WORD-LEVEL F0 RANGE IN MANDARIN CHINESE AND ITS APPLICATION TO INSERTING WORDS INTO A SENTENCE

Dawei Xu†, Hiroki Morii, and Hideki Kasuyai

† Graduate school of Engineering, Utsunomiya University, Japan
‡ Faculty of Engineering, Utsunomiya University, Japan
E-mail: {xu, hiroki, kasuya}@klab.ee.utsunomiya-u.ac.jp

ABSTRACT
This paper considers an automatic voice response application in which a word utterance is inserted into a fixed carrier sentence. An important task here is to adjust the $F_0$ contour of the inserted word according to the $F_0$ context of the carrier sentence. Instead of generating the $F_0$ contour on syllable basis, we employ an approach to adjust the $F_0$ contour of the whole word. In this approach, two questions arise: (a) how to evaluate the $F_0$ context and (b) how to adjust the $F_0$ contour suitably for the context. We have found that the $F_0$ contour of a word can be appropriately regulated in a tone-independent word-level $F_0$ range ($WF_R$). After estimating the $WF_R$s of the preceding and succeeding words, the $WF_R$ of the inserted word is set at the mean of these $WF_R$s. The $F_0$ contour of the inserted word is then mapped to the $WF_R$ taking into account the tone combination of the word. A perceptual evaluation experiment showed that the adjusted $F_0$ was coordinated well with the context.

1. INTRODUCTION
There are many applications of automatic voice response systems in which a single word utterance is inserted into a fixed carrier sentence, e.g. voice guided car navigation, telephone number consulting service on the phone, etc. In the car navigation, for instance, a large number of building names might be inserted into such a carrier sentence as “Please make a left turn, when you see *** building,” where *** is the name of the building. One of the important problems here is how to adjust the fundamental frequency ($F_0$) contour of the building name to be inserted into the $F_0$ context of the fixed carrier sentence.

In Mandarin Chinese, this task turns to that the $F_0$ contour of the inserted word should be arranged to sound natural while maintaining the tones in the word unchanged. In a syllable-based approach of composing the inserted word’s $F_0$ contour from the syllabic $F_0$ contours which are generated separately, there are many factors affecting naturalness of the synthetic sound. They include at least the syllabic position in a word, assimilation between the adjacent tones, and interaction between the segmental features and the tones. It seems difficult to model all these factors in an efficient way.

We employ a word-based method instead to generate the $F_0$ contour. All the needed words’ $F_0$ contours in a certain context are stored as a synthesis unit. Since it is not always probable to collect the words’ $F_0$ contours in all the possible contexts, the problem is then how to adjust a sample $F_0$ contour to suit for all the insertion positions in a sentence. In a similar way to regulating Chinese four tones of High, Rising Low, Falling (T1, T2, T3 and T4, respectively) in a relative $F_0$ range of the syllable, we regulate the word’s $F_0$ pattern in a tone-independent word-level $F_0$ range (D. Xu et al, 1999). By describing quantitatively the intonational effect on word’s $F_0$ range, we can adjust the sample $F_0$ contour to suit for every place in a sentence.

2. WORD-LEVEL F0 RANGE AND F0 CHANGE FIELD

2.1 Definition of Word-level $F_0$ Range and $F_0$ Change Field
Depending on the number of syllables in a word, the word-level $F_0$ range ($WF_R$) in a position of a sentence is defined as an $F_0$ range between the “highest” and the “lowest” $F_0$ values of the $F_0$ contours of all the tone combinations. A word’s $F_0$ change field is defined as a range between the maximum and minimum values of the $F_0$ contour. Figure 1 shows the relation between the $WF_R$ and the $F_0$ change fields. A word’s $F_0$ change field, which is between the high and low ends, is a sub-range comprised in the $WF_R$ that is designated as the range between the high and low edges. The $F_0$ change fields are expected to be stable in the $WF_R$ for an individual and to be scarcely effected by the word position in a sentence.

![Fig. 1: The concept of the WF0R and the F0 change field.](image)

2.2 Characteristics of $F_0$ Change Field within $WF_0R$
An experiment was performed to test whether the $F_0$ change field is stable in a $WF_R$ for an individual on different days and whether the word’s position in a carrier sentence affects $F_0$ change field.

Materials
Ten disyllabic words of city names were selected for each of the
16 tone combinations considering the balance of Chinese vowels. For the convenience of syllable segmentation, all the syllables of the words start with consonants. The 160 words were inserted into 3 short carrier sentences as follows:

Ini.) “shang4 hai3 chel zhan4 hen3 da4" (Shanghai’s railway station is very big);

Mid.) “qin3 dao4 shang4 hai3 xia4 chel" (Please get off the train at Shanghai);

Fin.) “xia4 zhan4 dao4 da2 shang4 hai3” (The next station is Shanghai);

in which the italic part is an example of the city name. The insertion position in there sentences are the initial, middle and final.

A native male adult read every three sentences once keeping the intonation as the same as possible. Words for each sentence are randomly selected. He also read the “Mid.” sentence on another day to investigate day-to-day variation of $F_0$ contours. The utterances were sampled at 11.025 kHz. The $F_0$ was analyzed every 10 ms and a 5-point median filter was applied to decrease the effect of consonants on the $F_0$ contours.

For each tone combination in the same position, the average of the maximum $F_0$ values of the 10 words was assigned to the high end of the averaged $F_0$ change field, while that of the minimum was to the low end. The maximum of the high ends of the averaged $F_0$ change fields of all the tone combinations was assigned to the high edge of the $WF_0R$, while the minimum of the low ends was to the low edge of the $WF_0R$. The high and low ends of the averaged $F_0$ change fields were further normalized into values between 0 and 1, which are corresponding to the low and high edge of the $WF_0R$.

**Stability of Normalized $F_0$ Change Fields**

A one-way ANOVA test (p<0.05) was carried out on each end of the normalized $F_0$ change fields that were calculated from the word utterances in the “Mid.” carrier sentence read on two separate days. Of all the 32 ends, only 3 ends were found to yield statistically significant differences. This means that the individual read the words in the same carrier sentence with almost the same normalized $F_0$ change fields.

**$F_0$ Change Fields in Different Positions**

The averaged $F_0$ change fields in the 3 positions are shown in Fig. 2 in semitone (ref. 1Hz), where each panel is for a tone combination. The high and low ends of the $F_0$ change field are shown as the high and the low end of the column, together with their standard deviations. There is an obvious trend of dropping down of the $F_0$ change fields over the 3 positions. A one-way ANOVA test (p<0.05) on the $F_0$ change fields in the 3 positions showed that the position had significant effect on most of the high and low ends. The $WF_0R$ for the initial, middle and final positions are [78.5, 89.0], [77.5, 88.1] and [77.1, 87.2] in semitone, respectively.

**Comparison of Normalized $F_0$ Change Fields in Different Positions**

The normalized $F_0$ change fields in each position are shown in Figure 3. A one-way ANOVA test (p<0.05) showed that the
in the middle and final positions. The average of the differences in the normalized high end of T1 between those fields dropping down. We evaluated the final lowering effect by final lowering instead causes the ends of normalized WF occupy big value of the individual’s normal voice, and that final T4s the minimum of T3 in the final word is almost the minimum on 1948). On the other hand, the data of this experiment shows that significant difference detected by ANOVA test (p<0.05). It is shown in Table 1, in which the boldfaced number refers to a significant difference. Therefore, there is little difference as a whole between the normalized $F_o$ changes of the initial and the middle position.

In summary, the normalized $F_o$ change field for each tone combination is constant within the $WF_o$R in the same carrier sentence. The normalized $F_o$ change fields in the initial position are almost the same as that in the middle position, while those in the final position can be approximated by making modification on those in the middle position.

### Table 1. The difference between the averaged normalized $F_o$ change fields of the final position and the middle position.

<table>
<thead>
<tr>
<th>Cat.</th>
<th>L</th>
<th>H</th>
<th>Cat.</th>
<th>L</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T1,T1)</td>
<td>-0.18</td>
<td>-0.16</td>
<td>(T3,T1)</td>
<td>-0.02</td>
<td>-0.11</td>
</tr>
<tr>
<td>(T1,T2)</td>
<td>-0.17</td>
<td>-0.16</td>
<td>(T3,T2)</td>
<td>-0.02</td>
<td>-0.14</td>
</tr>
<tr>
<td>(T1,T3)</td>
<td>0.04</td>
<td>-0.07</td>
<td>(T3,T3)</td>
<td>-0.02</td>
<td>-0.23</td>
</tr>
<tr>
<td>(T1,T4)</td>
<td>-0.23</td>
<td>-0.08</td>
<td>(T3,T4)</td>
<td>-0.09</td>
<td>-0.09</td>
</tr>
<tr>
<td>(T2,T1)</td>
<td>-0.09</td>
<td>-0.16</td>
<td>(T4,T1)</td>
<td>-0.13</td>
<td>-0.09</td>
</tr>
<tr>
<td>(T2,T2)</td>
<td>-0.20</td>
<td>-0.14</td>
<td>(T4,T2)</td>
<td>-0.10</td>
<td>-0.13</td>
</tr>
<tr>
<td>(T2,T3)</td>
<td>0.00</td>
<td>-0.20</td>
<td>(T4,T3)</td>
<td>-0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>(T2,T4)</td>
<td>-0.25</td>
<td>-0.04</td>
<td>(T4,T4)</td>
<td>-0.24</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

(T3, T1) is 0.13, in which the high ends of the $F_o$ change fields occur at T1.

We also found that T4 in the final position reached almost the low edge of $WF_o$R. Shih (1996) also incorporated this fact into his $F_o$ control rules in the text-to-speech system, where he assigned targets H and L to T4 in the final position, while targets H and M to that in the other positions.

It is thought to be efficient that a set of approximate normalized $F_o$ change fields in the final position could be obtained by modifying that in the middle position. The first modification is to change the low end of final T4 to 0.05. The second one is to decrease the other significantly different ends by 0.13 that is caused by final lowering. A one-way ANOVA test (p<0.05) between the $F_o$ change fields in the final position and the generated fields by the method mentioned above shows only 3 significant differences.

In summary, the normalized $F_o$ change field for each tone combination is constant within the $WF_o$R in the same carrier sentence. The normalized $F_o$ change fields in the initial position are almost the same as that in the middle position, while those in the final position can be approximated by making modification on those in the middle position.

### 3. METHOD TO INSERT WORDS INTO A SENTENCE

#### 3.1 Method

The normalized $F_o$ change fields mentioned above are used to compute $WF_o$R from the $F_o$ change field, and vice versa. If the $WF_o$R in the insertion position is specified, the $F_o$ contour can be arranged by mapping the normalized $F_o$ contour in the lexicon into the corresponding $F_o$ change field in logarithmic scale. It is assumed here that the $WF_o$R of the insertion position is the mean of those of the preceding and the succeeding words. Therefore, the method to insert a word into a carrier sentence is as follows:

1. Estimate the $WF_o$R of the preceding and the succeeding words using the $F_o$ change fields.
2. Determine the $WF_o$R in the insertion position as the mean of the $WF_o$R of the preceding and succeeding words.
3. Convert the $WF_o$R in the insertion position to the
In order to evaluate the word insertion method described above, a perceptual evaluation experiment was performed. A city name in each of the 16 disyllabic tone combinations was inserted into the following carrier sentence:

“qing3 zai4 liang3 dian3 zhil qian2 dao4 da2 ** zhuan3 che1” (Please come to ** before 2 o’clock and change train line).

A male read the carrier sentence in a sound proof room. He also read sixteen disyllabic city names inserted in the “Mid.” carrier sentence described in section 2.2. All the utterances were digitized at a sampling frequency of 11.025 kHz and were subjected to the analysis and re-synthesis experiments using the STRAIGHT method (Kawahara, 1996). The speech segment of the city name was isolated from the sentence utterance visually on the screen. Two sets of re-synthetic speech were generated: 1) the \( F_0 \) contour of an inserted word was generated by the proposed rule, and 2) the original \( F_0 \) contour of the isolated word was used with no adjustment. Three Chinese made a paired comparison test in terms of naturalness of re-synthetic speech. They evaluated ten times for each pair of the re-synthetic speech. They chose the speech generated by the proposed method as more natural at a rate of 88.1%.

### 3.3 Discussion

In order to get insight into the reason for the improvement of naturalness of a word inserted by the proposed method, the maximum and minimum values of the \( F_0 \) contours generated by the two methods were compared. Figure 4 illustrates the generated maximum (triangle) and minimum (circle) \( F_0 \) values subtracted from those of the original \( F_0 \) values against naturalness degree, which is the selected rate of the proposed method as more natural.

Among the words with the high naturalness degree, the circles in the lower-right corner show that the proposed method modified little of the \( F_0 \) minimum values. Meanwhile, the \( F_0 \) maximum values of these words were decreased at least by 4 semitones, which resulted in that the spans of the \( F_0 \) change fields were reduced. The other circles and triangles in the right half panel shows that the minimum values were adjusted as well as the maximum values. The spans of the \( F_0 \) change fields of these words were not modified much while the levels were shifted. The circles and triangles in the lower-left corner, in the tone combination of (T1, T3) and (T3, T3), show that the isolated \( F_0 \) contours had almost the same \( F_0 \) minimum values as the adjusted ones. Therefore, there was little perceived difference in the low tone of T3. The subjects did not show apparent preference for these word pairs.

The perceptual evaluation indicated that the proposed method was validated in adjusting both the level and span of the \( F_0 \) change fields when the isolated \( F_0 \) contours differed much from the expected ones. As for those isolated \( F_0 \) contours that gave reasonable level tone such as T3, the method did not affect significantly on naturalness.

### 4. SUMMARY

By investigating the \( F_0 \) range of disyllabic words of Mandarin at the word level, it has been found that the words’ \( F_0 \) change fields can be well regulated in the tone-independent Word-level \( F_0 \) Range. The intonational declination of \( F_0 \) in a sentence is described by the decreasing high and low edges of \( W_{F0}R \). Although the normalized \( F_0 \) change fields in the final position of a sentence differ from those in the middle position to some extent, they can be generated from that in the middle by rules. A method is proposed to control both the level and span of the \( F_0 \) change field of a word to be inserted into a carrier sentence. A perceptual evaluation experiment indicated that the proposed method can generate reasonable \( F_0 \) contours for the inserted word.

### 5. REFERENCES