T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

Title	An F0 modeling technique based on prosodic events for spontaneous speech synthesis
Authors	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
Citation	Proc. 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing, , , pp. 4589-4592
Pub. date	2012, 3
Copyright	(c) 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	http://dx.doi.org/10.1109/ICASSP.2012.6288940
Note	This file is author (final) version.

Powered by T2R2 (Tokyo Institute Research Repository)

AN F0 MODELING TECHNIQUE BASED ON PROSODIC EVENTS FOR SPONTANEOUS SPEECH SYNTHESIS

Tomoki Koriyama, Takashi Nose, Takao Kobayashi

Tokyo Institute of Technology
Interdisciplinary Graduate School of Science and Engineering
4259-G2-4 Nagatsuta-cho, Midori-ku, Yokohama-City, 226-8502, Japan

ABSTRACT

This paper proposes a technique for effective modeling of F0 contours using prosodic-event-based HMM units for HMM-based spontaneous speech synthesis. The modeling unit corresponds to one of prosodic event segments such as pitch falling by accent and pitch rising by boundary pitch movement (BPM). Since the prosodic events of one phrase are generally less frequent than the changes of phonemes, the proposed unit is expected to reduce the number of model parameters of F0, which leads to robust parameter estimation. The objective and subjective experiments using spontaneous conversational speech data show that the proposed technique can significantly reduce the number of model parameters while keeping the naturalness of the synthetic speech.

Index Terms— HMM-based speech synthesis, F0 modeling, Prosodic events, Spontaneous speech

1. INTRODUCTION

HMM-based parametric speech synthesis [1] has been studied as a technique which can give compact and flexible models for generating smooth and natural sounding speech. This enabled us to construct the model of expressive speech with a small number of parameters [2] and to perform speaker and style adaptation from average voice model [3]. Moreover it was applied to spontaneous conversational speech which has much more prosodic variability than simulated speech [4]. In [4], the naturalness of the synthetic speech was improved by incorporating a set of extended prosodic contexts, such as phone prolongation and tone information. However, the prosody generation of multiple speaking styles included in spontaneous speech has not been sufficient yet. For the realization of a diversity of prosody in the synthetic speech, we need another approach to prosody modeling in the HMM-based speech synthesis framework.

In the HMM-based speech synthesis, fundamental frequency (F0) is usually modeled using phone-unit-based HMMs and trained synchronously with spectral features. To model both voiced and unvoiced regions of the F0 pattern consistently, multi-space distribution HMM (MSD-HMM) [5] is utilized. Although this HMM is good at modeling of prosodic features of phone unit, it is not always suited for the F0 pattern of spontaneous speech well. This is because the positions of prosodic events such as accent and boundary pitch movement (BPM) do not always match those of phonetic moves. For instance, in a segment outside of prosodic events called *connection* in rise/fall/connection model [6], F0 features do not change so much as prosodic events even if the segment contains several phones. On the other hand, in a segment at the rise-fall pitch movement, F0 moves largely even if the segment has only one phone.

To alleviate this problem, there have been proposed different approaches to the F0 modeling, e.g., the use of hierarchical structures [7] and the use of longer units [8]. In this study, we propose an alternative approach to modeling F0 contour efficiently using prosodic-event-based HMM units. More specifically, we use components of prosodic events, such as the segment of pitch falling by accent and pitch rising by BPM, as the modeling units. Since the prosodic events of one phrase are less frequent than the changes of phonemes, the proposed unit is expected to reduce the number of model parameters of F0, which leads to robust parameter estimation. We examine the effectiveness of the proposed F0 modeling technique through both objective and subjective evaluation experiments.

2. F0 MODELING BASED ON PROSODIC EVENTS

2.1. Prosodic labels

In the proposed technique, the speech synthesis unit is defined using the prosodic label information related to F0 contours. Specifically, the label sequence represents not only the accent information, but also more precise form of F0 contour such as pitch rising or falling. In this study, we use the Corpus of Spontaneous Japanese [9] that includes the manually annotated speech data with prosodic labels using X-JToBI labeling scheme [10]. X-JToBI is an extension of J-ToBI [11] that is the Japanese version of ToBI [12]. By using X-JToBI labels, we can represent the prosodic variability of spontaneous speech more precisely. We use X-JToBI tone tier labels as the prosodic labels. Table 1 shows the labels used in this study. These labels include the timing information of the folding points of F0 contours. The type of label depends on the function in prosodic events. Phrasal tone and accent consist of "H−" and "A," respectively. Boundary pitch movements consist of "*%," "pH," and "pL." "L% (FBT: final boundary tone)" and "L% (LTBPM: low tone of BPM)" are distinguished by the function; the former expresses the end of the accent phrase with falling tone, and the latter is the start of BPM. Ordinary accent phrases can be expressed by the label sequence which starts with %L followed by other labels according with the label network shown in Fig. 1 and ends with the final boundary tone, "*%." Other phrases are prosodic fillers and prosodic word fragments. A prosodic filler is a filled pause which does not have neither a pitch rise nor a local pitch fall anywhere. If a speaker stops in or starts from the middle of accent phrase, the phrase is treated as a prosodic word fragment. This phenomenon is caused by disfluency of spontaneous speech.

2.2. Prosodic-unit HMM

The segment between X-JToBI tone tier labels can be regarded as the basic unit of a prosodic event. Hence, we adopt this segment as a unit of HMM for F0 modeling. We refer to the proposed prosodic-unit-based HMM as *prosodic-unit HMM*, whereas we refer to the conventional phone-unit-based HMM as *phone-unit HMM*. One prosodic unit is distinguished from others by the combination of the labels.

Table 1. X-JToBI tone tier labels

label	function	abbr.
%L	beginning of phrase	s
H−	end of pitch rise	m
A	beginning of pitch fall by accent	a
%LA	joint label of %L & A	b
L%(LTBPM)	low tone before BPM	t
L%(FBT)	end of phrase with fall tone	l
H%	end of phrase with rise tone	h
HL%	end of phrase with rise-fall tone	i
LH%	end of phrase with fall-rise tone	j
HLH%	end of phrase with rise-fall-rise tone	k
pH	high pointer in BPM	p
pL	low pointer in BPM	q
FL	filler with low pitch	FL
FH	filler with high pitch	FH

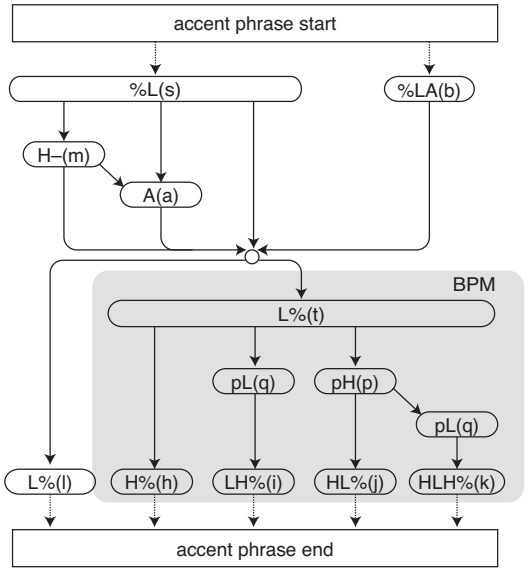


Fig. 1. Network of prosodic labels

For example, the segment between "%L" and "A" is labeled as "%L-A." To simplify the notation of prosodic unit, we use the combination of the single characters corresponding to the labels, which are shown in Table 1. The prosodic units are listed in Table 2. Here, "y" and "z" represent the beginning and end of the prosodic word fragments, respectively. The segment of filler is labeled as the label name, "FH" and "FL." "SP" means the prosodic space between accent phrases (e.g. the segment from "L%" to "%L,") and "PZ" and "SL" denote the pause and silence, respectively.

The training procedure of the prosodic-unit HMM is similar to that of the conventional phone-unit HMM. Log F0, its delta, and delta-delta coefficients are used as the features of HMM. To model voiced/unvoiced region, we use MSD in a manner similar to phone-unit HMM. An HMM of each prosodic unit is initialized by segmental K-means algorithm, and the parameters are refined by Baum-Weltch re-estimation. Then HMMs are clustered by their prosodic context. A context set for prosodic unit consists of quin-prosodic-unit and the information of the units of accent phrase, breath group, and utterance. By using the prosodic-unit HMM as the speech synthesis unit, we can model F0 patterns more efficiently with the prosodic label information compared to the case when using the conventional phone-unit HMM. Moreover the prosodic-unit HMM enables us to control the F0 contour more flexibly than phone-unit HMM because it is easy to manipulate the timing of the prosodic events.

Table 2. Prosodic units

type	unit names
normal segment	sm, sa, sl, st, ma, ml, mt al, at, bl, bt
BPM	th, tp, tq, pi, pq, qj, qk
prosodic filler	FH, FL
prosodic word fragment	yh, yl, az, bz, mz, yz
pause & silence	PZ, SL, SP

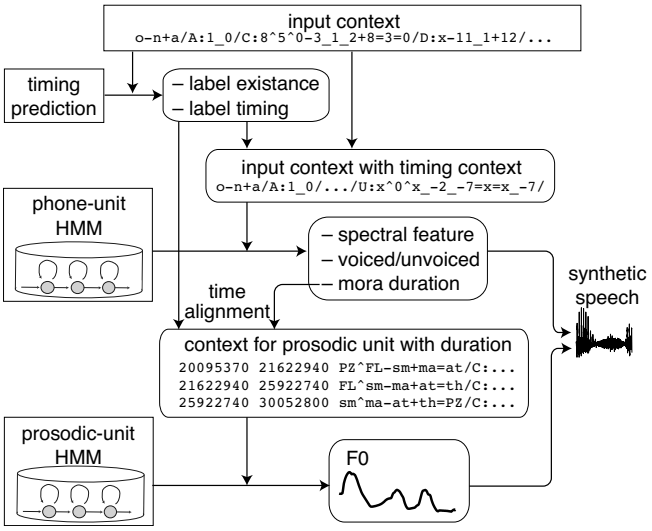


Fig. 2. An outline of speech synthesis using prosodic-unit HMM.

3. SPEECH SYNTHESIS USING PROSODIC-UNIT HMM

In this study, F0 is generated from the prosodic-unit HMM, whereas the other information such as spectral features is generated from phone-unit HMM. However, the positions of phones and prosodic events do not match when the speech parameters are generated separately. Accordingly, the time alignment between the phone-unit HMM and the prosodic-unit HMM is necessary to apply it to speech synthesis system. Especially, the position of pitch falling by accent is important for Japanese speech synthesis because some words with the same pronunciation are distinguished by it. For this purpose, we use timing prediction of prosodic label and apply it to the alignment. Figure 2 shows the procedure of the speech synthesis with the alignment.

In the training step, F0 is modeled by the prosodic-unit HMM. Spectral features, voiced/unvoiced feature, and duration are modeled by the phone-unit HMM. Here the phone-unit HMM is trained using the extended contexts proposed in [4] which include the mora positions of X-JToBI tone tier labels and the tone types, e.g., BPMs, prosodic fillers, and prosodic word fragments. In the timing prediction, the existence of "H−" and "A" and the timing information of each label in the accent phrase are predicted. In this study, we use a mora-normalized position shown in Fig. 3 as the label timing information. Mora-normalized position is a measure where the length of each mora is normalized into unity. In Fig. 3, the mora-normalized positions of the labels "%L", "H−", and "L%" are defined as about 0.1, 1.8, and 3.9, respectively. As explanatory variables for timing

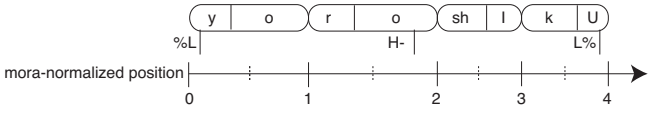


Fig. 3. An example of mora-normalized position. In this example, the accent phrase consists of 4 moras: "yo", "ro", "shi", and "ku". The mora-normalized positions of labels, "%L", "H−", and "L%" are defined as about 0.1, 1.8, and 3.9, respectively.

prediction, we use the information of accent phrase.

When synthesizing speech, we firstly construct the input context sequence from the word sequence with automatically or manually annotated prosodic event information such as accent and BPM, and then predict the label timing. Next, the contexts for the phone-unit HMM are constructed using predicted mora-normalized positions, and the spectral, voiced/unvoiced, and duration features are generated from the phone-unit HMMs. Then, the duration information of prosodic units are calculated from the generated mora durations and label timing by the time alignment, and the F0 contour is generated by the prosodic-unit HMMs with their durations. Here, to generate continuous F0 contour, we set a small value as a threshold of a voiced space weight of MSD-HMM. Finally, speech is synthesized using the generated spectral and F0 features.

4. EXPERIMENTS

4.1. Experimental conditions

Spontaneous conversational speech data was used for the evaluation experiments. We chose speech data of two female speakers (#19, #514) included in CSJ. Each speaker was non-professional speaker and uttered three sets of conversational speech: two interviews and a task-oriented dialog. The total length of speech samples of each speaker is approximately 25 minutes. Speech signals were sampled at a rate of 16 kHz. The spectral feature and F0 were extracted by STRAIGHT [13] with 5 ms frame shift. The feature vector of prosodic-unit HMM consisted of log F0, and their delta and delta-delta coefficients. 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 flag. We used hidden semi-Markov model (HSMM) [14] 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. In the case of F0, minimum number of observations [4] was also used to alleviate over-fitting. We set the minimum number of observations to 50 on the basis of a preliminary experimental result. We compared the proposed technique with the conventional HMM-based conversational speech synthesis technique described in [4]. In this technique, only the phone-unit HMM was used to model both of the spectral and prosodic features.

C4.5 was used for the prediction of the existence of the labels, and linear regression was used for the prediction of mora-normalized position. We chose these classifiers from preliminary experiments. For training and testing, the phonetic and prosodic contexts were automatically converted from the labels given in CSJ. Ten-fold cross-validation tests were performed in the evaluations.

4.2. Evaluation of F0 modeling

Performance of the proposed technique was evaluated both objectively and subjectively. To focus on the F0 modeling with the prosodic-unit HMM, F0 patterns were generated using the label timings annotated in the database. Tables 3 and 4 show the average F0 distortions, correlation coefficients, and tree sizes of log F0 of the proposed and conventional techniques. The average F0 distortion was calculated by RMS error between generated and original log F0s. Table 5 shows the result of subjective evaluation of reproducibility. In this test, to focus on the evaluation of F0 reproducibility, we used the acoustic features extracted from the original speech except F0. This test was performed by an XAB test. Six participants chose the sample more similar to the reference X. The reference sample was vocoded speech. If the participants could not determine the preference, "no preference" was chosen. Each participant evaluated 20 utterances randomly chosen from generated speech samples which were used for objective evaluation. We used speech samples having 10 or more moras. It is found from the results that, although average F0 distortions of the proposed technique were larger than those of the conventional technique, the scores of subjective evaluation were comparable. There was no significant difference between the subjective scores of the proposed and conventional techniques. The correlation coefficients were also comparable. It is noted that the number of leaf nodes of the proposed technique, which represents model complexity, was about 36% or 39% as many as the conventional technique. We will give a further discussion about the results in Section 5.

Table 3. F0 distortion and correlation using timing annotated in database

	RMSE of log F0[cent]		correlation coefficient	
	#19	#514	#19	#514
conventional	282.9	381.8	0.534	0.477
proposed	288.9	383.3	0.523	0.498

Table 4. Leaf node size

	number of leaf nodes	
	#19	#514
conventional	679	762
proposed	262	277

Table 5. Subjective evaluation of reproducibility

speaker	proposed	conventional	no preference
#19	36.7%	33.3%	30.0%
#514	36.7%	31.7%	31.7%

Table 6. F0 distortion and correlation using predicted timing

	RMSE of log F0[cent]		correlation coefficient	
	#19	#514	#19	#514
conventional	288.4	386.2	0.502	0.459
proposed	302.0	398.2	0.461	0.456

Table 7. Subjective evaluation of reproducibility of synthetic speech

speaker	proposed	conventional	no preference
#19	35.0%	26.7%	38.3%
#514	25.0%	28.3%	46.7%

4.3. Evaluation of synthetic speech

The performance of overall speech synthesis which uses the prosodic-unit HMM and timing prediction was evaluated objectively and subjectively. In this experiment, the F0 contour was generated using the technique explained in Section 3, i.e., the label timing was pre-

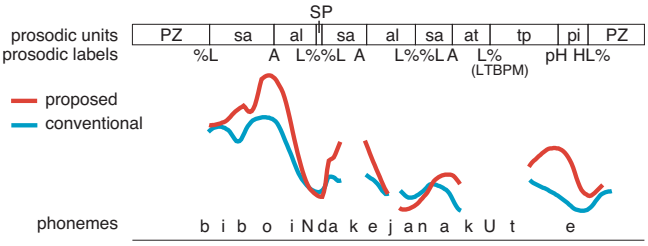


Fig. 4. Example of generated F0 contour of a Japanese phrase "biboin-dakeja-nakute". This phrase consists of 3 accent phrases: "bibo'in", "dake'ja" and "na'kute" and the last phrase ends with rise-fall BPM. ['] represents accent and pitch falling should be perceived after the accent.

Table 8. Standard deviations of generated log F0

	#19	#514
original	0.203	0.268
proposed	0.119	0.142
conventional	0.097	0.137

dicted and the time alignment between phone-unit and prosodic-unit HMMs was performed. Tables 6 and 7 show the F0 distortions, correlation coefficients, and the scores of subjective evaluation of reproducibility, respectively. Since the trained HMMs are the same as those of previous subsection, the actual leaf node size is also the same. The measurement and conditions of subjective evaluation were the same as the the experiment of Section 4.2. It can be found that the results were similar to those of the evaluation of F0 modeling. The scores of reproducibility were comparable between the proposed and conventional techniques and have no significant differences.

5. DISCUSSIONS

As seen in the above results, although the F0 distortions of the proposed technique were larger than those of the conventional technique, the reproducibility by subjective evaluation was comparable. A possible reason is the difference of the characteristics of generated F0, whose example is illustrated in Fig. 4. In this example, the proposed technique generated clear pitch falling and BPM which moves larger than the conventional technique. This tendency can be seen in the whole of generated F0 according to the standard deviations of generated log F0 shown in Table 8. This is because the prosodic-unit HMM models the prosodic events explicitly. When we carefully inspected the generated F0, we found that F0 moves more widely than the original F0 in some prosodic events. Although this leads the increase of F0 distortion, the listeners might perceive prosodic events clearly.

Although we examined only the speaker-dependent model in this study, the prosodic-unit HMM can be applied easily to speaker-independent one based on the average voice model [15].

6. CONCLUSION

In this paper, we proposed an F0 modeling technique based on the prosodic-unit HMM. The component of prosodic events was used as a unit of HMM in order to model F0 contour efficiently. The evaluation experiments were performed for both F0 modeling and speech synthesis. The results showed that the subjective reproducibility of the proposed technique was comparable to that of the conventional technique while reducing the leaf node size of F0 model to about 40%.

7. ACKNOWLEDGEMENTS

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063 and 23700195.

8. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[2] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, 2007.

[3] T. Nose, M. Tachibana, and T. Kobayashi, "HMM-based style control for expressive speech synthesis with arbitrary speaker's voice using model adaptation," *IEICE Trans. Inf. & Syst.*, vol. E92-D, no. 3, pp. 489–497, 2009.

[4] 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.

[5] 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.

[6] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1-2, pp. 169–186, 1994.

[7] M. Lei, Y. Wu, F.K. Soong, Z.H. Ling, and L. Dai, "A hierarchical F0 modeling method for HMM-Based speech synthesis," in *Proc. INTERSPEECH*, 2010, pp. 2170–2173.

[8] Y. Qian, Z. Wu, B. Gao, and F. K Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1702–1710, Aug. 2011.

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

[10] 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.

[11] N. Campbell and J. Venditti, "J-ToBI: An intonation labelling system for Japanese," in *Proc. the Autumn meeting of the Acoustical Society of Japan*, 1995, vol. 1, pp. 317–318.

[12] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *Second International Conference on Spoken Language Processing*, 1992.

[13] 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.

[14] 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.

[15] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational Spontaneous Speech Synthesis Using Average Voice Model," in *Proc. INTERSPEECH*, 2010, pp. 853–856.