STATISTICAL APPROACH TO FUJISAKI-MODEL PARAMETER ESTIMATION FROM SPEECH SIGNALS AND ITS QUANTITATIVE EVALUATION

Kota Yoshizato[†], Hirokazu Kameoka[†‡], Daisuke Saito[†], Shigeki Sagayama[†],

[†]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[‡] NTT Communication Science Laboratories, NTT Corporation, Japan
{yoshizato,kameoka,dsaito,sagayama}@hil.t.u-tokyo.ac.jp

Abstract

We have previously proposed a statistical model of speech F_0 contours, which is based on the discrete-time version of the Fujisaki model. One advantage of this model is that it allows us to introduce statistical methods to learn the Fujisaki-model parameters from speech F_0 contours. This paper proposes several modifications to our previous model and parameter inference algorithm, and quantitatively evaluates the performance of our modified parameter inference algorithm.

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

1. Introduction

The fundamental frequency (F_0) contours in speech contain various types of non-linguistic information such as the speaker's identity, emotion and level of attention. They also indicate intonation in pitch accent languages. Modeling F_0 contours of speech utterances is therefore potentially useful for many speech applications, in particular speech synthesis.

Thanks to the increasing availability of speech databases, speech synthesis systems based on statistical models such as hidden Markov models (HMMs) have attracted particular attention in recent years. Since unnatural F_0 contours result in a synthesis that sounds "emotionless" to human listeners, one of the primary challenges in speech synthesis technology is to create a natural-sounding F_0 contour for the utterance as a whole. The weakness of the current statistical text-to-speech systems is that they do not satisfactorily represent the macroscopic variations of F_0s in natural speech.

The Fujisaki model [1] is a well-founded mathematical model, which describes the process by which the whole F_0 contour of a speech utterance is generated. The notable feature of the Fujisaki model is that it consists of physiologically and physically meaningful parameters (called phrase and accent commands) and is able to fit F_0 contours of real speech well when they are chosen appropriately. Thus, one way of enabling statistical speech synthesizers to generate natural sounding F_0 contours would be to incorporate the Fujisaki model into the statistical model so that its parameters can be learned from a speech corpus in a unified manner. However, the Fujisaki model has an analytically complex form, making it difficult to incorporate it into statistical speech synthesis systems as is.

To this end, we have recently introduced a statistical model of the discrete-time version of the Fujisaki model [2]. This model makes the best use of powerful statistical methods such as the HMM and the Expectation-Maximization (EM) algorithm to estimate the Fujisaki model parameters. The aim of this paper is to (1) make several modifications to our previous model and parameter estimation algorithm, and (2) confirm the performance of our model through quantitative evaluations of its parameter estimation accuracy.

The rest of this paper is organized as follows. Section 2 briefly reviews the original Fujisaki model. Section 3 reviews our previously introduced model, i.e., a discrete counterpart of the Fujisaki model and its statistical model formulation. Section 4 presents an algorithm for estimating the Fujisaki model parameters from observed real speech F_0 contour data. Section 5 shows results of a quantitative evaluation obtained by conducting an experiment using real speech data excerpted from the ATR speech database. Section 6 concludes this paper.

2. Original Fujisaki Model

The Fujisaki model [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 $y_{\rm p}(t)$, an accent component $y_{\rm a}(t)$, and a base component $y_{\rm b}$. The phrase component consists of the major-scale pitch variations over the duration of the prosodic units, and 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 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). The baseline component 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. The log F_0 contour, $y(t)$, is thus expressed as

$$y(t) = y_{\rm p}(t) + y_{\rm a}(t) + y_b, \tag{1}$$

where

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

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

$$y_{\rm a}(t) = G_{\rm a}(t) * u_{\rm a}(t), \tag{4}$$

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

$*$ denotes convolution over time. α and β are natural angular frequencies of the two second-order systems, which are known

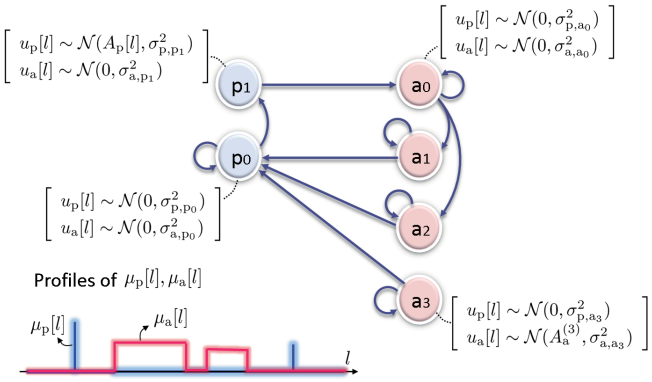

Figure 1: *Command function modeling with HMM.*

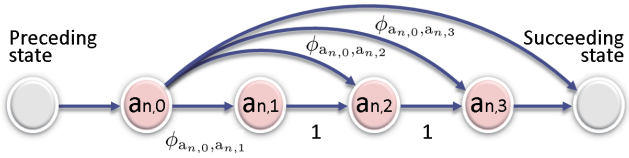

Figure 2: *The splitting of state* a_n *into 4 substates* $a_{n,0}$, $a_{n,1}$, $a_{n,2}$, *and* $a_{n,3}$. $\phi_{a_{n,0},a_{n,1}}$ *corresponds to the probability of staying at state* a_n *with 4 consecutive times.*

to be almost constant within an utterance as well as across utterances for a particular speaker. It has been shown that $\alpha = 3$ rad/s and $\beta = 20$ rad/s can be used as default values.

3. Statistical Speech F_0 Contour Model

This section describes a statistical model based on a discretized version of the Fujisaki model.

3.1. Discretization

To obtain a discrete-time version of the Fujisaki model, we apply a backward difference s-to-z transform to the phrase and accent control mechanisms. With this discretization, the relationships between the filter inputs and outputs are given as

$$u_p[k] = a_0 y_p[k] + a_1 y_p[k-1] + a_2 y_p[k-2], \quad (6)$$

$$a_2 = (\psi - 1)^2,\ a_1 = -2\psi(\psi - 1),\ a_0 = \psi^2, \quad (7)$$

$$u_a[k] = b_0 y_a[k] + b_1 y_a[k-1] + b_2 y_a[k-2], \quad (8)$$

$$b_2 = (\varphi - 1)^2,\ b_1 = -2\varphi(\varphi - 1),\ b_0 = \varphi^2, \quad (9)$$

where k is the discrete time index, $y_p[k]$, $u_p[k]$, $y_a[k]$ and $u_a[k]$ are the discrete-time versions of the phrase component, phrase command function, accent component and accent command function, respectively, $\psi = 1 + 1/(\alpha t_0)$, $\varphi = 1 + 1/(\beta t_0)$, and t_0 is the sampling period of the discrete-time representation.

3.2. Statistical formulation

In the original Fujisaki model, the phrase commands and accent commands are assumed to consist of Dirac deltas and rectangular pulses, respectively. In addition, they are not allowed to overlap each other. To incorporate these requirements, we find it convenient to model the $u_p[k]$ and $u_a[k]$ pair, i.e., $\boldsymbol{o}[k] = (u_p[k], u_a[k])^\mathsf{T}$, using a hidden Markov model (HMM). Specifically, we assume that $\{\boldsymbol{o}[k]\}_{k=1}^K$ is a sequence of outputs generated from an HMM with the specific topology illustrated in Figure 1. The output distribution of each state is a Gaussian distribution

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

$$\boldsymbol{\nu}[k] = \begin{bmatrix} \mu_p[k] \\ \mu_a[k] \end{bmatrix},\ \boldsymbol{\Upsilon}[k] = \begin{bmatrix} \upsilon^2_p[k] & 0 \\ 0 & \upsilon^2_a[k] \end{bmatrix}, \quad (11)$$

where the mean vector $\boldsymbol{\nu}[k]$ and variance matrix $\boldsymbol{\Upsilon}[k]$ are considered to evolve in time as a result of the state transition.

To parameterize the durations of the state transitions, each state is split into a certain number of substates such that they all have exactly the same emission densities. Figure 2 shows an example of the splitting of state a_n. The number of substates is set at a sufficiently large value and the transition probability from substate $a_{n,l}$ to substate $a_{n,l+1}$ is set at 1 for $l \neq 0$. This state splitting allows us to flexibly control the durations for which the process stays in state a_n through the settings of the transition probability. The transition probability from substate $a_{n,0}$ to substate $a_{n,l}$ $(l \geq 1)$ corresponds to the probability of the present HMM generating a rectangular pulse that has a particular duration. In the same way, we split states p_0 and a_0 to parameterize the probability of the spacing between phrase and accent commands. Henceforth, we use the notation $p_0 = \{p_{0,0}, p_{0,1}, \ldots\}$, $a_0 = \{a_{0,0}, a_{0,1}, \ldots\}$, and $a_n = \{a_{n,0}, a_{n,1}, \ldots\}$. The state splitting described above is one of the modifications we have made to our previous model [2]. The present HMM is now defined as follows:

Output sequence: $\{\boldsymbol{o}[k]\}_{k=1}^K$
Set of states: $\mathcal{S} = \{p_0, p_1, a_0, \ldots, a_N\}$
State sequence: $\{s_k\}_{k=1}^K$
Output distribution: $P(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k])$

$$\boldsymbol{\nu}[k] = \begin{cases} (0,0)^\mathsf{T} & (i \in p_0, a_0) \\ (A_p[k], 0)^\mathsf{T} & (i = p_1) \\ (0, A_a^{(n)})^\mathsf{T} & (i \in a_n) \end{cases}$$

$$\boldsymbol{\Upsilon}[k] = \begin{bmatrix} \sigma^2_{p,i} & 0 \\ 0 & \sigma^2_{a,i} \end{bmatrix}$$

Transition probability: $\phi_{i',i} = \log P(s_k = i | s_{k-1} = i')$

We further assume that the baseline component is normally distributed with mean μ_b, $y_b[k] \sim \mathcal{N}\left(\mu_b, \upsilon^2_b\right)$.

For real speech F_0 contours, we must take account of the uncertainty in the observed F_0 data, since observed data should not always be considered reliable. For example, F_0 estimates in unvoiced regions would be unreliable. Thus, we consider $y[k]$ to be the superposition of a latent component corresponding to the true F_0 and a noise component such that $y_n[k] \sim \mathcal{N}\left(0, \upsilon^2_n[k]\right)$. The variance $\upsilon^2_n[k]$ corresponds to the degree of uncertainty of the observed F_0 at time k. Therefore, $y[k] = y_p[k] + y_a[k] + y_b[k] + y_n[k]$.

Now, let us define

$$\begin{aligned} \boldsymbol{u}_p &= (u_p[1], \ldots, u_p[K])^\mathsf{T}, & \boldsymbol{u}_a &= (u_a[1], \ldots, u_a[K])^\mathsf{T}, \\ \boldsymbol{\mu}_p &= (\mu_p[1], \ldots, \mu_p[K])^\mathsf{T}, & \boldsymbol{\mu}_a &= (\mu_a[1], \ldots, \mu_a[K])^\mathsf{T}, \\ \boldsymbol{y}_p &= (y_p[1], \ldots, y_p[K])^\mathsf{T}, & \boldsymbol{y}_a &= (y_a[1], \ldots, y_a[K])^\mathsf{T}, \\ \boldsymbol{y}_b &= (y_b[1], \ldots, y_b[K])^\mathsf{T}, & \boldsymbol{y}_n &= (y_n[1], \ldots, y_n[K])^\mathsf{T}, \\ \boldsymbol{y} &= (y[1], \cdots, y[K])^\mathsf{T}. \end{aligned}$$

Then, we can write $\boldsymbol{u}_p$ and $\boldsymbol{u}_a$ as

$$\boldsymbol{u}_p = \boldsymbol{A}\boldsymbol{y}_p, \quad (12)$$

$$\boldsymbol{u}_a = \boldsymbol{B}\boldsymbol{y}_a, \quad (13)$$

where

$$\boldsymbol{A}=\begin{bmatrix} a_0 & & O \\ a_1 & a_0 & \\ a_2 & a_1 & a_0 \\ & \ddots & \ddots & \ddots \\ O & & a_2 & a_1 & a_0 \end{bmatrix},\quad \boldsymbol{B}=\begin{bmatrix} b_0 & & O \\ b_1 & b_0 & \\ b_2 & b_1 & a_0 \\ & \ddots & \ddots & \ddots \\ O & & b_2 & b_1 & b_0 \end{bmatrix}. \tag{14}$$

For simplicity, we treat $\phi_{i',i}$, μ_{b} $\sigma^2_{\mathrm{p},i}$, $\sigma^2_{\mathrm{a},i}$, υ^2_{b}, $\upsilon^2_{\mathrm{n}}[k]$, α, β as constants and $\Theta=\{\{A_{\mathrm{p}}[k],s[k]\}_{k=1}^{K},\{A_{\mathrm{a}}^{(n)}\}_{n=1}^{N}\}$ as the free parameters to be estimated. To sum up, the likelihood function of the Fujisaki model parameters Θ given $\boldsymbol{y}$ is given as

$$\begin{aligned} P(\boldsymbol{y}|\Theta) &= \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{K/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}, \\ \boldsymbol{\mu} &= \boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}}+\boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}}+\mu_{\mathrm{b}}\boldsymbol{1}, \\ \boldsymbol{\Sigma} &= \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}\left(\boldsymbol{A}^{\mathsf{T}}\right)^{-1}+\boldsymbol{B}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}\left(\boldsymbol{B}^{\mathsf{T}}\right)^{-1}+\boldsymbol{\Sigma}_{\mathrm{b}}+\boldsymbol{\Sigma}_{\mathrm{n}}, \end{aligned} \tag{15}$$

where

$$\boldsymbol{\Sigma}_{\mathrm{p}}=\begin{bmatrix}\upsilon^2_{\mathrm{p}}[1] & & O \\ & \ddots & \\ O & & \upsilon^2_{\mathrm{p}}[K]\end{bmatrix},\quad \boldsymbol{\Sigma}_{\mathrm{a}}=\begin{bmatrix}\upsilon^2_{\mathrm{a}}[1] & & O \\ & \ddots & \\ O & & \upsilon^2_{\mathrm{a}}[K]\end{bmatrix},$$
$$\boldsymbol{\Sigma}_{\mathrm{b}}=\begin{bmatrix}\upsilon^2_{\mathrm{b}} & & O \\ & \ddots & \\ O & & \upsilon^2_{\mathrm{b}}\end{bmatrix},\quad \boldsymbol{\Sigma}_{\mathrm{n}}=\begin{bmatrix}\upsilon^2_{\mathrm{n}}[1] & & O \\ & \ddots & \\ O & & \upsilon^2_{\mathrm{n}}[K]\end{bmatrix}. \tag{16}$$

4. Parameter Optimization Process

We present here an iterative algorithm that searches for the unknown parameters Θ by locally maximizing $P(\boldsymbol{y}|\Theta)$ given $\boldsymbol{y}$. By regarding $\boldsymbol{x}=(\boldsymbol{y}_{\mathrm{p}}^{\mathsf{T}},\boldsymbol{y}_{\mathrm{a}}^{\mathsf{T}},\boldsymbol{y}_{\mathrm{b}}^{\mathsf{T}},\boldsymbol{y}_{\mathrm{n}}^{\mathsf{T}})^{\mathsf{T}}$ as the complete data this problem can be viewed as an incomplete data problem, which can be dealt with using the EM algorithm. The log-likelihood function of Θ given $\boldsymbol{x}$ is written as

$$\begin{aligned} \log P(\boldsymbol{x}|\Theta) &\stackrel{\mathrm{c}}{=} \frac{1}{2}\log|\boldsymbol{\Lambda}^{-1}|-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{m}), \\ \boldsymbol{x} &= \begin{bmatrix}\boldsymbol{y}_{\mathrm{p}}\\ \boldsymbol{y}_{\mathrm{a}}\\ \boldsymbol{y}_{\mathrm{b}}\\ \boldsymbol{y}_{\mathrm{n}}\end{bmatrix},\ \boldsymbol{m}=\begin{bmatrix}\boldsymbol{A}^{-1}\boldsymbol{\mu}_{\mathrm{p}}\\ \boldsymbol{B}^{-1}\boldsymbol{\mu}_{\mathrm{a}}\\ \mu_{\mathrm{b}}\boldsymbol{1}\\ \boldsymbol{0}\end{bmatrix}, \\ \boldsymbol{\Lambda}^{-1} &= \begin{bmatrix}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{A} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{B} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{b}}^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{n}}^{-1}\end{bmatrix}. \end{aligned} \tag{17}$$

The auxiliary function is thus given by

$$\begin{aligned} Q(\Theta,\Theta') &\stackrel{\mathrm{c}}{=} \frac{1}{2}\Big[\log|\boldsymbol{\Lambda}^{-1}|-\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{y};\Theta']) \\ &\quad +2\boldsymbol{m}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta']-\boldsymbol{m}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\Big]+\log P(\Theta). \end{aligned} \tag{18}$$

$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{y};\Theta]$ are given explicitly as

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]=\boldsymbol{m}+\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{m}), \tag{19}$$

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\boldsymbol{y};\Theta] &= \boldsymbol{\Lambda}-\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} \\ &\quad +\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]^{\mathsf{T}}, \end{aligned} \tag{20}$$

by using the relationship $\boldsymbol{y}=\boldsymbol{H}\boldsymbol{x}$, where $\boldsymbol{H}=[\boldsymbol{I},\boldsymbol{I},\boldsymbol{I},\boldsymbol{I}]$. These are the values to be updated at the E step. Let $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]$ be partitioned into four $K\times 1$ blocks such that $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y};\Theta]=(\bar{\boldsymbol{x}}_{\mathrm{p}}^{\mathsf{T}},\bar{\boldsymbol{x}}_{\mathrm{a}}^{\mathsf{T}},\bar{\boldsymbol{x}}_{\mathrm{b}}^{\mathsf{T}},\bar{\boldsymbol{x}}_{\mathrm{n}}^{\mathsf{T}})^{\mathsf{T}}$.

The M step update formulas are given as follows.

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

$$\begin{aligned} \mathcal{I}(s) &:= -\frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{o}[k]-\boldsymbol{\nu}[k])^{\mathsf{T}}\boldsymbol{\Upsilon}[k]^{-1}(\boldsymbol{o}[k]-\boldsymbol{\nu}[k]) \\ &\quad +\log P(s_1)+\sum_{k=2}^{K}\log P(s_k|s_{k-1}), \end{aligned} \tag{21}$$

where $\boldsymbol{o}[k]:=([\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k,[\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k)^{\mathsf{T}}$. 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 $\mathcal{I}(s)$ can be solved efficiently using the Viterbi algorithm.

2) Magnitudes of phrase and accent commands: $Q(\Theta,\Theta')$ is maximized with respect to $A_{\mathrm{p}}[k]$ and $A_{\mathrm{a}}^{(n)}$ when

$$A_{\mathrm{p}}[k]=[\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k\ \ (k\in\mathcal{T}_{\mathrm{p}_1}),\ \ \mathcal{T}_{\mathrm{p}_1}=\{k|s_k=\mathrm{p}_1\}, \tag{22}$$

$$A_{\mathrm{a}}^{(n)}=\frac{1}{|\mathcal{T}_{\mathrm{a}_n}|}\sum_{k\in\mathcal{T}_{\mathrm{a}_n}}[\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k,\ \ \mathcal{T}_{\mathrm{a}_n}=\{k|s_k=\mathrm{a}_n\}. \tag{23}$$

The M step described above involves the computation of $\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ and $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$. This amounts to computing the filter inputs $\bar{\boldsymbol{u}}_{\mathrm{p}}$ and $\bar{\boldsymbol{u}}_{\mathrm{a}}$ from the estimates of the phrase and accent components, $\bar{\boldsymbol{x}}_{\mathrm{p}}$ and $\bar{\boldsymbol{x}}_{\mathrm{a}}$, using the "inverse filtering" matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Note that as $\bar{\boldsymbol{u}}_{\mathrm{p}}$ and $\bar{\boldsymbol{u}}_{\mathrm{a}}$ correspond to the estimates of command functions, they should be non-negative. However, the computations of $\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ and $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$ allow each element of the resulting vector to have a negative value. To effectively avoid allowing such unwanted estimates, we consider finding $\bar{\boldsymbol{u}}_{\mathrm{p}}$ and $\bar{\boldsymbol{u}}_{\mathrm{a}}$ that minimize $\|\boldsymbol{A}^{-1}\bar{\boldsymbol{u}}_{\mathrm{p}}-\bar{\boldsymbol{x}}_{\mathrm{p}}\|_2^2$ and $\|\boldsymbol{B}^{-1}\bar{\boldsymbol{u}}_{\mathrm{a}}-\bar{\boldsymbol{x}}_{\mathrm{a}}\|_2^2$ subject to non-negativity, instead of simply computing $\boldsymbol{A}\bar{\boldsymbol{x}}_{\mathrm{p}}$ and $\boldsymbol{B}\bar{\boldsymbol{x}}_{\mathrm{a}}$. In elementwise notation, they can be written as $\sum_k|G_{\mathrm{p}}[k]*\bar{u}_{\mathrm{p}}[k]-\bar{x}_{\mathrm{p}}[k]|^2$ and $\sum_k|G_{\mathrm{a}}[k]*\bar{u}_{\mathrm{a}}[k]-\bar{x}_{\mathrm{a}}[k]|^2$, where $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$ are the discrete-time versions of the impulse responses of the phrase and accent control mechanisms. Fortunately, this non-negative deconvolution problem can be solved efficiently by employing the method described in [3].

5. Experiment

One important contribution of our work is that the Fujisaki model has successfully been translated into a statistical model. We believe that this will open the door to combining our model and the HMM-based text-to-speech synthesis system [4] so that the Fujisaki-model parameters as well as the spectral parameter sequences can be learned from a speech corpus in a unified manner. In this regard, our model is already superior to conventional "non-statistical" methods such as [5]. However, it is not yet clear whether our statistical model is able to estimate the Fujisaki model parameters from real speech data as accurately as the state-of-the-art technique [5]. Thus, we quantitatively evaluated the parameter estimation accuracy of the present algorithm using real speech data, excerpted from the ATR Japanese speech database B-set [6]. This database consists of 503 phonetically balanced sentences. We selected speech samples of one male speaker (MHT). We used Fujisaki model parameters that had been manually annotated by an expert in the field of speech prosody as the ground truth data, where the baseline component was set constant at $\log 60$ Hz.

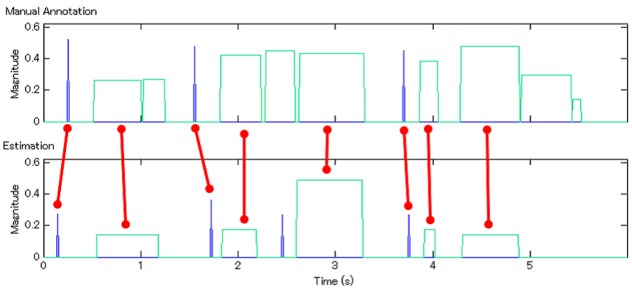

Figure 3: *An example of command sequence matching.*

Table 1: *Accuracy rate and Error rate* (S=0.3s).

All Commands	A	E_{I}	E_{D}
Init	0.688	0.088	0.224
Estimated	0.697	0.127	0.177

Phrase Commands	A	E_{I}	E_{D}
Init	0.647	0.109	0.244
Estimated	0.680	0.207	0.112

Accent Commands	A	E_{I}	E_{D}
Init	0.711	0.076	0.213
Estimated	0.708	0.083	0.207

F_0 contours were extracted using the method described in [7], from which the Fujisaki model parameters were estimated using the present algorithm. The constant parameters were fixed respectively at $N = 20$, $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $v_{\mathrm{p}}^2[k] = 0.2^2$, $v_{\mathrm{a}}^2[k] = 0.02^2$, $v_{\mathrm{b}}^2 = 0.001^2$, $v_{\mathrm{n}}^2[k] = 10^{15}$ for unvoiced regions and $v_{\mathrm{n}}^2[k] = 10^{-15}$ 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 the method described in [5]. The EM algorithm was then run for 20 iterations. The number of substates in the present 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 parameter estimation accuracy in the following manner: We performed matching between the estimated and ground truth command sequences as illustrated in Figure 3 on a command-by-command basis using dynamic programming. 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 the offsets of the estimated and ground truth accent commands. 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 baseline F_0 value, which was set differently in the present method and in the manual annotation. Let N_{E}, N_{A} be the total numbers of commands in the estimated and ground truth command sequences, N_{M} be the number of the matched commands between the two sequences, N_{Esum}, N_{Asum}, and N_{Msum} be the sum of N_{E}, N_{A}, N_{M} for all 303 sentences. We defined the insertion error rate E_{I} as $(N_{\mathrm{Esum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, the deletion error rate E_{D} as $(N_{\mathrm{Asum}} - N_{\mathrm{Msum}})/N_{\mathrm{Asum}}$, and the accuracy rate A as $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$. Table 1 shows the result of our quantitative evaluation with $S = 0.3$ s. The top, middle, and bottom tables show the accuracy and error rates of the phrase and accent commands, the phrase commands alone, and the accent commands alone, respectively. The "Init" row shows the accuracy and error rates of the initial command sequence (which was obtained with the method described in [5]), and the "Estimated" row shows that of the estimated command sequence after the EM iterations. As the result shows, our method performed slightly better than the state-of-the-art technique [5].

6. Conclusion

In this paper, we made several modifications to our previously reported model of speech F_0 contours and parameter estimation algorithm. We evaluated the parameter estimation accuracy of the present method using real speech data. Future work will include incorporating the present model into the HMM-based speech synthesis system (HTS) [4] in such a way that the Fujisaki-model parameters can be learned from a speech corpus in a unified manner. We are also currently exploring a reasonable model for the F_0 contours of a singing voice [8].

7. Acknowledgement

We thank Prof. Keikichi Hirose of the University of Tokyo, who kindly provided us with the manually annotated ground truth data associated with the ATR speech samples.

8. References

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

[2] H. Kameoka, J. L. Roux, and Y. Ohishi, "A statistical model of speech F_0 contours," in *Proc. ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, 2010, pp. 43–48.

[3] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *In Proc. 2008 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2009)*, 2009, pp. 45–48.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. 2000 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2000)*, 2000, pp. 1315–1318.

[5] 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. 2002 IEEE Intl. Conf. Acoust., Speech, Signal Process. (ICASSP 2002)*, 2002, pp. 509–512.

[6] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[7] H. Kameoka, "Statistical speech spectrum model incorporating all-pole vocal tract model and F_0 contour generating process model," in *Technical report of the Institute of Electronics, Information and Communication Engineers (IEICE)*, 2010, in Japanese.

[8] Y. Ohishi, H. Kameoka, D. Mochihashi, H. Nagano, and K. Kashino, "Statistical modeling of F_0 dynamics in singing voices based on Gaussian processes with multiple oscillation bases," in *Proc. 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, 2010, pp. 2598–2601.