A Statistical Model of Speech F_0 Contours

Hirokazu Kameoka, Jonathan Le Roux, Yasunori Ohishi[1]

[1]NTT Communication Science Laboratories, NTT Corporation, Japan

{kameoka,leroux,ohishi}@cs.brl.ntt.co.jp

Abstract

This paper proposes a statistical model of speech fundamental frequency (F_0) contours, based on the formulation of the discrete-time stochastic process version of the Fujisaki model, which is known as a well-founded mathematical model representing the control mechanism of vocal fold vibration. There are two important motivations for this statistical formulation. One is to derive a general parameter estimation framework for the Fujisaki model, allowing for the introduction of powerful statistical methods, and the other is to introduce a measure of speech naturalness in terms of an F_0 contour through a probability distribution assumption, that can be incorporated into many statistical speech processing problems such as speech analysis, synthesis, separation, denoising and dereverberation.

Index Terms: speech F_0 contour, statistical model

1. Introduction

The fundamental frequency (F_0) contour in normal speech contains various types of information including emotional and other non-linguistic information such as the speaker's identity, mood and level of attention, and plays as important a role in our daily speech communication as formants, through which we encode a phonemic sequence to convey linguistic information to the listener(s). An F_0 contour is a realization of the vocal fold oscillation with slowly varying frequencies, whose dynamics are governed by a combination of different factors, in particular the length and elasticity of vocal folds, laryngeal muscle tension, and subglottal air pressure. All possible F_0 contours produced by a particular speech apparatus should thus be characterized and constrained by the presumably small number of parameters governing the control mechanism of vocal fold vibration. Therefore, how well the F_0 contour of a certain sound matches the mechanical constraint is an important factor that determines how likely it is that the sound originates from a speech utterance. Accordingly, modeling the dynamics of the F_0 contour of speech can be potentially very beneficial for any speech applications that could be improved by taking account of the naturalness in terms of the F_0 contour.

The Fujisaki model [1–3] is a well-founded mathematical model consisting of a set of physiologically and physically meaningful parameters, which describes the process of generating F_0 contours by vocal folds in a reasonably simplified form. 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. On the other hand, several techniques have been proposed for solving the inverse problem of estimating the Fujisaki model parameters from raw F_0 contour observations [3–5], for the purpose of incorporating the extracted parameters in automatic speech/emotion recognition systems in some way to improve their performance, but so far with limited success due to the analytical complexity of the Fujisaki model.

In this paper, we formulate a discrete-time stochastic process version of the Fujisaki model. Our motivation behind this statistical formulation is twofold. Firstly, by making the best use of statistical techniques, we expect to be able to derive a powerful framework for the estimation of the Fujisaki model parameters, which has conventionally been considered a difficult task. Secondly, it should enable us to represent the notion of speech naturalness in terms of the F_0 contour through a probability distribution assumption. This will allow us to smoothly incorporate our probabilistic model as an additional speech naturalness measure into many statistical speech processing problems such as speech separation, denoising and dereverberation.

2. Original Fujisaki model

The Fujisaki model [1–3], shown in Fig. 1, assumes that an F_0 contour on a logarithmic scale, $y(t)$, where t is time, is the superposition of two contributions associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times, referred to as the phrase component, $y_{\rm p}(t)$, and the accent component, $y_{\rm a}(t)$, respectively [2]. The phrase component consists of the major-scale pitch variations over the duration of the prosodic units, which are characterized by a fast rise followed by a slower fall. The accent component 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 Dirac deltas (phrase commands), and the other with rectangular pulses (accent commands). The linear systems producing $y_{\rm p}(t)$ and $y_{\rm a}(t)$, namely phrase control and accent control mechanisms, are characterized by

$$G_{\rm p}(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{1}$$

$$S_{\rm a}(t) = \begin{cases} 1 - (1+\beta t)e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{2}$$

where $G_{\rm p}(t)$ is the impulse response of the phrase control mechanism and $S_{\rm a}(t)$ is the step response of the accent control mechanism. α and β are natural angular frequencies of the two second-order systems, which are known to be almost constant

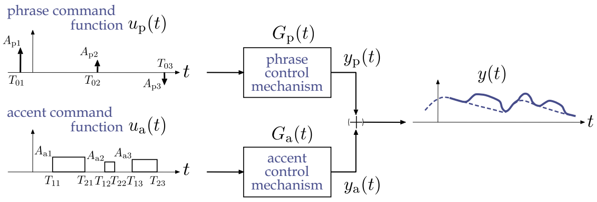

Figure 1: Block diagram of Fujisaki model

within an utterance as well as across utterances for a particular speaker. Although certain differences exist across speakers, it has been shown that $\alpha = 3$rad/s and $\beta = 20$rad/s can be used as default values. The two components $y_\mathrm{p}(t)$ and $y_\mathrm{a}(t)$ are further added by a constant value y_b related to the lower bound for the speaker's F_0, below which no regular vocal fold vibration can be maintained. The log F_0 contour, $y(t)$, is thus expressed as

$$\begin{aligned} y(t) = y_\mathrm{b} &+ \sum_i A_{\mathrm{p},i} G_\mathrm{p}(t - T_{0,i}) \\ &+ \sum_j A_{\mathrm{a},j} \big\{ S_\mathrm{a}(t - T_{1,j}) - S_\mathrm{a}(t - T_{2,j}) \big\}, \quad (3) \end{aligned}$$

where $T_{0,i}$ and $A_{\mathrm{p},i}$ denote the onset time and amplitude of the i-th phrase command, and $T_{1,j}$, $T_{2,j}$ and $A_{\mathrm{a},j}$ denote the onset time, offset time and amplitude of the j-th accent command.

Although the two functions $G_\mathrm{p}(t)$ and $S_\mathrm{a}(t)$ look different, it should be made clear that the impulse response of the accent control mechanism, $G_\mathrm{a}(t)$, has the same form as the phrase control mechanism, $G_\mathrm{p}(t)$, namely

$$G_\mathrm{a}(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \quad (4)$$

since $G_\mathrm{a}(t)$ is defined by the time differentiation of the step response $S_\mathrm{a}(t)$. The accent component can thus be seen as the convolution of $G_\mathrm{a}(t)$ and a command function composed of a set of stepwise functions.

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. 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

$$\mathcal{G}_\mathrm{p}(s) = \mathcal{L}\big[G_\mathrm{p}(t)\big] = \frac{\alpha^2}{(s+\alpha)^2}. \quad (5)$$

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

$$s \simeq \frac{1 - z^{-1}}{t_0}, \quad (6)$$

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

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

where

$$a_2 = (\psi - 1)^2, \quad (8)$$
$$a_1 = -2\psi(\psi - 1), \quad (9)$$
$$a_0 = \psi^2, \quad (10)$$

and $\psi = 1 + 1/(\alpha t_0)$. Here we use $u_\mathrm{p}[k]$ to denote the discrete-time version of the phrase command function where k indicates the discrete-time index. The discrete-time version of the phrase component, $y_\mathrm{p}[k]$, 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

$$u_\mathrm{p}[k] = a_0 y_\mathrm{p}[k] + a_1 y_\mathrm{p}[k-1] + a_2 y_\mathrm{p}[k-2]. \quad (11)$$

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

$$u_\mathrm{a}[k] = b_0 y_\mathrm{a}[k] + b_1 y_\mathrm{a}[k-1] + b_2 y_\mathrm{a}[k-2], \quad (12)$$
$$b_2 = (\varphi - 1)^2, \quad (13)$$
$$b_1 = -2\varphi(\varphi - 1), \quad (14)$$
$$b_0 = \varphi^2, \quad (15)$$

where $\varphi = 1 + 1/(\beta t_0)$. For the baseline F_0, we use $y_\mathrm{b}[k]$ to denote the discrete-time version of $y_\mathrm{b}(t)$. Altogether, the discrete-time version of the Fujisaki model can be expressed as the superposition of the three components:

$$y[k] = y_\mathrm{p}[k] + y_\mathrm{a}[k] + y_\mathrm{b}[k]. \quad (16)$$

4. Statistical formulation

4.1. Modeling phrase and accent command pair with HMM

We describe here how to model the command functions $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$. To explain the phrase and accend commands in a physiologically and linguistically meaningful way, it is important that they 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, $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$ are reciprocally constrained and so they should not simply be modeled separately. One challenge in 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 assumptions into the command functions, we propose modeling the $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$ pair using a hidden Markov model (HMM). The use of an HMM for modeling the command functions is our primary reason for considering modeling the discrete-time version of the Fujisaki model.

Let us arrange $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$ into a vector $\boldsymbol{o}[k]$. We assume that $\boldsymbol{o}[k]$ is a random vector driven by an additive white Gaussian noise $(\epsilon_\mathrm{p}[k], \epsilon_\mathrm{a}[k])^\mathrm{T}$ such that

$$\boldsymbol{o}[k] := \begin{bmatrix} u_\mathrm{p}[k] \\ u_\mathrm{a}[k] \end{bmatrix} = \begin{bmatrix} \mu_\mathrm{p}[k] \\ \mu_\mathrm{a}[k] \end{bmatrix} + \begin{bmatrix} \epsilon_\mathrm{p}[k] \\ \epsilon_\mathrm{a}[k] \end{bmatrix}. \quad (17)$$

Let $\epsilon_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$ and $\epsilon_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$ be mutually independent, then

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{\nu}[k], \boldsymbol{\Upsilon}\right) \tag{18}$$

$$\boldsymbol{\nu}[k] := \begin{bmatrix} \mu_{\rm p}[k] \\ \mu_{\rm a}[k] \end{bmatrix}, \ \boldsymbol{\Upsilon} := \begin{bmatrix} \sigma_{\rm p}^2 & 0 \\ 0 & \sigma_{\rm a}^2 \end{bmatrix}. \tag{19}$$

Equation (18) can be viewed as an HMM in which the output distribution of each state is a Gaussian distribution. The mean vector $\boldsymbol{\nu}[k]$ is thus considered to evolve in time as a result of the state transition. This way of thinking allows us to incorporate assumptions 1–3 into $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ by simply constraining the path of the state transitions, as illustrated in Fig. 2.

The present HMM consists of $N+3$ distinct states, $\rm p_0$, $\rm p_1$ and $\rm a_0, \cdots, a_N$. In state $\rm p_0$, $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ are both constrained to be zero. Given that the model is in state $\rm p_0$, it is only allowed either to stay in that state or to move to state $\rm p_1$. In state $\rm p_1$, $\mu_{\rm p}[k]$ can take a non-zero value as a function of time, $A_{\rm p}[k]$, whereas $\mu_{\rm a}[k]$ is still restricted to zero. It is important to note that in state $\rm p_1$, no self-transitions are allowed and only the transition to state $\rm a_0$ is possible. In state $\rm a_0$, $\mu_{\rm p}[k]$ and $\mu_{\rm a}[k]$ are again both assumed to be zero. Hence, this specific path constraint restricts $\mu_{\rm p}[k]$ to consisting of isolated deltas. State $\rm a_0$ leads to states $\rm a_1, \cdots, a_N$, in each of which $\mu_{\rm a}[k]$ can take a different non-zero value $A_{\rm a}^{(n)}$ assumed to be constant in time, whereas $\mu_{\rm p}[k]$ is forced to be zero. Direct state transitions from state $\rm a_n$ to state $\rm a_{n'}$ ($n \neq n'$, $1 \leq n \leq N$, $1 \leq n' \leq N$) without passing through state $\rm a_0$ are not allowed. This constraint restricts $\mu_{\rm a}[k]$ to consisting of rectangular pulses. It should also be noted that the present HMM ensures that no more than one command will be active at each point in time. To sum up, the proposed HMM is defined as follows:

$$\boxed{\begin{array}{rl}
\text{Output sequence:} & \{\boldsymbol{o}[k]\}_{k=1}^{K} \\
\text{Set of states:} & \mathcal{S} := \{\rm p_0, p_1, a_0, \cdots, a_N\} \\
\text{State sequence:} & \{s_k\}_{k=1}^{K} \\
\text{Output distribution:} & P(\boldsymbol{o}[k] | s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Upsilon}) \\
& \boldsymbol{c}_i[k] = \begin{cases} (0,0)^{\rm T} & (i = {\rm p_0}) \\ (A_{\rm p}[k], 0)^{\rm T} & (i = {\rm p_1}) \\ (0,0)^{\rm T} & (i = {\rm a_0}) \\ (0, A_{\rm a}^{(n)})^{\rm T} & (i = {\rm a}_n) \end{cases} \\
\text{Transition probability:} & \phi_{i',i} := \log P(s_k = i | s_{k-1} = i')
\end{array}}$$

For simplicity, we treat the transition probabilities $\phi_{i',i}$ as constant parameters in this paper, so that the free parameters to be determined in our command function model consist of the magnitude of the phrase command, $A_{\rm p}[k]$, the state sequence, s_k, the magnitude of the accent command, $\{A_{\rm a}^{(n)}\}_{n=1}^{N}$, and the variance of the output distribution, $\sigma_{\rm p}^2$, $\sigma_{\rm a}^2$. Hereafter we use θ_u to denote all these parameters:

$$\theta_u := \left\{\{A_{\rm p}[k], s[k]\}_{k=1}^{K}, \{A_{\rm a}^{(n)}\}_{n=1}^{N}, \sigma_{\rm p}^2, \sigma_{\rm a}^2\right\}. \tag{20}$$

Once the state sequence $\{s_k\}_{k=1}^{K}$ is specified, the mean sequences, $\{\mu_{\rm p}[k]\}_{k=1}^{K}$ and $\{\mu_{\rm a}[k]\}_{k=1}^{K}$, namely the phrase and accent command functions, are determined simultaneously by

$$\begin{bmatrix} \mu_{\rm p}[k] \\ \mu_{\rm a}[k] \end{bmatrix} = \boldsymbol{c}_{s_k}[k]. \tag{21}$$

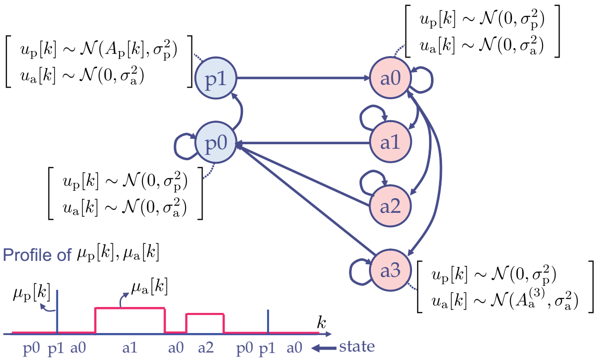
$u_{\rm p}[k] \sim \mathcal{N}(A_{\rm p}[k], \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
p1
a0
p0
a1
a2
a3
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(0, \sigma_{\rm a}^2)$
Profile of $\mu_{\rm p}[k], \mu_{\rm a}[k]$
$u_{\rm p}[k] \sim \mathcal{N}(0, \sigma_{\rm p}^2)$
$u_{\rm a}[k] \sim \mathcal{N}(A_{\rm a}^{(3)}, \sigma_{\rm a}^2)$
$\mu_{\rm p}[k]$
$\mu_{\rm a}[k]$
k
p0 p1 a0 a1 a0 a2 p0 p1 a0
state

Figure 2: Command function modeling with HMM.

4.2. Likelihood function and prior probabilities

In this subsection, we derive the probability density function of the F_0 contour, $y[1], \cdots, y[K]$, based on the statistical modeling of the command functions presented in the previous subsection. From Eqs. (18) and (19),

$$u_{\rm p}[k] | \theta_u \sim \mathcal{N}(\mu_{\rm p}[k], \sigma_{\rm p}^2), \tag{22}$$

$$u_{\rm a}[k] | \theta_u \sim \mathcal{N}(\mu_{\rm a}[k], \sigma_{\rm a}^2). \tag{23}$$

Since $u_{\rm p}[k]$ and $u_{\rm a}[k]$ are both assumed to be driven by white noise, it can be written in vector notation as

$$\boldsymbol{u}_{\rm p} | \theta_u \sim \mathcal{N}(\boldsymbol{\mu}_{\rm p}, \boldsymbol{\Sigma}_{\rm p}), \ \ \boldsymbol{\Sigma}_{\rm p} = \sigma_{\rm p}^2 \boldsymbol{I}, \tag{24}$$

$$\boldsymbol{u}_{\rm a} | \theta_u \sim \mathcal{N}(\boldsymbol{\mu}_{\rm a}, \boldsymbol{\Sigma}_{\rm a}), \ \ \boldsymbol{\Sigma}_{\rm a} = \sigma_{\rm a}^2 \boldsymbol{I}, \tag{25}$$

where $\boldsymbol{u}_{\rm p} := (u_{\rm p}[1], \cdots, u_{\rm p}[K])^{\rm T}$, $\boldsymbol{u}_{\rm a} := (u_{\rm a}[1], \cdots, u_{\rm a}[K])^{\rm T}$, $\boldsymbol{\mu}_{\rm p} := (\mu_{\rm p}[1], \cdots, \mu_{\rm p}[K])^{\rm T}$, $\boldsymbol{\mu}_{\rm a} := (\mu_{\rm a}[1], \cdots, \mu_{\rm a}[K])^{\rm T}$. By using the linear equation given in Section 3, the phrase component $\boldsymbol{y}_{\rm p} := (y_{\rm p}[1], \cdots, y_{\rm p}[K])^{\rm T}$ and the accent component $\boldsymbol{y}_{\rm a} := (y_{\rm a}[1], \cdots, y_{\rm a}[K])^{\rm T}$ can be written in terms of $\boldsymbol{u}_{\rm p}$ and $\boldsymbol{u}_{\rm a}$, respectively, such that

$$\boldsymbol{u}_{\rm p} = \boldsymbol{A} \boldsymbol{y}_{\rm p}, \tag{26}$$

$$\boldsymbol{u}_{\rm a} = \boldsymbol{B} \boldsymbol{y}_{\rm a}, \tag{27}$$

where

$$\boldsymbol{A} := \begin{bmatrix} a_0 & & & & O \\ a_1 & a_0 & & & \\ a_2 & a_1 & a_0 & & \\ & \ddots & \ddots & \ddots & \\ O & & a_2 & a_1 & a_0 \end{bmatrix}, \ \boldsymbol{B} := \begin{bmatrix} b_0 & & & & O \\ b_1 & b_0 & & & \\ b_2 & b_1 & a_0 & & \\ & \ddots & \ddots & \ddots & \\ O & & b_2 & b_1 & b_0 \end{bmatrix}. \tag{28}$$

Hence, it follows from Eqs. (24) and (25) that

$$\boldsymbol{y}_{\rm p} | \theta_u, \alpha \sim \mathcal{N}\big(\boldsymbol{A}^{-1} \boldsymbol{\mu}_{\rm p}, \boldsymbol{A}^{-1} \boldsymbol{\Sigma}_{\rm p} (\boldsymbol{A}^{-1})^{\rm T}\big), \tag{29}$$

$$\boldsymbol{y}_{\rm a} | \theta_u, \beta \sim \mathcal{N}\big(\boldsymbol{B}^{-1} \boldsymbol{\mu}_{\rm a}, \boldsymbol{B}^{-1} \boldsymbol{\Sigma}_{\rm a} (\boldsymbol{B}^{-1})^{\rm T}\big). \tag{30}$$

As for the base component $y_{\rm b}[k]$, we assume that it is also a random variable driven by additive white Gaussian noise $\epsilon_{\rm b}[k] \sim \mathcal{N}(0, \sigma_{\rm b}^2)$ such that

$$y_{\rm b}[k] = \mu_{\rm b} + \epsilon_{\rm b}[k], \tag{31}$$

and hence

$$\boldsymbol{y}_{\rm b} | \mu_{\rm b} \sim \mathcal{N}(\mu_{\rm b} \boldsymbol{1}, \boldsymbol{\Sigma}_{\rm b}), \ \ \boldsymbol{\Sigma}_{\rm b} = \sigma_{\rm b}^2 \boldsymbol{I}. \tag{32}$$

We use $\theta_{\rm b} := \{\mu_{\rm b}, \sigma_{\rm b}^2\}$ to denote the parameters related to the base component. Let $\epsilon_\xi[j]$ and $\epsilon_{\xi'}[j']$ be mutually independent when $(\xi, j) \neq (\xi', j')$, then $\boldsymbol{y}_{\rm p}$, $\boldsymbol{y}_{\rm a}$ and $\boldsymbol{y}_{\rm b}$ are assumed to be statistically independent. Care must be taken that this does not mean that $\boldsymbol{\mu}_{\rm p}$ and $\boldsymbol{\mu}_{\rm a}$ are independent. It therefore follows from Eqs. (29), (30) and (32) that their sum $\boldsymbol{y} = \boldsymbol{y}_{\rm p} + \boldsymbol{y}_{\rm a} + \boldsymbol{y}_{\rm b}$ will also be normally distributed such that

$$\begin{aligned}\boldsymbol{y}|\Theta \sim \mathcal{N}\big(&\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\rm p} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\rm a} + \mu_{\rm b}\boldsymbol{1},\\ &\boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\rm p}(\boldsymbol{A}^{-1})^{\rm T} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\rm a}(\boldsymbol{B}^{-1})^{\rm T} + \boldsymbol{\Sigma}_{\rm b}\big), \quad (33)\end{aligned}$$

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

$$\begin{aligned}P(\boldsymbol{y}|\Theta) &= \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\rm T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},\\ \boldsymbol{\mu} &= \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\rm p} + \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\rm a} + \mu_{\rm b}\boldsymbol{1}, \quad (34)\\ \boldsymbol{\Sigma} &= \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\rm p}\big(\boldsymbol{A}^{\rm T}\big)^{-1} + \boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\rm a}\big(\boldsymbol{B}^{\rm T}\big)^{-1} + \boldsymbol{\Sigma}_{\rm b}.\end{aligned}$$

As for the prior probability of Θ, we assume that the parameters are independent of each other, the parameters other than the state sequence, $\{s[k]\}_{k=1}^K$, phrase control parameter, ψ, and the accent control parameter, φ, are uniformly distributed, and $\{s[k]\}_{k=1}^K$ is a first-order Markov chain:

$$P(\Theta) \propto P(\psi)P(\varphi)P(s_1)\prod_{k=2}^{K} P(s_k|s_{k-1}). \quad (35)$$

5. Practical problem setting

In this section, we present some practical problems into which the proposed statistical F_0 contour model can be incorporated.

5.1. Parameter estimation from a raw F_0 contour

The first problem involves estimating the Fujisaki model parameters from a raw F_0 contour. Here the raw F_0 contour refers to F_0 data assumed to have been extracted using some F_0 detection method from a speech signal of interest. Looking back at Eq. (34), it can be seen that we have thus far implicitly assumed that we are given a set of F_0 observations on the whole sample period. However, F_0s should only exist in the voiced region, which means that generally the observable data should be “incomplete” in the sense that the F_0 data in the unvoiced regions are missing. Hence, when talking about the estimation of the Fujisaki model parameters from a raw F_0 contour, it is generally necessary to deal with such incomplete data. In a statistical framework, this can simply be viewed as a missing data imputation problem, which can be effectively dealt with using the Expectation-Maximization (EM) algorithm [7].

Let us define $\boldsymbol{y} \in \mathbb{R}^K$ as the “complete” data, which consist of the observed raw F_0 data and the missing data. If we could estimate which region was the missing region in advance, for example by using a voicing determination algorithm such as the one incorporated in YIN [8], the relationship between the observed raw F_0 data, $\boldsymbol{y}_{\rm obs} \in \mathbb{R}^{K'}$ such that $K' \leq K$, and the complete data, $\boldsymbol{y} \in \mathbb{R}^K$, could be written explicitly as $\boldsymbol{y}_{\rm obs} = \boldsymbol{M}\boldsymbol{y}$, where $\boldsymbol{M}$ is a K'-by-K binary matrix that has exactly one entry 1 in each row and 0’s elsewhere. Owing to space limitations, we omit all the mathematical details and only provide the procedure needed in practice: in the expectation step, the complete data $\boldsymbol{y}$ (strictly speaking, the expectation of the complete data) are updated according to

$$\boldsymbol{y} \leftarrow \boldsymbol{\mu} + \boldsymbol{\Sigma}\boldsymbol{M}^{\rm T}(\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^{\rm T})^{-1}(\boldsymbol{y}_{\rm obs} - \boldsymbol{M}\boldsymbol{\mu}), \quad (36)$$

and in the maximization step, the model parameters are updated using the data hypothetically completed in the expectation step

$$\Theta \leftarrow \underset{\Theta}{\rm argmax} \log P(\boldsymbol{y}|\Theta)P(\Theta), \quad (37)$$

where $P(\boldsymbol{y}|\Theta)$ and $P(\Theta)$ are defined as Eqs. (34) and (35), respectively. Though trivial, when $\boldsymbol{M} = \boldsymbol{I}$ for example, which means there is no missing region, Eq. (36) becomes $\boldsymbol{y} \leftarrow \boldsymbol{y}_{\rm obs}$, implying that we can treat the raw F_0 data as is as the complete data.

5.2. Proposed model as an F_0 contour prior

Given some statistical model of speech features (e.g., waveform, spectrum or others) which takes F_0 values as free parameters, let us consider the problem of estimating the unknown parameters of the given model by incorporating the present statistical F_0 contour model as a prior distribution over the F_0 parameters. In this case the Fujisaki model parameters play the role of hyperparameters in the entire system under analysis.

Suppose that we are given a sequence of observed speech features on the whole sample period, $\mathcal{D} = \{d_k\}_{k=1}^K$, which is assumed to be generated according to a statistical model $P(\mathcal{D}|\boldsymbol{y}, \Xi)$, where $\boldsymbol{y}$ denotes the F_0 parameters and Ξ contains all other parameters independent of $\boldsymbol{y}$. By incorporating Eq. (34) as a prior distribution over $\boldsymbol{y}$, and assuming $\mathcal{D}$ is not directly dependent on the hyperparameter Θ, the Maximum A Posteriori (MAP) estimation problem can be formalized as

$$\{\hat{\boldsymbol{y}}, \hat{\Theta}, \hat{\Xi}\} = \underset{\boldsymbol{y},\Theta,\Xi}{\rm argmax} \log P(\mathcal{D}|\boldsymbol{y}, \Xi)P(\boldsymbol{y}|\Theta)P(\Theta)P(\Xi).$$

Although it may depend on how $P(\mathcal{D}|\boldsymbol{y}, \Xi)$ is defined, in many cases we need to employ an iterative method in which each iteration comprises the following conditional maximization steps:

$$\begin{aligned}\Xi &\leftarrow \underset{\Xi}{\rm argmax} \log P(\mathcal{D}|\boldsymbol{y}, \Xi)P(\Xi), & (38)\\ \boldsymbol{y} &\leftarrow \underset{\boldsymbol{y}}{\rm argmax} \log P(\mathcal{D}|\boldsymbol{y}, \Xi)P(\boldsymbol{y}|\Theta), & (39)\\ \Theta &\leftarrow \underset{\Theta}{\rm argmax} \log P(\boldsymbol{y}|\Theta)P(\Theta). & (40)\end{aligned}$$

Equation (39) updates the F_0 contour estimate such that the statistical speech model best explains the observed speech features under the prior distribution defined by the Fujisaki model parameters obtained at the previous iteration. Eq. (40) then updates the Fujisaki model parameters such that the present statistical F_0 contour model best explains the updated F_0 contour estimate $\boldsymbol{y}$.

6. Parameter optimization process

We can notice that the algorithms in either of the two examples shown in the previous section consist of performing the same optimization step, that is, Eqs. (37) and (40). 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 a set consisting of the phrase, accent and base components, $\boldsymbol{x} := (\boldsymbol{y}_{\rm p}^{\rm T}, \boldsymbol{y}_{\rm a}^{\rm T}, \boldsymbol{y}_{\rm b}^{\rm T})^{\rm T}$, as the complete data, this problem can be viewed as an incomplete data problem, which can be dealt with again using the EM algorithm. In this case, the log-likelihood of Θ given the complete data is given as

$$\log P(\boldsymbol{x}|\Theta) \overset{\mathrm{c}}{=} \frac{1}{2}\log|\boldsymbol{\Lambda}^{-1}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{m}),$$

$$\boldsymbol{x} := \begin{bmatrix}\boldsymbol{y}_{\mathrm{p}}\\ \boldsymbol{y}_{\mathrm{a}}\\ \boldsymbol{y}_{\mathrm{b}}\end{bmatrix},\ \boldsymbol{m} := \begin{bmatrix}\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}}\\ \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}}\\ \mu_{\mathrm{b}}\boldsymbol{1}\end{bmatrix}, \tag{41}$$

$$\boldsymbol{\Lambda}^{-1} := \begin{bmatrix}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\end{bmatrix}.$$

Taking the conditional expectation of Eq. (41) with respect to $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\Theta = \Theta'$, and then adding $\log P(\Theta)$ to both sides, we obtain the Q function

$$Q(\Theta,\Theta') \overset{\mathrm{c}}{=} \frac{1}{2}\Big[\log|\boldsymbol{\Lambda}^{-1}| - \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']) + 2\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta'] - \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\Big] + \log P(\Theta). \tag{42}$$

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

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{m}), \tag{43}$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathrm{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} + \mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]^{\mathrm{T}}. \tag{44}$$

The expectation step computes $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']$ according to Eqs. (43) and (44) by substituting the current parameter estimate into Θ'.

Now if we let $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']$ be partitioned into three $K\times 1$ blocks and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta']$ into nine $K\times K$ blocks such that

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta'] = \begin{bmatrix}\bar{\boldsymbol{x}}_{\mathrm{p}}\\ \bar{\boldsymbol{x}}_{\mathrm{a}}\\ \bar{\boldsymbol{x}}_{\mathrm{b}}\end{bmatrix},\ \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}|\boldsymbol{y};\Theta'] = \begin{bmatrix}\boldsymbol{R}_{\mathrm{p}} & * & *\\ * & \boldsymbol{R}_{\mathrm{a}} & *\\ * & * & \boldsymbol{R}_{\mathrm{b}}\end{bmatrix}, \tag{45}$$

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

$$\begin{aligned}Q(\Theta,\Theta') \overset{\mathrm{c}}{=} \frac{1}{2}\Big[&\log|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A}| + \log|\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B}| + \log|\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}|\\ &- \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A}\boldsymbol{R}_{\mathrm{p}}) + 2\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}\\ &- \mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B}\boldsymbol{R}_{\mathrm{a}}) + 2\boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}\\ &- \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\boldsymbol{R}_{\mathrm{b}}) + 2\boldsymbol{\mu}_{\mathrm{b}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\bar{\boldsymbol{x}}_{\mathrm{b}}\\ &- \boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} - \boldsymbol{\mu}_{\mathrm{b}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\boldsymbol{\mu}_{\mathrm{b}}\Big]\\ &+ \log P(\Theta).\end{aligned} \tag{46}$$

The update formula for each parameter in the maximization step can be derived using Eq. (46).

1) State sequence $s_1,\cdots,s_K$: Leaving only the terms in $Q(\Theta,\Theta')$ that depend on $s := \{s_k\}_{k=1}^{K}$, we have

$$\mathcal{I}_1(s) := -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{o}[k]-\boldsymbol{c}_{s_k}[k])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k]-\boldsymbol{c}_{s_k}[k]) + \log P(s_1) + \sum_{k=2}^{K}\log P(s_k|s_{k-1}), \tag{47}$$

where $\boldsymbol{o}[k] := ([\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k, [\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k)^{\mathrm{T}}$. Here the notation $[\cdot]_k$ is used to denote the k-th element of a vector. The state sequence $\{s_k\}_{k=1}^{K}$ maximizing $\mathcal{I}_1(s)$ can be solved efficiently using the Viterbi algorithm as follows. We first set $\delta_1(i)$ at

$$\delta_1(i) = -\frac{1}{2}(\boldsymbol{o}[1]-\boldsymbol{c}_i[1])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[1]-\boldsymbol{c}_i[1]). \tag{48}$$

for all the hidden states i. We can then compute $\delta_k(i)$ for $k = 2,\cdots,K$ recursively via

$$\delta_k(i) = \max_{i'}\Big[\delta_{k-1}(i') - \frac{1}{2}(\boldsymbol{o}[k]-\boldsymbol{c}_i[k])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k]-\boldsymbol{c}_i[k]) + \phi_{i',i}\Big]. \tag{49}$$

The most likely transition for each state should be registered at each recursion $\Psi_k(i) = \mathrm{argmax}_{i'}[\delta_{k-1}(i') + \phi_{i',i}]$, so that the most likely state sequence can be traced at the end of the recursion with $s_{k-1} = \Psi_k(s_k)$ $(k = K,\cdots,2)$, where $s_K = \mathrm{argmax}_i\,\delta_K(i)$. Substituting the updated state sequence $\{s_k\}$ into Eq. (21), we finally obtain the updated $\boldsymbol{\mu}_{\mathrm{p}}$ and $\boldsymbol{\mu}_{\mathrm{a}}$.

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

$$A_{\mathrm{p}}[k] = [\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k\ \ (k\in\mathcal{T}_{\mathrm{p}_1}),\ \ \mathcal{T}_{\mathrm{p}_1} = \{k|s_k = \mathrm{p}_1\}. \tag{50}$$

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

$$A_{\mathrm{a}}^{(n)} = \frac{1}{|\mathcal{T}_{\mathrm{a}_n}|}\sum_{k\in\mathcal{T}_{\mathrm{a}_n}}[\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k,\ \ \mathcal{T}_{\mathrm{a}_n} = \{k|s_k = \mathrm{a}_n\}. \tag{51}$$

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

$$\psi \sim \mathcal{N}(\mu_\psi, 1/\nu_\psi^2). \tag{52}$$

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

$$\mathcal{I}_2(\psi) = \log|\boldsymbol{A}| - \frac{1}{2}\mathrm{tr}\big(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A}\boldsymbol{R}_{\mathrm{p}}\big) + \boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A}\bar{x}_{\mathrm{p}} - \frac{1}{2}\nu_\psi^2(\psi-\mu_\psi)^2. \tag{53}$$

Now, let

$$\boldsymbol{U}_2 := \begin{bmatrix}1 & & & & O\\ -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ O & & 1 & -2 & 1\end{bmatrix},\ \boldsymbol{U}_1 := \begin{bmatrix}0 & & & & O\\ 2 & 0 & & & \\ -2 & 2 & 0 & & \\ & \ddots & \ddots & \ddots & \\ O & & -2 & 2 & 0\end{bmatrix},$$

$$\boldsymbol{U}_0 := \begin{bmatrix}0 & & & & O\\ 0 & 0 & & & \\ 1 & 0 & 0 & & \\ & \ddots & \ddots & \ddots & \\ O & & 1 & 0 & 0\end{bmatrix}, \tag{54}$$

then from Eqs. (8)–(10), $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{U}_2\psi^2 + \boldsymbol{U}_1\psi + \boldsymbol{U}_0. \tag{55}$$

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

$$\begin{aligned}&2\mathrm{tr}(\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2\boldsymbol{R}_{\mathrm{p}})\psi^4+3\mathrm{tr}(\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_1\boldsymbol{R}_{\mathrm{p}})\psi^3\\&+\{\mathrm{tr}((2\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_0+\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1)\boldsymbol{R}_{\mathrm{p}})-2\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{U}_2\bar{\boldsymbol{x}}_{\mathrm{p}}+\sigma_{\mathrm{p}}^2\nu_\psi^2\}\psi^2\\&+\{\mathrm{tr}(\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_0\boldsymbol{R}_{\mathrm{p}})-\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{U}_1\bar{\boldsymbol{x}}_{\mathrm{p}}-2\sigma_{\mathrm{p}}^2\nu_\psi^2\mu_\psi\}\psi-2K\sigma_{\mathrm{p}}^2,\end{aligned}$$

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) Mean of base component μ_{b}: $Q(\Theta,\Theta')$ is maximized with respect to μ_{b} when $\mu_{\mathrm{b}}=\boldsymbol{1}^{\mathrm{T}}\bar{\boldsymbol{x}}_{\mathrm{b}}/K$.

7) Variances of additive noises σ_{p}^2, σ_{a}^2 and σ_{b}^2: $Q(\Theta,\Theta')$ is maximized with respect to σ_{p}^2, σ_{a}^2 and σ_{b}^2 when

$$\sigma_{\mathrm{p}}^2=\big(\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{R}_{\mathrm{p}})-2\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}+\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{p}}\big)/K, \quad (56)$$

$$\sigma_{\mathrm{a}}^2=\big(\mathrm{tr}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}_{\mathrm{a}})-2\boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}+\boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\mu}_{\mathrm{a}}\big)/K, \quad (57)$$

$$\sigma_{\mathrm{b}}^2=\big(\mathrm{tr}(\boldsymbol{R}_{\mathrm{b}})-2\mu_{\mathrm{b}}\boldsymbol{1}^{\mathrm{T}}\bar{\boldsymbol{x}}_{\mathrm{b}}\big)/K+\mu_{\mathrm{b}}^2. \quad (58)$$

7. Experiment

Before applying our model to practical problems such as those introduced in Section 5, we focus solely on testing the behavior of the optimization algorithm presented in Section 6. This is important because the devised algorithm is simply a local search algorithm, which is not guaranteed to solve the global optimum, and how seriously the local optima can affect the result of the parameter estimation is not yet understood. For this reason, we only present the result of a numerical analysis that we ran on artificial F_0 contour data created using the original (continuous-time) Fujisaki model. There are two points that we want to investigate using this experiment. One is the possibility of avoiding undesirable solutions that may be caused by local optima. The other is the influence of the approximation errors originating from the discretization.

The test data, depicted in Fig. 3, were created with the following settings. We set the total duration at 10s, the sampling period at 10ms, $\alpha=3$, $\beta=20$, $y_{\mathrm{b}}=4$, $T_{0,1}=1$, $T_{0,2}=3$, $T_{0,3}=4.5$, $T_{0,4}=6.5$, $A_{\mathrm{p},1}=0.7$, $A_{\mathrm{p},2}=0.6$, $A_{\mathrm{p},3}=0.3$, $A_{\mathrm{p},4}=0.4$, $T_{1,1}=1.5$, $T_{2,1}=1.8$, $A_{\mathrm{a},1}=0.5$, $T_{1,2}=2.1$, $T_{2,2}$=2.5, $A_{\mathrm{a},2}$=0.4, $T_{1,3}$=3.4, $T_{2,3}$=3.8, $A_{\mathrm{a},3}$=0.6, $T_{1,4}$=5.0, $T_{2,4}$=5.5, $A_{\mathrm{a},4}$=0.3, $T_{1,5}$=6.8, $T_{2,5}$=7.2, $A_{\mathrm{a},5}$=0.6, $T_{1,6}$=7.5, $T_{2,6}$=7.9, and $A_{\mathrm{a},6}$=0.3. The conditions for the present algorithm were as follows. The iteration was run for 10 iterations. N was set at 10, so that the total number of hidden states was 13. The state transition probabilities were set respectively at $\phi_{\mathrm{p}_0,\mathrm{p}_0}=\log(0.999)$, $\phi_{\mathrm{p}_0,\mathrm{p}_1}=\log(0.001)$, $\phi_{\mathrm{p}_1,\mathrm{a}_0}=\log(1.0)$, $\phi_{\mathrm{a}_0,\mathrm{a}_0}$=log(0.999), $\phi_{\mathrm{a}_n,\mathrm{a}_0}$=log(0.001), $\phi_{\mathrm{a}_0,\mathrm{a}_n}$=log(0.0001), $\phi_{\mathrm{a}_n,\mathrm{a}_n}$=log(0.899), $\phi_{\mathrm{a}_n,\mathrm{p}_0}$=log(0.1), with $1\leq n\leq 10$.

The parameter estimates of $\boldsymbol{\mu}$, $\boldsymbol{\mu}_{\mathrm{p}}$ and $\boldsymbol{\mu}_{\mathrm{a}}$ obtained using the present algorithm are shown in Fig. 4. By comparing Fig. 4 with Fig. 3, one can confirm that the parameter estimates are extremely close to the correct values. Similar results were obtained when using initial parameters that were randomized differently. This fact indicates that we can solve the optimization problem given by Eq. (37) or Eq. (40) quite satisfactorily with the present algorithm, which strongly motivates us to deal with the practical problems introduced in Section 5. Another important conclusion drawn from the results was that the approximation error due to the discretization did not appear to be a crucial matter.

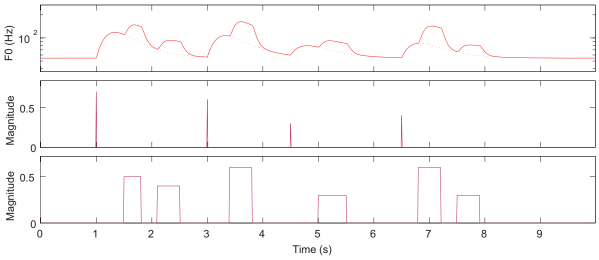

Figure 3: Test data: F_0 contour in solid line and phrase component in dotted line (top), phrase commands (middle), and accent commands (bottom).

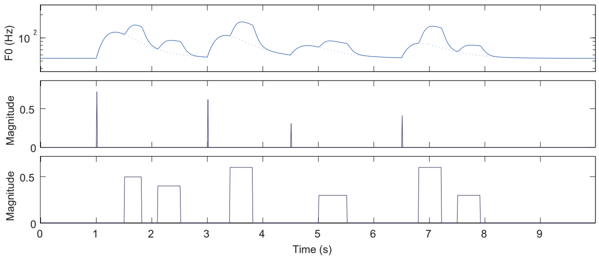

Figure 4: Parameter estimates: $\boldsymbol{\mu}$ in solid line and $\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}}$+$\boldsymbol{\mu}_{\mathrm{b}}$ in dotted line (top), $\boldsymbol{\mu}_{\mathrm{p}}$ (middle), and $\boldsymbol{\mu}_{\mathrm{a}}$ (bottom).

8. Conclusion

This paper proposed a statistical model of speech F_0 contours, based on the formulation of a discrete-time stochastic process version of the Fujisaki model. Some examples of practical problems into which the proposed model could be incorporated were mentioned. A parameter estimation framework for the proposed model based on the EM approach was derived, and the behavior of the devised algorithm was tested on artificially created data.

9. References

[1] H. Fujisaki, S. Nagashima, “A model for synthesis of pitch contours of connected speech,” *Annual Report of Engineering Research Institute, University of Tokyo*, Vol. 28, pp. 53–60, 1969.

[2] H. Fujisaki, “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*, (O. Fujimura, ed.) Raven Press, pp. 347–355, 1988.

[3] H. Fujisaki, 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.

[4] H. Mixdorf, “A novel approach to the fully automatic extraction of Fujisaki model parameters,” *Proc. Intl. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vol. 3, pp. 1281–1284, 2000.

[5] S. Narusawa, N. Minematsu, K. Hirose, H. Fujisaki, “A method for automatic extraction of model parameters from fundamental frequency contours of speech,” *Proc. Intl. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vol. 1, pp. 509–512, 2002.

[6] Y. Ohishi, H. Kameoka, K. Kashino, K. Takeda, “Parameter estimation method of F0 control model for singing voices,” *Proc. Intl. Conf. Spoken Language Process. (ICSLP)*, pp. 139–142, 2008.

[7] A.P. Dempster, N.M. Laird, D.B. Rubin, “Maximum likelihood from incomplete aata via the EM algorithm,” *J. of Royal StatisticalSociety Series B*, Vol. 39, pp. 1–38, 1977.

[8] A. de Cheveigné, H. Kawahara, “YIN, a fundamental frequency estimator for speech and music,” *J. Acoust. Soc. Am.*, Vol. 111, No. 4, pp. 1917–1930, 2002.