PAPER

Statistical examination of invariance of relative F_0 change field for Chinese disyllabic words

Dawei Xu[1,*], Hiroki Mori[2,†] and Hideki Kasuya[2,‡]

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

(*Received 23 July 2001, Accepted for publication 27 May 2002*)

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. In Mandarin Chinese, it is required that the effects of tone and intonation on F_0 contours be represented separately. We proposed a scheme to solve the problem in terms of a word-level F_0 range (WF_0R) and a set of relative F_0 change fields. WF_0R in any position of a sentence is a tone-independent general F_0 range to represent the intonational effect; whereas F_0 change field (F_0CF) is an F_0 range that accounts for the result of both the tone combination of words and the intonation. Relative F_0CF is regulated in reference to WF_0R and represents tonal effect on F_0. In this paper, we statistically examine the invariance of the relative F_0CF with various speakers' speech data. From an analysis of four native speakers' utterances of 160 disyllabic words in the initial, middle and final parts of three carrier sentences, which were recorded on 2 or 3 days, it is found that: (1) Chinese speakers read words in the same sentence position with stable relative F_0CFs, even on different days; and (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.

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

PACS number: 43.70.Fq, 43.72.Ja

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 target words are inserted as required. In car navigation, many street names may be inserted into such a fixed sentence as "Please make a left turn at $***$," where $***$ is the name of a street. Instead of synthesizing these sentences by current text-to-speech systems where the robot-like intonation is little improved, a preferable way of generating more natural-sounding speech is to record the fixed sentence utterance and street name separately and to compose the target sentence by inserting the required street name. One of the most important problems in such a process is the adjustment of 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, the F_0 contour of an utterance is determined by the lexical tone of the syllables in the sentence in the first instance. Chinese has four lexical tones: high, rising, low, and falling (T1, T2, T3 and T4). There is also a neutral tone T0 or T5, which lacks fixed pitch specification. The syllable in T0 is often of shorter spoken duration with lower F_0 after T1, T2 and T4, but higher after T3. The neutral tone T0 is not addressed in this study. Besides the tones, the F_0 contour of an utterance is affected by the intonation, which yields global declination, intonational phrase reset and final lowering etc. Intonation has a global effect, but tone affects the local F_0 contour of words. Therefore, the process of word substitution requires that the F_0 contour of the inserted word be arranged to meet the sentence intonation so that the word sounds natural within the sentence.

In a syllable-based approach to composing the F_0

[*] e-mail: xu@klab.jp
[†] e-mail: hiroki@klab.jp
[‡] e-mail: kasuya@klab.jp

contour of inserted words from separately generated syllabic F_0 contours, numerous factors affect the naturalness of the synthetic sound, including the syllabic position in a word, assimilation between adjacent tones, and interaction between the segmental features and tones. It appears difficult to model all these factors precisely in an efficient manner.

In this study, we employ a word-based method for generating the F_0 contour. The F_0 contours of all words required for the insertion are stored as synthesis units. As it is not feasible to collect the F_0 contours of a 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 method by which Chinese is regulated in terms of four tones in the relative F_0 range of the syllable [1], we have developed a scheme in which the F_0 ranges of words are regulated into tone-dependent relative values in respect of a tone-independent word-level F_0 range. By quantitatively describing the effects of intonation on the F_0 range of a word, we are able to adjust the sample F_0 contour to suit any position in a sentence [2]. In a later study, we verified the validity of the method using the speech of one individual, comparing the regulated F_0 range of identical words uttered on different days and in different positions of carrier sentences [3]. We also demonstrated the validity of this method through perceptual experiments on the naturalness of the re-synthesized words [3]. However, there remains a degree of uncertainty as to whether this approach is applicable to various speakers. Based on materials spoken by four individuals, therefore, we attempt to answer the following questions in this paper: (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?

2. WORD-LEVEL F_0 RANGE AND RELATIVE 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 the 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 word F_0 contour. Figure 1 shows the relationship between WF_0R and F_0CF. The two curves show the F_0 patterns of words in tone combinations (T1, T3) and (T2, T4). The WF_0R is the overall range of the F_0CFs of all the tone combinations and the F_0CF of a word is a sub-range of WF_0R. The relative F_0CF is defined in relation to WF_0R, which is the relative value of the F_0CF of a tone combination in reference to WF_0R. An example of calculated relative F_0CF is shown in Fig. 2, where the inner circle of the radar chart represents 0, the outer circle represents 1, and every axis in the chart represents a tone combination. It was calculated from one speaker's utterances of 10 words for each tone combination, which were inserted into the middle position of a carrier sentence.

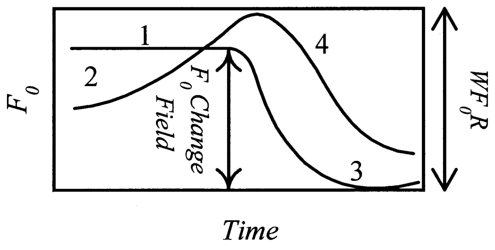


Fig. 1 WF_0R and F_0 change field.

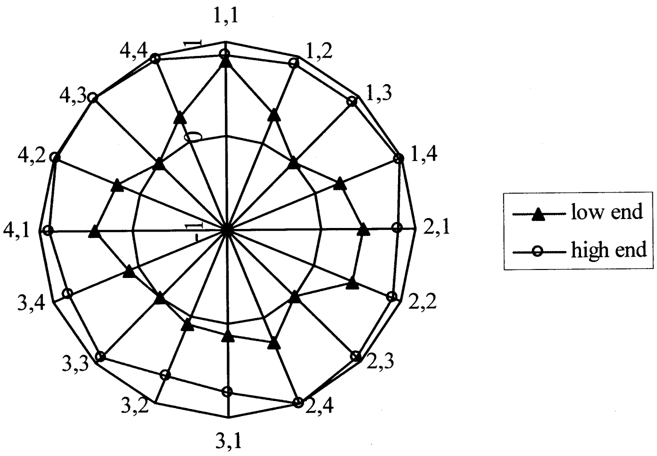


Fig. 2 Relative F_0 change field for the middle position for subject MXZ.

3. EXPERIMENT

To answer the two questions mentioned in the introduction, we compare the relative F_0CFs of four native speakers on two or three different days, and for words inserted in the initial, middle and final positions of a sentence.

3.1. Materials

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 began with a consonant. The test involved the insertion of 160 words into three short carrier sentences, as follows:

Init.) "*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. Subjects

Four Chinese students, two males and two females, who were all born and brought up in Beijing and native speakers of Mandarin Chinese without obvious accent, served as subjects. One of them (MXZ) had lived in Japan for about 6 years. They had not received any special training of speaking.

3.3. Data Analysis

All four subjects read each of the three sentences once, maintaining the same pattern of intonation as much as possible. Words for each sentence were randomly selected. The subjects also read the same sentences on different days in order to investigate the day-to-day variation of F_0 contours. Subjects MJJ, FYS and FYH read the second time the following day. The second reading of subject MXZ was made after about 40 days and the third after about 400 days. With the data of the 3 days of utterances for subject MXZ, we investigate whether the interval between the recording days would affect relative F_0CF. The utterances were sampled at 11.025 kHz. The F_0 was analyzed every 10 ms and a five-point median smoother was applied in order to reduce errors in the transition 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 maximal high end of the average F_0CF for all tone combinations was used as the high edge of the WF_0R, and the minimal low end was used as the low edge. The high and low ends of the averaged F_0CFs for each tone combination were normalized as relative F_0CF of values between 0 and 1. An example of the normalized relative F_0CF for the middle position for subject MXZ is shown above in Fig. 2. We can see that the high and low ends of the tone combinations differ from each other. In a tone combination such as (T4, T3), the span of relative F_0CF is almost 1, whereas in that for (T2, T1), the span is less than 0.5.

3.4. Results

3.4.1 Day-to-day variation 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). As for the subject of MXZ, the two days are the first and second days. 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 WF_0Rs of the other two subjects 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

For the middle position of sentence, all the low and high ends for the 4 subjects on both days are shown in Fig. 3. As a whole, the two sets calculated from the two separate days are quite similar. Although the spans of WF_0Rs for MXZ and FYS differ slightly between the two days (Table 1), 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 utterances in the ''Mid.'' carrier sentence for the two separate days. Table 2 shows the number of ends found to yield statistically significant differences. Of 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 specified carrier sentence with almost the same relative F_0CF.

There arises a question whether the relative F_0CF would change after a long period rather than just a day or a month. For subject MXZ, a further inspection was made for the relative F_0CFs calculated from the speech data of the third day that was more than 1 year from the first and the second days. The number of the ends with significant differences for the third day with the first and the second day is shown in Table 3. These are also small numbers considering all the 16 tone combinations.

3.4.2 Sentence position dependence of relative F_0CF

The WF_0Rs in the three positions of the three carrier sentences read on the second day are shown in Fig. 4 in semitones, where each panel represents a single subject. 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 revealed that 65% of the relative F_0CF ends are dependent on the position of the word in a sentence. However, different results were obtained when the positions were examined two at a time.

A comparison of the relative F_0CFs for the initial and middle positions, giving the number of differing low and high ends, is shown in Table 4. For all the subjects except FYH, only a few ends were found to be significantly different. A comparison of the relative F_0CFs for the

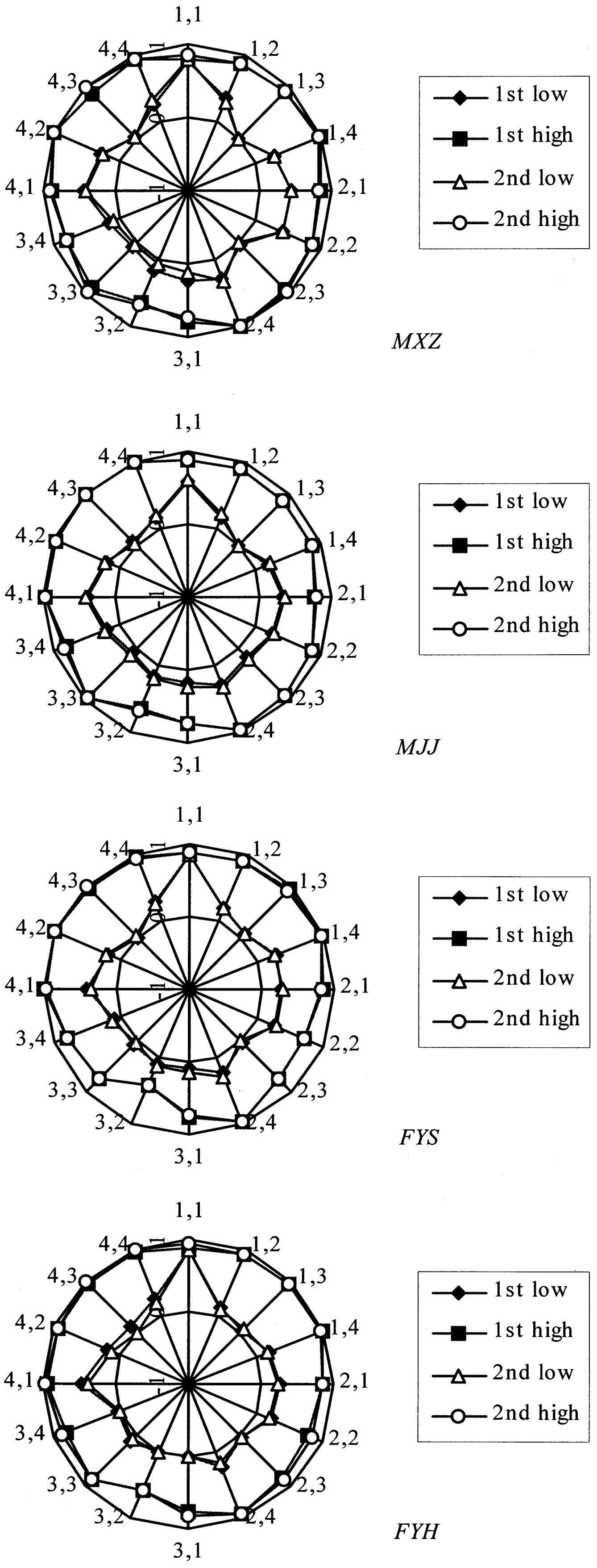


Fig. 3 The relative F_0CFs for the middle position for subjects MXZ, MJJ, FYS, FYH on two separate days.

Table 2 Number (Num.) and percentage (Per.) of relative F_0CF ends found to change significantly between the two days.

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

Table 3 Number (Num.) and percentage (Per.) of the ends that significantly changed on the first and the second days comparing with those of the third day for subject MXZ.

Recording	Low end		High end	
day	Num.	Per. (%)	Num.	Per. (%)
First	3	18.8	1	6.3
Second	2	12.5	2	12.5

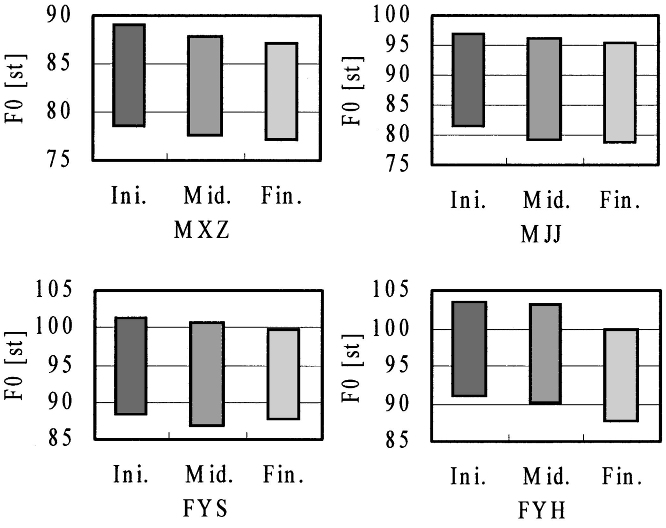


Fig. 4 The WF_0Rs in the initial, middle and final position for the 4 subjects.

Table 4 Number (Num.) and percentage (Per.) of the ends that significantly changed between words in the initial and middle positions.

Subject	Low end		High 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	9	56.3	3	18.8

Table 5 Number of the ends that significantly changed between words in the middle and final positions.

Subject	Low end		High 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

middle and final positions is shown in Table 5. For all the subjects, the ends are significantly different by more than 37%, and among these ends, almost all the values in the final position are smaller than those in the middle position, as shown in Fig. 4. This can be attributed to final lowering phenomena [4], by which a final syllable or word in a sentence is read lower than the former part of the sentence.

4. DISCUSSION

After introducing the concept of WF_0R, F_0CF and relative F_0CF, we see that the tonal effects on the F_0 pattern for all subjects are stable when the F_0 ranges of the words are represented in terms of relative F_0CF. Using relative F_0CF, we can answer the two questions posed in the introduction as follows: (1) Subjects speak the words without any significant variance in relative F_0CF for all the 16 tone combinations on different days. The WF_0R for the same position in the same sentence may vary from day to day, as seen for MXZ and FYS (see Table 1), whereas the relative F_0CF changes little. This means that when words are recorded on different days, although the F_0CF of the words in the same tone combination may vary due to the change of WF_0R, the relative F_0CF remains constant. A long interval of even one year does not introduce significant difference into relative F_0CF either, as we see from subject MXZ. (2) The relative F_0CF is independent of F_0 declination in the intonational phrase in the initial and middle positions of the sentences. Although the WF_0Rs in the initial, middle and final positions decline (see Fig. 4), the relative F_0CFs of the words in the initial and middle positions are stable for all subjects in this study except FYH, for whom the low ends of the words in the initial position tended to be higher than those in the middle position. Of the 9 low ends that differed between the initial and middle positions (see Table 4), 8 are higher in the initial position than in the middle. However, the perceptual impression of sentences with words in these tone combinations is the same as the others, suggesting that larger changes in F_0 than the statistical difference standard may be perceptually acceptable. An investigation of the perceptual acceptability will be conducted in the near future.

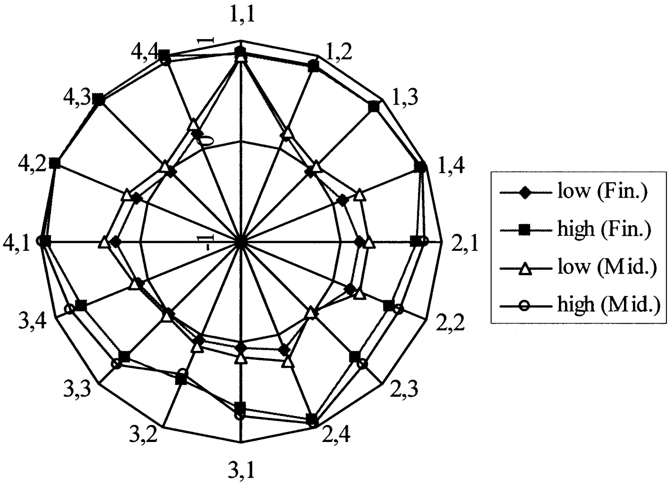


Fig. 5 Relative F_0 change field of the words in the middle and final position for FYS.

In contrast, many of the relative F_0CFs of words differed between the middle and final positions. In the final part of a sentence, extra final lowering [4–6] affects the F_0 contour in addition to the intonational declination of F_0. In an intonational language such as English, final lowering is used to indicate that the F_0 peak of the final accent is lower than the target predicted according to global F_0 declination [5]. In Mandarin Chinese, final lowering occurs at the final syllable [6]. Our data shows that the F_0 contours of the first syllable in disyllabic words of certain tone combinations are also lower than the other words. Figure 5 illustrates an example of the relative F_0CFs of words in the final and middle position for subject FYS. In tone combinations (T2, T2) and (T2, T3), the high ends of relative F_0CFs were lower in the final position than in the middle position. This lowering occurred in the first syllable of tone 2. However, the F_0 of first syllables with T1 and T4 did not lower to the same extent as for the tone combinations (T2, T2) and (T2, T3). It appears that the final lowering is not simply a word-level linear addition of phrasal declination.

It was even expected that the relative F_0CFs would be the same among individuals. However, the investigation showed that due to individuality, the values varied noticeably. A one-way ANOVA test ($p < 0.05$) revealed that among the total of 32 high and low ends, 20 ends were found to be significantly different, and most of these were low ends. However, after scrutinizing the relative F_0CFs of the four subjects, it was found that the effects of tones on the relative F_0CFs followed the same pattern among the individuals. For example, all the high ends in tone combinations (T4, T1), (T4, T2), (T4, T3), and (T4, T4) for the four subjects were approximately 1, as illustrated in Fig. 6. The low ends of these tone combinations are illustrated in Fig. 7. For these subjects, the low end in (T4, T1) was the highest, those in (T4, T4) and (T4, T2) were

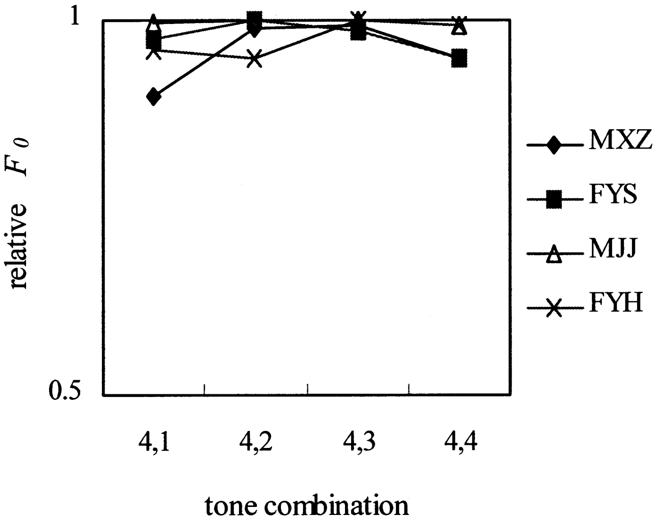


Fig. 6 High ends in tone combinations of (T4, T1), (T4, T2), (T4, T3), (T4, T4).

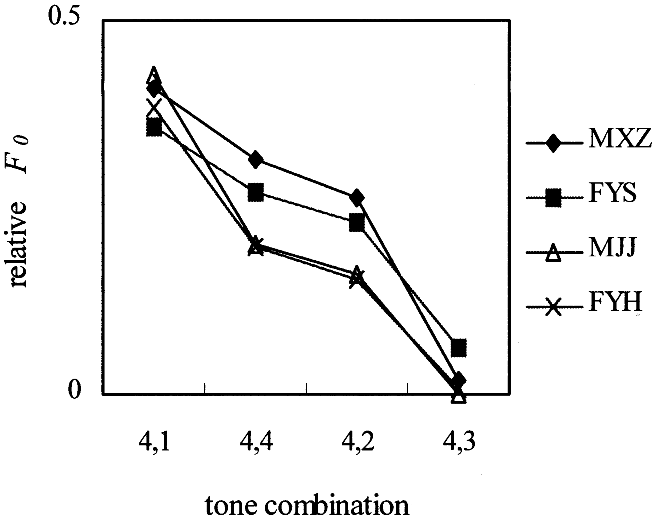


Fig. 7 Low ends in tone combinations of (T4, T1), (T4, T4), (T4, T2), (T4, T3).

lower, and that in (T4, T3) was the lowest at close to 0. This pattern was common to all four subjects. Therefore, although the values of relative F_0CF were found to be statistically different in most cases, tonal effects shared a universal pattern for all subjects.

Xu and Wang [7] proposed a framework for tonal targets in Chinese tones in which tones 1 and 3 contain high and low static targets, whereas tones 2 and 4 contain rising and falling dynamic targets. The present authors approach tonal effects on the F_0 contour in a different manner, considering the F_0 range only; the alignment and speed of F_0 change are not considered. For the applications considered in this research, the data shows that both the transference of the rising or falling of F_0 and F_0 span are important for tones 2 and 4.

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 into fixed sentences. The effects of intonation were reflected in the word-level F_0 range, and tonal effects were indicated in the normalized relative F_0 change field. We demonstrated 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) the relative F_0 change fields for all subjects were stable from day to day, (2) most subjects maintained the same relative F_0 change fields when words were read in the initial and middle positions of a sentence, (3) due to final lowering, the relative F_0 change fields for words in the final position were lower than those for words in the middle, and (4) although the values of F_0 change fields might vary somewhat between individuals, the pattern of the relative F_0 change field variation according to tonal combinations was individual-independent. In conclusion, the independence of the relative F_0 change field from the intonation throughout the subjects suggests that the proposed framework for separating intonational and tonal effects is applicable to unspecific speakers. We are expecting the statement to be strengthened by more data and practical applications.

ACKNOWLEDGEMENTS

We would like to thank Professor Yongbao He (Fudan University, Shanghai, China) for his encouragement and Professor Zongge Li (Fudan University, Shanghai, China) and his colleagues for their support in speech recording made at Fudan University, China. Part of this research was supported by a Japanese government scholarship.

REFERENCES

[1] Y.-R. Chao, *Mandarin Primer* (Harvard University Press, Cambridge, MA, 1948).

[2] D. Xu, H. Mori and H. Kasuya, "The prosody control of Chinese speech considering the F_0 range in word level," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 319–320 (1999).

[3] D. Xu, H. Mori and H. Kasuya, "Word-level F_0 range in Mandarin Chinese and its application to inserting words into a sentence," *Proc. ICSLP 2000*, Vol. 3, pp. 338–441 (2000).

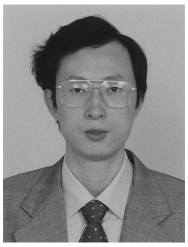
[4] D. Ladd, *Intonational Phonology* (Cambridge University Press, Cambridge, 1996).

[5] M. Liberman and J. B. Pierrehumbert, "Intonational invariance under changes in pitch range and length," in *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, M. Aronoff, R. T. Oehrle, Eds. (MIT Press, Cambridge, MA, 1984), pp. 157–233.

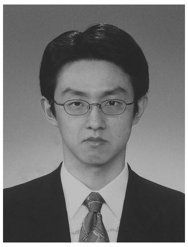
[6] K. L. Pike, *Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion* (University of Michigan Press, Ann Arbor, 1948).

[7] Y. Xu and Q. E. Wang, "Pitch targets and their realization:

Evidence from Mandarin Chinese," *Speech Commun.*, **33**, 319–337 (2001).



Dawei Xu received the B.E. degree in Computer Engineering from Tianjin University, China, in 1993, and the M.S. degree in Computer Science from Fudan University, China, in 1996, respectively. In 1996, he joined the Department of Computer Science as a research associate at Fudan University, China. He is currently a Ph.D. candidate in Production and Information Sciences at Utsunomiya University, Japan. His current research interests include Text-To-Speech system and prosody modeling for speech synthesis. He is a member of the International Speech Communication Association and the Acoustical Society of Japan.



Hiroki Mori received the B.E., M.E. and Ph.D. degrees from Tohoku University, in 1993, 1995 and 1998, respectively. He was with the Graduate School of Engineering, Tohoku University in 1998. He is now a Research Associate of Utsunomiya University. His research interests include speech recognition, speech synthesis, spoken dialogue systems, and natural language processing. He is a member of the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers, the Information Processing Society of Japan, and the Japan Society of Logopedics and Phoniatrics.

Hideki Kasuya received the B.S., M.S., and Ph.D. degrees in Electrical Communication Engineering all from Tohoku University, Sendai, Japan, in 1963, 1965, and 1970, respectively. In 1968, he joined the Research Institute of Electrical Communication as a research associate at Tohoku University, where he was primarily engaged in speech analysis, perception and recognition. From 1974 to 1977 he was a visiting researcher at the Speech Communications Research Laboratory, Inc., California, U.S.A., working on speech recognition. Since 1978 he has been with the Faculty of Engineering, Utsunomiya University, where he is now a professor in the Department of Electrical and Electronic Engineering. His research interests include various areas of speech science and technology, digital signal processing, and image processing.