論文 / 著書情報
Article / Book Information

Title	Discontinuous observation HMM for prosodic-event-based F0 generation
Authors	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
Citation	Proc. 13th Annual Conference of the International Speech Communication Association, , , pp. 462-465,
Pub. date	2012, 9
Copyright	(c) 2012 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Discontinuous Observation HMM for Prosodic-Event-Based F0 Generation

Tomoki Koriyama, Takashi Nose, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Japan

koriyama.t.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper examines F0 modeling and generation techniques for spontaneous speech synthesis. In the previous study, we proposed a prosodic-unit HMM where the synthesis unit is defined as a segment between two prosodic events represented by a ToBI label framework. To take the advantage of the prosodic-unit HMM, continuous F0 sequences must be modeled from discontinuous F0 data including unvoiced regions. The conventional F0 models such as the MSD-HMM and the continuous F0 HMM are not always appropriate for such demand. To overcome this problem, we propose an alternative F0 model named discontinuous observation HMM (DO-HMM) where the unvoiced frames are regarded as missing data. We objectively evaluate the performance of the DO-HMM by comparing it with the conventional F0 modeling techniques and discuss the results.

Index Terms: HMM-based speech synthesis, F0 modeling, prosody generation, discontinuous observation HMM, spontaneous speech.

1. Introduction

With diversification of speech synthesis applications, such as a human-like spoken dialog system, it is essential to develop a technique which can model variability of speech. In reality, however, since spontaneous speech has a lot of prosodic variability owing to turn-taking, speech acts, and other factors, it is not an easy task to model its variability.

Toward solving this problem, we have proposed an F0 modeling technique using prosodic-event-based HMM in an HMM-based speech synthesis framework [1]. Prosodic-event-based HMM uses segments, such as pitch falling by accent and rising by boundary pitch movement (BPM), as the modeling units of HMMs. We refer to it as *prosodic-unit* HMM whereas the ordinary phone-unit-based HMM as *phone-unit* HMM. We showed that incorporation of the prosodic-unit HMM enabled us to reduce the number of model parameters of F0 significantly while keeping the naturalness of the generated F0.

In the prosody modeling using the prosodic-unit HMM, it is assumed that one prosodic unit has a certain continuous F0 pattern. However, in real speech, F0 observations are often discontinuous even in one prosodic unit. To model discontinuous F0 observation, several approaches have been proposed for HMM-based speech synthesis [2–4]. A widely accepted approach is the use of multi-space probability distribution HMM (MSD-HMM) [2]. Although it has been shown that MSD-HMM can be applied successfully to phone-unit HMMs, it is not always appropriate for prosodic-unit HMM because its assumption of continuous F0 observations does not always meet. Another approach is the use of continuous F0 HMMs [3, 4]. In this approach, the model parameters are trained using continuous observation sequences in which unvoiced regions are replaced by the best candidates of F0 extraction or interpolated using a certain method. One of the issues of this approach is that the modeling performance could depend on the interpolation performance.

To overcome these problems in the conventional approaches, we propose an alternative F0 modeling technique for prosodic-unit HMM which optimizes the model parameters using only actually observed F0 sequences. We apply a method which utilizes missing data of unvoiced region [5] to the modeling of prosodic-unit HMM. We also examine two F0 generation approaches in which F0 values are generated in whole region or voiced regions only.

2. F0 modeling for prosodic-unit HMM

2.1. Speech synthesis using prosodic-unit HMM

In the prosodic-unit HMM, synthesis unit is defined as a speech segment between two prosodic events. For the prosodic events of Japanese speech we focus on in this study, we employ X-JToBI [6], an extension of ToBI, which includes tone tier labels with timing information of the folding points of F0 contours. It is noted that the prosodic-unit HMM of other languages can be constructed in a similar way by preparing the annotations of prosodic events. In the previous study [1], spontaneous speech was modeled and synthesized using both the prosodic-unit and the phone-unit HMMs. The prosodic unit was used for modeling continuous F0 sequences, and the phone-unit HMM was used for modeling spectral features and voice/unvoiced regions. Though a time alignment between two HMMs is required in the parameter generation process, the prosodic variability is well modeled using the prosodic-unit HMM with a smaller number of parameters than the conventional F0 modeling technique using the phone-unit HMM.

2.2. F0 modeling problem in prosodic-unit HMM

Let S_v and S_u be sets of voiced and unvoiced frame indexes, respectively. $\boldsymbol{O}_v$ represents discontinuous F0 sequence of voiced frames and is dependent on $S = (S_v, S_u)$. Let $\boldsymbol{A}$ be a set of transition matrices of the prosodic-unit HMM and $\boldsymbol{B}$ be a set of output probability density functions (pdfs) for the voiced space of the prosodic-unit HMM. To represent the prosodic-unit HMM, we can use either MSD-HMM or continuous F0 HMM. However, there are some problems. When the MSD-HMM is applied to the prosodic-unit HMM, it is necessary to use a set of weight parameters, $\boldsymbol{w}$, for the voiced/unvoiced space. Although $\boldsymbol{w}$ is optimized in the model training, it is ignored in the F0 generation step. Therefore the other parameter sets, $\boldsymbol{A}$ and $\boldsymbol{B}$, are not optimized appropriately when we generate continuous F0 sequence in the parameter generation step. The continuous F0 HMM maximizes the likelihood of continuous observation sequence $\boldsymbol{O}_c$ which is obtained by interpolating unvoiced regions. However the observations of the unvoiced regions are not always reliable and some parameters are optimized using such unreliable observations in the model training.

3. Discontinuous observation HMM

For the MSD-HMM and continuous F0 HMM, it is difficult to model the continuous F0 sequences without the influence of the unobserved data in unvoiced frames. In contrast, the advantage of discontinuous observation HMM (DO-HMM) described in the following is that the likelihood calculation in the DO-HMM depends only on the observed F0 data $\boldsymbol{O}_v$ in voiced frames and is not affected by the unobserved data.

3.1. Definition

We utilize the idea of the F0 modeling proposed in [5] where the values of unvoiced regions are regarded as missing data. This enables us to deal with continuous F0 sequence. Let the missing F0 sequence be $\boldsymbol{O}_u$ and the whole F0 sequence be $\boldsymbol{O}_c = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T)$ which is determined by $(\boldsymbol{O}_u, \boldsymbol{O}_v, S)$. The DO-HMM is represented by a model parameter set $\lambda = (\boldsymbol{A}, \boldsymbol{B})$, and likelihood is given by

$$\begin{aligned} P(\boldsymbol{O}_v|S,\lambda) &= \int P(\boldsymbol{O}_u, \boldsymbol{O}_v|S,\lambda) d\boldsymbol{O}_u \\ &= \int P(\boldsymbol{O}_c|\lambda) d\boldsymbol{O}_u. \end{aligned} \tag{1}$$

3.2. Parameter estimation algorithm

By introducing the missing observation $\boldsymbol{O}_u$ into the parameter estimation of an EM algorithm, Q-function is defined by

$$Q(\lambda, \tilde{\lambda}) = \mathbb{E}\left[\log P(\boldsymbol{q}, \boldsymbol{O}_u, \boldsymbol{O}_v|S,\tilde{\lambda})|\boldsymbol{O}_v, S, \lambda\right] \tag{2}$$

and decomposed into Q-functions of the state transition probability a_{ij} and the output pdf $b_i(\boldsymbol{o})$:

$$\begin{aligned} Q(\lambda,\tilde{a}) &= \sum_{\boldsymbol{q}} \int P(\boldsymbol{q}, \boldsymbol{O}_u|\boldsymbol{O}_v, S,\lambda) \log P(\boldsymbol{q}|S,\tilde{\lambda}) d\boldsymbol{O}_u \\ &= \sum_{\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{O}_v, S,\lambda) \log P(\boldsymbol{q}|S,\tilde{\lambda}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T-1} \xi_t^{(v)}(i,j) \log \tilde{a}_{ij}, \end{aligned} \tag{3}$$

$$\begin{aligned} Q(\lambda,\tilde{b}) &= \sum_{\boldsymbol{q}} \int P(\boldsymbol{q}, \boldsymbol{O}_u|\boldsymbol{O}_v, S,\lambda) \log P(\boldsymbol{O}_u, \boldsymbol{O}_v|S,\tilde{\lambda}) d\boldsymbol{O}_u \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} \int P(q_t = i, \boldsymbol{O}_u|\boldsymbol{O}_v, S,\lambda) \log \tilde{b}_i(\boldsymbol{o}_t) d\boldsymbol{O}_u \\ &= \sum_{i=1}^{N} \Bigg(\sum_{t\in S_v} P(q_t = i|\boldsymbol{O}_v, S,\lambda) \log \tilde{b}_i(\boldsymbol{o}_t) \\ &\quad + \sum_{t\in S_u} \int P(\boldsymbol{o}_t = \boldsymbol{x}|q_t = i, \boldsymbol{O}_v, S,\lambda) \\ &\qquad P(q_t = i|\boldsymbol{O}_v, S,\lambda) \log \tilde{b}_i(\boldsymbol{x}) d\boldsymbol{x} \Bigg) \\ &= \sum_{i=1}^{N} \Bigg(\sum_{t\in S_u} \gamma_t^{(v)}(i) \int b_i(\boldsymbol{x}) \log \tilde{b}_i(\boldsymbol{x}) d\boldsymbol{x} \\ &\quad + \sum_{t\in S_v} \gamma_t^{(v)}(i) \log \tilde{b}_i(\boldsymbol{o}_t) \Bigg). \end{aligned} \tag{4}$$

Here, N is the number of states, T is the total number of frames, and $\boldsymbol{q} = [q_1, \ldots, q_T]$ is a state sequence. $\gamma_t^{(v)}(i)$ represents the state occupation probability given the discontinuous observation $\boldsymbol{O}_v$. $\gamma_t^{(v)}(i)$ and $\xi_t^{(v)}(i,j)$ are calculated by

$$\begin{aligned} \gamma_t^{(v)}(i) &= P(q_t = i|\boldsymbol{O}_v, S,\lambda) \\ &= \frac{\alpha_t^{(v)}(i)\beta_t^{(v)}(i)}{\sum_{k=1}^{N} \alpha_t^{(v)}(k)\beta_t^{(v)}(k)}, \end{aligned} \tag{5}$$

$$\begin{aligned} \xi_t^{(v)}(i,j) &= P(q_t = i, q_{t+1} = j|\boldsymbol{O}_v, S,\lambda) \\ &= \frac{\alpha_t^{(v)}(i) a_{ij} b'_j(\boldsymbol{o}_{t+1}) \beta_{t+1}^{(v)}(j)}{P(\boldsymbol{O}_v|S,\lambda)}, \end{aligned} \tag{6}$$

$$b'_i(\boldsymbol{o}_t) = \begin{cases} b_i(\boldsymbol{o}_t) & (t \in S_v) \\ 1 & (t \in S_u) \end{cases} \tag{7}$$

where $\alpha_t^{(v)}(i)$ is the forward probability defined by $\alpha_t^{(v)}(i) = P(\boldsymbol{O}_v^{(\alpha)}, q_t = i|S,\lambda)$. $\boldsymbol{O}_v^{(\alpha)}$ is the voiced observation sequence before the frame t. In a similar manner, $\beta_t^{(v)}(i)$ is the backward probability defined by $\beta_t^{(v)}(i) = P(\boldsymbol{O}_v^{(\beta)}|q_t = i, S,\lambda)$ and $\boldsymbol{O}_v^{(\beta)}$ is the voiced observation sequence after the frame t. $\alpha_t^{(v)}(i)$ and $\beta_t^{(v)}(i)$ are calculated using a forward-backward algorithm as

$$\alpha_1^{(v)}(i) = \pi_i b'_i(\boldsymbol{o}_T), \tag{8}$$

$$\beta_T^{(v)}(i) = 1, \tag{9}$$

$$\alpha_t^{(v)}(i) = \left(\sum_{j=1}^{N} \alpha_{t-1}^{(v)}(j) a_{ji} \right) b'_i(\boldsymbol{o}_t), \tag{10}$$

$$\beta_t^{(v)}(i) = \sum_{j=1}^{N} a_{ij} b'_j(\boldsymbol{o}_{t+1}) \beta_{t+1}^{(v)}(j). \tag{11}$$

Here, we assume that the output probability $b_i(\boldsymbol{o})$ is expressed by a single Gaussian pdf $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{V}_i)$ to simplify the description. In this case, an integral included in Eq. (4) becomes

$$\begin{aligned} \int b_i(x) \log \tilde{b}_i(\boldsymbol{x}) d\boldsymbol{x} = -\frac{1}{2} \Big(& d\log(2\pi) + \log|\tilde{\boldsymbol{V}}_i| \\ & + \mathrm{T}r(\boldsymbol{V}_i \tilde{\boldsymbol{V}}_i^{-1}) + (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{V}}_i^{-1} (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}_i) \Big). \end{aligned} \tag{12}$$

The model parameters are updated by maximizing Q-functions Eqs. (3) and (4). The updating equations are derived as follows:

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t^{(v)}(i,j)}{\sum_{k=1}^{N} \sum_{t=1}^{T-1} \xi_t^{(v)}(i,k)}, \tag{13}$$

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{t\in S_v} \gamma_t^{(v)}(i) \boldsymbol{o}_t + \sum_{t\in S_u} \gamma_t^{(v)}(i) \boldsymbol{\mu}_i}{\sum_{t=1}^{T} \gamma_t^{(v)}(i)}, \tag{14}$$

$$\begin{aligned} \tilde{\boldsymbol{V}}_i = \frac{1}{\sum_{t=1}^{T} \gamma_t^{(v)}(i)} \Bigg\{ & \sum_{t\in S_v} \gamma_t^{(v)}(i) (\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_i)^\top \\ & + \sum_{t\in S_u} \gamma_t^{(v)}(i) \left(\boldsymbol{V}_i + (\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i - \tilde{\boldsymbol{\mu}}_i)^\top \right) \Bigg\}. \end{aligned} \tag{15}$$

Consequently, the transition probabilities are updated in a way similar to the ordinary HMM. On the other hand, the parameters of output pdfs are updated by the weighted sum of the observed values and the previous parameter set before update. When we use hidden semi-Markov model (HSMM) [7] which models the state duration explicitly, the similar estimation equations are derived in the same way.

3.3. Decision tree-based context clustering for DO-HMM

The tree-based context clustering is performed under assumptions that the assignment of states to observations does not change during the clustering process and that the likelihood can be approximated by the sum of log output probability weighted by the state occupation probability [8]. It follows that the likelihood of $\boldsymbol{O}_v$, $\mathcal{L}$, is approximated by

$$\mathcal{L} = \sum_{m=1}^{M} \sum_{i \in C(m)} \sum_{k=1}^{K} \sum_{t \in S_{k,v}} \gamma_{k,t}^{(v)}(i) \log b_i(\boldsymbol{o}_{k,t}) \tag{16}$$

where M and K are the total numbers of leaf nodes and speech samples, respectively. $C(m)$ represents a set of states included in a leaf node m and $S_{k,v}$ is a set of unvoiced frames of a speech sample k. $\gamma_{k,t}^{(v)}(i)$ is the state occupation probability of a state i at a frame t in a speech sample k. The difference from the conventional HMM is that the sum is calculated only for voiced frames. Therefore, using the state occupation count

$$\Gamma_m = \sum_{i \in C(m)} \sum_{k=1}^{K} \sum_{t \in S_{k,v}} \gamma_{k,t}^{(v)}(i) \tag{17}$$

and covariance matrix $\boldsymbol{V}_m$, the change of likelihood $\Delta\mathcal{L}$ by dividing a leaf node m_p into m_y and m_n is given by

$$\Delta\mathcal{L} = \frac{1}{2}(\Gamma_{m_y} \log |\boldsymbol{V}_{m_y}| + \Gamma_{m_n} \log |\boldsymbol{V}_{m_n}| - \Gamma_{m_p} \log |\boldsymbol{V}_{m_p}|). \tag{18}$$

We choose the leaf node and question for contexts which maximize $\Delta\mathcal{L}$ when dividing the node. The mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{V}_m$ of a clustered leaf node are calculated under the assumption that the state occupation probability does not change as follows:

$$\boldsymbol{\mu}_m = \frac{\sum_{i \in C(m)} \sum_{k=1}^{K} \sum_{t \in S_{k,v}} \gamma_{k,t}^{(v)}(i) \boldsymbol{o}_{k,t}}{\Gamma_m}, \tag{19}$$

$$\boldsymbol{V}_m = \frac{\sum_{i \in C(m)} \sum_{k=1}^{K} \sum_{t \in S_{k,v}} \gamma_{k,t}^{(v)}(i) (\boldsymbol{o}_{k,t} - \boldsymbol{\mu}_m)(\boldsymbol{o}_{k,t} - \boldsymbol{\mu}_m)^\top}{\Gamma_m}. \tag{20}$$

4. Generating discontinuous observation sequence from HMMs

For the parameter generation of F0, we can consider two kinds of likelihood as shown in Table 1. One method is *whole region generation* which maximizes the likelihood of F0 feature sequence of the whole frames $\boldsymbol{O}_c$ and generates continuous F0 sequence $\boldsymbol{C}_c$. After the parameter generation, the F0 values of voiced frames are used for synthesizing speech. The other is *voiced region generation* which maximizes only the likelihood of discontinuous sequence of the voiced frames where the voiced/unvoiced information for each frame is given by the phone-unit HMM.

For the whole region generation, the conventional synthesis method for the HMM-based speech synthesis can be used. Hence, we only describe the algorithm for the voiced region generation. The output sequence $\boldsymbol{C}_v$ is estimated by maximizing the likelihood of voiced F0 frames $\boldsymbol{O}_v$ given the voiced/unvoiced information S as follows:

$$\boldsymbol{C}_v^* = \underset{\boldsymbol{C}_v}{\operatorname{argmax}} \log P(\boldsymbol{O}_v | \boldsymbol{q}, S, \lambda). \tag{21}$$

Table 1: *The likelihoods in each generation method.*

Generation method	Likelihood
Whole region generation	$P(\boldsymbol{O}_c \| \boldsymbol{A}, \boldsymbol{B})$
Voiced region generation	$P(\boldsymbol{O}_v \| S, \boldsymbol{A}, \boldsymbol{B})$

Here, let $\boldsymbol{L}$ and $\boldsymbol{L}'$ be the mapping matrices which satisfy

$$\boldsymbol{C}_v = \boldsymbol{L}\boldsymbol{C}_c, \tag{22}$$

$$\boldsymbol{O}_v = \boldsymbol{L}'\boldsymbol{O}_c. \tag{23}$$

Let $\boldsymbol{W}_c$ be a window matrix for whole region generation, $\boldsymbol{M}_c = [\mu_{q_1}^\top, \ldots, \mu_{q_T}^\top]^\top$, and $\boldsymbol{V}_c = \operatorname{diag}[\boldsymbol{V}_{q_1}, \ldots, \boldsymbol{V}_{q_T}]$. We define $\boldsymbol{W}_v = \boldsymbol{L}'\boldsymbol{W}_c\boldsymbol{L}^\top$, $\boldsymbol{V}_v = \boldsymbol{L}'\boldsymbol{V}\boldsymbol{L}'^\top$, and $\boldsymbol{M}_v = \boldsymbol{L}'\boldsymbol{M}$, then we have

$$\boldsymbol{O}_v = \boldsymbol{W}_v\boldsymbol{C}_v, \tag{24}$$

$$\log P(\boldsymbol{O}_v | \boldsymbol{q}, S, \lambda) = -(1/2) \log |\boldsymbol{V}_v| - \frac{1}{2}(\boldsymbol{W}_v\boldsymbol{C}_v - \boldsymbol{M}_v)^\top \boldsymbol{V}_v^{-1}(\boldsymbol{W}_v\boldsymbol{C}_v - \boldsymbol{M}_v) + const. \tag{25}$$

As a result, the optimal F0 sequence $\boldsymbol{C}_v^*$ is given by

$$\boldsymbol{C}_v^* = (\boldsymbol{W}_v^\top \boldsymbol{V}_v^{-1} \boldsymbol{W}_v)^{-1} \boldsymbol{W}_v^\top \boldsymbol{V}_v^{-1} \boldsymbol{M}_v. \tag{26}$$

5. Experiments

5.1. Experimental conditions

Spontaneous speech data with rich prosodic labels was used for the evaluation experiments and X-JToBI tone tier labels [6] are used as prosodic labels. We chose speech data of two non-professional female speakers (#19, #514) included in the Corpus of Spontaneous Japanese (CSJ) [9]. Two speech sets, lecture and conversation, were used. Conversational speech consists of two interviews and a task-oriented dialog. The total length of speech samples of each speaker and set is approximately 25 minutes. Speech signals were sampled at a rate of 16 kHz. F0 was extracted by SWIPE [10] and spectral feature was extracted by STRAIGHT [11] with 5-ms frame shift. The feature vector of prosodic-unit HMM consisted of log F0, and their delta and delta-delta coefficients. In the case of continuous F0 HMM, the F0 sequence of unvoiced region was made by linear interpolation and smoothed. The feature vector of phone-unit HMM consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, their delta and delta-delta coefficients, and a voiced/unvoiced information. We used hidden semi-Markov model (HSMM) which has explicit duration distributions for both prosodic-unit and phone-unit HMM. The model topology was 5-state left-to-right context-dependent HSMM without skip paths. Each state had a single Gaussian distribution with a diagonal covariance matrix. MDL was used for the stopping criterion and minimum number of observations [12] was also used to alleviate over-fitting. We set the minimum number of observations to 50 from the result of a preliminary experiment.

For training and testing, the phonetic and prosodic contexts were automatically converted from the labels given in CSJ. Although speech synthesis using prosodic-unit HMM needs alignment of label timings with phone-unit HMM [1], we used the F0 patterns generated from the annotated label timings to focus on F0 models in this study. Five-fold cross-validation tests were performed in the evaluations.

Table 2: *Average F0 distortions and correlation coefficients of synthetic speech and leaf node sizes of F0 model.*

Training method	Generation method	F0 RMSE [ms]	Correlation coefficient	Ave. # of leaf nodes
Phone-unit HMM		281.2	0.609	563.7
MSD	Whole region	279.7	0.613	277.9
	Voiced region	284.2	0.595	
Continuous	Whole region	279.3	0.621	309.0
	Voiced region	286.3	0.595	
DO	Whole region	280.3	0.616	241.8
	Voiced region	288.0	0.594	

5.2. Results

The three training methods and F0 generation methods for prosodic-unit HMM described in Sec. 2.2 were evaluated objectively. The conventional phone-unit HMM was also evaluated. The measurements for evaluation were average F0 distortion and correlation coefficient. The average F0 distortion was calculated by RMS error between generated and original log F0s.

Table 2 shows the results with the average number of leaf nodes of F0 decision trees. Each RMS error and correlation coefficient are the average values of all data sets. RMSEs using voiced region generation were larger than that using whole region generation. A possible reason is that the generated F0 contour using voiced region generation was not smoothed well between the adjacent voiced regions, because each contour of a voiced region is determined using only the parameters included in that region rather than considering influence of preceding and succeeding voiced regions. In the case using the whole region generation, the distortion and correlation of DO-HMM were comparable to that of the MSD-HMM and the continuous F0 HMM. However some differences are seen in the number of leaf nodes of F0 decision trees which implies how compact the model is. We can see that the DO-HMM has smaller F0 trees than the MSD-HMM and the continuous F0 HMM. This is because the total number of frames of the DO-HMM are calculated using only voiced frames and the MDL criterion changed. Consequently, DO-HMM expressed F0 patterns with a more compact parameter set than the other methods. Since it is not always easy to prepare a sufficient amount of spontaneous speech data manually labeled with rich prosodic information, this advantage of DO-HMM is important for estimating reliable parameters in the model training or the model adaptation.

6. Conclusion

We proposed the DO-HMM to train the continuous F0 model from the discontinuous F0 data including unvoiced regions for spontaneous speech synthesis using the prosodic-unit HMM. In the training of DO-HMM, the model parameters were optimized using only actually observed F0 sequence which avoid the influence of the unobserved data in unvoiced frames. The preliminary experimental results showed that the DO-HMM gives comparable performance to the conventional models using a smaller number of model parameters. In addition, we examined suitable parameter generation methods for the prosodic-unit HMM. In the future work, we will incorporate trajectory model in order to improve generated F0 since the current HMM framework is frame-based modeling. The dependency of observations of frames defined by the trajectory model is expected to enhance the usefulness of missing data.

7. Acknowledgments

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071 and 23700195.

8. References

[1] T. Koriyama, T. Nose, and T. Kobayashi, "An F0 modeling technique based on prosodic events for spontaneous speech synthesis," in *Proc. ICASSP*, 2012, pp. 4589–4592.

[2] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[3] Q. Zhang, F. Soong, Y. Qian, Z. Yan, J. Pan, and Y. Yan, "Improved modeling for f0 generation and v/u decision in hmm-based tts," in *Proc. ICASSP*. IEEE, 2010, pp. 4606–4609.

[4] K. Yu and S. Young, "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 99, pp. 1071–1079, 2011.

[5] K.N. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 295–309, 1999.

[6] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: an extended J-ToBI for spontaneous speech," in *Proc. 7th ICSLP*, 2002, pp. 1545–1548.

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. 90, no. 5, pp. 825–834, 2007.

[8] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Queen's College, 1995.

[9] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[10] A. Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, Ph.D. thesis, University of Florida, 2007.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[12] T. Koriyama, T. Nose, and T. Kobayashi, "On the use of extended context for HMM-based spontaneous conversational speech synthesis," in *Proc. INTERSPEECH*, 2011, pp. 2657–2660.