Probabilistic speech F_0 contour model incorporating statistical vocabulary model of phrase-accent command sequence

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

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

10.21437/Interspeech.2013-110

Abstract

We have previously proposed a generative model of speech F_0 contours, based on the discrete-time version of the Fujisaki model (a model of the mechanisim for controlling F_0s through laryngeal muscles). One advantage of this model is that it allows us to apply statistical methods to estimate the Fujisaki-model parameters from speech F_0 contours. This paper proposes a new generative model of speech F_0 contours incorporating a vocabulary model of intonation patterns. A parameter inference algorithm for the present model is derived. We quantitatively evaluated the performance of our parameter inference algorithm.

1. Introduction

All human speech has expression. We recognize it as part of the humanness of speech, and it is a quality listeners expect to find in daily human communication. Without expression speech sounds lifeless and artificial. Our future goal is to quantify the expression of speech in such a way that we can analyze and synthesize it based on an engineering approach.

Fundamental frequency (F_0) contours in normal speech convey various types of non-linguistic information such as speaker's identity, emotion, attitude and intention. Modeling the F_0 contours of speech utterances can thus be potentially useful for many speech applications, including emotion recognition, speaker recognition, speech synthesis, and dialogue systems, where prosodic features play an important role. F_0 contours consist of two major components: major-scale pitch variations over the duration of the prosodic units (called the *phrase component*) and smaller-scale pitch variations in accented syllables (called the *accent component*). This is justified by the fact that the thyroid cartilage involves two mutually independent types of movement with different muscular reaction times. Specifically, the phrase and accent components respectively correspond to contributions associated with the translation and rotation movements of the thyroid cartilage. The Fujisaki model [1] is a well-founded mathematical model that describes an F_0 contour as the sum of these two contributions. 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.

Since a prosodic feature in speech is predominantly characterized by the levels and timings of the phrase and accent components, one important challenge is to solve an inverse probelm of estimating the Fujisaki-model parameters automatically from a raw F_0 contour.

However, this problem has been a difficult task. Several techniques have already been developed [2, 3, 4], but so far with limited success due to the ill-posedness of the inverse problem and 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, 7]. This reformulation has successfully allowed us to derive an efficient algorithm based on the expectation-maximization (EM) algorithm for estimating the Fujisaki-model parameters from a raw F_0 contour.

A key idea in our previous model is that the sequence of the phrase and accent command pair (i.e., the parameters that we want to estimate) is modeled as a path-restricted hidden Markov model (HMM) so that estimating the state transition of the HMM directly amounts to estimating the Fujisaki-model parameters. Generally speaking, the better top-down knowledge is incorporated, the better the solution to an inverse problem is obtained. In speech recognition, for example, designing an appropriate state transition network of an HMM allows us to effectively search for a linguistically likely phoneme sequence that best explains an observed sequence. With the same strategy, we would want to design an appropriate state transition network of our HMM to effectively reduce the solution space of the phrase and accent command sequences. Here, a question is how to design a state transition network for our model. In normal speech, many phrases or sentences share the same intonation pattern. This is because an intonation pattern is usually determined by the grammatical structure of an uttered sentence or the accent type each phrase is associated with. Thus, it may be natural to hypothesize that phrase and accent command sequences are governed by a vocabulary model. More specifically, we assume that we have a dictionary consisting of a finite number of left-to-right HMM templates and a sequence of the phrase and accent command pair is generated according to a concatenation of those templates. In this paper, we formulate a probabilistic F_0 contour model by designing a state transition network of our HMM based on the above assumption and derive a parameter optimization algorithm.

2. Probabilistic Pitch Contour Model

2.1. Original Fujisaki Model

The Fujisaki model [1] assumes that an F_0 contour on a logarithmic scale, $x(t)$, where t is time, is the superposition of three components: a phrase component $x_{\mathrm{p}}(t)$, an accent component

Copyright © 2013 ISCA

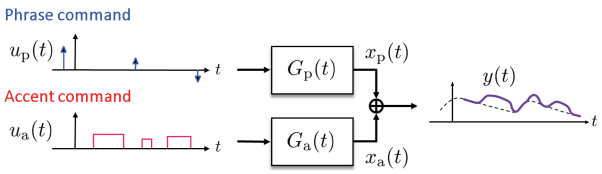

Figure 1: *Original Fujisaki model [1].*

$x_\mathrm{a}(t)$, and a base component x_b:

$$x(t) = x_\mathrm{p}(t) + x_\mathrm{a}(t) + x_b. \tag{1}$$

The phrase component $x_\mathrm{p}(t)$ consists of the major-scale pitch variations over the duration of the prosodic units, and the accent component $x_\mathrm{a}(t)$ 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 $u_\mathrm{p}(t)$ consisting of Dirac deltas (phrase commands), and the other with $u_\mathrm{a}(t)$ consisting of rectangular pulses (accent commands):

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

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

$$x_\mathrm{a}(t) = G_\mathrm{a}(t) * u_\mathrm{a}(t), \tag{4}$$

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

where $*$ denotes convolution over time. The baseline component x_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 [1].

2.2. Probabilistic speech F_0 contour model

Here, we briefly review our probabilistic pitch contour model based on the discrete-time version of the Fujisaki model [5, 6, 7].

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 $u_\mathrm{p}[k]$ and $u_\mathrm{a}[k]$ pair, i.e., $\boldsymbol{o}[k] = (u_\mathrm{p}[k], u_\mathrm{a}[k])^\mathsf{T}$, using a hidden Markov model (HMM). In [5, 6, 7], we have assumed that $\{\boldsymbol{o}[k]\}_{k=1}^K$ is a sequence of outputs generated from an HMM with the specific topology illustrated in Fig. 2. The output distribution of each state is assumed to be a Gaussian distribution

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

where s_k indicates the state variable. Namely, the mean vector $\boldsymbol{\mu}[k] = (\mu_\mathrm{p}[k], \mu_\mathrm{a}[k])^\mathsf{T} = \boldsymbol{c}_{s_k}$ and covariance matrix $\boldsymbol{\Sigma}[k] = \boldsymbol{\Upsilon}_{s_k}$ are considered to evolve in time as a result of the state transition $s_1, \ldots, s_K$. The definition of the above HMM can be summarized as follows:

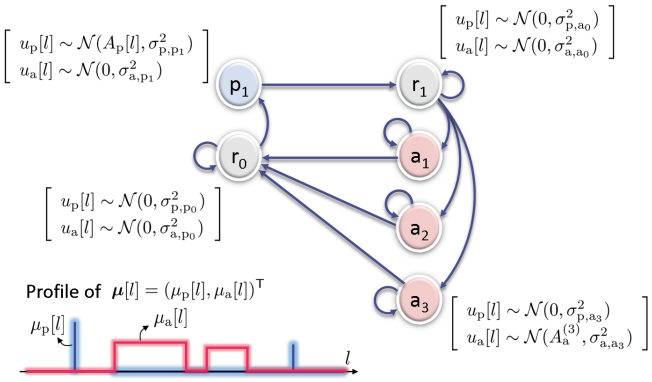

Figure 2: *Previous HMM topology for phrase-accent command sequence modeling [5, 6, 7]. In state* r_0, $\mu_\mathrm{p}[k]$ *and* $\mu_\mathrm{a}[k]$ *are both constrained to be zero. In state* p_1, $\mu_\mathrm{p}[k]$ *can take a non-zero value,* $A_\mathrm{p}[k]$, *whereas* $\mu_\mathrm{a}[k]$ *is still restricted to zero. In state* r_1, *no self-transitions are allowed. In state* r_1, $\mu_\mathrm{p}[k]$ *and* $\mu_\mathrm{a}[k]$ *again become zero. This path constraint restricts* $\mu_\mathrm{p}[k]$ *to consisting of isolated deltas. State* a_0 *leads to states* $\mathrm{a}_1, \ldots, \mathrm{a}_N$, *in each of which* $\mu_\mathrm{a}[k]$ *can take a different non-zero value* $A_\mathrm{a}^{(n)}$, *whereas* $\mu_\mathrm{p}[k]$ *is forced to be zero. Direct state transitions from state* a_n *to state* $\mathrm{a}_{n'}$ *without passing through state* r_1 *are not allowed. This constraint restricts* $\mu_\mathrm{a}[k]$ *to consisting of rectangular pulses.*

Output sequence:	$\{\boldsymbol{o}[k]\}_{k=1}^K$
State sequence:	$\{s_k\}_{k=1}^K$
Output distribution:	$P(\boldsymbol{o}[k]\|s_k) = \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{c}_{s_k}, \boldsymbol{\Upsilon}_{s_k})$
Mean sequence:	$\boldsymbol{\mu}[k] = (\mu_\mathrm{p}[k], \mu_\mathrm{a}[k])^\mathsf{T} = \boldsymbol{c}_{s_k}$
Transition probability:	$\phi_{i',i} = \log P(s_k = i \| s_{k-1} = i')$

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

$$x_\mathrm{p}[k] = u_\mathrm{p}[k] * G_\mathrm{p}[k], \tag{7}$$

$$x_\mathrm{a}[k] = u_\mathrm{a}[k] * G_\mathrm{a}[k], \tag{8}$$

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

$$x[k] = x_\mathrm{p}[k] + x_\mathrm{a}[k] + x_\mathrm{b}, \tag{9}$$

where x_b denotes the baseline value.

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, observed F_0 values in unvoiced regions must not be trusted. 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 $x_\mathrm{n}[k] \sim \mathcal{N}\left(0, \upsilon^2_\mathrm{n}[k]\right)$, where $\upsilon^2_\mathrm{n}[k]$ 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

$$y[k] = x[k] + x_\mathrm{n}[k]. \tag{10}$$

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

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

Recall from (6) that given a state sequence $\boldsymbol{s} = \{s_k\}_{k=1}^{K}$, $\boldsymbol{o}$ is generated according to $P(\boldsymbol{o}|\boldsymbol{s}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{c}_{s_k}[k], \boldsymbol{\Upsilon}_{s_k})$ where $\boldsymbol{\theta}$ denotes the mean vectors and covariance matrices of the state emission distributions. $P(\boldsymbol{s})$ is given by the product of the state transition probabilities: $P(\boldsymbol{s}) = \phi_{s_1} \prod_{k=2}^{K} \phi_{s_k, s_{k-1}}$.

3. Proposed model

3.1. Vocabulary model of phrase-accent command sequence

A key idea in the above model is that the sequence of the phrase and accent command pair is modeled as an HMM. In this paper, we are concerned with designing an HMM topology (a state transition network). The previous HMM topology shown in 2.2 allows for any state transitions unless they break the Fujisaki model's constraint. However, most of the state transitions are not appropriate in a linguistic sense. If we can effectively constrain the solution space of the phrase and accent command sequence through an HMM topology design, we may expect to obtain a better solution to the inverse problem. Here, a question is how to design a state transition network for our model.

In normal speech, we use many intonation patterns when we speak. However, the types of intonation patterns are limited. We use a common intonation pattern to utter a large set of phrases (or sentences). As for Japanese, phrases or sentences can be categorized according to accent types. Phrases with the same accent type show the same pitch pattern. For example, when we utter the phrases "arayuru genjitsu wo" and "ashita wa rinko da", the intonation patterns should be almost the same since the corresponding accent types are the same. Thus, it may be natural to hypothesize that all phrase and accent command sequences are drawn from a vocabulary consisting of relatively small number of intonation pattern templates. Namely, we assume that we have a dictionary consisting of a finite number of left-to-right HMM templates and a sequence of the phrase and accent command pair is generated according to a concatenation of those templates. This can be modeled by an HMM topology described in Fig. 3. In order to parameterize the duration of the self-transition (except for state $\mathrm{p}_1, \mathrm{p}_2, \ldots$), as with [6, 7], we split each state into a certain number of substates such that they all have exactly the same emission densities. Fig. 4 shows an example of the splitting of state $\mathrm{a}_{1,1}$.

The proposed method consists of two stages: training and recognition. In the training stage, the mean vectors of the state emission distributions and the state transition probabilities are learned from a training data set. This corresponds to learning the intonation pattern templates. In the recognition stage, we search for the optimal state sequence with fixed templates and state transition probabilities, given a test data. Both stages are based upon the same optimization algorithm, which will be described in the next section. The only difference is that the HMM parameters (the mean parameters of the state emission distributions and the state transition probabilities) are fixed during the recognition stage.

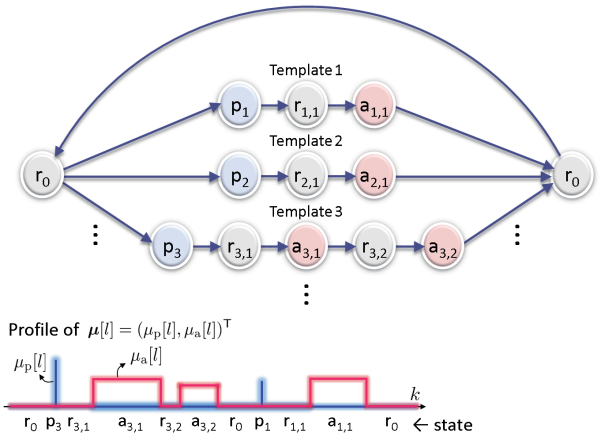

Figure 3: *Proposed HMM topology for phrase-accent command sequence modeling based on the vocabulary model of pitch pattern templates.*

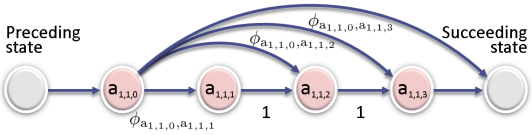

Figure 4: *The splitting of state* $\mathrm{a}_{1,1}$ *into 4 substates* $\mathrm{a}_{1,1,0}$, $\mathrm{a}_{1,1,1}$, $\mathrm{a}_{1,1,2}$, *and* $\mathrm{a}_{1,1,3}$. $\phi_{\mathrm{a}_{1,1,0},\mathrm{a}_{1,1,1}}$ *corresponds to the probability of staying at state* $\mathrm{a}_{1,1}$ *with 4 consecutive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

3.2. Parameter Optimization Algorithm

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 [8]. By using the Jensen's inequality we obtain an inequality

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

Table 1: *Accuracy rates* (S=0.3s). *The left, middle, and right columns show the accuracy rates of the phrase and accent commands, the phrase commands alone, and the accent commands alone, respectively. The "Init" row shows the accuracy rates obtained with the θ before the EM iterations (henceforth Narusawa's method [4]), the "N templates" row shows that of the estimated command sequence after the EM iterations with the number N of the templates, and the "Conv." row shows those obtained with our previous method [6, 7].*

	All	Phrase commands	Accent commands
Init [4]	65.9%	67.4%	65.1%
$T = 5$	**70.2**%	**70.3**%	70.0%
$T = 10$	**70.0**%	**71.0**%	69.5%
$T = 15$	67.3%	**72.5**%	64.7%
Conv. [6, 7]	69.3%	63.8%	72.1%

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 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.

4. Experimental Evaluation

Accuracy of phrase and accent commands estimation should depend on the size of the vocabulary model (i.e., the number of the intonation pattern templates). Thus, we evaluated the present method with different settings of the number of the templates in terms of estimation accuracy.

In the training stage, For each utterance in the first 50 sentences of ATR speech database [9] spoken by a male speaker (MHT), F_0 contours were extracted using the method described in [10], from which the model parameters were estimated. The constant parameters were fixed respectively at $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $\upsilon_{\mathrm{n}}^2[k] = 10^{15}$ for unvoiced regions and $\upsilon_{\mathrm{n}}^2[k] = 0.2^2$ for voiced regions. u_{b} was set at the minimum $\log F_0$ value in the voiced regions. Numbers of substates for states a and a were 250.

In the recognition stage, for each F_0 contour extracted in the same way as training stage from the last 53 sentences in the ATR speech database, the Fujisaki model parameters were estimated from which commands estimation accuracies were calculated. In both stage, the initial values of θ were set at the values obtained with the method described in [4]. We tested the present method with different settings of the number of the templates T: 5, 10, and 15. The number of accent commands in each template was set at 1 to 3. The numbers of templates with 1, 2, 3 or 4 accent commands were $(2, 2, 1, 0)$, $(4, 4, 2, 0)$ and $(5, 5, 3, 2)$ for $T = 5, 10, 15$, respectively.

Evaluation procedure is same as our previous works [6, 7], where a pair of commands are matched in the dynamic programming if the difference between the estimated and ground truth was shorter than S seconds. The magnitudes of the phrase and accent commands not taken account because the magnitude estimation was very sensitive to the baseline F_0 value, which were set differently in the present method and in the manual annotation.

Tab. 1 shows the result of our quantitative evaluation with

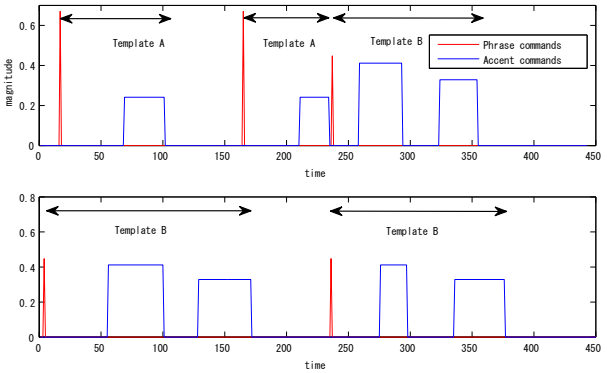

Figure 5: *An example of results from Viterbi decoding in the recognition stage. Commands sequences are extracted from japanese utterances "watashi wa sore o ryokan ni motte kaetta" (above) and "ippyo no kakusa wa sarani hirogaru daro" (below.) In the first command sequence, there are three templates. First and second templates are the same and are classified as "template A," while the third is classified as "template B." In the second command sequence, there are two templates and both are classified as "template B."*

$S = 0.3$ s. The accuracy rate is slightly improved in $T = 5$ and 10. This is probably because the present method is a template-based method which made a best use of information about the levels and timings of accent commands to accurately estimate at which position each phrase command should appear. In $T = 15$, although the accuracy rate of the phrase commands was improved, unnecessary templates with 4 accent commands seemed to make the accuracy of accent commands worse.

5. Discussion

The proposed method is a template-based method, thus it can be considered that the method classifies commands sequences into finite kind of templates, as shown in Fig. 5. This side effects suggests possibility to combine our method with clustering-based approach, namely speaker recognition [11], voice conversion [12], or text to speech synthesis [13]. Such classification can worsen accuracies of command estimation, however, as shown in the preceding sections, accuracy rates are at least equivalent to previous methods.

6. Conclusion

This paper dealt with the problem of estimating the Fujisaki-model parameters (phrase-accent command sequence) from a raw F_0 contour, and proposed a probabilistic generative model of speech F_0 contours incorporating a vocabulary model of pitch pattern templates. The present method was evaluated in terms of phrase and accent commands estimation accuracy with different settings of the number of the templates. The experimental results revealed that the reduction in the solution space of the phrase and accent sequence had an effect on improving the accuracy of command estimation.

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. 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. Le 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] K. Yoshizato, H. Kameoka, D. Saito, and S. Sagayama, "Hidden Markov convolutive mixture model for pitch contour analysis of speech," in Proc. The 13th Annual Conference of the International Speech Communication Association (Interspeech 2012), Sep. 2012.

[8] 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.

[9] 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.

[10] 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.

[11] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," in *Proc. ICASSP*, 1985, pp. 387–390.

[12] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.