Skeletonising Chinese Fundamental Frequency Contours with A Functional Model and Its Evaluation

Jinfu Ni & Hisashi Kawai

ATR Spoken Language Translation Research Laboratories, Japan
{jinfu.ni; hisashi.kawai}@atr.co.jp

Abstract

This paper presents a method for skeletonising a fundamental frequency ($F_0$) contour with its underlying $F_0$ peaks and valleys, without losing the linguistic and para-linguistic information that it conveys. The $F_0$ peaks and valleys are mainly associated with underlying lexical tones, and can be easily converted into other features, such as the response time and amplitude of local $F_0$ rise/fall movements. Consequently, the exact shape of the $F_0$ contour can be then recovered by the use of a functional $F_0$ model, given the $F_0$ peaks and valleys. Experiments were conducted on 668 Chinese utterances (around 1.4 hours of speech) from two native speakers. The validity of the proposed method is consistently proved by a three-fold evaluation: error formations processing and intonation [1][2][3][4]. Skeletonising analyses, perceptual similarity between the re-synthesised tone and intonation and the original, and a listening test of the naturalness of synthetic speech with incorporation of the recovered $F_0$ contours into the unit selection process for synthesis.

1. Introduction

Perception tests and instrumental analyses of the past have yielded a consensus that the fundamental frequency ($F_0$) contour of an utterance can multiply manifest lexical tones, stress and intonation [1][2][3][4]. Skeletonising $F_0$ contours is thus desirable in prosodic analysis and its application to speech information processing. The first reason for this is that the $F_0$ peaks and valleys play a prominent role in anchoring the tone and intonation patterns. Pitch targets, basically comprising high and low, are commonly used to describe the intonating of accent languages, such as English and Japanese [5]. In Chinese, however, there exist four lexical tones, named Tones 1 to 4, and a neutral tone named Tone 0. If the range of a speaker’s voice is divided into four equal intervals, marked by five points, 1 low, 2 half-low, 3 middle, 4 half-high, and 5 high, Tones 1 to 4 are represented by 55, 35, 214, 51, respectively [1]. Because both the actual intervals and the absolute pitch are relative to the individual voice and the mood at the moment of speaking, the pitch targets are usually measured as $F_0$ peaks and valleys. Reliable analysis and labeling of the prosody must be capable of dealing with the tone variability under various conditions.

The second reason is related to the necessity of combining a statistical model with knowledge-based techniques to synthesise natural tone and intonation, arising from the development of text-to-speech conversion systems. Because the pitch targets can capture the interaction of the tone, stress and intonation [1], skeletonising $F_0$ contours shall be a key step in approaching such an aim. In this paper, we propose an efficient data-driven method upon our previous work to shrink an $F_0$ contour into the $F_0$ peaks and valleys that makes use of a functional $F_0$ model [6][7]. This model bridges the gap between linguistic and acoustic $F_0$ features, and creates constraints to reduce speaker-dependent effects, thus facilitating the data-driven learning and parameter estimation.

The remainder of the paper explains this method. Section 2 includes a description of the model and the algorithm for skeletonising $F_0$ contours. Experimental results are described in Section 3, and remarks and future work are given in Section 4.

2. Outline of the method

It is commonly assumed that the $F_0$ contour of an utterance is the physical implementation of a sequence of discrete speech events or pitch targets through which the linguistic and para-linguistic information is conveyed. Because the vocal cords are a physical system, the $F_0$ contour produced by vocal cord vibrations is predictable to a certain extent, given the pitch targets. To bridge the gap between the acoustic and the linguistic features, a model is helpful for analysing and skeletonising the $F_0$ contours.

2.1. A functional model of the $F_0$ contours

In this paper, we use a functional model [6] to represent the observed $F_0$ contours in a parametric form. An advantage of this model, compared to the Fujisaki model [8], is that it supports automatic analysis of the $F_0$ contours [7]. According to the model, the voice register (a frequency register of utterances) of a speaker is first transposed to a so-called RONDO scale (similar to a log-scale). The RONDO-$F_0$ contour is then expressed in concatenated mountain-shaped patterns lined up in series at the time axis. The $F_0$ contour $F_0(t)$ is given as follows:

\[
\frac{\ln F_0(t) - \ln f_{0b}}{\ln f_{0t} - \ln f_{0b}} = \frac{A(\Lambda(t)) - A(\lambda_0)}{A(\Lambda_1) - A(\lambda_0)}, \quad t \geq 0 \tag{1}
\]

where

\[
A(\lambda) = \frac{1}{(1 - (1 - 2\mu^2)\lambda)^2 + 4\mu^2(1 - 2\mu^2)\lambda}, \quad \lambda \geq 1, \tag{2}
\]

and

\[
\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \min(\Lambda_{r_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_1}(t) \tag{3}
\]

\[
\min(z_1, z_2) = \begin{cases} z_1 & \text{if } z_1 \leq z_2, \\ z_2 & \text{otherwise}. \end{cases}
\]

Equations (1) and (2) jointly indicate the transposition of the voice register, while Eq. (3) expresses the RONDO-$F_0$ contour $\Lambda(t)$, where $\Lambda_{r_1}(t)$ and $\Lambda_{f_1}(t)$ indicate the rise and fall components of the mountain-shaped pattern, respectively. Particularly,

\[
\Lambda_{r_1}(t) = \begin{cases} \lambda_{p_i} + \Delta \lambda_{r_1}(1 - D_{r_1}(t - p_i)), & t \leq t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}
\]

\[
\Lambda_{f_1}(t) = \begin{cases} \lambda_{p_i} + \Delta \lambda_{f_1}(1 - D_{f_1}(t - t_{p_i})), & t > t_{p_i}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}
\]

where

\[
D_{r_1}(t) = \left(1 + \frac{4.8t}{\Delta t_{r_1}}\right)e^{-\frac{4.8t}{\Delta t_{r_1}}}, \quad t \geq 0. \tag{6}
\]
Parameters $\zeta, \lambda_1$ and $\lambda_2$ can be commonly fixed at 0.237, 1 and 2, respectively [6]. There are then two speaker-dependent but utterance-independent parameters in the frequency domain, 

$[f_{0u}, f_{0l}]$: top and bottom frequencies of the voice register,

and five utterance-dependent but speaker-independent parameters in the RONDO-time space,

$n$: number of mountain-shaped patterns,

$\Delta t_{xi}$: response time for the $i$th rise/fall component,

$\Delta \lambda_{xi}$: amplitude of the $i$th rise/fall component, $x \in \{r, f\}$

$(t_{pi}, \lambda_{pi})$: peak of the $i$th mountain-shaped pattern, $i = 1, ..., n$.

Figure 1 shows the tone modeling with the mountain-shaped patterns and the association of the model parameters with the mountain-shaped pattern, where H, R, L and F indicate Tones 1 to 4. Because the close correlation of the model parameters with the peaks and valleys of a tone, it is reasonable to use the functional model while skeletonising an $F_0$ contour with its underlying $F_0$ peaks and valleys.

2.2. Outline of the algorithm

Let us take the example shown in Fig. 2 to describe the process of skeletonising an $F_0$ contour and demonstrate the performance of this method. Given the observed $F_0$ contour (the “+” sequence) shown in Fig. 2 (a), it is first represented in a parametric form based on the functional model using the method in [7]. Consequently, the peak resulting from a part of the set of model parameters, and a model-approximated contour is available, given the estimated parameters. Figure 2 (b) shows the model-approximated contours (the solid lines). Then, a valley is searched for the rise/fall components around the peak in the RONDO-time space. Figure 2 (c) shows the peak (the solid circles) and the valleys (the empty circles), giving the skeleton of the $F_0$ contour. The $F_0$ peaks and valleys can be then converted into the model parameters to recover the $F_0$ contour; a copy is shown in Fig. 2 (d) (the solid lines).

2.2.1. Improved parameter estimation

A method has been proposed to extract the tone peak and gliding features from observed $F_0$ contours that makes use of the functional model [7]. According to this method, the tone peaks are first determined by adjusting several baseline tone patterns to fit the $F_0$ contour fragment of a syllable with the analysis-by-synthesis-based pattern matching technique. Tone gliding features are then re-estimated after the determination of tone peaks with the criterion of minimising the error between the model-approximated contours and the observed ones.

To reliably skeletonise the $F_0$ contours, a few constraints were newly incorporated into the algorithm for re-estimation of the model parameters relative to the tone gliding features, namely, $\Delta t_{xi}$ and $\Delta \lambda_{xi}$. Let $(\lambda_{xi}, t_{xi})$ denote the observed $F_0$ valleys between the $i$th peak $(\lambda_{pi}, t_{pi})$ and the next, taking into account the voice frame probability to suppress the effect of potential $F_0$ extraction errors. The constraints for re-estimation of $\Delta t_{xi}$ and $\Delta \lambda_{xi}$ are listed below:

$\lambda_{pi} + \Delta \lambda_{fi} \leq \lambda_{xi} \times 1.1$ \hspace{1cm} (7)

$\lambda_{pi} + \Delta \lambda_{fi} = \lambda_{pi+1} + \Delta \lambda_{ri+1}$ \hspace{1cm} (8)

$\Delta \lambda_{fi} \geq 0.02$ \hspace{1cm} (9)

$\Delta \lambda_{ri+1} \geq 0.02$ \hspace{1cm} (10)

$0.05 \leq \Delta t_{xi} \leq (t_{xi} - t_{pi}) \times 1.1$ \hspace{1cm} (11)

$0.05 \leq \Delta t_{ri+1} \leq (t_{pi+1} - t_{xi}) \times 1.1$ \hspace{1cm} (12)

2.2.2. Valley search

Search of an $F_0$ valley $(t_{xi}, \lambda_{xi})$ for either of the $i$th rise and fall components is performed on the RONDO-Contours around the $i$th peak. The candidate for a valley, for example, $(t_{xi}, \lambda_{xi})$ is first set at the valley $(t_{pi}, \lambda_{pi})$ of the RONDO-Contours between the $i$th peak and the next. Then, the valley candidate is moved toward the $i$th peak through decreasing $t_{xi}$ with a very short interval (e.g., 0.005 sec) along the RONDO-contour until

$\lambda_{xi} - \lambda_{pi} \leq (\lambda_{xi} - \lambda_{pi}) \times 0.95$, \hspace{1cm} (13)

or $t_{xi} = t_{pi}$. In Eq. (13), the constant 0.95 is determined by considering the definition of $\Delta t_{xi}$ as the response time required for unit decay from 1 to 0.05, i.e.,

$D_a(\Delta t_{xi}) = (1 + \alpha\Delta t_{xi})e^{-\alpha\Delta t_{xi}} = 1 - 0.95$. \hspace{1cm} (14)

The relationship between $\alpha$ and $\Delta t_{xi}$ can be expressed as

$\alpha = \frac{\Delta t_{xi}}{\Delta t_{xi}}$. \hspace{1cm} (15)

It is noted that if the difference between $t_{xi}$ and $t_{xi+1}$ is less than a threshold, as the two valleys lie between them.

2.2.3. Parameter conversion

Given the peaks and valleys, the other model parameters necessary for recovering an $F_0$ contour are calculated as follows:

$\Delta t_{xi} = \max(0.05, t_{si} - t_{xi})$. \hspace{1cm} (16)

$\Delta \lambda_{xi} = \max(0.02, (\lambda_{xi} - \lambda_{pi}) \times 1.05)$, \hspace{1cm} (17)

$\Delta t_{fi} = \max(0.05, t_{si} - t_{pi})$, \hspace{1cm} (18)

$\Delta \lambda_{fi} = \max(0.02, (\lambda_{fi} - \lambda_{pi}) \times 1.05)$. \hspace{1cm} (19)

3. Experimental evaluation

Two experiments were conducted on 668 Chinese utterances to test the effectiveness of this method. Experiment 1 rates the perceptual similarity between the recovered tone and intonation patterns and the original. Experiment 2 judges the naturalness of synthetic speech with the effects of the modeling (i.e., automatic parameter estimation) and the skeletonising on the prosodic properties by comparing them with the original.

For reference, the average errors between the model-generated contours and the original were also calculated.
A: Types 1, 2, 3 in statements
B: Types 1, 2, 3 in lexically and grammatically unmarked yes-no questions
C: Types 1, 2, 3 in yes-no questions with interrogative particle "ma" in sentence-final positions
D0: Type 2 in yes-no questions with "shi4-bu2/shi4 structures
D1: Type 2 in yes-no questions with "X-mei2-X" structures
D2: Type 2 in yes-no questions with "X-le0/mei2-X" structures
E0: Type 2 in alternative questions with "X-hai2/shi4-Y" structures
E1: Type 2 in questions with "shi4-X-hai2/shi4-Y" structures
E2: Type 2 in questions with "hai2/shi4-X-hai2/shi4-Y" structures
F0: Type 2 in why "we4/she2/me0/shi2/ho4" questions
F1: Type 2 in when "she2/me0/shi2/ho4" questions
F2: Type 2 in what "she2/me0" questions

We recorded the 72 sentences twice in a sound-proofed room with a female speaker without expressive emotion. The voice register of the speaker [f0a, f0b] was fixed at [110 Hz, 500 Hz] for the parameter estimation. The 144 observed F0 contours were first automatically analysed using the method, after which the F0 peaks and valleys were manually checked with a visual inspection of the F0 contours, taking into account the underlying tones. The number of F0 peaks for each tone was basically determined according to the tone modeling shown in Figure 1. With model-generated F0 contours, we re-synthesised the 144 utterances for perceptual experiments using a tool called STRAIGHT [10].

3.1.2. Results

Table 1 shows the statistical results of the tone-related samples measured from the speech material, where μμ and σσ indicate the mean and variance of these manually checked model parameters (checked parameters), respectively; μμp and σσp indicate those predicted by the F0 peaks and valleys (prediction parameters). The columns μμ and σσ list the mean and variance of the errors between the checked and prediction parameters.

<table>
<thead>
<tr>
<th>Contour</th>
<th>Count</th>
<th>μμ</th>
<th>σσ</th>
<th>μμp</th>
<th>σσp</th>
<th>μμ</th>
<th>σσ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δε1</td>
<td>360</td>
<td>0.140</td>
<td>0.047</td>
<td>0.122</td>
<td>0.043</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>Δλ1</td>
<td>366</td>
<td>0.224</td>
<td>0.147</td>
<td>0.215</td>
<td>0.143</td>
<td>0.013</td>
<td>0.035</td>
</tr>
<tr>
<td>Δε1</td>
<td>382</td>
<td>0.139</td>
<td>0.047</td>
<td>0.134</td>
<td>0.055</td>
<td>0.019</td>
<td>0.027</td>
</tr>
<tr>
<td>Δλ1</td>
<td>382</td>
<td>0.196</td>
<td>0.129</td>
<td>0.188</td>
<td>0.122</td>
<td>0.007</td>
<td>0.015</td>
</tr>
</tbody>
</table>

The average errors between the model-generated contours and the observed ones were 6.38 Hz (1.64 Hz per 100 Hz) for those with the prediction parameters and 5.94 Hz (1.52 Hz per 100 Hz) for those with the checked parameters, respectively.

To test the similarity between the model-generated tone and intonation patterns and the original, we performed a perceptual experiment with 288 stimulus pairs, including 144 re-synthesised utterances with the checked parameters and 144 utterances with the prediction parameters. The stimuli were presented to two native speakers through headphones in a silent room. After hearing each pair of stimuli, the listener rated the similarity of the tone and intonation between them with a three-point scale, 0 (very different), 1 (similar), 2 (no different). The listeners were allowed to hear the same stimuli several times before making a judgment. The average scores for the checked and prediction parameters were 1.93 and 1.89, respectively, and no “very different” samples occurred. The experimental result indicated that the pitch targets, i.e., the F0 peaks and valleys over time, suffice to capture the nature of the tone and intonation patterns.

3.2. Experiment 2: Application of the recovered F0 contours to the unit selection for speech synthesis

Experiment 2 was conducted on 524 utterances from another speaker; the prosodic and spectral features extracted from these utterances were used as the targets to guide the unit selection for synthesis of the speech samples used in this experiment. The voice register of the speaker [f0a, f0b] was fixed at [120 Hz, 420 Hz]. The speech samples were prepared in five steps.

Step 1: Extracting the F0 contours from the 524 utterances and parameterising them based on the functional model.

Step 2: Skeletonising these F0 contours with the F0 peaks and valleys using the proposed method.

Step 3: Converting the F0 peaks and valleys into the model parameters using Eqs. (16)-(19).

Step 4: Recovering the F0 contours using these parameters.

Step 5: Corpus-based synthesis with the recovered F0 contours.

An example of the skeleton of F0 contours and F0’s recovered contours are shown in Figure 3. This example and all of the analyzed samples showed that the recovered F0 contours closely...
matched the original. The average errors are 6.01 Hz (2.0 Hz per 100 Hz) for those with the prediction model parameters, and 3.63 Hz (1.21 Hz per 100 Hz) for those re-synthesised by the auto-estimated model parameters.

The speech corpus used for the speech synthesis consists of 20-hour speech data from one speaker, and the unit selection algorithm is an updated version of the suggestion [11]; no diphone unit was used here. There exist five sub-costs to rate the difference between a candidate and the target. This experiment only focused on the effect of the $F_0$ contours on the naturalness of synthetic speech, taking one of the three tokens in turn: the recovered $F_0$ contours (hereafter, skeleton), those with auto-estimated model parameters (modelling), and the observed $F_0$ contours (original). As a result, we obtained 524 stimuli; each consists of the three synthetic speech samples in a random order. The stimuli were presented to the two natives through headphones in a silent room. After hearing a set of stimuli, the listener was asked to rate the difference in naturalness among them and answer the two following questions.

Is there any difference in naturalness among the three samples? If different, which is the best or the worst?

The experimental results are summarised in Table 2. According to the output of the unit selection module, there are 466 sets of samples, in which there exists at least one different unit candidate among them. On the other hand, there are 302 pairs of samples with identical unit candidates for each pair. According to the result shown in Table 2, human perception may perceive identical samples with different perceptual impressions of naturalness: improved 6.4% and degraded 2.65%. Taking into account the perceptual errors, the results obtained from 466 sets of samples indicate that the recovered $F_0$ contours can capture the essential properties of the observed $F_0$ contours, as proved in Experiment 1.

4. Remarks and future work

This paper presents a method for skeletonising an $F_0$ contour with its underlying $F_0$ peaks and valleys that makes use of a functional $F_0$ model. Several analysis and perceptual experiments were conducted on the speech material designed for studying Chinese tone and intonation patterns and speech synthesis. Experimental results indicated that the pitch targets play a prominent role in anchoring the tone and intonation patterns; the exact $F_0$ contours can be predicted from the $F_0$ peaks and valleys using the functional model, without losing the primary linguistic and para-linguistic information that it conveys.

Future work will include applying this $F_0$ skeletonising method to speech information processing, such as investigation of a pitch-target-based method for analysing and synthesising the tone and intonation patterns to improve the naturalness of the synthetic speech.

Acknowledgement This research was supported in part by the Telecommunications Advancement Organization of Japan.

5. References