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

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 $F_0$ model, Continuous $F_0$ 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 phrase 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 $F_0$ 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 $\lambda$ (frequency ratio square) and $\alpha$ (angle related to damping ratio) [13], hereafter referred to as $\lambda = f(\alpha)$, is determined according to Eq. (1).

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

where $A(\frac{\lambda}{\alpha}) = \frac{1}{\sqrt{1 + \lambda^2 \cos^2 2\alpha - 2 \lambda \cos^2 2\alpha}}$ (2) which indicates a resonance transformation [13]. For convenience, let $\alpha = f^{-1}(\lambda)$ be the inverse mapping. When $\lambda$ runs from 0 to 1, $\alpha$ takes values from $\frac{\pi}{4}$ to 0 in falling order.

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

$$\lambda_{f_0} := \frac{\ln f_0 - \ln f_{0b}}{\ln f_{0t} - \ln f_{0b}},$$

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 $\sqrt{0.5\lambda f_{0t}^3}$.
- Map the cubic volumes to $\alpha$, $\alpha_{f_0} := f^{-1}(\sqrt{0.5\lambda f_{0t}^3})$.
- Map a reference $F_0$, $f_{0r} \in [f_{0b}, f_{0t}]$, to $\alpha$ similarly.
- $\alpha_{f_0} := f^{-1}(\sqrt{0.5\lambda f_{0t}^3})$.
- Calculate $\alpha_{f_0} - \alpha_{f_0}$, mirror symmetry with respect to $\alpha_{f_0}$, thus $\alpha_{f_0} - \alpha_{f_0}$, having rising order.
- Define a spherical object having volume

$$\phi_{f_0:f_0} = \frac{4\pi}{3} \times (\alpha_{f_0} - \alpha_{f_0}).$$

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

$$\ln f_0 = \ln f_{0b} + 2f^2(\alpha_{f_0} - \alpha_{f_0})(\ln f_{0t} - \ln f_{0b}).$$

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 $C_p(t)$ on phrase component $C_{\alpha}(t)$.

$$\ln F_{0}(t) = \ln f_{0b} + 2f^2(\alpha(t))(\ln f_{0t} - \ln f_{0b}),$$

$$\alpha(t) = f^{-1}\left(\frac{C_p(t) - \ln f_{0b}}{2(\ln f_{0t} - \ln f_{0b})}\right) - \frac{C_{\alpha}(t) - 0.5}{10 \times 4\pi/3}.$$
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 $\gamma_{ai}$ 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]. $f_{o1}$ and $f_{o2}$ 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 and phrase targets, the automatically estimated phrase components (the thin curves) are assumed and the corresponding accent components are superimposed on the bottom.

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.

Figure 3 shows 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.

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].
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 then 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.

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-cepstral 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.

Table 3: $F_0$ error comparison between Original and Proposed.

<table>
<thead>
<tr>
<th>Method</th>
<th>Closed (400 utterances)</th>
<th>Open (10 utterances)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>19.60 Hz (SD: 16.66)</td>
<td>21.51 Hz (SD: 17.34)</td>
</tr>
<tr>
<td>Proposed</td>
<td>21.77 Hz (SD: 18.32)</td>
<td>22.53 Hz (SD: 18.52)</td>
</tr>
</tbody>
</table>

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.

Table 3: $F_0$ error comparison between Original and Proposed.
8. References


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