boldsymbol{x}$ where $\boldsymbol{H} := [\boldsymbol{I}, \boldsymbol{I}, \boldsymbol{I}]$, $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta]$ are given explicitly as

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^\top)^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{m}), \tag{54}$$

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^\top)^{-1}\boldsymbol{H}\boldsymbol{\Lambda} \\ + \mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]^\top. \end{aligned} \tag{55}$$

It can be shown that an iterative procedure consisting of maximizing $Q(\Theta,\Theta')$ with respect to Θ (the maximization step) and substituting Θ into Θ' (the expectation step) locally maximizes the posterior density $P(\Theta|\boldsymbol{y})$. The expectation step computes $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta\prime]$ according to Eqs. (54) and (55) by substituting the current parameter estimate into Θ'.

Now, if we partition $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ into three $K\times 1$ blocks and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta']$ into nine $K\times K$ blocks such that

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta'] = \begin{bmatrix}\bar{\boldsymbol{x}}_{\mathrm{p}}\\ \bar{\boldsymbol{x}}_{\mathrm{a}}\\ \bar{\boldsymbol{x}}_{\mathrm{b}}\end{bmatrix}, \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top|\boldsymbol{y};\Theta'] = \begin{bmatrix}\boldsymbol{R}_{\mathrm{p}} & * & *\\ * & \boldsymbol{R}_{\mathrm{a}} & *\\ * & * & \boldsymbol{R}_{\mathrm{b}}\end{bmatrix}, \tag{56}$$

where $*$ stands for blocks that we can ignore hereafter, the auxiliary function can be rewritten in a more convenient form:

$$\begin{aligned} Q(\Theta,\Theta') \overset{\Theta}{=} \frac{1}{2}[&\log|\boldsymbol{F}_{\mathrm{p}}^\top\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}| + \log|\boldsymbol{F}_{\mathrm{a}}^\top\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}| + \log|\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}| \\ &- \mathrm{tr}(\boldsymbol{F}_{\mathrm{p}}^\top\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}\boldsymbol{R}_{\mathrm{p}}) + 2\boldsymbol{\mu}_{\mathrm{p}}^\top\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}} \\ &- \mathrm{tr}(\boldsymbol{F}_{\mathrm{a}}^\top\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}\boldsymbol{R}_{\mathrm{a}}) + 2\boldsymbol{\mu}_{\mathrm{a}}^\top\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}} \\ &- \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\boldsymbol{R}_{\mathrm{b}}) + 2\mu_{\mathrm{b}}\boldsymbol{1}^\top\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\bar{\boldsymbol{x}}_{\mathrm{b}} \\ &- \boldsymbol{\mu}_{\mathrm{p}}^\top\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_{\mathrm{a}}^\top\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} - \mu_{\mathrm{b}}^2\boldsymbol{1}^\top\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\boldsymbol{1}] \\ &+ \log P(\Theta). \end{aligned} \tag{57}$$

The update formula for each parameter in the maximization step can be derived using Eq. (57).

1) **State sequence** $s_0,\ldots,s_{K-1}$: Leaving only the terms in $Q(\Theta,\Theta')$ that depend on $s := \{s_k\}_{k=0}^{K-1}$, we have

$$\begin{aligned} Q(\Theta,\Theta') \overset{s}{=} &-\frac{1}{2}\sum_{k=0}^{K-1}(\boldsymbol{o}[k]-\boldsymbol{c}_{s_k}[k])^\top\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k]-\boldsymbol{c}_{s_k}[k]) \\ &+\phi_{s_0} + \sum_{k=1}^{K-1}\phi_{s_{k-1},s_k}, \end{aligned} \tag{58}$$

where $\boldsymbol{o}[k] := ([F_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k, [F_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k)^\top$. Here the notation $[\cdot]_k$ is used to denote the k-th element of a vector. The state sequence $\{s_k\}_{k=0}^{K-1}$ maximizing $Q(\Theta,\Theta')$ can be solved efficiently using the Viterbi algorithm as follows. We first set $\delta_0(i)$ at

$$\delta_0(i) = -\frac{1}{2}(\boldsymbol{o}[0]-\boldsymbol{c}_i[0])^\top\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[0]-\boldsymbol{c}_i[0]) + \phi_i. \tag{59}$$

for all the hidden states i. If we consider state $\mathrm{r}_{0,0}$ to be the starting state, we shall set ϕ_i at

$$\phi_i = \begin{cases} 0 & (i = \mathrm{r}_{0,0}) \\ -\infty & (i \neq \mathrm{r}_{0,0}) \end{cases}. \tag{60}$$

We can compute $\delta_k(i)$ for $k = 1,\ldots,K-1$ recursively via

$$\begin{aligned} \delta_k(i) = &\max_{i'}[\delta_{k-1}(i') + \phi_{i',i}] \\ &-\frac{1}{2}(\boldsymbol{o}[k]-\boldsymbol{c}_i[k])^\top\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k]-\boldsymbol{c}_i[k]). \end{aligned} \tag{61}$$

The most likely transition for each state should be registered at each recursion $\Psi_k(i) = \arg\max_{i'}[\delta_{k-1}(i') + \phi_{i',i}]$, so that the most likely state sequence can be traced at the end of the recursion with $s_{k-1} = \Psi_k(s_k)(k = K-1,\ldots,1)$, where $s_K = \arg\max_i\delta_K(i)$. Substituting

the updated state sequence $\{s_k\}$ into Eq. (20), we finally obtain the updated $\boldsymbol{\mu}_{\mathrm{p}}$ and $\boldsymbol{\mu}_{\mathrm{a}}$.

2) **Magnitude of phrase command** $C^{(p)}[k]$: $Q(\Theta,\Theta')$ is maximized with respect to $C^{(p)}[k]$ when

$$C^{(\mathrm{p})}[k] = [\boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k (k \in \mathcal{K}_{\mathrm{p}_0}), \quad \mathcal{K}_{\mathrm{p}_0} = \{k | s_k = \mathrm{p}_0\}. \quad (62)$$

3) **Magnitude of accent command** $C_n^{(a)}$: $Q(\Theta,\Theta')$ is maximized with respect to $C_n^{(a)}$ when

$$C_n^{(\mathrm{a})} = \frac{1}{|\mathcal{K}_{\mathrm{a}_n}|} \sum_{k \in \mathcal{K}_{\mathrm{a}_n}} [\boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k, \quad \mathcal{K}_{\mathrm{a}_n} = \{k | s_k \in \mathrm{a}_n\}. \quad (63)$$

4) **Phrase control parameter** ψ: Let us assume a Gaussian prior distribution over ψ such that

$$\psi \sim \mathcal{N}(\mu_\psi, 1/\nu_\psi^2). \quad (64)$$

Leaving only the terms in $Q(\Theta,\Theta')$ that depend on ψ, we have

$$Q(\Theta,\Theta') \stackrel{\psi}{=} \log|\boldsymbol{F}_{\mathrm{p}}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{F}_{\mathrm{p}}^\top \boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}\boldsymbol{R}_{\mathrm{p}}) + \boldsymbol{\mu}_{\mathrm{p}}^\top \boldsymbol{\Sigma}_{\mathrm{p}}^{-1} F_{\mathrm{p}}\bar{x}_{\mathrm{p}} - \frac{1}{2}\nu_\psi^2(\psi-\mu_\psi)^2. \quad (65)$$

Now, let

$$U_2 := \begin{bmatrix} 1 & & & & O \\ -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ O & & 1 & -2 & 1 \end{bmatrix}, U_1 := \begin{bmatrix} 0 & & & & O \\ 2 & 0 & & & \\ -2 & 2 & 0 & & \\ & \ddots & \ddots & \ddots & \\ O & & -2 & 2 & 0 \end{bmatrix},$$

$$U_0 := \begin{bmatrix} 0 & & & & O \\ 0 & 0 & & & \\ 1 & 0 & 0 & & \\ & \ddots & \ddots & \ddots & \\ O & & 1 & 0 & 0 \end{bmatrix}, \quad (66)$$

then from Eqs. (9)–(11), F_{p} can be written as

$$\boldsymbol{F}_{\mathrm{p}} = \boldsymbol{U}_2\psi^2 + \boldsymbol{U}_1\psi + \boldsymbol{U}_0. \quad (67)$$

The partial derivative of $\mathcal{I}_2(\psi)$ (or $Q(\Theta,\Theta')$) with respect to ψ is a quartic function, equal up to a constant factor to

$$2\mathrm{tr}(U_2^\top \boldsymbol{U}_2\boldsymbol{R}_{\mathrm{p}})\psi^4 + 3\mathrm{tr}(\boldsymbol{U}_2^\top \boldsymbol{U}_1 R_{\mathrm{p}})\psi^3 + \{\mathrm{tr}((2\boldsymbol{U}_2^\top \boldsymbol{U}_0 + \boldsymbol{U}_1^\top \boldsymbol{U}_1)\boldsymbol{R}_{\mathrm{p}}) - 2\boldsymbol{\mu}_{\mathrm{p}}^\top \boldsymbol{U}_2\bar{\boldsymbol{x}}_{\mathrm{p}} + \sigma_{\mathrm{p}}^2\nu_\psi^2\}\psi^2 + \{\mathrm{tr}(\boldsymbol{U}_1^\top \boldsymbol{U}_0\boldsymbol{R}_{\mathrm{p}}) - \boldsymbol{\mu}_{\mathrm{p}}^\top \boldsymbol{U}_1\bar{\boldsymbol{x}}_{\mathrm{p}} - 2\sigma_{\mathrm{p}}^2\nu_\psi^2\mu_\psi\}\psi - 2K\sigma_{\mathrm{p}}^2, \quad (68)$$

and its roots, namely the stationary points of $Q(\Theta,\Theta')$, can be solved algebraically, from which we can find the optimal ψ.

5) **Accent control parameter** φ: Let us again assume a Gaussian prior distribution over φ such that $\varphi \sim \mathcal{N}(\mu_\varphi, 1/\nu_\varphi^2)$. As the derivation follows in exactly the same manner as above, we shall omit it.

6) **Baseline value** μ_{b}: $Q(\Theta,\Theta')$ is maximized with respect to μ_{b} when

$$\mu_{\mathrm{b}} = \frac{\mathbf{1}^\top \boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\bar{\boldsymbol{x}}_{\mathrm{b}}}{\mathbf{1}^\top \Sigma_{\mathrm{b}}^{-1}\mathbf{1}} = \frac{\sum_k [\bar{\boldsymbol{x}}_{\mathrm{b}}]_k/\sigma_{\mathrm{n}}[k]^2}{\sum_k 1/\sigma_{\mathrm{n}}[k]^2}. \quad (69)$$

7) **Variances of state emission densities** $\sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2$: $Q(\Theta,\Theta')$ is maximized with respect to $\sigma_{\mathrm{p},i}^2$ and $\sigma_{\mathrm{a},i}^2$ when

$$\sigma_{\mathrm{p}}^2 = (\mathrm{tr}(\boldsymbol{F}_{\mathrm{p}}^\top \boldsymbol{F}_{\mathrm{p}}\boldsymbol{R}_{\mathrm{p}}) - 2\boldsymbol{\mu}_{\mathrm{p}}^\top \boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}} + \boldsymbol{\mu}_{\mathrm{p}}^\top \boldsymbol{\mu}_{\mathrm{p}})/K, \quad (70)$$

$$\sigma_{\mathrm{a}}^2 = (\mathrm{tr}(\boldsymbol{F}_{\mathrm{a}}^\top \boldsymbol{F}_{\mathrm{a}}\boldsymbol{R}_{\mathrm{a}}) - 2\boldsymbol{\mu}_{\mathrm{a}}^\top \boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}} + \boldsymbol{\mu}_{\mathrm{a}}^\top \boldsymbol{\mu}_{\mathrm{a}})/K. \quad (71)$$

To summarize, we obtain the following iterative algorithm that guarantees monotonic convergence to a local maximum of the posterior density $P(\Theta|\boldsymbol{y})$:

a) (E-step) Update $\bar{\boldsymbol{x}}_{\mathrm{p}}$, $\bar{\boldsymbol{x}}_{\mathrm{a}}$, $\bar{\boldsymbol{x}}_{\mathrm{b}}$, $\boldsymbol{R}_{\mathrm{p}}$, $\boldsymbol{R}_{\mathrm{a}}$ and $\boldsymbol{R}_{\mathrm{b}}$ via Eqs. (54) and (55).

b) (M-step) Increase $Q(\Theta,\Theta')$ w.r.t. Θ through the following updates:
   a) Update $\boldsymbol{s}$ by using the Viterbi algorithm.
   b) Update θ via Eqs. (62), (63), (70) and (71).
   c) Update ψ (and φ) by solving the root of Eq. (68)
   d) Update μ_{b} via Eq. (69).

   Return to 1) until convergence.

The complexity of this algorithm is $O(|\mathcal{S}|^2 K)$, where $|\mathcal{S}|$ is the number of hidden states.

B. Parameter Inference Under Non-negativity Constraints

It has been shown that phrase and accent commands must be non-negative in many non-tonal languages such as Japanese, English, German and Spanish [6]–[9]. In V-A, we treated $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ as latent variables (i.e., parameters to be marginalized out), and did not explicitly take the non-negativity constraints on $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ into consideration. While the method described in V-A can be generally applied even for such languages as Scandinavian, Portuguese and Chinese in which phrase and accent commands can be negative [10]–[13], this subsection focuses on parameter estimation under the non-negativity constraint. To impose the non-negativity constraint explicitly, it is convenient to treat $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ as model parameters instead of latent variables.

Throughout this subsection, let us assume for simplicity that ψ and φ are constants. Now, we first describe an expanded version of the generative process of $\boldsymbol{y}$:

$$\boldsymbol{y}|\boldsymbol{u}_{\mathrm{p}}, \boldsymbol{u}_{\mathrm{a}}, \mu_{\mathrm{b}} \sim \mathcal{N}(\boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{u}_{\mathrm{p}} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{u}_{\mathrm{a}} + \mu_{\mathrm{b}}\mathbf{1}, \Sigma_{\mathrm{b}}), \quad (72)$$

$$\boldsymbol{u}_{\mathrm{p}}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{p}}, \boldsymbol{\Sigma}_{\mathrm{p}}), \quad (73)$$

$$\boldsymbol{u}_{\mathrm{a}}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{a}}, \boldsymbol{\Sigma}_{\mathrm{a}}). \quad (74)$$

Note that it can be readily verified that marginalizing out $\boldsymbol{u}_{\mathrm{p}}$ and $\boldsymbol{u}_{\mathrm{a}}$ reduces Eqs. (72)–(74) to Eq. (48). Hereafter, we use $\boldsymbol{o}$ to denote the set consisting of $\boldsymbol{u}_{\mathrm{p}}$ and $\boldsymbol{u}_{\mathrm{a}}$. Instead of $\boldsymbol{o}$, we consider treating $\boldsymbol{s}$ as a latent variable. Namely, we are concerned with maximizing the posterior density $P(\boldsymbol{o},\theta,\mu_{\mathrm{b}}|\boldsymbol{y}) \propto P(\boldsymbol{o},\theta,\mu_{\mathrm{b}},\boldsymbol{y})$. We can obtain the joint probability density $P(\boldsymbol{o},\theta,\mu_{\mathrm{b}},\boldsymbol{y})$ by marginalizing

$$P(\underbrace{\boldsymbol{s},\theta,\mu_{\mathrm{b}}}_{\Theta},\boldsymbol{o},\boldsymbol{y}) \propto \underbrace{P(\boldsymbol{y}|\boldsymbol{o},\mu_{\mathrm{b}})}_{\mathrm{Eq.}(72)} \underbrace{P(\boldsymbol{o}|\boldsymbol{s},\theta)}_{\mathrm{Eqs.}(73)\&(74)} \underbrace{P(\boldsymbol{s})}_{\mathrm{Eq.}(51)}, \quad (75)$$

with respect to $\boldsymbol{s}$. We notice that the only difference from the joint probability density $P(\Theta, \boldsymbol{y})$ that we wanted to maximize in V-A is that $\boldsymbol{s}$ is replaced with $\boldsymbol{o}$. For convenience of notation, we use $\tilde{\Theta}$ to denote the set consisting of $\boldsymbol{o}$, θ and μ_{b}. Hence, $\{\Theta, \boldsymbol{o}, \boldsymbol{y}\} = \{\tilde{\Theta}, \boldsymbol{s}, \boldsymbol{y}\}$.

Here we describe an iterative algorithm that searches for the maximum a posteriori estimates of $\tilde{\Theta}$ by locally maximizing $P(\tilde{\Theta}|\boldsymbol{y})$ given $\boldsymbol{y}$, using the generalized EM algorithm. An auxiliary function for the posterior density of interest can be written as

$$\begin{aligned} Q(\tilde{\Theta}, \tilde{\Theta}') &= \sum_{\boldsymbol{s} \in \mathcal{S} \times \cdots \times \mathcal{S}} P(\boldsymbol{s}|\boldsymbol{y}, \tilde{\Theta}') \log P(\tilde{\Theta}, \boldsymbol{y}, \boldsymbol{s}) \\ &\stackrel{\tilde{\Theta}}{=} \log P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}}) + \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y}, \tilde{\Theta}') \log P(\boldsymbol{o}|\boldsymbol{s}, \theta) \\ &\stackrel{\tilde{\Theta}}{=} \log P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}}) \\ &\quad + \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y}, \tilde{\Theta}') \sum_{k=1}^{K} \log P(\boldsymbol{o}[k]|\theta, s_k) \\ &= \log P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}}) \\ &\quad + \sum_{k} \sum_{s_k \in \mathcal{S}} P(s_k|\boldsymbol{y}, \tilde{\Theta}') \log P(\boldsymbol{o}[k]|\theta, s_k), \end{aligned} \tag{76}$$

where

$$\log P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}}) \stackrel{\tilde{\Theta}}{=} - \sum_{k} \frac{\left(y[k] - x_{\mathrm{p}}[k] - x_{\mathrm{a}}[k] - \mu_{\mathrm{b}}\right)^2}{2\sigma_{\mathrm{n}}[k]^2},$$

$$\log P(\boldsymbol{o}[k]|\theta, s_k) \stackrel{\tilde{\Theta}}{=} - \frac{\left(u_{\mathrm{p}}[k] - \mu_{\mathrm{p}}[k]\right)^2}{2\sigma_{\mathrm{p}}^2} - \frac{\left(u_{\mathrm{a}}[k] - \mu_{\mathrm{a}}[k]\right)^2}{2\sigma_{\mathrm{a}}^2},$$

and

$$\begin{aligned} x_{\mathrm{p}}[k] &= \sum_{l} G_{\mathrm{p}}[k-l] u_{\mathrm{p}}[l], \\ x_{\mathrm{a}}[k] &= \sum_{l} G_{\mathrm{a}}[k-l] u_{\mathrm{a}}[l], \\ \mu_{\mathrm{p}}[k] &= [\boldsymbol{c}_{s_k}[k]]_1, \quad \mu_{\mathrm{a}}[k] = [\boldsymbol{c}_{s_k}[k]]_2. \end{aligned}$$

Note that $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$ stand for the discrete-time versions of $G_{\mathrm{p}}(t)$ and $G_{\mathrm{a}}(t)$, respectively. It can be shown that an iterative algorithm that consists of computing $\gamma = \{\gamma_{q,k}\}_{q \in \mathcal{S}, 1 \le k \le K}$ where $\gamma_{q,k} := P(s_k = q|\boldsymbol{y}, \tilde{\Theta}')$ (via the Forward-Backward algorithm), increasing $Q(\tilde{\Theta}, \tilde{\Theta}')$ with respect to $\tilde{\Theta}$, and then substituting $\tilde{\Theta}$ into $\tilde{\Theta}'$ locally maximizes the posterior $P(\tilde{\Theta}|\boldsymbol{y})$. Here, we must ensure that increasing $Q(\tilde{\Theta}, \tilde{\Theta}')$ with respect to $\boldsymbol{o}$ is performed subject to non-negativity. This can be accomplished by invoking the idea described in [24] as follows.

By using Jensen's inequality we obtain

$$\begin{aligned} &\left(x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k]\right)^2 \\ &= \left(\sum_{l} G_{\mathrm{p}}[k-l] u_{\mathrm{p}}[l] + \sum_{l} G_{\mathrm{a}}[k-l] u_{\mathrm{a}}[l] \right)^2 \\ &= \left(\sum_{l} \lambda_{k,l}^{(\mathrm{p})} \frac{G_{\mathrm{p}}[k-l] u_{\mathrm{p}}[l]}{\lambda_{k,l}^{(\mathrm{p})}} + \sum_{l} \lambda_{k,l}^{(\mathrm{a})} \frac{G_{\mathrm{a}}[k-l] u_{\mathrm{a}}[l]}{\lambda_{k,l}^{(\mathrm{a})}} \right)^2 \\ &\le \sum_{l} \lambda_{k,l}^{(\mathrm{p})} \left(\frac{G_{\mathrm{p}}[k-l] u_{\mathrm{p}}[l]}{\lambda_{k,l}^{(\mathrm{p})}} \right)^2 + \sum_{l} \lambda_{k,l}^{(\mathrm{a})} \left(\frac{G_{\mathrm{a}}[k-l] u_{\mathrm{a}}[l]}{\lambda_{k,l}^{(\mathrm{a})}} \right)^2 \\ &= \sum_{l} \frac{G_{\mathrm{p}}[k-l]^2 u_{\mathrm{p}}[l]^2}{\lambda_{k,l}^{(\mathrm{p})}} + \sum_{l} \frac{G_{\mathrm{a}}[k-l]^2 u_{\mathrm{a}}[l]^2}{\lambda_{k,l}^{(\mathrm{a})}}, \end{aligned} \tag{77}$$

where $\lambda_{k,l}^{(p)}, \lambda_{k,l}^{(a)} \ge 0$ are auxiliary variables satisfying

$$\sum_{l} \left(\lambda_{k,l}^{(\mathrm{p})} + \lambda_{k,l}^{(\mathrm{a})} \right) = 1. \tag{78}$$

We can verify that the equality in this inequality holds when $\lambda_{k,l}^{(p)}$ and $\lambda_{k,l}^{(a)}$ are given by

$$\hat{\lambda}_{k,l}^{(\mathrm{p})} = \frac{G_{\mathrm{p}}[k-l] u_{\mathrm{p}}[l]}{\sum_{l'} G_{\mathrm{p}}[k-l'] u_{\mathrm{p}}[l'] + \sum_{l'} G_{\mathrm{a}}[k-l'] u_{\mathrm{a}}[l']}, \tag{79}$$

$$\hat{\lambda}_{k,l}^{(\mathrm{a})} = \frac{G_{\mathrm{a}}[k-l] u_{\mathrm{a}}[l]}{\sum_{l'} G_{\mathrm{p}}[k-l'] u_{\mathrm{p}}[l'] + \sum_{l'} G_{\mathrm{a}}[k-l'] u_{\mathrm{a}}[l']}. \tag{80}$$

We can use this inequality to construct a lower bound function for $Q(\tilde{\Theta}, \tilde{\Theta}')$:

$$Q(\tilde{\Theta}, \tilde{\Theta}') \ge \breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda), \tag{81}$$

where

$$\begin{aligned} &\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda) \\ &\stackrel{\Theta}{=} - \sum_{k} \frac{1}{2\sigma_{\mathrm{n}}[k]^2} \Bigg\{ (y[k] - \mu_{\mathrm{b}})^2 - 2(y[k] - \mu_{\mathrm{b}})(x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k]) \\ &\quad + \sum_{l} \left(\frac{G_{\mathrm{p}}[k-l]^2 u_{\mathrm{p}}[l]^2}{\lambda_{k,l}^{(\mathrm{p})}} + \frac{G_{\mathrm{a}}[k-l]^2 u_{\mathrm{a}}[l]^2}{\lambda_{k,l}^{(\mathrm{a})}} \right) \Bigg\} \\ &\quad + \sum_{k} \sum_{s_k} p(s_k|\boldsymbol{y}, \tilde{\Theta}') \log p(\boldsymbol{o}[k]|\theta, s_k). \end{aligned} \tag{82}$$

Note that we have used λ to denote $\{\lambda_{k,l}^{(p)}, \lambda_{k,l}^{(a)}\}_{0 \le k \le K, 0 \le l \le K}$. We can use this lower bound function to derive an update equation for each element of $\tilde{\Theta}$ that guarantees a certain increase of $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}')$.

As mentioned above, the maximization of this lower bound function with respect to λ can be achieved when it is given by Eqs. (79) and (80). With λ fixed, the maximization of $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda)$ with respect to other parameters can be achieved analytically as follows. First, $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda)$ is maximized with respect to $\boldsymbol{o}$ when

$$\hat{u}_{\mathrm{p}}[l] = \frac{\sum_{k} \frac{(y[k] - \mu_{\mathrm{b}}) G_{\mathrm{p}}[k-l]}{\sigma_{\mathrm{n}}[k]^2} + \sum_{q} \frac{\gamma_{q,l} [\boldsymbol{c}_q[l]]_1}{\sigma_{\mathrm{p}}^2}}{\sum_{k} \frac{G_{\mathrm{p}}[k-l]^2}{\sigma_{\mathrm{n}}[k]^2 \lambda_{k,l}^{(\mathrm{p})}} + \frac{1}{\sigma_{\mathrm{p}}^2}}, \tag{83}$$

$$\hat{u}_{\mathrm{a}}[l] = \frac{\sum_{k} \frac{(y[k] - \mu_{\mathrm{b}}) G_{\mathrm{a}}[k-l]}{\sigma_{\mathrm{n}}[k]^2} + \sum_{q} \frac{\gamma_{q,l} [\boldsymbol{c}_q[l]]_2}{\sigma_{\mathrm{a}}^2}}{\sum_{k} \frac{G_{\mathrm{a}}[k-l]^2}{\sigma_{\mathrm{n}}[k]^2 \lambda_{k,l}^{(\mathrm{a})}} + \frac{1}{\sigma_{\mathrm{a}}^2}}, \tag{84}$$

where

$$\gamma_{q,l} = p(s_l = q|\boldsymbol{y}, \tilde{\Theta}'). \tag{85}$$

Next, $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda)$ is maximized with respect to μ_{b} when

$$\hat{\mu}_{\mathrm{b}} = \frac{\sum_{k} (y[k] - x_{\mathrm{p}}[k] - x_{\mathrm{a}}[k]) / \sigma_{\mathrm{n}}[k]^2}{\sum_{k} 1/\sigma_{\mathrm{n}}[k]^2}. \tag{86}$$

Finally, $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda)$ is maximized with respect to $C^{(p)}[k]$, $C_n^{(a)}$, σ_{p}^2 and σ_{a}^2 when

$$\hat{C}^{(\mathrm{p})}[k] = u_{\mathrm{p}}[k], \tag{87}$$

$$\hat{C}_n^{(\mathrm{p})} = \frac{\sum_{q \in \mathrm{a}_n} \sum_k \gamma_{q,k} u_{\mathrm{a}}[k] / \sigma_{\mathrm{n}}[k]^2}{\sum_{q \in \mathrm{a}_n} \sum_k \gamma_{q,k} / \sigma_{\mathrm{n}}[k]^2}, \tag{88}$$

$$\hat{\sigma}_{\mathrm{p}}^2 = \frac{\sum_k \sum_{q \in \mathcal{S}} \gamma_{q,k} (u_{\mathrm{p}}[k] - [\boldsymbol{c}_q[k]]_1)^2}{\sum_k \sum_{q \in \mathcal{S}} \gamma_{q,k}}, \tag{89}$$

$$\hat{\sigma}_{\mathrm{a}}^2 = \frac{\sum_k \sum_{q \in \mathcal{S}} \gamma_{q,k} (u_{\mathrm{a}}[k] - [\boldsymbol{c}_q[k]]_2)^2}{\sum_k \sum_{q \in \mathcal{S}} \gamma_{q,k}}. \tag{90}$$

The above equations can be obtained by setting the partial derivatives of $\breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \lambda)$ with respect to $u_{\mathrm{p}}[l]$, $u_{\mathrm{a}}[l]$, μ_{b}, $C^{(p)}[k]$ and $C_n^{(a)}$ at zero, respectively. Here, it is important to note that when Eq. (83) (or Eq. (84)) becomes negative, $u_{\mathrm{p}}[l] = 0$ (or $u_{\mathrm{a}}[l] = 0$) is the maximizer of $\breve{Q}$ subject to non-negativity, since $\breve{Q}$ is a quadratic (strictly convex) function of $u_{\mathrm{p}}[l]$ (or $u_{\mathrm{a}}[l]$). Thus, under the non-negativity constraint, the update rules of $u_{\mathrm{p}}[l]$ and $u_{\mathrm{a}}[l]$ shall be written as

$$u_{\mathrm{p}}[l] \leftarrow \max(\hat{u}_{\mathrm{p}}[l], 0), \tag{91}$$

$$u_{\mathrm{a}}[l] \leftarrow \max(\hat{u}_{\mathrm{a}}[l], 0). \tag{92}$$

It can be shown that $Q(\tilde{\Theta}, \tilde{\Theta}')$ is non-decreasing under the updates of λ and $\tilde{\Theta}$ with the above update equations since

$$Q(\tilde{\Theta}, \tilde{\Theta}') = \breve{Q}(\tilde{\Theta}, \tilde{\Theta}', \hat{\lambda}) \leq \breve{Q}(\hat{\tilde{\Theta}}, \tilde{\Theta}', \hat{\lambda}) \leq Q(\hat{\tilde{\Theta}}, \tilde{\Theta}'). \tag{93}$$

To summarize, we obtain the following algorithm that guarantees the convergence to a local maximum of the posterior density of interest and the non-negativity of u_{p} and u_{a}:

1) Update γ using the Forward-Backward algorithm.
2) Increase $Q(\tilde{\Theta}, \tilde{\Theta}')$ w.r.t. $\tilde{\Theta}$ through the following updates:
   a) Update λ via Eqs. (79) and (80).
   b) Update $\boldsymbol{o}$ via Eqs. (91) and (92).
   c) Update θ via Eqs. (87)–(90).
   d) Update μ_{b} via Eq. (86).

   Return to 1) until convergence.

As with the algorithm in V-A, the complexity of this algorithm is $O(|\mathcal{S}|^2 K)$. After convergence, we search for the optimal state sequence $\boldsymbol{s}$ from the output sequence $\boldsymbol{o} = \{\boldsymbol{o}[k]\}_{k=1}^K$ by using the Viterbi algorithm.

VI. Evaluation of Parameter Estimation Accuracy

A. Parameter Estimation Using Real Speech Data

To evaluate the parameter estimation accuracy of the algorithms proposed in V-A and V-B, we conducted an experiment using real speech data, excerpted from the ATR Japanese speech database B-set [25]. This database consists of 503 phonetically balanced sentences. We selected speech samples of one male speaker (MHT). The ground truth data of the Fujisaki model parameters had been manually annotated by an expert in the speech prosody field. In these ground truth data, the baseline values were all set at log 60 Hz. We chose the Fujisaki model parameter extractor developed by Narusawa [15] as a baseline method for comparison.

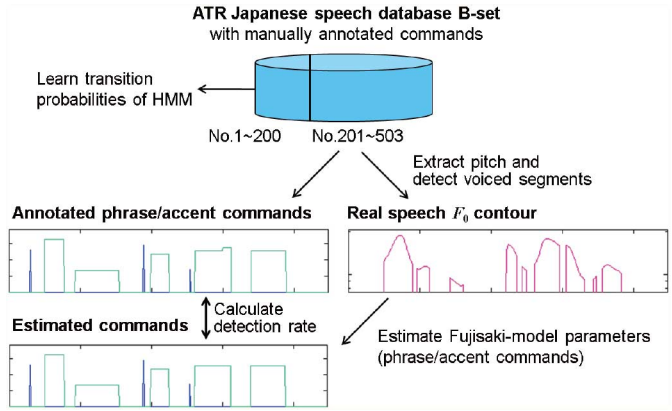


Fig. 4. Overview of the experiment in VI-A.

Fig. 4 shows the experimental scheme of the evaluation. F_0 contours were extracted using a method we had previously developed [26], from which the Fujisaki model parameters were estimated using the present algorithm. V/UV segments were obtained by simple energy thresholding. The constant parameters were fixed at $N = 10$, $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $v_{\mathrm{p}}^2 = 0.2^2$, $v_{\mathrm{a}}^2 = 0.1^2$, $v_{\mathrm{n}}^2[k] = 10^{15}$ for unvoiced regions and $v_{\mathrm{n}}^2[k] = 0.2^2$ for voiced regions. μ_{b} was set at the minimum log F_0 value in the voiced regions. The initial values of Θ were set at the values obtained with Narusawa's method [15]. The EM algorithm was then run for 20 iterations. 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: (1) the detection rate of the phrase and accent commands, and (2) the root mean squared error (RMSE) between an observed log F_0 contour, $y[k]$, and an estimated model, $G_{\mathrm{p}}[k] * \mu_{\mathrm{p}}[k] + G_{\mathrm{a}}[k] * \mu_{\mathrm{a}}[k] + \mu_{\mathrm{b}}$ over the voiced regions. The aim of this experiment was to confirm whether the present method is able to achieve accurate model fitting while ensuring that the estimated parameters are linguistically reasonable. The log F_0 RMSE indicates how well the estimated Fujisaki model fits an observed F_0 contour. The detection rate of the phrase and accent commands indicates how linguistically reasonable the estimated parameters are. The detection rate of the phrase and accent commands was calculated by matching the estimated and ground truth command sequences on a command-by-command basis using a dynamic programming algorithm. If the time difference between the estimated and ground truth phrase commands was shorter than S seconds, the estimated phrase command was considered "matched" and the local distance was set at zero. Otherwise the local distance was set at 1. As for the accent commands, we took the average of the time difference between the onsets of the estimated and ground truth accent commands and the time difference between their offsets. In the same way, when the average time difference was shorter than S seconds, the estimated accent command was considered matched. The magnitudes of the phrase and accent commands were not taken into account in our evaluation. This is because the magnitude estimation was very sensitive to the

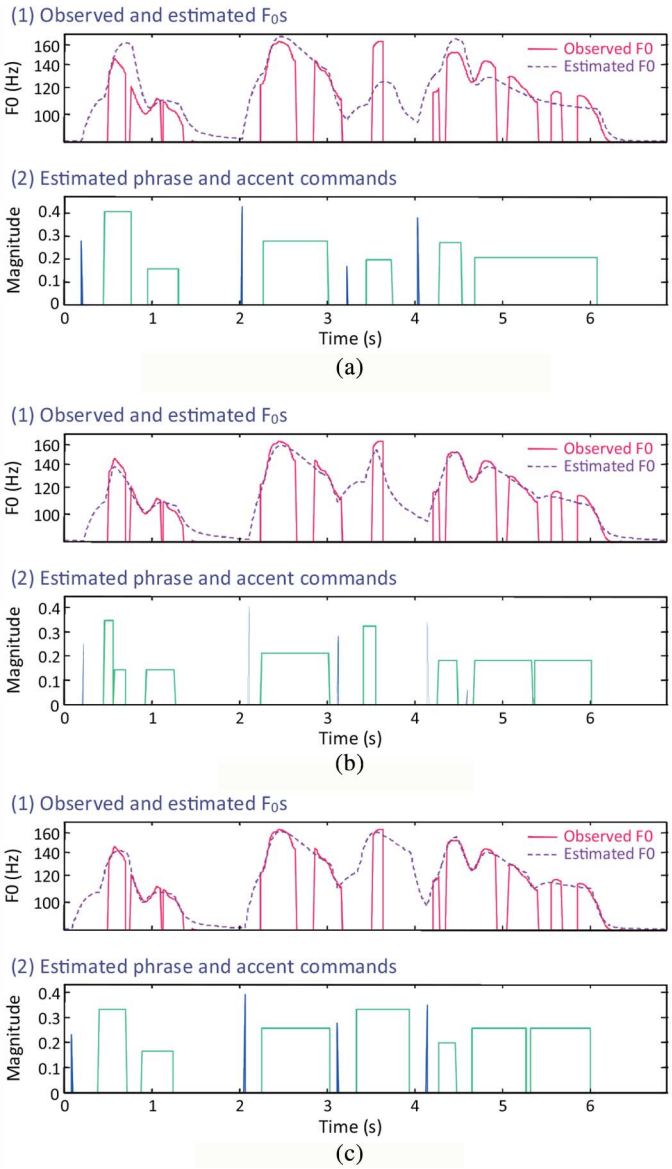


Fig. 5. Example of command detection (1 of 4). (1) An observed F_0 contour in voiced regions (solid line) and the estimated F_0 contours (dotted line) along with (2) the estimated phrase and accent commands. (a) Narusawa's method [15]. (b) Algorithm proposed in V-A. (c) Algorithm proposed in V-B.

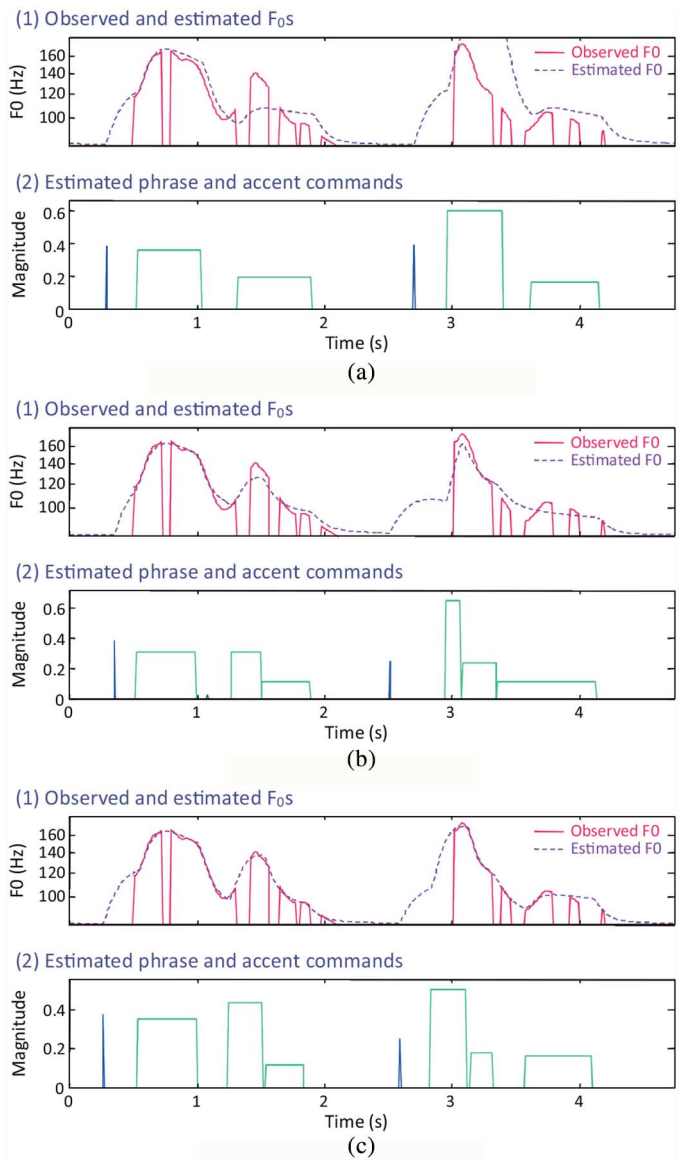


Fig. 6. Example of command detection (2 of 4). (1) An observed F_0 contour in voiced regions (solid line) and the estimated F_0 contours (dotted line) along with (2) the estimated phrase and accent commands. (a) Narusawa's method [15]. (b) Algorithm proposed in V-A. (c) Algorithm proposed in V-B.

baseline F_0 value, which was set differently in the present method and in the manual annotation. Let N_{A} be the number of commands in the ground truth command sequences and $N_{\mathrm{A}sum}$ be the sum of N_{A} for all 303 sentences. We defined the detection rate as

$$\frac{N_{\mathrm{Asum}} - N_{\mathrm{Isum}} - N_{\mathrm{Ssum}} - N_{\mathrm{Dsum}}}{N_{\mathrm{Asum}}} \times 100(\%), \qquad (94)$$

where N_{Isum}, N_{Ssum} and N_{Dsum} are the total numbers of insertion, substitution and deletion errors, respectively.

Table I shows the detection rate results for phrase and accent commands with $S = 0.3, 0.2, 0.1[s]$. "C," "P1" and "P2" refer to Narusawa's method [15], and the algorithms proposed in V-A and V-B, respectively. The results showed that the proposed algorithms "P1" and "P2" were superior to Narusawa's method, and "P2" was slightly superior to "P1" in terms of detection rate. The left graph of Fig. 9 shows the log F_0 RMSEs. As the results show, "P2" yielded the highest model fitting accuracy. Figs. 5–8 show some examples of observed F_0 contours and the estimated F_0 contours obtained with the conventional and present methods, from which we can confirm that the present methods (especially "P2") were able to fit the Fujisaki model to observed F_0 contours fairly well.

It should be noted that the detection rate tended to drop significantly when $S = 0.1[s]$. Considering the fact that average syllable durations are typically about 0.2 [s], deviations longer than 0.1 [s] from the true positions are not negligible. This implies that the proposed methods still have plenty of room for improvement.

B. Parameter Estimation Using Synthetic F_0 Contours

To evaluate the pure behavior of the present parameter estimation algorithms, we also conducted a command estimation experiment using synthetic F_0 contours. The synthetic F_0

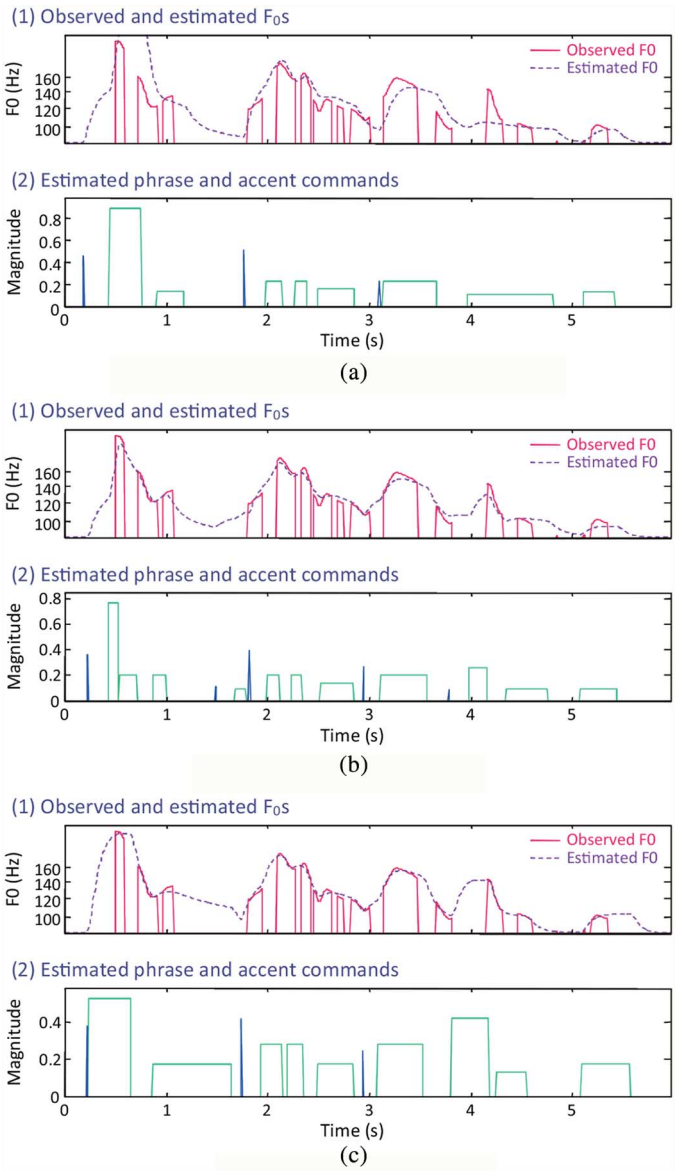


Fig. 7. Example of command detection (3 of 4). (1) An observed F_0 contour in voiced regions (solid line) and the estimated F_0 contours (dotted line) along with (2) the estimated phrase and accent commands. (a) Narusawa's method [15]. (b) Algorithm proposed in V-A. (c) Algorithm proposed in V-B.

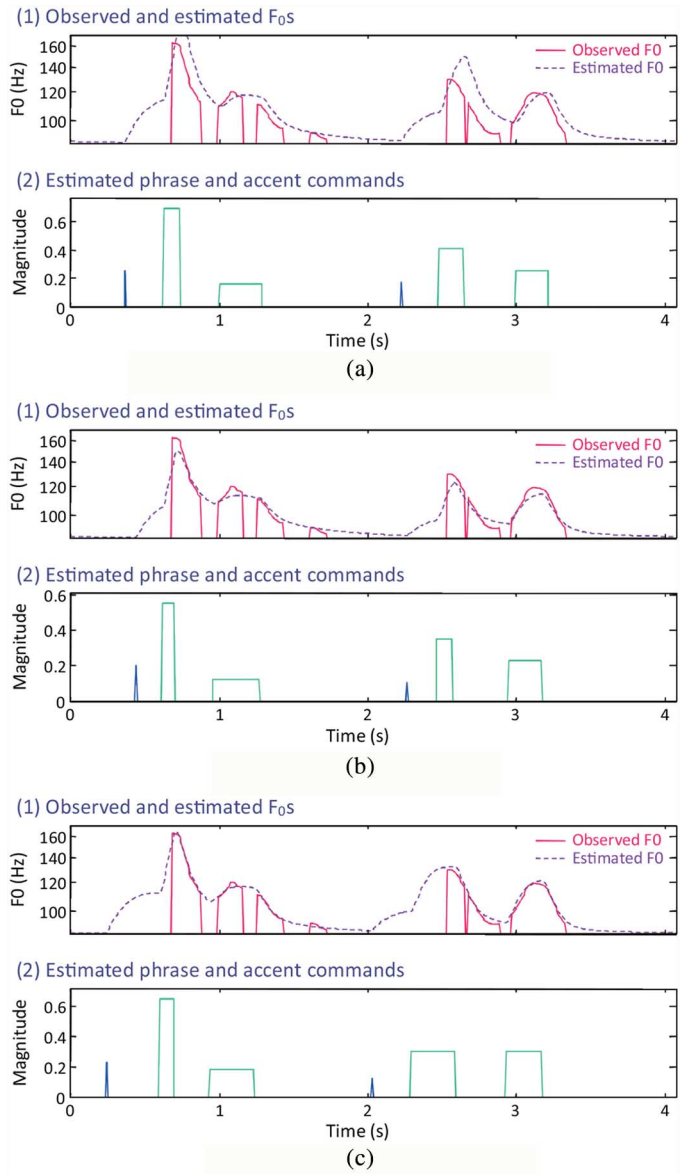


Fig. 8. Example of command detection (4 of 4). (1) An observed F_0 contour in voiced regions (solid line) and the estimated F_0 contours (dotted line) along with (2) the estimated phrase and accent commands. (a) Narusawa's method [15]. (b) Algorithm proposed in V-A. (c) Algorithm proposed in V-B.

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. Fig. 10 provides an overview of this experiment.

Table II shows detection rate results for command sequences with different S settings. The log F_0 RMSEs are shown in the right graph of Fig. 9. As the results show, the proposed algorithms were again significantly superior to the conventional method in terms of both the detection rate of the command sequences and the model fitting accuracy. However, as regards the accent command detection rate, the proposed methods were outperformed by the conventional method. The conventional method uses the fact that the maxima and minima of the first derivative of the F_0 contour of the Fujisaki model correspond to the onsets and offsets of accent commands with a constant delay of $1/\beta$ if the contributions of the phrase components can be disregarded. One reason why the conventional method was

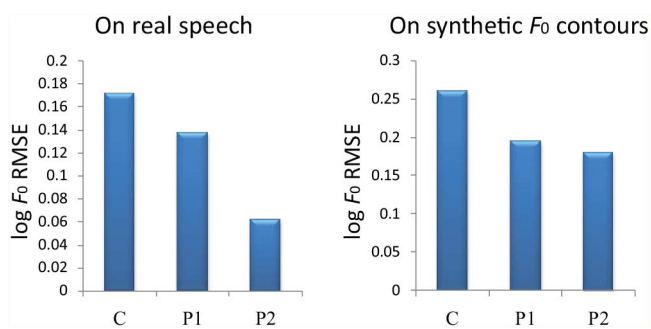


Fig. 9. The root mean squared errors (RMSEs) between observed log F_0 contours and estimated models.

able to accurately detect accent commands might be that this experiment used synthetic F_0 contours created using the Fujisaki model as the test data, which agrees particularly well with the above assumption.

TABLE I
DETECTION RATE (%) OF PHRASE AND ACCENT COMMANDS WITH DIFFERENT S SETTINGS EVALUATED ON REAL SPEECH DATA

	$S = 0.3$[s]			$S = 0.2$[s]			$S = 0.1$[s]		
	C	P1	P2	C	P1	P2	C	P1	P2
All commands	89.1	91.2	**93.2**	84.2	85.0	**88.0**	64.5	58.3	**67.9**
Phrase commands	87.4	99.0	90.3	84.9	94.9	88.1	69.0	83.6	72.4
Accent commands	90.0	87.1	94.7	83.8	79.8	88.0	62.7	54.8	65.6

TABLE II
DETECTION RATE (%) OF PHRASE AND ACCENT COMMANDS WITH DIFFERENT S SETTINGS EVALUATED ON SYNTHETIC F_0 CONTOURS

	$S = 0.3$[s]			$S = 0.2$[s]			$S = 0.1$[s]		
	C	P1	P2	C	P1	P2	C	P1	P2
All commands	88.6	**94.4**	93.8	85.6	**93.2**	91.7	75.7	**88.1**	84.2
Phrase commands	71.5	99.3	95.5	66.2	98.4	91.0	52.1	97.0	85.9
Accent commands	99.3	91.8	92.8	97.5	90.5	92.1	89.9	83.5	83.4

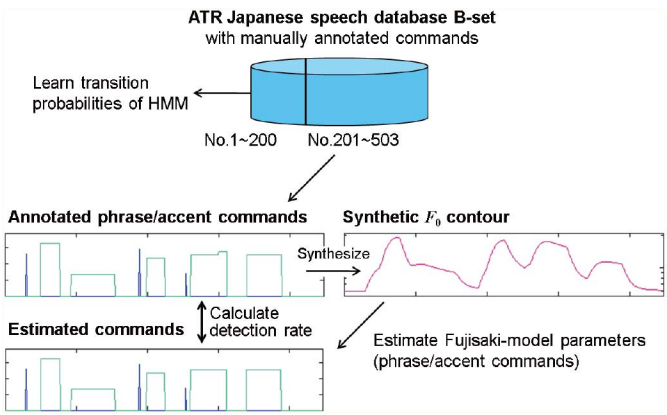


Fig. 10. Overview of the experiment described in VI-B.

VII. CONCLUSION

This paper proposed introducing a generative model of voice F_0 contours for estimating prosodic features from raw speech data. We formulated the present F_0 contour model 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. There were two motivations behind this formulation. One was to derive a general parameter estimation framework for the Fujisaki model that allows the introduction of powerful algorithms such as the Viterbi algorithm, forward-backward algorithm and EM algorithm. The other was to construct an automatically trainable version of the Fujisaki model that we can naturally incorporate into statistical speech synthesis and conversion frameworks. We quantitatively evaluated the performance of the proposed Fujisaki model parameter extractor using real speech data. Experimental results revealed that our method was superior to a state-of-the-art Fujisaki model parameter extractor. The application of the present F_0 contour model to prosody generation for text-to-speech synthesis is one of our ongoing projects. A preliminary study is presented in [27].

ACKNOWLEDGMENT

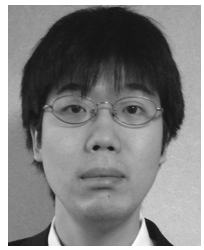
We thank Prof. Keikichi Hirose (The University of Tokyo) for kindly providing us with the manually annotated data associated with the ATR speech samples. We also thank Prof. Shigeki Sagayama (Meiji University) and Dr. Daisuke Saito (The University of Tokyo) for fruitful discussions.

REFERENCES

[1] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F_0 contours," in *Proc. SCA Workshop Statist. Percept. Audition (SAPA'10)*, 2010, pp. 43–48.
[2] 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, pp. 175–178.
[3] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden Markov convolutive mixture model for pitch contour analysis of speech," in *Proc. 13th Annu. Conf. Int. Speech Commun. Association (Interspeech'12)*, 2012.
[4] H. Kameoka, K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama, "Generative modeling of speech F_0 contours," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech'13)*, 2013, pp. 1826–1830.
[5] H. Fujisaki, , O. Fujimura, Ed., "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal Physiology: Voice Production, Mechanisms and Functions*. New York, NY, USA: Raven, 1988.
[6] 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.
[7] H. Fujisaki and S. Ohno, "Analysis and modeling of fundamental frequency contours of English utterances," in *Proc. 4th Eur. Conf. Speech Commun. Technol. (EUROSPEECH'95)*, 1995, vol. 2, pp. 985–988.
[8] H. Mixdorff and H. Fujisaki, "Analysis of voice fundamental frequency contours of German utterances using a quantitative model," in *Proc. 3rd Int. Conf. Spoken Lang. Process. (ICSLP'94)*, 1994, vol. 4, pp. 2231–2234.
[9] H. Fujisaki, S. Ohno, K. Nakamura, M. Guirao, and J. Gurlekian, "Analysis of accent and intonation in Spanish based on the command-response model," in *Proc. 3rd Int. Conf. Spoken Lang. Process. (ICSLP'94)*, 1994, vol. 1, pp. 355–358.
[10] H. Fujisaki, M. Ljungqvist, and H. Murata, "Analysis and modeling of word accent and sentence intonation in Swedish," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'93)*, 1993, pp. 211–214.
[11] H. Fujisaki, S. Narusawa, S. Ohno, and D. Freitas, "Analysis and modeling of F_0 contours of Portuguese utterances based on the command-response model," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (EUROSPEECH'03)*, 2003, vol. 3, pp. 2317–2320.
[12] C. Wang, H. Fujisaki, R. Tomana, and S. Ohno, "Analysis of fundamental frequency contours of standard Chinese in terms of a command-response model and its application to synthesis by rule of intonation," in *Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP'00)*, 2000, vol. 3, pp. 326–329.

[13] H. Fujisaki, W. Gu, and K. Hirose, "The command-response model for the generation of F_0 contours of Cantonese utterances," in *Proc. 7th Int. Conf. Signal Process. (ICSP'04)*, 2004, vol. 1, pp. 655–658.
[14] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'00)*, 2000, vol. 3, pp. 1281–1284.
[15] 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. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, pp. 509–512.
[16] P. S. Rossi, F. Palmieri, and F. Cutugno, "A method for automatic extraction of Fujisaki-model parameters," in *Proc. Speech Prosody'02*, 2002, pp. 615–618.
[17] P. S. Rossi, F. Palmieri, and F. Cutugno, "Inversion of F_0 model for natural-sounding speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, pp. 520–523.
[18] H. R. Pfitzinger, H. Mixdorff, and J. Schwarz, "Comparison of Fujisaki-model extractors and F_0 stylizers," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech'09)*, 2009, pp. 2455–2458.
[19] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. Applicat. Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.
[20] M. Russell and R. Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'85)*, 1985, pp. 5–8.
[21] S. Levinson, "Continuously variable duration hidden Markov models for speech analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'86)*, 1986, pp. 1241–1244.
[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[23] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.
[24] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, 2009, pp. 45–48.
[25] 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 Commun.*, vol. 27, pp. 187–207, 1999.
[26] H. Kameoka, "Statistical approach to multipitch analysis," Ph.D. dissertation, The Univ. of Tokyo, Tokyo, Japan, 2007.
[27] K. Kadowaki, T. Ishihara, N. Hojo, and H. Kameoka, "Speech prosody generation for text-to-speech synthesis based on generative model of F_0 contours," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech'14)*, 2014, pp. 2322–2326.

Hirokazu Kameoka received B.E., M.S. and Ph.D. degrees all from the University of Tokyo, Japan, in 2002, 2004 and 2007, respectively. He is currently a Distinguished Researcher at NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, and is also an Adjunct Associate Professor at the University of Tokyo. His research interests include audio, speech, and music signal processing and machine learning. He has served as a publicity chair of ISMIR 2009 and a program committee member of ISMIR 2014 and ISMIR 2015. Since 2015, he has been an associate editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING. He is a member of IEEE, the Institute of Electronics, Information and Communication Engineers (IEICE), the Information Processing Society of Japan (IPSJ) and the Acoustical Society of Japan (ASJ). He received 13 awards over the past 10 years, including the IEEE Signal Processing Society 2008 SPS Young Author Best Paper Award. He is the author or co-author of about 90 articles in journal papers and peer-reviewed conference proceedings.



Kota Yoshizato received the B.E. and M.S. degrees from the University of Tokyo, Japan, in 2011 and 2013, respectively. Since then, he has been a Software Engineer at DeNA Corp.

Tatsuma Ishihara received the B.E. and M.S. degrees from the University of Tokyo, Japan, in 2012 and 2014, respectively. Since then, he has been a Research Engineer with the Toshiba Corporate Research & Development Center.

Kento Kadowaki received the B.E. degree from Osaka University, Japan, in 2013 and the M.S. degree from the University of Tokyo, Japan, in 2015. He is currently with Nomura Research Institute Co., Ltd., Japan.

Yasunori Ohishi received his Ph.D. degree in information science in 2009 from Nagoya University, Japan. In 2009, he joined NTT Communication Science Laboratories. His research interests generally concern statistical acoustic signal processing, music information retrieval, and unsupervised machine learning with audio applications. In 2009, he received the 36th Awaya Kiyoshi Science Promotion Award from Acoustical Society of Japan (ASJ).

Kunio Kashino received B.E., M.S., and Ph.D. degrees all from the University of Tokyo, Japan, in 1990, 1992, and 1995, respectively. He is currently a Senior Distinguished Researcher at Nippon Telegraph and Telephone Corporation, Executive Manager of Media Information Laboratory, NTT Communication Science Laboratories, and a Visiting Professor at National Institute of Informatics. His research interests include media information processing and audio information processing. He has been involved in the research projects on computational auditory scene analysis since 1990 and robust media search since 1996. He is a senior member of the IEEE and the Institute of Electronics, Information and Communication Engineers (IEICE), and a member of the Information Processing Society of Japan (IPSJ), the Acoustical Society of Japan (ASJ), and the Japanese Society for Artificial Intelligence (JSAI). He received more than 10 awards over the past 20 years, including the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award.