A Targets-based Superpositional Model of Fundamental Frequency Contours Applied to HMM-based Speech Synthesis

Jinfu Ni, Yoshinori Shiga, Chiori Hori, and Yutaka Kidawara

Spoken Language Communication Laboratory, Universal Communication Research Institute
National Institute of Information and Communications Technology, Kyoto, Japan
{jinfu.ni, yoshinori.shiga, chiori.hori, kidawara}@nict.go.jp

10.21437/Interspeech.2013-117

Abstract

Superpositional model of fundamental frequency (F_0) contours as suggested by the Fujisaki model can well represent F_0 movements of speech keeping a clear relation with linguistic information of utterances. Therefore, improvement of HMM-based speech synthesis is expected by using the merit of superpositional model. In this paper, a targets-based superpositional model is proposed in the light of the Fujisaki model. Here, both accent and phrase components are parameterized by respectively defined low and high targets which allow flexible interaction between accent and phrase components. Due to the flexible interaction, the new method consistently treats such complex F_0 movements as low digging, varying declination, and final lowering by simply adjusting parameter values. This facilitates extraction of the model parameters from observed F_0 contours, which is one of major problems preventing the use of the Fujisaki model. Extraction of the target parameters is evaluated for a Japanese speech corpus and the F_0 contours generated by the model are used for HMM training instead of the original. Listening test of synthetic speech indicates significant improvements in speech quality. Micro-prosodic effects are also investigated. Results show that adding the micro-prosody to the generated F_0 contours does not significantly improve speech quality.
Index Terms: Prosody modeling, Superpositional F0 model, Continuous F0 modeling, HMM-based speech synthesis

1. Introduction

Modeling of fundamental frequency (F_0) for HMM-based speech synthesis is critical for achieving good naturalness and communicative functions. F_0 contours observed from a speech corpus usually are discontinuous. The multi-space probability distribution (MSD) HMM [1] is widely used to model the discontinuous F_0 observation. Recent research [2] indicates that continuous F_0 HMM leads to better F_0 trajectory than MSD-HMM which produces over-smoothed F_0 contours in a frame-by-frame manner. To solve the issue of over-micro F_0 modeling, several different methods have been proposed to capture the F_0 movements related to different prosodic layers [3][4][5]. An explicit formulation of different prosodic layers is the Fujisaki model [6]. The model represents a sentence F_0 contour in logarithmic scale as superposition of accent components on phase components. These components are known to have clear correspondences with linguistic and para-linguistic information that is conveyed by prosody [7]. One of major problems preventing the use of the model, for example, in HMM-based speech synthesis is that the performance of automatic extraction of the model parameters from observed F_0 contours of a speech corpus is still rather limited [7][8][9].

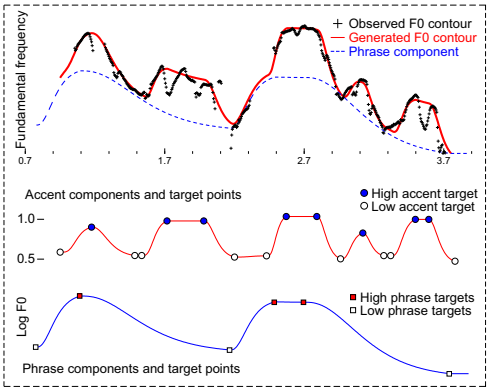

Figure 1: *Schematic diagram of decomposing F_0 contours into accent and phrase components represented by target points.*

It is more straightforward to capture prosodic contributions of linguistic information to F_0 contours of utterances by a series of target points [10][11]. The target points are relatively easy to be detected from observed F_0 contours [11], and the transitions between target points can be well represented by Poisson process-based interpolation [12]. Towards automatic fitting of F_0 contours of a speech corpus for HMM-based speech synthesis, this paper proposes a target-based method to formulate both accent and phrase components with the Poisson process-based interpolation in the light of the Fujisaki model.

The rest of the paper is organized as follows. Section 2 describes the proposed method. Section 3 outlines an algorithm for extraction of the model parameters with experiment results in section 4. Section 5 presents the use of F_0 contours generated by the proposed model for HMM training in HMM-based speech synthesis, followed by a discussion on representing F_0 contours in HMMs in section 6. Section 7 concludes this paper.

2. Modeling of F0 contours using two-level control mechanisms

Following the Fujisaki model [6], we decompose F_0 contours into accent and phrase components but represent them by using respective low and high targets (Fig. 1). Basically, each of accent and phrase components is defined by three (or four) targets and the two high targets, if necessary, for each component are assumed to be identical in magnitude. The motivation of using targets is to deal with non-linear interactions between accent and phrase components by relatively defining accent and phrase targets. To deal with non-linear interactions between accent and phrase components, the two components have to be

further treated at a higher level. Therefore, we model F_0 contours by two-level mechanisms. At the first level, a Poisson process-based mechanism [12] is used to generate both accent and phrase components. At the second level, a resonance-based mechanism [13] coherently unifies them to form F_0 contours.

2.1. A resonance-based F_0 decomposition

F_0 results from vocal-cord vibrations. It is effective to use a resonance mechanism to manipulate F_0 contours [14]. Here, a resonance-based mapping [13] is applied to deal with latent interactions between accent and phrase components, which are particularly treated as a kind of topology deformations.

The resonance-based mapping between λ (frequency ratio square) and α (angle related to damping ratio) [13], hereafter referred to as $\lambda = f(\alpha)$, is defined according to Eq. (1).

$$\frac{\lambda}{1} = \frac{A(\lambda,\alpha)-1}{A(1,\alpha)-1}, \ 0 \le \lambda < 1, \tag{1}$$

$$\text{where, } A(\lambda,\alpha) = \frac{1}{\sqrt{1+\lambda^2\cos^2 2\alpha - 2\lambda\cos^2 2\alpha}} \tag{2}$$

which indicates a resonance transformation [13]. For convenience, let $\alpha = f^{-1}(\lambda)$ be the inverse mapping. When λ runs from 0 to 1, α takes values from $\frac{1}{3}$ to 0 in falling order.

Let f_0 be any F_0 in a voice range specified by bottom frequency f_{0_b} and top frequency f_{0_t}. With normalizing f_0 to $[0, 1]$

$$\lambda_{f_0} := \frac{\ln f_0 - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}}, \tag{3}$$

a topological deformation between cubic and spherical objects as described in [13] is applied to f_0. More specifically,

- Define a cubic object with volume $\sqrt{(0.5\lambda_{f_0})^3}$.
- Map the cubic volumes to α, $\alpha_{f_0} := f^{-1}(\sqrt{(0.5\lambda_{f_0})^3})$.
- Map a reference F_0, $f_{0_r} \in [f_{0_b}, f_{0_t}]$, to α similarly.

$$\alpha_{f_{0_r}} := f^{-1}(\sqrt{(0.5\lambda_{f_{0_r}})^3}).$$

- Calculate $\alpha_{f_{0_r}} - \alpha_{f_0}$, mirror symmetry with respect to $\alpha_{f_{0_r}}$, thus $\alpha_{f_{0_r}} - \alpha_{f_0}$ having rising order.
- Define a spherical object having volume

$$\phi_{f_0|f_{0_r}} := \frac{4\pi \times (\alpha_{f_{0_r}} - \alpha_{f_0})}{3}. \tag{4}$$

$\phi_{f_0|f_{0_r}}$ is spherical because $\alpha_{f_{0_r}} - \alpha_{f_0}$ is cubic.

Equation (4) indicates a decomposition of $\ln f_0$ over time. More particularly, $\alpha_{f_{0_r}}$ is used to represent phrase components (treated as a baseline) and $\phi_{f_0|f_{0_r}}$ accent components. On the other hand, when giving accent components by $\phi_{f_0|f_{0_r}}$ and phrase components by $\alpha_{f_{0_r}}$, $\ln f_0$ can be calculated by

$$\ln f_0 = \ln f_{0_b} + 2f^{\frac{2}{3}}(\alpha_{f_{0_r}} - \frac{\phi_{f_0|f_{0_r}}}{4\pi/3})(\ln f_{0_t} - \ln f_{0_b}). \tag{5}$$

Accordingly, the resonance-based mechanism can be utilized to deal with non-linear interactions between accent and phrase components while unifying them to give F_0 contours.

2.2. A resonance-based superpositional F_0 model

A model of F_0 contours as a function of time t, $F_0(t)$, in logarithmic scale is represented as resonance-based superposition of accent component $C_a(t)$ on phrase component $C_p(t)$.

$$\ln F_0(t) = \ln f_{0_b} + 2f^{\frac{2}{3}}(\alpha(t))(\ln f_{0_t} - \ln f_{0_b}), \tag{6}$$

$$\alpha(t) = f^{-1}\left((\frac{C_p(t) - \ln f_{0_b}}{2(\ln f_{0_t} - \ln f_{0_b})})^{\frac{3}{2}}\right) - \frac{C_a(t) - 0.5}{10 \times 4\pi/3}, \tag{7}$$

$$C_p(t) = \sum_{i=0}^{I_p} \gamma_{p_{i-1}} + (\gamma_{p_i} - \gamma_{p_{i-1}})P(t - t_{p_{i-1}}, t_{p_i} - t_{p_{i-1}}),$$

$$C_a(t) = \sum_{i=0}^{I_a} \gamma_{a_{i-1}} + (\gamma_{a_i} - \gamma_{a_{i-1}})P(t - t_{a_{i-1}}, t_{a_i} - t_{a_{i-1}}),$$

$$P(t,\Delta t) = 1 - \sum_{j=0}^{k} \frac{[\frac{c(k)t}{\Delta t}]^j}{j!} e^{-\frac{c(k)t}{\Delta t}}, \ t \ge 0. \tag{8}$$

The model parameters for representing F_0 contours of utterances are described as follows.

f_{0_t}:	The top F_0 of a speaker's voice range.
f_{0_b}:	The bottom F_0 of the voice range.
$I_p + 1$:	The number of phrase targets for an utterance.
(t_{p_i}, γ_{p_i}):	The ith phrase target; t_{p_i} is time and γ_{p_i} magnitude.
$I_a + 1$:	The number of accent targets for the utterance.
(t_{a_i}, γ_{a_i}):	The ith accent target; t_{a_i} is time and γ_{a_i} magnitude.

The other symbols from Eq. (6) to Eq. (8) are described below.

$F_0(t)$:	Generated F_0 contours as a function of time t.
$f(x)$:	Resonance-based mapping by Eqs. (1) and (2).
$f^{-1}(x)$:	Inverse mapping of $f(x)$.
$C_p(t)$:	Phrase components generated by the phrase targets.
$C_a(t)$:	Accent components generated by the accent targets.
$\alpha(t)$:	Superposition of the accent and phrase components.
$P(t,\Delta t)$:	A Poisson process-based filter [12].
k:	Sustaining a target [12]. Basically $k = 2$,$c(2) = 6.3$.
$c(k)$:	Coefficients by solving $\sum_{j=0}^{k} \frac{[c(k)]^j}{j!} e^{-c(k)} = 0.05$.
Factor 10:	In Eq. (7), it scales $C_a(t)$ into the α domain $(0, \frac{1}{3})$.

Phrase target γ_{p_i} is defined by F_0 in $[f_{0_b}, f_{0_t}]$ in logarithmic scale and accent target γ_{a_i} in $(0, 1.5)$ with reference zero 0.5. When $\gamma_{a_i} < 0.5$, part of the accent components digs into under the phrase components, thus achieving low digging and final lowering in F_0 as observed in natural speech.

3. Model parameter estimation

An algorithm is developed for estimating the target parameters from observed F_0 contours of utterances in Japanese, given accentual phrase boundary information. Parameters f_{0_b} and f_{0_t} are set to the F_0 range of a set of observed F_0 contours. In Japanese, an accentual phrase basically has an accent (accent type 0, 1, 2 ...). The algorithm is described as follows.

- Preprocessing: Convert F_0 contours into $\phi_{f_0|f_{0_r}}$ with $f_{0_r} = f_{0_b}$ and then smooth them jointly using varying window sizes (short term: 10 points, and long term: 80 points) to suppress effects of micro-prosody (the modification of F_0 by phonetic segments) taking into account the general rise-(flat)-fall characteristics of Japanese accents. The smoothed contours are converted back to F_0 using Eq. (5) for parameter extraction.
- Parameter extraction: A segment between pauses longer than 0.3 s is regarded as a breath group and a breath group may be further divided into two groups determined by its long-term smoothed F_0 contours. The following processes are conducted for each group with the criterion of minimizing absolute F_0 errors. (a) Initialize a three-target phrase component having two low targets and a high target (Fig. 1): Setting the timing of the high target to the start of the second mora and shifting 0.3 s earlier for the timing of the first low target, and setting the timing of the other low target to the end of the breath group. The initial values for γ_{p_i} are determined by using the long-term

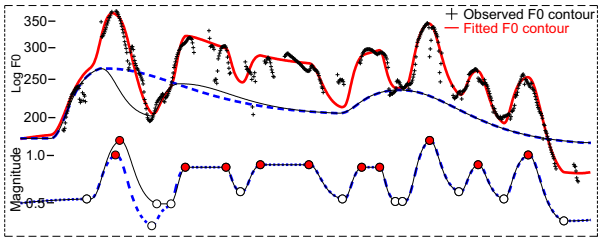

Figure 2: *Examples of fitting observed F_0 contours using the model. Two phrase components (the dashed curves) and three phrase components (the thin curves) are assumed and the corresponding accent components are superimposed on the bottom.*

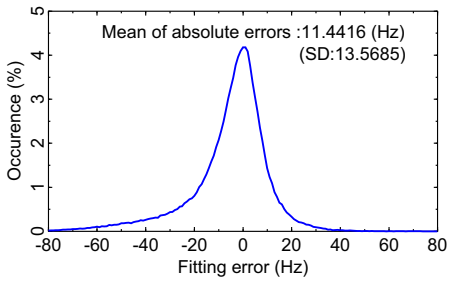

Figure 3: *Distribution of mismatch errors between the generated and observed F_0 contours.*

smoothed F_0 contours. (b) Calculate accent components by using Eq. (4) with both the smoothed F_0 contours and the current phrase components and then estimate accent targets from the current accent components. (c) Adjust γ_{a_i} into $[0.9, 1.1]$ for all the high accent targets and $[0.4, 0.6]$ for all the low accent targets and re-calculate the accent components using the adjusted accent targets. (d) Re-estimate phrase targets taking into account the current accent components. (e) Go to (b) with pre-defined times (eg., 3). (f) Insert a high phrase target if absolute errors between the generated and smoothed F_0 contours decrease over a pre-defined threshold, and go to (b).

• Parameter optimization: The accent targets are optimized by minimizing the mismatch errors between the generated and observed F_0 contours, given the estimated phrase components.

4. Experimental evaluation

Experiments of extracting model parameters are conducted for 503 utterances (ATR503set) of a female narrator. The F_0 contours are extracted with 5 ms frame shift by using the get f_0 module in the Snack Sound Toolkit [15]. f_{0_b} and f_{0_t} are set to 120 Hz and 420 Hz, respectively. The accent and phrase targets for fitting the F_0 contours are automatically estimated by using the algorithm mentioned above. In the process of parameter estimation, the phonetic boundary information of accentual phrases is given and at most two high accent targets are assumed within an accentual phrase. To investigate the general figures of accent and phrase targets, the automatically estimated phrase targets are manually checked with a graphic user interface. The accent targets are then automatically optimized again with the phrase components generated by the checked phrase targets.

Figure 2 shows examples of using the targets to flexibly treat interactions between accent and phrase components. As illustrated in this example, the model has a merit of using two-level decomposition (accent and phrase components) to implement three levels of phrases: accentual phrase, intermediate phrase, and intonational phrase [16]. An intermediate phrase boundary is achieved by making some low accent targets to drop under the reference zero line ($C_a(t) = 0.5$). Also, the phenomenon of final lowering [17] can be handled in the same way, adjusting the last low accent target downward as shown in Fig. 2. Using the Fujisaki model, however, additional phrase commands must be used for these situations, consequently leading difficulty in extracting the model parameters.

Table 1: Statistical results for the phrase targets.

Target position	The 1st low	High	The last low
Mean (in semitone)	41.1	46.5	39.6
SD	1.45	1.24	2.32

Table 2: Statistical results for the high accent targets.

Accent	Type 0	1	2	3	4	5
Mean	0.950	1.034	1.034	1.039	1.015	1.014
SD	0.097	0.134	0.142	0.120	0.089	0.089

Figure 3 shows error distributions for model-based fitting of the observed F_0 contours. The average of absolute F_0 errors is 11.44 Hz with SD (standard deviation) 13.57 Hz. The results indicate that the proposed model with a few parameters can represent F_0 contours well. Because of ignoring the effects of micro-prosody, the effects of phonetic segments (micro-prosody) and F_0 extraction errors are observable from the biased bell-shape in Fig. 3 in comparison of the left side with the right.

Table 1 shows the statistics of phrase targets in magnitude in semitone defined by $12ln(F_0/16.35)/ln2$ (base frequency 16.36 Hz indicating the lowest C note in piano.) The average number of phrase components per utterance is 2.78 (SD: 1.15). The average target number per phrase component is 3.45 (SD: 0.64). The results show that there exist at least two kinds of phrase patterns (Fig. 1) in natural speech. Complex combinations of phrase components need when using Fujisaki model.

Table 2 shows the statistical results for the high accent targets. The average of the low accent targets in magnitude for the speech corpus is 0.53 (SD: 0.08). The results indicate that the estimated accent components have good regularity across accent types. They also imply that the phrase components basically act as a baseline anchoring the low tones of accents.

5. HMM-based speech synthesis

Speech synthesis experiments are conducted using the same continuous speech corpus of 503 sentences as used in section 4. HTS-2.1 [18] is used to train HMMs. Out of 503 sentences, 490 sentences are used for HMM training, the rest sentences are used for testing. Speech signals are sampled at 16 kHz sampling rate and the spectral envelopes are extracted by STRAIGHT analysis [19] with 5 ms frame shift. The feature vector consists of 40 mel-cepstral coefficients including the 0th coefficient, log F_0, and their delta and delta-delta coefficients. A five-state left-to-right model topology is used.

Four versions of F_0 contours are prepared to train HMMs.

• F_0 contours extracted from speech waveforms (*Original*).

• These generated by the proposed F_0 model (*Proposed*).

• These combining both the Original voiced F_0's and the Proposed at the unvoiced regions (*Prop.+MP* (micro-prosody)).

• These combining both the Original voiced F_0's and spline-based interpolation for the unvoiced regions (*Spl.+MP*) [2].

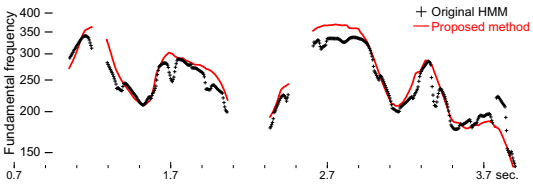

Figure 4: *Comparison of F_0 contours for a Japanese sentence generated by the Original HMM and the proposed method.*

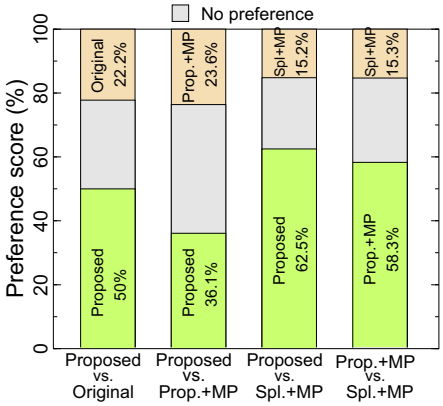

Figure 5: *Comparison between four pairs of versions in a preference test on synthetic speech naturalness.*

The last three versions use continuous F_0 contours. Note that the Proposed excludes both micro-prosody and F_0 extraction errors, but the others include both of them.

The Original as usual takes MSD-HMM [1], but the others are respectively trained by adding the continuous F_0 contours (including their delta and delta-delta) as the 5th stream while training the MSD-HMM; the 5th stream weight is set to 0. Consequently, continuous F_0 HMMs result for the last three versions. At the phase of speech synthesis, continuous F_0 contours are first synthesized by the continuous F_0 HMMs and their voiced/unvoiced decision is then taken from the MSD-HMM.

Figure 4 compares F_0 contours generated by the Original HMM and the proposed method. Compared to the Original, the F_0 contours by the proposed method are smooth and the peak portions are raised, significantly improving over micro-prosody effects. Table 3 compares F_0 errors in HMM-based prediction between the Original and Proposed. Higher errors in the Proposed are due to the ignorance of micro-prosody.

To evaluate the proposed method and the micro-prosody effects, four pair-wise preference listening tests are conducted: Original *vs.* Proposed, Proposed *vs.* Prop.+MP, Proposed *vs.* Spl.+MP, and Prop.+MP *vs.* Spl.+MP. Five natives participate in a listening test of synthetic speech in naturalness. Nine sentences (open test) make up a test set for each listener. The nine wave file pairs are duplicated and the order of two versions in a pair is swapped. The final 72 ($= 4 \times 9 \times 2$) wave file pairs are provided to the listeners in random order. Each listener was asked to select which is preferable or no preference.

The results are shown in Fig. 5. The proposed method outperforms the use of observed F_0 contours (Proposed *vs.* Original). Adding micro-prosody to the proposed method does not improve speech naturalness (Proposed *vs.* Prop.+MP). The proposed method also outperforms the spline-based interpolation of observed F_0 contours for continuous F_0 HMMs [2] (Proposed *vs.* Spl.+MP). The last two observations are re-confirmed by the result for Prop.+MP *vs.* Spl.+MP as shown in Fig. 5.

Table 3: F_0 error comparison between Original and Proposed.

Method	Closed (400 utterances)	Open (10 utterances)
Original	19.60 Hz (SD: 16.66)	21.51 Hz (SD: 17.34)
Proposed	21.77 Hz (SD: 18.32)	22.53 Hz (SD: 18.52)

6. Discussion

The F_0 contours observed from a speech corpus usually are quasi-continuous. MSD-HMM has been widely used for modeling the quasi-continuous F_0 contours [1]. The method has a merit that F_0 of each frame can be used directly as the training data and thus is good at synchronization of both mel-cepstrual and prosodic features automatically. Although the method can achieve good performance even using a rather limited size of speech, it has a rather limited ability to track long-term F_0 patterns against the effects of over micro-prosody and F_0 extraction errors on the resultant HMM. To cope with this issue, several different methods using hierarchical and/or adding structures have been proposed [3][4][5]. Compared to these methods, our method is of the merit of Fujisaki model keeping a clear relation of the underlying linguistic information, which is expected to further improve HMM-based speech synthesis. Compared to the Fujisaki model, our method allows to consistently treat such complex F_0 movements as low digging, varying long-term upward/downward movements, and final lowering by simply adjusting the targets (Fig. 2). This feature strengthens automatic extraction of the model parameters from observed F_0 contours of a speech corpus, which is one of the major problems preventing the use of the Fujisaki model [7].

F_0 contours generated by the proposed method do not cover the full F_0 movements including deviation caused by phonetic segments. A lot of research work in the literature (eg., [20]) has pointed out that micro-prosody affects speech quality. Typically, high vowels tend to have a higher F_0 than low vowels. However, our results and informal listening of re-synthesized speech excluding the component of micro-prosody indicate that having micro-prosody or not does not significantly affect synthetic speech quality. Two reasons are considered for this observation. One is the use of high-quality spectra analyzed by STRAIGHT [19]. The other is that the intrinsic F_0 differences in vowels probably are captured by the individual targets.

A few caveats in the work need to be mentioned. The number of listeners is quite limited and the experiment is only conducted with a female speaker. Also, we do not perform MOS (mean opinion score) evaluation. Further work is needed.

7. Conclusions

This paper proposed a new superpositional model of F_0 contours to strengthen automatic extraction of the model parameters from observed F_0 contours. The proposed model is of the merit of the Fujisaki model: a limited number of model parameters can well represent F_0 contours of speech keeping a clear relation of linguistic information of utterances. By using F_0 contours generated by the proposed model for HMM training instead of original F_0's, an improvement in synthetic speech quality was achieved. The effects of micro-prosody on HMM-based speech synthesis were also investigated with the proposed model. The results show that having micro-prosody or not does not significantly affect synthetic speech quality.

Acknowledgements We would like to thank Dr. Toda and Dr. Tokuda for their valuable discussions.

8. References

[1] Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999), "Hidden Markov models based on multispace probability distribution for pitch pattern modeling," *Proc. of ICASSP1999*, 229–232.

[2] Yu, K. and Young, S. (2011), "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Trans. on ASLP*, vol. 19, No. 5, 1071-1079.

[3] Sakai, S. (2005), "Fundamental frequency modeling for speech synthesis based on a statistical learning technique," *IEICE Trans. on Inf. and Syst.*, Vol E88D, No.3, 489-495.

[4] Zen, H. and Braunschweiler, N. (2010), "Context-dependent additive log F_0 modeling for HMM-based speech synthesis," Proc. of Interspeech2010, 889-892.

[5] Wu, Y. J. and Soong, F. (2012), "Modeling pitch trajectory by hierarchical HMM with minimum generation error training," *Proc. of ICASSP2012*, 4017-4020.

[6] Fujisaki, H., and Hirose, K. (1984), "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Jpn.*, 5, 233-242.

[7] Hashimoto, H., Hirose, K., and Minematsu, N. (2012), "Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis," *Proc. of Interspeech2012*.

[8] Sakurai, A. and Hirose, K. (1996), "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," *Proc. of ICSLP1996*, 817-820.

[9] Mixdorff, H., Hu, Y., and Chen, G. (2003), "Towards the automatic extraction of Fujisaki model parameters for Mandarin," *Proc. of Eurospeech2003*, 873-876.

[10] Hirst, D., Cristo, A. D., and Espesser, R. (2000), "Levels of representation and levels of analysis for the description of intonation systems," In *Prosody: Theory and Experiment*, edited by M. Horne (Kluwer, Dordrecht), 51-87.

[11] Ni, J. and Hirose, K. (2006), "Quantitative and structural modeling of voice fundamental frequency contours of speech in Mandarin," *Speech Communication*, **48** (8), 989-1008.

[12] Ni, J. and Nakamura, S. (2007), "Use of Poisson processes to generate fundamental frequency contours", *Proc. of ICASSP2007*, 825–828.

[13] Ni, J, Shiga, Y., Kawai, H., and Kashioka, H. (2012), "Resonance-based spectral deformation in HMM-based speech synthesis," *Proc. of ISCSLP2012*, 88-92.

[14] Ni, J., Kawai, H., and Hirose, K. (2006), "Constrained tone transformation technique for separation and combination of Mandarin tone and intonation," *J. Acoust. Soc. Am.*, 119 (3), 1764-1782.

[15] http://www.speech.kth.se/snack

[16] Beckman, M. E. and Pierrehumbert, J. B. (1986), "Intonational structure in Japanese and English," *Phonology Yearbook 3*, 255–309.

[17] Liberman, M. and Pierrehumbert, J. B. (1984), "Intonational invariance under changes in pitch range and length," In M. Aronoff and R. T. Oehrle (eds.) *Language sound structure*, Cambridge, Mass.: MIT Press. 157-233.

[18] http://hts.sp.nitech.ac.jp/

[19] Kawahara, H., Ikuyo, M. K., and Cheneigne, A. (1999), "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*, 27, 187-207.

[20] Pierrehumbert, J. B. (1981), "Synthesizing intonation," *J. Acoust. Soc. Am.*, 70 (4), 985-995.