Generative model of $F_0$ change field for Mandarin trisyllabic words

Dawei Xu1,*, Hiroki Mori2,† and Hideki Kasuya2,‡

1Graduate School of Engineering, Utsunomiya University, 7–1–2, Yoto, Utsunomiya, 321–8585 Japan
2Faculty of Engineering, Utsunomiya University, 7–1–2, Yoto, Utsunomiya, 321–8585 Japan

(Received 5 September 2003, Accepted for publication 6 November 2003)

Keywords: Speech synthesis, Word prosody, $F_0$ range, Mandarin Chinese

PACS number: 43.70.Fq, 43.72.Ja [DOI: 10.1250/ast.25.210]

1. Introduction

Mandarin is a tonal language, namely, every syllable has a lexical tone (high, rising, low, falling or neutral: T1, T2, T3, T4 or T0). The fundamental frequency ($F_0$) contour of Mandarin is largely affected by the lexical tone besides other factors such as intonation. However, the $F_0$ contour for a tone is also substantially affected by its neighboring tone, a phenomenon which is called tonal coarticulation [1,2]. With a view to adjusting a word’s $F_0$ contour to fit into a certain fixed sentence, the authors investigate the $F_0$ contour of Mandarin at the word level. In applications such as speech guidance for car navigation systems, there are some fixed sentences such as “qing3 zai4 he2 ping2 fan4 dian4 you4 zhu4an3 wan1 (Please turn right at the Peace Hotel)”, in which only some of the words (here, the place name shown in italics) need to be changed from case to case. Recording all the sentence combinations for all words would be time-consuming and take up much storage space. One reasonable solution is to insert the variable words into the fixed sentences. However, for these two tonal coarticulations, it should be possible to generate trisyllabic $F_0CF$. In this study, a generative model to predict trisyllabic $F_0CF$ was proposed and two experiments were conducted to test its validity. The first experiment investigated the prediction error and the second one evaluated whether this error would affect the perceived naturalness.

2. Model

A trisyllabic word here is treated as overlapped disyllabic units, where the first unit contains the first and second syllables and the second one contains the second and third syllables. Given a disyllabic $WF_0R$, disyllabic $F_0$ of the two units can be calculated from $WF_0R$ and disyllabic $RF_0CF$ using the following equations:

$$F_{L\text{di}} = F_L + \alpha \left(F_H - F_L\right) \quad (1.1)$$
$$F_{H\text{di}} = F_L + \beta \left(F_H - F_L\right) \quad (1.2)$$

where $F_{L\text{di}}$ and $F_{H\text{di}}$ are the low and high ends of disyllabic $F_0CF$, $F_L$ and $F_H$ are the low and high edges of the disyllabic $WF_0R$, and $\alpha$ and $\beta$ are the low and high ends of disyllabic $RF_0CF$ of a certain tone combination, respectively.

Assuming that the tonal coarticulations in the first and
second disyllabic units are independent, the trisyllabic \( F_{0}CF \) can be modeled as a union of the two disyllabic \( F_{0}CF \) as shown in Fig. 3. The high and low ends of trisyllabic \( F_{0}CF \), \( F_{h,tri} \) and \( F_{l,tri} \), are predicted as follows:

\[
\hat{F}_{h,tri} = \text{Max}(F_{h,di}^{(1)}, F_{h,di}^{(2)})
\]

\[
\hat{F}_{l,tri} = \text{Min}(F_{l,di}^{(1)}, F_{l,di}^{(2)})
\]

where \( F_{h,di}^{(1)}, F_{h,di}^{(2)}, F_{l,di}^{(1)} \) and \( F_{l,di}^{(2)} \) are the low and high ends of the first and second disyllabic units.

3. Experiment 1

This experiment investigated the prediction error of the proposed generative model. In this letter, trisyllabic words with only T1, T2, T3 and T4 are considered; those with a neutral tone are left to a future study. For the tone combinations containing successive T3, the former T3 is changed to T2 according to the tone sandhi rule in Mandarin.

Except for this kind of 7 tone combinations, 57 tone combinations were investigated.

3.1. Speech material and subject

Among disyllabic tone combinations, (T4, T3) occupies a maximum span in \( RF_{0}CF \) as can be seen in Fig. 2, which implies that the words in (T4, T3) are appropriate for estimating \( WF_{0}R \) with little error. In this experiment, ten disyllabic town names in this tone combination were prepared. For each of the 57 trisyllabic tone combinations, six trisyllabic town names were prepared. All these 342 trisyllabic words and two tokens of the 10 disyllabic words were inserted randomly into the carrier sentence of “qing3 dao4 shan1 hai3 guan1 xia4 che1 (Please get off the train at shan hai guan)”, where the italicized syllables corresponded to a town name. In total, 362 sentences were prepared.

A Mandarin-speaking native male adult, who was born and brought up in Beijing, served as the subject.

3.2. Results

The utterances were sampled at 11.025 kHz and \( F_{0} \) was analyzed every 10 ms. The averages of minima and maxima \( F_{0} \) of the six words were assigned as a measured \( F_{0}CF \) \([F_{l,tri}, F_{h,tri}]\). Estimated by the average of the \( F_{0}CFs \) of the 20 disyllabic words, disyllabic \( WF_{0}R \) was \([77.30, 88.63]\) in semitone. With this \( WF_{0}R \) and the \( RF_{0}CF \) shown in Fig. 2, predicted trisyllabic \( F_{0}CF \) ends \( \hat{F}_{l,tri} \) and \( \hat{F}_{h,tri} \) were calculated according to the generative model for all the 57 tone combinations, respectively. The prediction errors for \( \hat{F}_{l,tri} \) and \( \hat{F}_{h,tri} \) were defined as their difference from \( F_{l,tri} \) and \( F_{h,tri} \).

The prediction errors of low and high ends for all tone combinations are shown in Table 1, which are classified by whether the predictions were on the first or second disyllabic units. On the low end, for the cases in which the trisyllabic ends were predicted by the first disyllabic units, predicted values tended to be lower than the measured values; the negative maximum prediction error reached \(-2.94\) st. In contrast, for the cases in which the trisyllabic ends were predicted by the second disyllabic units, prediction errors were not so large. On the high end, although the averages of prediction errors were small, the positive and negative maxima were over 1 st. Whether the prediction errors affected the naturalness of synthesized speech was evaluated through the following perceptual experiment.

4. Experiment 2

The experiment evaluated the perceptual significance of the prediction errors. Town names synthesized by the ARX method [5] were inserted into the sentence mentioned above. Sentences with words in tone combinations having the largest

| Table 1 Prediction errors for the tone combinations according to whether the prediction is on the first or second unit. |
|-------------------------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                                   | Low end [st]    | High end [st]   |
| First                                            | Second          | First           | Second          |
| Average                                         | -1.74           | -0.26           | 0.18            | 0.37            |
| Positive maximum                                 | 0.00            | 0.49            | 1.15            | 1.53            |
| Negative maximum                                 | -2.94           | -1.03           | -1.49           | -1.44           |
and next-largest prediction errors were evaluated as to whether they sounded natural.

4.1. Speech material and method

For the low end and high end, tone combinations with the largest and next-largest prediction errors are shown in Table 2. For each tone combination, two words were used for the experiment. For these eight words, two categories of speech were synthesized and then evaluated. In Category 1, there were re-synthesized words with the original $F_0$CFs. In Category 2, there were words with the predicted $F_0$CF. If the subjects were only required to evaluate these two categories of test data, there would be a risk of overestimating the differences of these two categories. But after the subjects had heard all the speech data in the first session of evaluation and had the impression that all the data were nearly natural, the two categories of data in the succeeding sessions might be evaluated as having the same score. To avoid these two kinds of risk, speeches of a third category with different perceptual impressions were prepared. In Category 3, words were synthesized with both the low and high ends being 3 st higher than the original $F_0$CFs, where the scale of 3 st is about the same as the largest prediction error of $(T1, T2, T1)$, as shown in Table 2.

Four male native adults served as subjects. On a 5-grade scale, they evaluated the naturalness of the sentences which appeared five times randomly.

4.2. Results

The average naturalness scores for the three categories are shown in Table 3. Although the words in Category 2 were synthesized with the predicted $F_0$CFs with maximum prediction errors, they achieved nearly the same high scores as those in Category 1. Since the highest score was 5, the two categories were both evaluated as being reasonably natural. Therefore, the prediction errors do not affect the perceived naturalness of the generated sentences. On the other hand, the sentences in Category 3 were evaluated as being rather unnatural. This implies that when words are inserted into a sentence, it is important to generate their $F_0$ with an appropriate $WF_0R$ not only in its span, but also in its level.

5. Discussion

Lu [6] and Feng [7] indicated that in Mandarin, syllables in trisyllabic words tend to be of shorter duration than those in disyllabic words. This could explain why the measured low ends were higher than the predicted ones. For the first disyllabic unit in the trisyllabic word, when the $F_0$ started from the same value as those disyllabic words, the syllables in trisyllabic words with shorter duration produced higher low ends than those in disyllabic words. Similar phenomena were observed on high ends in some tone combinations which were led by T3. Since the $F_0$ started in a low tone, the succeeding syllables in T4 turned out to have lower high ends. However, since the difference was not perceptually important, this kind of compensation was omitted in the model.

6. Summary

Under the assumption that the two tonal coarticulations in trisyllabic words are independent, the $F_0$CF of trisyllabic words was modeled as the superior set of the two $F_0$CFs of the overlapped first and second disyllabic units. The magnitudes of the prediction errors were within perceptual tolerance. Experiments with more data from various speakers are now under way.

References


---

Table 2  Tone combinations with the largest and next-largest prediction errors.

<table>
<thead>
<tr>
<th>Tone combination</th>
<th>Error [st]</th>
<th>Tone combination</th>
<th>Error [st]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T1, T2, T1)</td>
<td>−2.94</td>
<td>(T3, T4, T2)</td>
<td>1.53</td>
</tr>
<tr>
<td>(T2, T3, T4)</td>
<td>−2.54</td>
<td>(T4, T1, T3)</td>
<td>−1.49</td>
</tr>
</tbody>
</table>

Table 3  Average naturalness score.

<table>
<thead>
<tr>
<th>Naturalness score</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.49</td>
<td>4.59</td>
<td>1.83</td>
</tr>
</tbody>
</table>