FAST ALGORITHM FOR STATISTICAL PHRASE/ACCENT COMMAND ESTIMATION BASED ON GENERATIVE MODEL INCORPORATING SPECTRAL FEATURES

Ryotaro Sato[1], *Hirokazu Kameoka*[2], *Kunio Kashino*[2]

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

ABSTRACT

An important challenge in speech processing involves extracting non-linguistic information from a fundamental frequency (F_0) contour of speech. We propose a fast algorithm for estimating the model parameters of the Fujisaki model, namely, the timings and magnitudes of the phrase and accent commands. Although a powerful parameter estimation framework based on a stochastic counterpart of the Fujisaki model has recently been proposed, it still had room for improvement in terms of both computational efficiency and parameter estimation accuracy. This paper describes our two contributions. First, we propose a hard expectation-maximization (EM) algorithm for parameter inference where the E step of the conventional EM algorithm is replaced with a point estimation procedure to accelerate the estimation process. Second, to improve the parameter estimation accuracy, we add a generative process of a spectral feature sequence to the generative model. This makes it possible to use linguistic or phonological information as an additional clue to estimate the timings of the accent commands. The experiments confirmed that the present algorithm was approximately 16 times faster and estimated parameters about 3% more accurately than the conventional algorithm.

Index Terms— voice fundamental frequency contour, Fujisaki model, prosodic information processing, EM algorithm

1. INTRODUCTION

Various speech recognition and synthesis systems have become widespread in our daily lives, and the importance of non-linguistic information in spoken communication has been highlighted in recent years [1]. Voice fundamental frequency (F_0) contours constitute one of the most important physical entities correlated with socially essential information, such as the speaker's identity, intention, attitude, and mood.

An F_0 contour typically consists of relatively long-term frequency variations over the duration of a phrase and short-term variations such as those found in accented syllables. The Fujisaki model [2] describes a log F_0 contour as the sum of these two components, which are called phrase and accent components, respectively. These components are modeled as the responses of critically damped second-order linear systems to impulse and step inputs, called commands. The basic underlying parameters consist of the timing and magnitudes of the impulse inputs (phrase commands) and the onset/offset times and magnitudes of the step inputs (accent commands).

The Fujisaki model allows us to approximate actual F_0 contours fairly well using a relatively small number of parameters when the parameters are appropriately chosen and has been applied to various languages in multiple application areas. Although the parameters are generally difficult to estimate automatically from a raw F_0 contour [3], a powerful parameter estimation framework based on a stochastic counterpart of the Fujisaki model has recently been proposed [4]. This method has been reported to yield reasonably good results by taking advantages of powerful statistical inference and training techniques for the estimation. However, when we consider the various applications that must be run in real time or faster under realistic circumstances, its computational efficiency and accuracy need to be further improved.

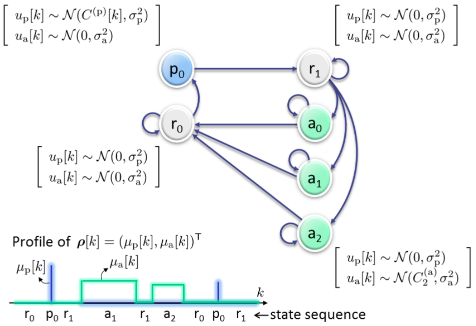
$\begin{bmatrix} u_\mathrm{p}[k] \sim \mathcal{N}(C^{(\mathrm{p})}[k], \sigma_\mathrm{p}^2) \\ u_\mathrm{a}[k] \sim \mathcal{N}(0, \sigma_\mathrm{a}^2) \end{bmatrix}$
$\begin{bmatrix} u_\mathrm{p}[k] \sim \mathcal{N}(0, \sigma_\mathrm{p}^2) \\ u_\mathrm{a}[k] \sim \mathcal{N}(0, \sigma_\mathrm{a}^2) \end{bmatrix}$
p0
r1
r0
a0
$\begin{bmatrix} u_\mathrm{p}[k] \sim \mathcal{N}(0, \sigma_\mathrm{p}^2) \\ u_\mathrm{a}[k] \sim \mathcal{N}(0, \sigma_\mathrm{a}^2) \end{bmatrix}$
a1
a2
Profile of $\boldsymbol{\rho}[k] = (\mu_\mathrm{p}[k], \mu_\mathrm{a}[k])^\mathsf{T}$
$\mu_\mathrm{p}[k]$
$\mu_\mathrm{a}[k]$
$\begin{bmatrix} u_\mathrm{p}[k] \sim \mathcal{N}(0, \sigma_\mathrm{p}^2) \\ u_\mathrm{a}[k] \sim \mathcal{N}(C_2^{(\mathrm{a})}, \sigma_\mathrm{a}^2) \end{bmatrix}$
k
r0 p0 r1 a1 r1 a2 r0 p0 r1 ←state sequence

Fig. 1. HMM generating phrase/accent command functions.

On the basis of the generative model described above, we propose an efficient algorithm for parameter inference where the E step of the conventional EM algorithm is replaced with a point estimation procedure. We also propose using spectral information as an additional clue to accurately estimate the parameters, assuming that accents appearing in F_0 contours are related to linguistic or phonological information. In the rest of this paper, Section 2 describes the addressed problem, and Section 3 introduces the proposed algorithm. Section 4 reports the experimental results, and Section 5 concludes this paper.

2. GENERATIVE MODEL APPROACH TO F_0 PARAMETER ESTIMATION

2.1. Generative Model of F_0 Command Functions

Several researchers have addressed the automatic parameter estimation of the Fujisaki model [3]. Our method is based on the generative model approach introduced by Kameoka et al. [4]. We briefly review its framework.

The Fujisaki model assumes that an F_0 contour $y[t]$ on a loga-

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

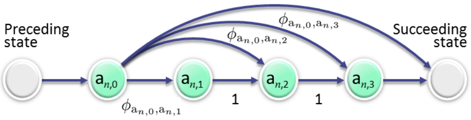
Preceding state
$\mathrm{a}_{n,0}$
$\mathrm{a}_{n,1}$
$\mathrm{a}_{n,2}$
$\mathrm{a}_{n,3}$
Succeeding state
$\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,3}}$
$\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,2}}$
$\phi_{\mathrm{a}_{n,0},\mathrm{a}_{n,1}}$
1
1

Fig. 2. State splitting to flexibly assign probabilities to state durations.

rithmic scale is given as the sum of three components:

$$y[k] = x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + \mu_{\mathrm{b}}, \tag{1}$$

where $x_{\mathrm{p}}[k]$ and $x_{\mathrm{a}}[k]$ respectively are a phrase component and an accent component at time frame k, and μ_{b} is a constant value. The phrase and accent components are assumed to be the outputs of different second-order critically damped filters, $G_{\mathrm{p}}[k]$ and $G_{\mathrm{a}}[k]$, excited with pulse sequence $u_{\mathrm{p}}[k]$ (phrase commands) and rectangular pulse sequence $u_{\mathrm{a}}[k]$ (accent commands), respectively:

$$x_{\mathrm{p}}[k] = G_{\mathrm{p}}[k] * u_{\mathrm{p}}[k], \tag{2}$$

$$x_{\mathrm{a}}[k] = G_{\mathrm{a}}[k] * u_{\mathrm{a}}[k], \tag{3}$$

where $*$ is convolution over time. The key idea of Kameoka's approach [4] is that a phrase/accent command pair sequence is modeled as an output sequence of the path-restricted hidden Markov model (HMM) with Gaussian state emission densities shown in Fig. 1.

Here the states p_0, $\mathrm{a}_n(n = 0, \cdots, N-1)$, and $\mathrm{r}_i(i = 0, 1)$, respectively, are states where only the phrase command is being activated, i.e., $\mu_{\mathrm{p}}[k] = \boldsymbol{E}[u_{\mathrm{p}}[k]] = C^{(p)}[k], \mu_{\mathrm{a}}[k] = \boldsymbol{E}[u_{\mathrm{a}}[k]] = 0$, only the accent command is being activated, i.e., $\mu_{\mathrm{p}}[k] = 0, \mu_{\mathrm{a}}[k] = C_n^{(a)}$, and neither command is being activated, i.e., $\mu_{\mathrm{p}}[k] = 0, \mu_{\mathrm{a}}[k] = 0$. The path constraint shown in Fig. 1 restricts $\mu_{\mathrm{p}}[k] = \boldsymbol{E}[u_{\mathrm{p}}[k]]$ to consisting of isolated deltas and $\mu_{\mathrm{p}}[k] = \boldsymbol{E}[\mu_{\mathrm{a}}[k]]$ to consisting of rectangular pulses. Hence, estimating the state transition of the HMM directly amounts to estimating the Fujisaki-model parameters. Furthermore, we can directly assign probabilities to the durations of self-transition by splitting each state (except for p_0) into a certain number of sub states, as shown in Fig. 2. Given a state sequence $\boldsymbol{s} = (s_0, \ldots, s_{K-1})$, this HMM emits $\boldsymbol{o}[k] = [u_{\mathrm{p}}[k], u_{\mathrm{a}}[k]]^{\mathsf{T}}$ according to the Gaussian state emission density:

$$\boldsymbol{o}[k] \sim \mathcal{N}(\boldsymbol{o}; \boldsymbol{\rho}[k], \boldsymbol{\Upsilon}), \tag{4}$$

where $\boldsymbol{\rho}[k] = [\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k]]^{\mathsf{T}}$ and $\boldsymbol{\Upsilon} = \mathrm{diag}[\sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2]$.

2.2. Conventional Parameter Estimation Algorithm

By using the conditional density $P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}})$ reflecting the constraints (1)-(3) of the Fujisaki model, and the HMM likelihood $P(\boldsymbol{o}|\boldsymbol{s}, \theta)$, the joint probability density function is written as

$$P(\boldsymbol{s}, \theta, \mu_{\mathrm{b}}, \boldsymbol{o}, \boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{o}, \mu_{\mathrm{b}})P(\boldsymbol{o}|\boldsymbol{s}, \theta)P(\boldsymbol{s})P(\mu_{\mathrm{b}})P(\theta), \tag{5}$$

where $\theta = \left\{\{C^{(p)}[k]\}_{k=0}^{K-1}, \{C_n^{(a)}\}_{n=0}^{N-1}, \sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2\right\}$ contains all the HMM parameters except for the transition probabilities, and $\boldsymbol{s} = \{s_k\}_{k=0}^{K-1}$ represents the state sequence. $\boldsymbol{y} = \{y[k]\}_{k=0}^{K-1}$ is the observed F_0 contour. The statistical estimation is then summarized to maximize $P(\boldsymbol{o}, \theta, \mu_{\mathrm{b}}|\boldsymbol{y})$ for a given $\boldsymbol{y}$. For parameter inference, we previously proposed an algorithm based on an auxiliary function and the EM algorithm [5] by treating $\boldsymbol{s}$ as a latent variable to be marginalized out and θ and μ_{b} as the model parameters to be estimated. The procedure is as follows:

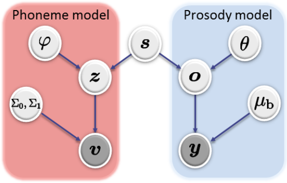

Fig. 3. Graphical representation of proposed model.

- E step: Update $P(s_k = q|\boldsymbol{y}, \theta, \mu_{\mathrm{b}}, \boldsymbol{o}) = P(s_k = q|\theta, \boldsymbol{o})$ for each frame k and each state q using the Forward-Backward algorithm.
- M step: Update $\boldsymbol{o}$, θ, μ_b using the auxiliary function.

After convergence, MAP estimation for the state sequence $\boldsymbol{s}$ is performed using the Viterbi algorithm to obtain the estimated command sequence $\boldsymbol{o}_{\mathrm{est}} = \{[\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k]]\}_{k=0}^{K-1}$.

3. PROPOSED METHOD

We propose two improvements to the conventional method described above: the use of the hard EM algorithm and the introduction of spectral features.

3.1. Modified EM Algorithm

In the method described in Sec. 2.2, our analysis revealed that more than 90% of the computation time was spent by the Forward-Backward algorithm when computing the posterior state probabilities at each frame in the E step. Each output distribution of the HMM is a Gaussian distribution, and so a logarithm and an exponential must be computed for every addition at each frame, and this is computationally expensive. To sidestep these computations, we propose using the joint probability density function $P(\boldsymbol{s}, \theta, \mu_b, \boldsymbol{o}|\boldsymbol{y})$ instead of the marginal distribution $P(\theta, \mu_b, \boldsymbol{o}|\boldsymbol{y})$ as the objective function. This results in a parameter estimation algorithm in which the Forward-Backward procedure in the conventional algorithm is replaced with a state decoding procedure using the Viterbi algorithm. Since the Viterbi algorithm is generally faster than the Forward-Backward algorithm, we expect the entire computation time to be reduced. This is called a hard EM algorithm.

The variables involved in this method are summarized at the right half of Fig. 3. The estimation process using the modified EM algorithm can be written as follows:

- E step (hard EM): Update $\boldsymbol{s}$ by maximizing $P(\boldsymbol{o}|\boldsymbol{s}, \theta)P(\boldsymbol{s})$ using the Viterbi algorithm.
- M step:
  Update $\boldsymbol{o}$, θ, μ_b (as with the conventional method).

We conjecture that this method can also improve the estimation accuracy. This can be explained in terms of the objective functions considered in the conventional and proposed algorithms. The conventional method can be interpreted as the optimization of $P(\theta, \mu_{\mathrm{b}}, \boldsymbol{o}|\boldsymbol{y})$, with $\boldsymbol{s}$ being integrated out. However, the parameters we need to infer are $\mu_{\mathrm{p}}[k]$ and $\mu_{\mathrm{a}}[k]$ for each frame k, which are uniquely determined by $(\theta, \boldsymbol{s})$. Therefore, the objective function that we should actually maximize is $P(\theta, \mu_{\mathrm{b}}, \boldsymbol{o}, \boldsymbol{s}|\boldsymbol{y})$ rather than that mentioned above. The conventional method [4] performs a MAP estimation for $\boldsymbol{s}$ only at the final step, so the finally inferred parameters $(\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k])$ are not guaranteed to correspond to the local maximum of $P(\theta, \mu_b, \boldsymbol{o}, \boldsymbol{s}|\boldsymbol{y})$.

In contrast, our method maximizes $P(\theta, \mu_b, \boldsymbol{o}, \boldsymbol{s}|\boldsymbol{y})$ directly. This objective function can be factorized as Eq. (5) (ignoring a constant factor) and so the modified E step can be interpreted as the conditional maximization of this function with respect to $\boldsymbol{s}$ while keeping the other parameters fixed.

3.2. Introduction of Spectral Features

Several researchers have addressed the use of linguistic information to automatically estimate F_0 contour parameters [6, 7, 8]. The basic idea behind these approaches is the assumption that accent command functions are somewhat related to linguistic information. More specifically, the onsets of the accent commands tend to occur when phoneme transitions occur. Since phoneme transitions are usually accompanied by rapid changes in spectral feature values, we can expect the accuracy of accent command estimation to be improved if we assume that the temporal variations of the spectral feature values are likely to be large when an accent command is activated.

Fig. 3 shows the relationships between the variables in the proposed probabilistic model comprising both spectral variations and F_0 commands.

$$\text{Phoneme transition probability} : \boldsymbol{z} = \{z[k]\}_{k=0}^{K-1} \quad (0 \le z[k] \le 1) \tag{6}$$

$$\text{Output sequence} : \boldsymbol{v} = \{\boldsymbol{v}[k]\}_{k=0}^{K-1} \tag{7}$$

$$P(z[k]|s_{k-1}, s_k) = \begin{cases} f(z[k], \varphi) & (s_{k-1}=r_1, s_k=a_n) \\ 1 & (\text{otherwise}) \end{cases} \tag{8}$$

$$\begin{aligned} &P(\boldsymbol{v}[k]|z[k], \Sigma_0, \Sigma_1) \\ &\quad = (1 - z[k])\mathcal{N}(\boldsymbol{v}[k]|\mathbf{0}, \Sigma_0) + z[k]\mathcal{N}(\boldsymbol{v}[k]|\mathbf{0}, \Sigma_1) \end{aligned} \tag{9}$$

$z[k]$ represents the probability that the kth frame corresponds to the phoneme boundary. The prior of $z[k]$ is given by the function f and the parameter φ. The distribution of the spectral feature vector $\boldsymbol{v}[k]$ is expressed by a two-component Gaussian mixture distribution where $z[k]$ is the weight. One component of this mixture is intended to express the distribution of the spectral features at phoneme boundaries and the other is intended to express that within other segments. The norm of a Δ-spectral feature vector tends to be large at a phoneme boundary, and our proposed probabilistic model is designed to take this characteristic into account.

The F_0 estimation process based on the EM algorithm is summarized as follows:

- E step (hard EM): Update $\boldsymbol{s}$ by maximizing $P(\boldsymbol{z}|\boldsymbol{s})P(\boldsymbol{o}|\boldsymbol{s}, \theta)P(\boldsymbol{s})$ using the Viterbi algorithm.
- M step:
  1. Update the variables of the prosody model $\boldsymbol{s}$, θ, μ_b (the same as the entire M step of the conventional method).
  2. For each $k \in [1, K-1]$ that satisfies $s_{k-1} = r_1$ and $s_k = a_i$, update $z[k]$ to 1.
  3. For each $k \in [0, K-1]$ where the kth frame is judged to be unvoiced, update $z[k]$ to 0.
  4. Update Σ_0, Σ_1 using the update function of the Gaussian Mixture Model $P(\boldsymbol{v}|\boldsymbol{z}, \Sigma_0, \Sigma_1)$
     (Here, we assume that Σ_0 and Σ_1 are diagonalized).
  5. Update $\boldsymbol{z}$ by maximizing $P(\boldsymbol{z}|\Sigma_0, \Sigma_1, \boldsymbol{v})$.

4. EXPERIMENTS

To evaluate the speed and accuracy of the methods, experiments were conducted using real speech data, excerpted from the ATR Japanese sentence database B-set [9]. We used 503 sentences spoken by one male speaker (MHT). For the ground truth data, we used manually annotated Fujisaki model command function parameters.

F_0 contours were estimated by using Nakatani's method [10], and the baseline values μ_b were set for each sentence by using the minimum value of F_0 in the voiced segments. The initial values of the command functions were extracted by using Narusawa's method [3].

We estimated the command functions using three methods: (1) the conventional method (C) [4], (2) the method introduced in Section 3.1 (P1), and (3) the method described in Section 3.2 (P2). For the distribution of $\boldsymbol{z}$ in P2, we used $f(z[k], a) = C\exp(az[k])$, and we empirically chose $a = 10.0$. For the spectral features, we used ΔMFCC (to 12th order), ΔLPC (to 20th order), Δpower, ΔLPC+Δpower.

Throughout the experiments, we fixed $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $N = 10$, $t_0 = 8$ ms, $v_{\rm p}^2 = 0.03^2$, $v_{\rm a}^2 = 0.03^2$, $v_{\rm n}^2 = 100^2$ for unvoiced segments and $v_{\rm n}^2 = 0.03^2$ for voiced segments. We divided the 503 manually annotated sentences into a training set and a test set; that is, the transition probabilities of the HMM were obtained in the training step using the test set comprising the first 200 sentences from the corpus, and then the evaluation test was performed using the remaining 303 sentences.

We used two criteria as measures of estimation accuracy: (1) the root mean squared error (RMSE) between the observed logF_0 and the synthesized value $G_{\rm p}[k] * \mu_{\rm p}[k] + G_{\rm a}[k] * \mu_{\rm a}[k] + \mu_{\rm b}$ using the estimated $\mu_{\rm p}$ and $\mu_{\rm a}$ and the impulse responses $G_{\rm p}[k], G_{\rm a}[k]$, over the voiced segments, and (2) the deletion and insertion rates of the phrase and accent commands. The second criterion must be used because the objective of the estimation is not only to minimize the RMSE but also to obtain a compact description of the F_0 contours. The deletion/insertion rates are defined as follows. First, we match the estimated and ground truth command sequences on a command-by-command basis with a dynamic programming algorithm. By using a predefined time difference tolerance S, the estimated commands that found a match were judged "matched." With the accent commands, we compare the mean value of the onset time difference and the offset time difference with S. Let N, $N_{\rm est}$ and $N_{\rm match}$ be the number of commands in the ground truth data, the number of commands in the estimated result, and the number of commands judged to be matched, respectively. By using these values, the deletion/insertion rates are defined as $p_{\rm del} = (N - N_{\rm match})/N$ and $p_{\rm ins} = (N_{\rm est} - N_{\rm match})/N$. The numerators in this definition mean the number of deleted and inserted commands in the estimated results compared with the ground truth data. The relationship between these rates and the detection rate $p_{\rm det}$ in [4] is written as $p_{\rm det} = 1 - \max(p_{\rm del}, p_{\rm ins})$. Note that the magnitudes of the commands are not introduced in the second criterion because the aim is to evaluate appropriateness of estimation in terms of the number of correctly estimated commands. We performed the evaluation using $S = 0.1$ sec throughout our experiments.

Fig. 4 shows the RMSE and insertion/deletion rates of command functions estimated with the conventional and proposed methods described above. There were 20 iterations in the EM algorithm in this experiment.

With all of the proposed methods, the RMSE values are improved compared with the conventional method. The total deletion/insertion rate, shown in the right panel, stands for the mean

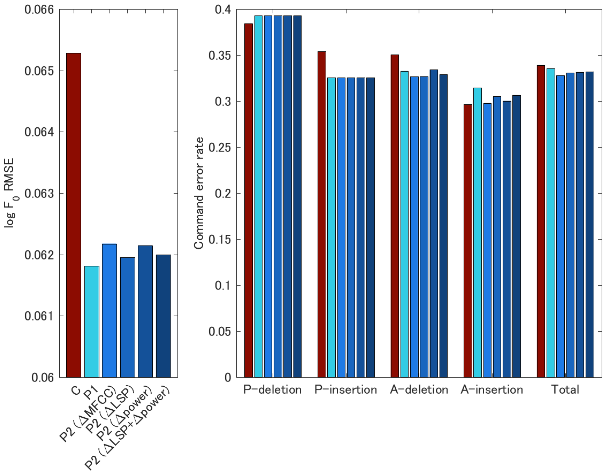

Fig. 4. RMSE and deletion/insertion rate.

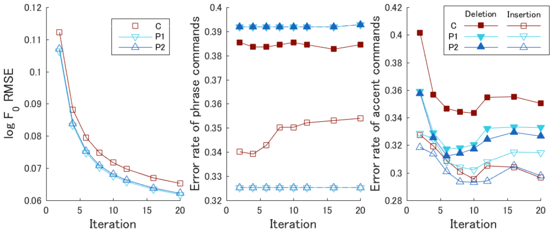

Fig. 5. Number of iterations vs. RMSE and insertion/deletion rates.

value of phrase/accent deletion and insertion rate weighted by the numbers of commands in the ground truth data. The total rates are also improved, and the technique using ΔMFCC achieved the best rate among the methods tested here, reducing the total insertion/deletion error by 3 %.

The computation times are listed in Table 1, showing that the proposed methods greatly increased the speed. For example, the computation time in E step is 70 times shorter in the hard EM method (P1) than in the conventional method (C), and the total computation time is 16 times shorter for the proposed ΔMFCC method (P2) than for C. This means that the real time factor of the estimation process is less than 1 with the proposed method (P2), if there are fewer than 36 iterations.

Fig. 5 shows the dependence of the RMSE and the insertion/deletion rates for phrase/accent commands on the number of iterations. The RMSE decreases monotonically for all algorithms. Nevertheless, the insertion/deletion rate obtains a local minimum at around eight iterations. This phenomenon is presumed to occur because our experiments were not conducted to fit the F_0 contours generated by the ground truth command functions but to fit the observed F_0 contours.

Fig. 6 shows an example of the observed F_0 contour, the ground truth of the commands, the estimated commands (P1 and P2 with ΔMFCC), and the ΔMFCC values. The dotted purple curve in the graph second from the bottom visualizes the $\boldsymbol{z}$ estimated in P2. This shows that the phonological information successfully prevents the unnecessary switching of accent commands right after $t = 3$ sec.

Table 1. Computation time for command function estimation.

Method	E step [s/iteration]	M step [s/iteration]
C [4]	1.610	0.039
P1 (hard EM)	0.023	0.042
P2 (ΔMFCC)	0.049	0.050

CPU: Core i7-6700K 4.0GHz, RAM:32GB
Windows 7 SP1, MATLAB R2016a
Length of sound data: 3.62 seconds

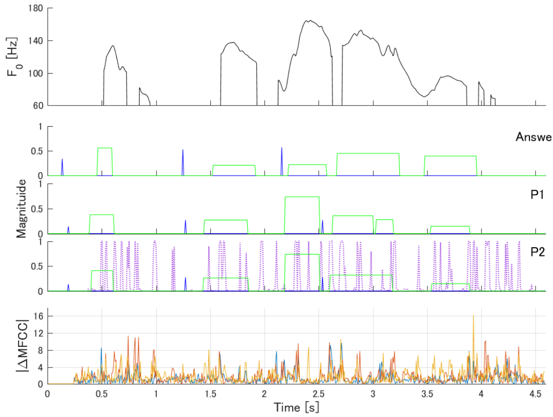

Fig. 6. Example of processing results. Purple dotted line overlapping estimated value of command functions obtained with method P2 represents $\boldsymbol{z}$ at each frame.

5. CONCLUSIONS

We described a way of improving the Fujisaki model parameter estimation algorithm proposed in [4] in terms of computational efficiency and parameter estimation accuracy. To accelerate the algorithm, we proposed replacing the E step in the conventional algorithm with a state decoding procedure using the Viterbi algorithm. To improve the estimation accuracy, we proposed adding a Δ-spectral feature generative process to the original generative model, which made it possible to estimate the F_0 commands from F_0 and spectral features. Experiments using real speech data showed that the proposed method achieved an approximately 70 times shorter computational time in Estep and an approximately 16 times shorter whole calculation time than the conventional method. The spectral features were also shown to be effective. The detection rate was the best when using ΔMFCC as the spectral features, and the number of wrong F_0 command insertions or deletions can be reduced by 3 % compared with the case where no spectral features were used. As future work, we plan a further study of the use of spectral or phonological information to improve the robustness of the estimation.

6. ACKNOWLEDGMENTS

We thank Prof. Keikichi Hirose for providing us with the manually annotated data associated with the ATR speech samples. We also thank Dr. Tomohiro Nakatani (NTT), who kindly provided us with the source code of the F_0 extractor used in the experiment.

7. REFERENCES

[1] K. Hirose, J. Tao (eds.): "Speech prosody in speech synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis," Springer (2015).

[2] H. Fujisaki, O. Fujimura (eds.): "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, Machanisms and Functions*, Raven (1988).

[3] S. Narusawa, N. Minematsu, K. Hirose, H. Fujisaki: "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Proc. (ICASSP)*, pp. 509–512 (2002).

[4] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, K. Kashino: "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Trans. on Audio, Speech and Language Proc.*, vol.23, no.6, pp.1042–1053 (2015).

[5] A. P. Dempster, N. M. Laird, D. B. Rubin: "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. soc. Ser. B*, vol.39, no.1, pp.1–38 (1977).

[6] K. Hirose, Y. Furuyama, S. Narusawa, N. Minematsu, H. Fujisaki: "Use of linguistic information for automatic extraction of F_0 contour generation process model parameters," in *Proc. Eurospeech* pp.141–144 (2003).

[7] H. M. Torres, J. A. Gurlekian, H. Mixdorff, H. Pfitzinger: "Linguistically motivated parameter estimation methods for a superpositional intonation model," *EURASIP Journal on Audio, Speech, and Music Processing* (2014).

[8] K. Hirose, Y. Furuyama, N. Minematsu: "Corpus-based extraction of F_0 contour generation process model parameters," in *Proc. Interspeech* pp. 3257–3260 (2005).

[9] A. Kurematsu, K. Takeda, Y. Sagisaka, S. katagiri, H. Kuwabara, K. Shikano: "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Comm.*, vol.9, no.4, pp.357–363 (1990).

[10] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, T. Kondo: "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Comm.*, vol.50, no.3, pp. 203-?214 (2008).