Hidden Markov Convolutive Mixture Model for Pitch Contour Analysis of Speech

Kota Yoshizato[1], Hirokazu Kameoka[1,2], Daisuke Saito[1], Shigeki Sagayama[1],

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

Abstract

This paper proposes a stochastic model of speech F_0 contours, based on the stochastic formulation of the Fujisaki model. Our motivation for the stochastic formulation is twofold. Firstly, it allows us to derive a well-behaved algorithm for estimating the Fujisaki model parameters from a raw F_0 contour. Secondly, it will open the door to incorporating the well-founded F_0 contour model into various statistical speech processing problems. We quantitatively evaluated the performance of our method in terms of an Fujisaki-model parameter estimation accuracy using real speech data. Experimental results revealed that our method was superior to a state-of-the-art Fujisaki model parameter extractor.

Index Terms: speech F0 contours, statistical model, Fujisaki model, hidden Markov model, EM algorithm

1. Introduction

The fundamental frequency (F_0) contours in normal speech contains various types of non-linguistic information such as speaker's identity, emotion and attention. Modeling the F_0 contours of speech utterances can thus be potentially very useful for many speech applications, including speech recognition, speaker recognition, speech synthesis, and dialogue systems.

The Fujisaki model [1] is a well-founded mathematical model, which describes the generating process of the whole F_0 contour of a speech utterance. The remarkable feature of the Fujisaki model is that it consists of physiologically and physically meaningful parameters (called the phrase and accent commands) and is able to fit F_0 contours of real speech well when they are chosen appropriately. For this reason, the Fujisaki model has often been used to manually design an F_0 contour for synthesizing natural speech. Thus, to enable speech synthesizers to automatically generate natural-sounding F_0 contours, one way would be to incorporate the Fujisaki model into the generative model of phonemic sequences so that its parameters can be learned from a speech corpus in a unified manner. However, estimating (learning) the Fujisaki model parameters from raw F_0 contour observations has been a difficult task. Several techniques have already been developed [2, 3, 4], for the purpose of incorporating the extracted parameters in automatic speech/emotion recognition systems to improve their performance, but so far with limited success due to the analytical complexity of the Fujisaki model.

We have previously derived a stochastic model of speech F_0 contours by translating the Fujisaki model into a probabilistic generative model [5, 6]. The stochastic reformulation of the Fujisaki model has allowed us to derive an efficient algorithm for automatically estimating the Fujisaki-model parameters from raw F_0 contours. In this paper, we present an improved version of our previous models described in [5, 6]. The rest of this paper is organized as follows. Section 2 briefly reviews the original Fujisaki model. Section 3 formulates a probabilistic generative model of speech F_0 contours based on the Fujisaki model. Section 4 derives an algorithm for finding the maximum a posteriori (MAP) estimates of the Fujisaki model parameters from an observed F_0 contour. Section 5 presents results of a quantitative evaluation conducted 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 $y_{\rm b}$:

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

The phrase component $y_{\rm p}(t)$ consists of the major-scale pitch variations over the duration of the prosodic units, and the accent component $y_{\rm a}(t)$ 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):

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

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

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

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

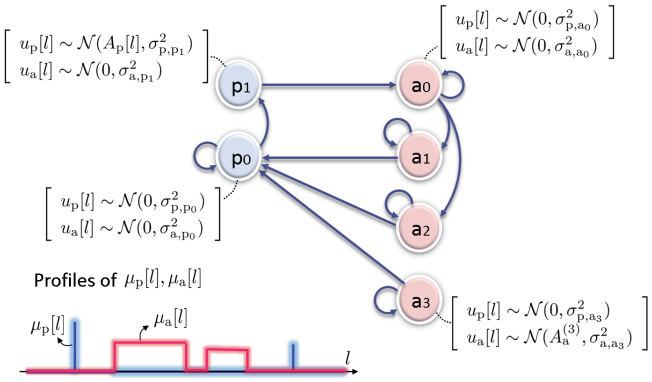


Figure 1: *Command function modeling with HMM.*

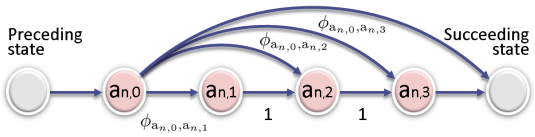


Figure 2: *The splitting of state* a_n *into 4 substates* $\mathrm{a}_{n,0}$, $\mathrm{a}_{n,1}$, $\mathrm{a}_{n,2}$, *and* $\mathrm{a}_{n,3}$. $\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,1}}$ *corresponds to the probability of staying at state* a_n *with 4 consecutive times.*

where $*$ denotes convolution over time. The baseline component $y_{\rm b}$ 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.

3. Stochastic model of speech F_0 contours

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

We first describe the process for generating the phrase and accent command functions, $u_{\rm p}[k]$ and $u_{\rm a}[k]$, where k denotes the discrete-time index. In the original Fujisaki model, it is required that the phrase commands must consist of Dirac deltas and the accent commands must consist of rectangular pulses. In addition, they are not allowed to overlap each other. To incorporate these requirements, we find it convenient to model the $u_{\rm p}[k]$ and $u_{\rm a}[k]$ pair, i.e., $\boldsymbol{o}[k] = (u_{\rm p}[k], u_{\rm a}[k])^{\mathsf{T}}$, using a Hidden Markov Model (HMM). Specifically, we assume that $\{\boldsymbol{o}[k]\}_{k=1}^K$ is a sequence of outputs emitted from an HMM with the specific topology illustrated in Fig. 1. The output distribution of each state is a Gaussian distribution

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k]\right), \tag{6}$$

$$\boldsymbol{\nu}[k] = \begin{bmatrix} \mu_{\rm p}[k] \\ \mu_{\rm a}[k] \end{bmatrix}, \quad \boldsymbol{\Upsilon}[k] = \begin{bmatrix} v^2_{\rm p}[k] & 0 \\ 0 & v^2_{\rm a}[k] \end{bmatrix}, \tag{7}$$

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 self transitions, each state is split into a certain number of substates such that they all have exactly the same emission densities. Fig. 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 $\mathrm{a}_{n,m}$ to substate $\mathrm{a}_{n,m+1}$ is set at 1 for $m \neq 0$. This state splitting allows us to flexibly control the durations for which the process stays in state a_n through the settings of the transition probability. The transition probability from substate $\mathrm{a}_{n,0}$ to substate $\mathrm{a}_{n,m}$ $(m \geq 1)$ corresponds to the probability of the present HMM generating a rectangular pulse that has a particular duration. In the same way, we split states p_0 and a_0 to parameterize the probability of the spacing between phrase and accent commands. Henceforth, we use the notation $\mathrm{p}_0 = \{\mathrm{p}_{0,0}, \mathrm{p}_{0,1}, \ldots\}$, $\mathrm{a}_0 = \{\mathrm{a}_{0,0}, \mathrm{a}_{0,1}, \ldots\}$, and $\mathrm{a}_n = \{\mathrm{a}_{n,0}, \mathrm{a}_{n,1}, \ldots\}$.

The present HMM is now defined as follows:

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

Given the state sequence $\boldsymbol{s} = \{s_k\}_{k=1}^K$, the above HMM generates the $u_{\rm p}[k]$ and $u_{\rm a}[k]$ pair. From (2) and (4), $u_{\rm p}[k]$ and $u_{\rm a}[k]$ are then fed through different critically damped filters, $G_{\rm p}[k]$ and $G_{\rm a}[k]$, to generate the phrase and accent components, $x_{\rm p}[k]$ and $x_{\rm a}[k]$:

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

$$x_{\rm a}[k] = u_{\rm a}[k] * G_{\rm a}[k], \tag{9}$$

where $*$ denotes convolution over k. The entire F_0 contour is then given by

$$x[k] = x_{\rm p}[k] + x_{\rm a}[k] + u_{\rm b}, \tag{10}$$

where $u_{\rm b}$ denotes the baseline component. In non-tonal languages such as standard Japanese, the pharse and accent commands should be non-negative. In our previous model [5, 6], we treated $u_{\rm p}[k]$, $u_{\rm a}[k]$ and $u_{\rm b}$ as latent variables (i.e., parameters to be marginalized out), and did not explicitly take the non-negativity constraints on $u_{\rm p}[k]$ and $u_{\rm a}[k]$ into consideration. To explicitly impose the non-negativity constraints, it is convenient to treat these variables as model parameters instead of latent variables.

For real speech F_0 contours, observed F_0s should not always be considered reliable. For example, F_0 estimates obtained with a pitch extractor in unvoiced regions would be totally unreliable. When performing parameter inference, we would want to trust only reliable observations and neglect unreliable ones. To incorporate the degree of uncertainty of F_0 observations, we consider modeling an observed F_0 contour $y[k]$ as a superposition

of the "ideal" F_0 contour $x[k]$ and a noise component $x_{\mathrm{n}}[k] \sim \mathcal{N}\left(0, \upsilon_{\mathrm{n}}^2[k]\right)$, where $\upsilon_{\mathrm{n}}^2[k]$ represents the degree of uncertainty of the F_0 observation at time k, which is assumed to be given.

Overall, an observed F_0 contour $y[k]$ is described as $y[k] = x[k] + x_{\mathrm{n}}[k]$. For simplicity, we henceforth treat $\phi_{i',i}$, u_{b}, $\sigma_{\mathrm{p},i}^2$, $\sigma_{\mathrm{a},i}^2$, $\upsilon_{\mathrm{n}}^2[k]$, α, and β as constants. By marginalizing $x_{\mathrm{n}}[k]$ out, we obtain the probability density function of $\boldsymbol{y} = \{y[k]\}_{k=1}^K$, given $\boldsymbol{o} = \{\boldsymbol{o}[k]\}_{k=1}^K$, as

$$P(\boldsymbol{y}|\boldsymbol{o}) = \prod_{k=1}^{K} \mathcal{N}(y[k]; x[k], \upsilon_{\mathrm{n}}^2[k]),$$
$$x[k] = G_{\mathrm{p}}[k] * u_{\mathrm{p}}[k] + G_{\mathrm{a}}[k] * u_{\mathrm{a}}[k] + u_{\mathrm{b}}. \quad (11)$$

Recall from (6) that given a state sequence $\boldsymbol{s} = \{s_k\}_{k=1}^K$ and $\boldsymbol{\theta} = \{\{A_{\mathrm{p}}[k]\}_{k=1}^K, \{A_{\mathrm{a}}^{(n)}\}_{n=1}^N, \}$, $\boldsymbol{o}$ is generated according to $P(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{\theta}) = \prod_{k=1}^K \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{\nu}[k], \boldsymbol{\Upsilon}[k])$. $P(\boldsymbol{s})$ is given by the product of the state transition probabilities: $P(\boldsymbol{s}) = \phi_{s_1} \prod_{k=2}^K \phi_{s_k, s_{k-1}}$. Furthermore, we assume that $\boldsymbol{\theta}$ is uniformly distributed.

4. Parameter Optimization Process

In this section, we describe an iterative algorithm that searches for the maximum a posteriori estimates of $\boldsymbol{o}$ and $\boldsymbol{\theta}$ by locally maximizing $P(\boldsymbol{o}, \boldsymbol{\theta}|\boldsymbol{y})$ given $\boldsymbol{y}$ using the generalized Expectation-Maximization (EM) algorithm. We treat $\boldsymbol{s}$ as a latent variable and consider marginalizing $P(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{o})P(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{\theta})P(\boldsymbol{s})$ with respect to $\boldsymbol{s}$ to obtain the objective $P(\boldsymbol{o}, \boldsymbol{\theta}|\boldsymbol{y})$. The auxiliary function (as known as the "Q-function") can be written as

$$Q(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{o}', \boldsymbol{\theta}') = \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{o}', \boldsymbol{\theta}') \log P(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{s}|\boldsymbol{y})$$
$$\stackrel{c}{=} \log P(\boldsymbol{y}|\boldsymbol{o}) + \sum_{\boldsymbol{s}} P(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{o}', \boldsymbol{\theta}') \log P(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{\theta})P(\boldsymbol{s}),$$

where $\stackrel{c}{=}$ denotes equality up to constant terms. An iterative algorithm that consists of computing $P(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{o}', \boldsymbol{\theta}')$ (via the Forward-Backward algorithm), increasing $Q(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{o}', \boldsymbol{\theta}')$ with respect to $\boldsymbol{o}$ and $\boldsymbol{\theta}$, and then substituting $\boldsymbol{o}$ and $\boldsymbol{\theta}$ into $\boldsymbol{o}'$ and $\boldsymbol{\theta}'$ locally maximizes the posterior $P(\boldsymbol{o}, \boldsymbol{\theta}|\boldsymbol{y})$. Here, care must be taken that increasing $Q(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{o}', \boldsymbol{\theta}')$ with respect to $\boldsymbol{o}$ must be performed subject to non-negativity. This can be done by invoking the idea of [7]. By using the Jensen's inequality we obtain an inequality

$$-\Big(\sum_{i\in\{\mathrm{p,a,b}\}} \sum_{l} G_i[k-l]u_i[l]\Big)^2$$
$$\geq -\sum_{i\in\{\mathrm{p,a,b}\}} \sum_{l} \frac{G_i^2[k-l]u_i^2[l]}{\lambda_{i,k,l}}, \quad (12)$$

where $G_{\mathrm{b}}[k] = \delta[k]$ (Kronecker's delta), $\lambda_{i,k,l} \geq 0$ is an auxiliary variable satisfying $\sum_i \sum_l \lambda_{i,k,l} = 1$. We can use this inequality to construct a lower bound function

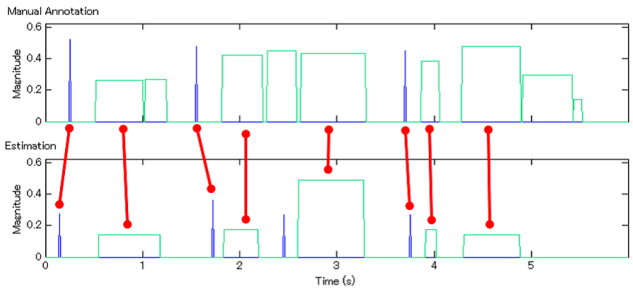


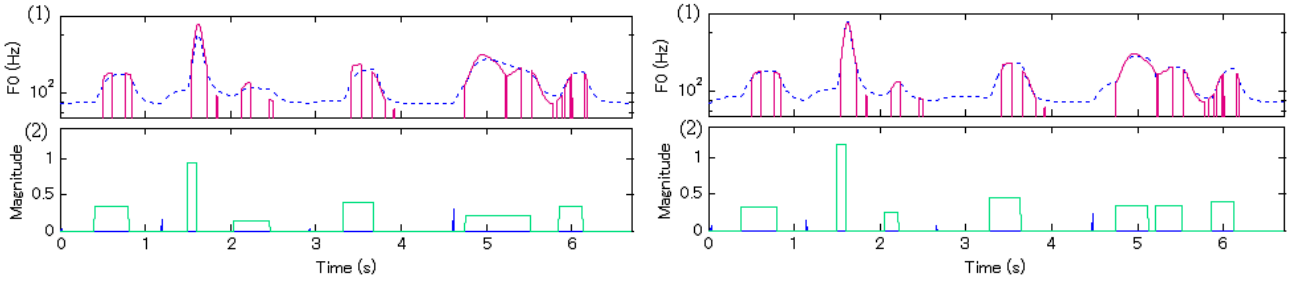
Figure 3: *An example of command sequence matching.*

for $Q(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{o}', \boldsymbol{\theta}')$. The maximization of this lower bound function w.r.t. $\boldsymbol{o}$ (subject to non-negativity) and λ can be achieved analytically, which guarantees a certain increase of $Q(\boldsymbol{o}, \boldsymbol{\theta}, \boldsymbol{o}', \boldsymbol{\theta}')$.

After convergence, we search for the optimal state sequence $\boldsymbol{s}$ by using the Viterbi algorithm.

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 various statistical speech applications 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 [4]. 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. 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 [8]. 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 at $\log 60$ Hz. F_0 contours were extracted using a method we had previously developed [9], from which the Fujisaki model parameters were estimated using the present algorithm. The constant parameters were fixed respectively at $N = 10$, $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $\upsilon_{\mathrm{p}}^2[k] = 0.2^2$, $\upsilon_{\mathrm{a}}^2[k] = 0.1^2$, $\upsilon_{\mathrm{b}}^2 = 0.001^2$, $\upsilon_{\mathrm{n}}^2[k] = 10^{15}$ for unvoiced regions and $\upsilon_{\mathrm{n}}^2[k] = 0.2^2$ 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 the non-statistical method [4]. 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: $\log F_0$ RMSE (root mean squared error) and detection rates. Our aim was to confirm whether the present model and algorithm can achieve high model reconstruction accuracy



(a): The previous model (b): The proposed model

Figure 4: *(1) An observed F_0 contour in voiced regions (in solid line) and the estimated F_0 contours (in dotted line) along with (2) the estimated phrase and accent commands.*

Table 1: *Detection rates and* $\log F_0$ *RMSE* (S=0.3s).

	Detection rates	$\log F_0$[Hz] RMSE
Init	0.688	0.1719
Estimated	0.695	0.0611

while keeping the meaningfulness of the model parameters. $\log F_0$ RMSE was used to evaluate the reconstruction accuracy, which measures the root mean squared error between an observed F_0 contour and the estimated F_0 contour. The detection rate was used to evaluate the meaningfulness of the parameter estimates, which was calculated in the following way: We performed matching between the estimated and ground truth command sequences as illustrated in Fig. 3 on a command-by-command basis by using an 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 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 detection rate D as $1 - E_{\mathrm{I}} - E_{\mathrm{D}}$.

Tab. 1 shows the result of our quantitative evaluation with $S = 0.3$ s. The "Init" row shows the detection rate and $\log F_0$ RMSE of the initial command sequence (which was obtained with the non-statistical method [4]), and the "Estimated" row shows that of the estimated command sequence after the EM iterations. From the results, we confirmed that our method was comparable to a state-of-the-art Fujisaki model extractor in terms of the detection rate. On the other hand, our method was superior to the conventional method in terms of the model reconstruction accuracy. We can also confirm from Fig. 4 that the present model is able to fit an observed F_0 contour more accurately than our previous model.

6. Conclusion

In this paper, we proposed a statistical model of speech F_0 contours and parameter estimation algorithm. We evaluated the parameter estimation accuracy of the proposed method using real speech data, and confirm the advantage of the proposed method. Future work will include incorporating the present model into the statistical speech applications such as the HMM-based speech synthesis system (HTS) in such a way that the Fujisaki-model parameters can be learned from a speech corpus in a unified manner.

7. References

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

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

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

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

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

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

[7] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. ICASSP*, 2009, pp. 45–48.

[8] 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.

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