Invariance of Relative F_0 Change Field of Chinese Disyllabic Words

Dawei Xu†, Hiroki Mori‡, and Hideki Kasuya‡

† 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

In automatic voice response systems where a large number of words are inserted into fixed sentences, such as in voice-guided car navigation systems, one of the most important problems is the adjustment of the fundamental frequency (F_0) contour of the inserted word to suit the F_0 context of the fixed sentence. The effects of intonation and tone on the F_0 contours of Chinese words can be described in terms of a word-level F_0 range (WF_0R) and an F_0 change field (F_0CF). WF_0R in any position of a sentence is a tone-independent general F_0 range, whereas F_0CF is an F_0 range taking the tone combination of words into account. Relative F_0CF is regulated in reference to WF_0R. If WF_0R is used to represent the declination of a sentence, the relative F_0CF should be invariant but dependent on the tone combination of a word. This paper examines the invariance of the relative F_0CF among individuals. From an analysis of four native speakers' utterances of 160 words in the initial, middle and final parts of three carrier sentences, conducted over 2 days, we show that: (1) Chinese speakers read words in the same sentence position with stable relative F_0 change; (2) the relative F_0CFs in the middle position of a sentence are generally the same as those in the initial position, but slightly different from those in the final position; and (3) the relative F_0CFs reveal that the effects of tone on F_0 contour is individual independent.

1. Introduction

In many applications of automatic voice response systems, such as voice-guided car navigation and telephone-based financial report services, sentences are fixed and key words are inserted as required. In car navigation, a many street names may be inserted into such a fixed sentence, i.e. "Please make a left turn at ***," where *** is the name of a street. One of the most important problems here is how to adjust the fundamental frequency (F_0) contour of the inserted street name so that it fits the F_0 context of the fixed sentence.

In Mandarin Chinese, this task requires that the F_0 contour of the inserted word be arranged so that it sounds natural while the tones in the word remain unchanged. In a syllable-based approach to composing the F_0 contour of inserted words from the syllabic F_0 contours that are generated separately, there are numerous factors affecting the naturalness of the synthetic sound. As a minimum, these factors include the syllabic position in a word, assimilation between adjacent tones, and interaction between segmental features and tones. It appears difficult to model all these factors in an efficient way.

In this study, we employ a word-based method for generating the F_0 contour. The F_0 contours of all words in a certain context are stored as a synthesis unit. As it is not feasible to collect the F_0 contours of a single word in all possible contexts, the problem is then how to adjust a sample F_0 contour to suit all possible insertion positions in a sentence. Similar to the way in which Chinese is regulated in terms of four tones: High, Rising, Low, Falling (T1, T2, T3 and T4) in the relative F_0 range of the syllable, we have developed a scheme in which the F_0 patterns of words are regulated within a tone-independent word-level F_0 range [1]. By describing the effects of intonation on the F_0 range of a word quantitatively, we can adjust the sample F_0 contour to suit any position in a sentence. In the preceding study, we verified the validity of this method using the speech of one individual, comparing the regulated F_0 range of identical words uttered on different days and in different positions within carrier sentences. We also demonstrated the validity of this method through perceptual experiments on the naturalness of the re-synthesized words [2]. There remains a certain degree of uncertainty concerning whether this approach is applicable for general speakers. In the current paper, we clarify this issue by investigating whether the regulated F_0 range of a word is invariant between individuals. Based on material spoken by four individuals, we answer the following questions: (1) Do individuals maintain the same regulated F_0 range for the same tone combination on different days? (2) Is the regulated F_0 range of a word invariant when the word is inserted into the initial, middle or final part of a sentence? (3) Are there commonalities between the regulated F_0 ranges of given words between individuals?

2. Word-level F_0 Range and F_0 Change Field

Depending on the number of syllables in a word, a word-level F_0 range (WF_0R) in a given position in a sentence is defined as the F_0 range between the "highest" and "lowest" F_0 values, termed the high and low edges, of F_0 contours of all tone combinations. A word's F_0 change field (F_0CF) is defined as the range between the maximum and minimum values, termed the high and low ends, of the F_0 contour.

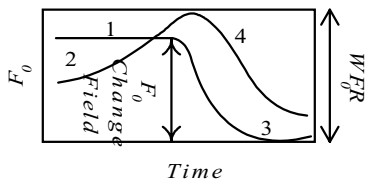

Figure 1: WF_0R and F_0 change field

Figure 1 shows the relationship between WF_0R and F_0CF.

The F_0CF of a word is a sub-range of the WF_0R, and the relative F_0CF is defined in relation to WF_0R, taking a value between 0 and 1. Relative F_0CFs are expected to be stable between all individual speakers, regardless of the WF_0R, and to scarcely be affected by the position of a word within a sentence.

3. Experiment

To answer the three questions considered in this study, we compare the relative F_0CFs of four native speakers on two different days, for words inserted in the initial, middle and final position of a sentence. We further compare the relative F_0CFs of four narrators mutually.

3.1 Material

Ten disyllabic city names were selected for each of the 16 tone combinations, taking the balance of Chinese vowels into account. For the convenience of syllable segmentation, all the initial syllables begin with a consonant. Then, 160 words were inserted into three short carrier sentences as follows:

Ini.) "*shang4 hai3* che1 zhan4 hen3 da4" (*Shanghai*'s railway station is very big);

Mid.) "qin3 dao4 *shang4 hai3* xia4 che1" (Please get off the train at *Shanghai*);

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

where the italicized syllables correspond to a city name. These disyllabic words were inserted into the carrier sentences in the initial, middle and final positions.

3.2 Data analysis

Two native males and females read each of three sentences once, maintaining the same pattern of intonation as far as possible. Words for each sentence were randomly selected. The subjects also read the same sentences 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 nonlinear double smoothing filter was applied to reduce the errors at transference from a consonant to a vowel.

For each tone combination in the same position, the average of the maximum F_0 values of the 10 words was used as the high end of the averaged F_0CF, and the average minimum was used as the low end. Then, the maximum high end of the average F_0CF for all tone combinations was used as the high edge of the WF_0R, and the minimum low end was used as the low edge of the WF_0R. The high and low ends of the averaged F_0CFs were normalized to values between 0 and 1. An example of the normalized relative F_0CF of subject MXZ is shown in figure 2; the inner circle of the radar chart represents 0, the outer circle represents 1, and every axis in the chart represents a tone combination. We can see that the high and low ends of tone combinations differ from each other. In a tone combination such as (T4, T3), the span of relative F_0CF is almost 1, whereas for that for (T2, T1), the span is smaller than 0.5.

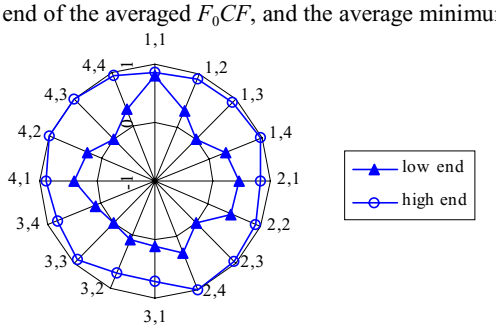

Figure 2: Relative F_0 change field for MXZ.

3.3 Results

3.3.1 Day-to-day invariance of relative F_0CF

The span and median of WF_0Rs for the middle position for all four subjects on two separate days are shown in table 1 in semitones (ref. 1 Hz). We see that the span of MXZ's WF_0R on the second day is reduced by about 8%, while that of FYS's is extended by about 9%. The other two subjects' WF_0Rs remain relatively unchanged over the two days.

Table 1. Span and median of WF_0R over two days

Subject	Span (semitone)		Median (semitone)	
	1st day	2nd day	1st day	2nd day
MXZ	11.0	10.2	83	82.7
MJJ	16.6	17.0	87.8	87.7
FYS	12.5	13.8	93.4	93.6
FYH	13.6	13.0	96.5	96.8

All the low and high ends for the 4 subjects are shown in Figure 3. We can see that as a whole, the two sets calculated from the two separate days are quite similar. We further notice that although the spans of WF_0Rs for MXZ and FYS on two days differ to a certain degree, the relative F_0CFs are quite similar.

A one-way ANOVA test ($p<0.05$) was carried out on each end of the relative F_0CFs calculated from the word utterances in the "Mid" carrier sentence read on two separate days. Table 2 shows the numbers of ends found to yield statistically significant differences. Of all the 16 tone combinations, only a few ends exhibit a significant difference for all four subjects. From these results we can conclude that the individuals read the given words in the same carrier sentence with almost the same relative F_0CF.

Table 2. Number of relative F_0CF ends found to change significantly between the two days

Subject	High end		Low end	
	Num.	Per.(%)	Num.	Per.(%)
MXZ	1	6.3	0	*0.0*
MJJ	*2*	*12.5*	*3*	*18.8*
FYS	2	12.5	0	*0.0*
FYH	1	6.3	3	*18.8*

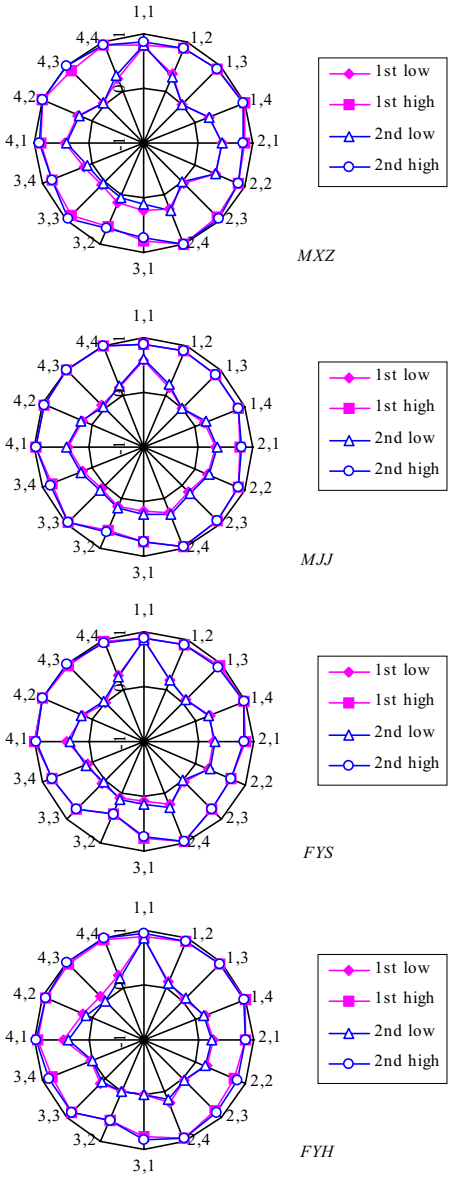

Figure 3: The relative F_0CFs of subject MXZ, MJJ, FYS, FYH on two separate days.

3.3.2 Sentence position independency of relative F_0CF

The WF_0Rs in the three positions of the three carrier sentences read on the second day are shown in figure 4 in semitones, where each panel represents a single subject. We can see that the high and low WF_0R edges for the four subjects become lower in the order of position.

A one-way ANOVA test ($p<0.05$) on the relative F_0CFs for all the three positions showed that the relative F_0CFs are dependent on the word position in a sentence. However, this result was not sustained when the positions were examined two at a time.

A comparison of the relative F_0CFs for the initial and middle positions, revealing the number of differing low and high ends, is shown in table 3. For three of the subjects (excluding FYH), only a few ends were found to be significantly different. Therefore, there is in general little difference between the relative F_0CFs for the initial and middle positions.

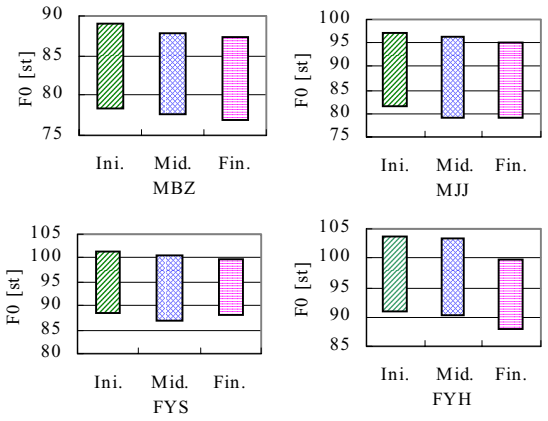

Figure 4: The WF_0Rs in the initial, middle and final position for the 4 subjects.

A comparison of the relative F_0CFs for the middle and final positions is shown in table 4. For all 4 subjects, about 50% of ends are significantly different, and among these ends, we find that almost all the values in the final position are smaller than those in the middle position. We attribute this to final lowering phenomena, where a final syllable or word in a sentence is read lower than the former part of the sentence.

Table 3. Number of ends that significantly changed between words in the initial and middle positions

subject	High end		Low end	
	Num.	Per.(%)	Num.	Per.(%)
MXZ	0	0.0	1	6.3
MJJ	1	6.3	2	12.5
FYS	0	0.0	1	6.3
FYH	3	18.8	9	56.3

Table 4. Number of ends that significantly changed between words in the middle and final positions

Subject	High end		Low end	
	Num.	Per.(%)	Num.	Per.(%)
MXZ	10	62.5	11	68.8
MJJ	12	75.0	12	75.0
FYS	12	75.0	6	37.5
FYH	7	43.8	9	56.3

3.3.3 Individual-independency of relative F_0CFs

The last question to be addressed is whether the same set of relative F_0CFs is shared among individuals, and if it is not, which part is common and which part is individual-dependent?

A one-way ANOVA test ($p<0.05$) revealed that among the

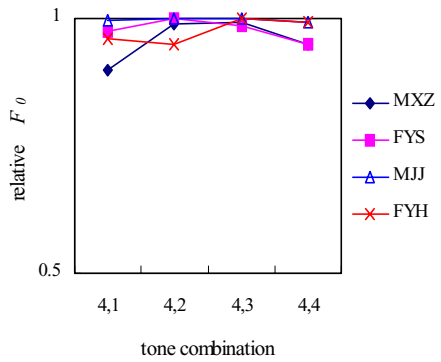

Figure 5: High ends in tone combinations of (T4, T1), (T4, T2) (T4,T2), (T4, T4)

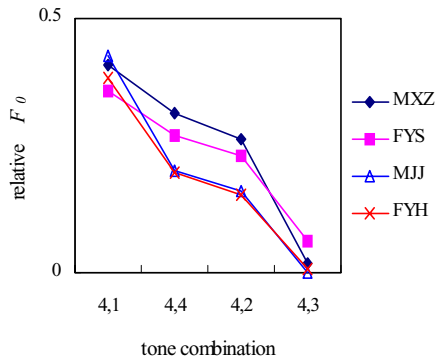

Figure 6: Low ends in tone combinations of (T4, T1), (T4, T4) (T4,T2), (T4, T3)

total of 32 high and low ends, 20 ends are found to be significantly different, and most of these were low ends. This means that we can not simply apply the set of relative F_0CFs for one speaker to another. But this assertion remains a further study by perceptual experiment.

However, after scrutinizing the relative F_0CFs of the four subjects it is found that the effect of tones on the relative F_0CFs shares the same pattern among individuals. For example, all the high ends in tone combinations such as (T4, T1), (T4, T2) , (T4, T3), and (T4, T4) for the four subjects are about 1, as illustrated in figure 5. The low ends of these tone combinations are illustrated in figure 6. For these subjects, the low end in (T4, T1) is highest, those in (T4, T4) and (T4, T2) are lower, and that in (T4, T3) is lowest, near 0. This pattern is common to all four subjects.

Therefore, we can conclude that although the values of relative F_0CFs differ individually to some degree, the effect of tone on relative F_0CFs shares the same pattern among individuals.

4. Discussion

After introducing the concepts of WF_0R, F_0CF and relative F_0CF, the effects of intonation and tone on F_0 contour realization can be represented separately. The WF_0R reveals the intonational declination of F_0 along a sentence, and relative F_0CF indicates the tonal effect. The WF_0R for the same position in the same sentence can also vary between days, as seen for MXZ and FYS, whereas the relative F_0CF changes little. This means that when words are recorded on different days, the F_0CF of the words in the same tone combination may vary because the WF_0R on different days may vary. This can be checked and any necessary modifications made by referring to the standard of the relative F_0CF, which varies only negligibly.

In the debate of what tonal targets Chinese tones contain, Y. Xu *et al.* (2001) proposed a framework in which tones 1 and 3 contain high and low static targets, whereas tones 2 and 4 contain rising and falling dynamic targets. The current study differs with his work in that we consider the F_0 range only; alignment and F_0 changing speed were not considered. However, our study does not conflict with the framework of Y. Xu *et al.* in that the maxima and minima usually occur among static targets or dynamic targets in duration.

5. Summary

In this study, we defined a word-level F_0 range and a relative F_0 change field in order to separate the effects of intonation and tone on the F_0 contour of words inserted in fixed sentences. The intonational effect is reflected in the word-level F_0 range, and the tonal effect is indicated in the normalized relative F_0 change field. We showed the generality of this framework by investigating the invariance of the relative F_0 change fields using the utterances of four native speakers. We found that (1) all subjects have the same relative F_0 change fields on different days, (2) most subjects maintain the same relative F_0 change fields when words are read in the initial and middle parts of a sentence, (3) due to final lowering, the relative F_0 change fields for words in the final position are lower than those for words in the middle, and (4) although the value of F_0 change fields may vary somewhat between individuals, the pattern of the relative F_0 change fields determined by tone is individual-independent. The invariance of the F_0 change fields between individuals suggests that the proposed framework for separating intonational and tonal effects is applicable to unspecific speakers.

6. References

[1] Xu, D., Mori H. and Kasuya H., "The Prosody Control of Chinese Speech Considering the F_0 Range in Word Level," *Proceedings of the 1999 autumn meeting of the Acous. Soc. of Japan*, 319-320, 1999 (in Japanese).

[2] Xu, D., Mori, H. and Kasuya H., "Word-level F_0 Range in Mandarin Chinese and Its Application to Inserting Words into a Sentence", *Proc. of ICSLP2000* 3:338-441, 2000.

[3] Xu, Y. and Wang, Q.E., "Pitch Targets and Their Realization: Evidence from Mandarin Chinese", *Speech Communication*, 33(4):319-337, 2001.

[4] Ladd, D., *Intonational Phonology*, Cambridge University Press, Cambridge, 1996.