Acoustic variations of focused disyllabic words in Mandarin Chinese: Analysis, synthesis and perception

Zhenglai Gu, Hiroki Mori, Hideki Kasuya

Graduate school of Engineering
Utsunomiya University, Japan
{gzl, hiroki, Kasuya}@klab.jp

Abstract

The focus effects on acoustic correlates include both prosodic and segmental modifications. Analysis of 35 focused words in a carrier sentences uttered by 2 male and 3 female speakers has shown that: (1) there is a significant asymmetry of vowel duration as well as F0 range between the pre-stressed and post-stressed syllables, implying that different strategies are employed in the task of focusing disyllabic words, i.e., emphasizing the first syllable as well as weakening the second syllable for the former, but emphasizing the second syllable only for the latter; (2) the tonal combinations significantly affect the variations of both the vowel duration and F0 range; (3) the formant frequencies (F1, F2) are changed systematically in a way that that the formants of the vowels plotted in the (F1, F2) plane were stretched outwards. Perceptual validation of the relative importance of these acoustic cues for signaling a focal word has been accomplished. Results of the perception experiment indicate that F0 is the dominant cue closely related to the judgment of focused word and the other two cues, duration and formant frequencies contribute less to the judgment.

1. Introduction

It is widely known that the focally stressed syllable is accompanied by the changes of fundamental frequency (F0), duration, intensity, formant and spectral tilt [1-7]. F0 movements, moreover, are generally seen as the reliable prosodic cues to focus. As a tonal language, the F0 movements in Mandarin Chinese are generally described on the basis of a top- and bottom-line intonation model [8-9].

There have been several reports on the nature of double–stressed focal words and their surroundings in very short sentences, including: (1) a radical asymmetry of fundamental frequency (F0) range occurs around the focused word [6-7], i.e., the F0 range of on-focus words is significantly increased and the F0 range of post-focus words is decreased, while the F0 range of the pre-focus words is little changed; and (2) a remarkable increase in duration of on-focus words is observed, which is associated with small changes in the duration of both the pre- and post-focus syllables [7], a phenomenon we call symmetry of duration. Despite these accumulated findings, however, little attention has been paid to the nature of single-stressed focal words, which are fundamental components of fluent Chinese speech. The stress patterns of disyllabic Chinese words are defined conventionally as follows [10]: stress placement on the first syllable (Sp10), the second syllable (Sp01) and on both syllables (Sp11). Hereafter, Sp10 and Sp01 are referred to as "single stress" and Sp11 as "double stress" patterns. Meanwhile studies on perception experiments with modifications on the spectral characters of the focally stressed syllable are rare.

The present study attempts to answer the following questions: (1) How the three aforementioned stress patterns and the tone combinations affect the F0 range and duration. (2) How focally stressed syllables change in terms of the formant frequencies of vowels. (3) Whether and to what degree the F0, duration and formant frequencies contribute to the perception of the focused words.

We deal with disyllabic words in the paper. Concerning *disyllabic words* it is stated by Yao and Pan [11], that they are the most frequently used, since his report indicates that there are 50% of disyllabic words in lexical and 70% of disyllabic prosodic words in speech. Disyllabic words are the fundamental constituents of other polysyllabic words of three syllables or more.

2. Speech analysis experiments

2.1 Materials and analytical procedure

To control the stress patterns of the focal word, we used the carrier sentence, "Zhe4 shi4 TW er2 bu2 shi4 CW. (This is TW but not CW.)", into which target and contrastive words were inserted.

For example, given the word "gang1cai2" ("steel material"), we have the following four derivatives.

"Zhe4 shi4 gang1cai2 er2 bu2 shi4 mu4cai2." ("This is steel material but not wood.")
"Zhe4 shi4 gang1cai2 er2 bu2 shi4 gang1chan3ping3." ("This is steel material but not a steel product.")
"Zhe4 shi4 gang1cai2 er2 bu2 shi4 mu4chan3ping3." ("This is steel material but not a wood product.")
"Zhe4 shi4 gang1cai2 er2 bu2 shi4 qi2ta1." ("This is steel material but not anything else.")

In the above sentences, the stressed syllables are underlined. The target words reflected all 15 tonal combinations (Chinese 4 tones: H, R, L, F), excluding tone L-L (tone L will be changed into tone R when it precedes another tone L). Two words were selected for each tone combination. Five additional monosyllabic target words with the same tone H and the same consonant /d/ were selected for the analysis of formant frequencies. Five native Chinese speakers born in Beijing, two males and three females, read one sentence twice, with an additional 4 times reading of the sentences including the last five monosyllabic target words. Thus, 1350 (= 30 words × 3 stress patterns and 1 neutral × 2 repetitions × 5 speakers + 5 words × 6 repetitions × 5 speakers) utterances were recorded on DAT in a soundproof room before being digitized at a sampling rate of 11.025 kHz. The syllable and vowel segmentation was carried out manually by observing the waveform and sound spectrogram of an utterance. F0 contours in semitone (ref. 1 Hz) were automatically analyzed at 5-ms intervals and erroneous F0 values, if any, were corrected manually. The largest and smallest F0 values of the F0 contour

of the vowel segment of a syllable were extracted as the F0 maximum and F0 minimum values of the syllable, respectively. The F0 range was defined as the difference between the two extreme values. Formant frequencies (F1, F2) of all the simple vowels in the five monosyllabic target words were extracted by the ARX analysis [12], under the conditions of ARX equation orders p=12 and q=0, frame length of 35 ms, and frame shift of 5 ms. Formant frequencies of a target vowel were defined as the mean estimated values of 10 frames at the middle part of the vowel.

2.2 Results of prosodic variations

2.2.1 Effects of stress patterns

The values of the prosodic correlates of each tone combination were averaged across the two target words of the same tone and across both repetitions of the target words in the same sentence uttered with the same stress pattern. Thus, each stress pattern has a total of 75 (= 15 tone combinations × 5 speakers) mean values for the representation of the variation of vowel duration and F0 range.

Vowel Duration

Table 1 shows the mean values $\overline{\alpha}_1$ and $\overline{\alpha}_2$ of the vowel duration ratios in syllables 1 and 2 (α_1, α_2), respectively. As can be seen in Table 5, $\overline{\alpha}_2$ is decreased significantly to an average of 85% for Sp10 and increased remarkably to an average of 149% for Sp01. This is in contrast to the increase in $\overline{\alpha}_1$ to an average of 115% and 143% for Sp01 and Sp10, respectively. In addition, the difference between the mean vowel duration ratio of pre-stressed syllable ($\overline{\alpha}_1$=115%) and mean vowel duration ratio of post-stressed syllable ($\overline{\alpha}_2$=85%) is significant ($p<0.05$), indicating that single stress on the syllable of disyllabic words not only lengthens the stressed syllable but also systematically affects the length of the unstressed syllable.

Table 1: Mean values $\overline{\alpha}_1$ and $\overline{\alpha}_2$ of vowel duration ratios in syllables 1 and 2 (α_1, α_2), respectively.

Stress pattern	$\overline{\alpha}_1$	$\overline{\alpha}_2$
Sp11	1.39	1.32
Sp10	1.43	0.85
Sp01	1.15	1.49

F0 Range

Table 2 shows the mean values $\overline{\beta}_1$ and $\overline{\beta}_2$ of F0 range differences in syllables 1 and 2 (β_1, β_2), respectively. As can be seen, the mean F0 range differences increase more for the single stress patterns ($\overline{\beta}_1$=2.65, $\overline{\beta}_2$=3.31) than those for the double stress pattern ($\overline{\beta}_1$=1.86, $\overline{\beta}_2$=2.18), with significant differences at the level of $p<0.05$. Another substantial difference is that the mean F0 range difference of the pre-stressed syllable 1 ($\overline{\beta}_1$=0.51) is larger than that of the post-stressed syllable 2 ($\overline{\beta}_2$=-0.45) ($p<0.05$).

The results illustrate that the stressed syllables were consistently produced with significantly increased vowel duration and F0 range, but with regard to the pre- and post-stressed syllable in the words, there was an overall asymmetric variation in vowel duration and F0 range, i.e., a slight increase in pre-stressed syllable and a considerable decrease in post-stressed syllable.

Table 2: Mean values $\overline{\beta}_1$ and $\overline{\beta}_2$ of F0 range differences in syllables 1 and 2 (β_1, β_2), respectively.

Stress pattern	$\overline{\beta}_1$ [semitone]	$\overline{\beta}_2$ [semitone]
Sp11	1.86	2.18
Sp10	2.65	-0.45
Sp01	0.51	3.31

2.2.2 Effects of tone combinations

We performed an extended CHAID (Chi-square Automatic Interaction Detection) analysis [13], with the standardized values of α_1, α_2, β_1 and β_2 as dependent variables and the identities of the tones and speakers as independent variables (Table 3). The standardization was made on the basis of their means and standard deviations, where the standardized values were denoted as $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The CHAID algorithm recursively splits the dependent variable space into two or more subspaces. An exhaustive search is made to find the best split resulting in the most statistically significant difference among all possible splits with regard to the independent variables. The splitting process is repeated until insignificant difference is found after splitting. The one-way ANOVA is used to indicate the degree of the difference. Thus the larger the F-value, the greater the difference becomes. We also calculated the reduction ratio of variance (RR) after the split. Tables 4 and 5 show the results for vowel duration and F0 range, respectively. As can be seen, most of the CHAID-based splitting process underwent at least one split, except for the process for the dependent variable $\hat{\alpha}_2$ under stress pattern Sp10. Moreover, the identities of the tones were mostly selected as the major independent variable for the dependent variables $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$ and $\hat{\beta}_2$, indicating that the tone combinations significantly affect the variations of both durations and F0 range.

Vowel Duration

The tonal categories of the second syllables (T_F) under stress pattern Sp01 strongly affect not only the value of $\hat{\alpha}_2$ but also that of $\hat{\alpha}_1$. As can be seen in Table 4, the standardized vowel duration ratio of the stressed tone L ($\hat{\alpha}_2$=2.17) was much larger than that of the other tones ($\hat{\alpha}_2$=0.01), with a significant difference (F=97, RR=57%), which even anticipatively affects the standardized vowel duration ratios of the pre-stressed syllables (F=11, RR=15%). The results for the stress pattern Sp10 less clearly show the tonal effects with a difference (F=8, RR=10%).

F0 Range

As can be seen in Table 5, the tonal categories strongly affect the variations of the F0 range of both stressed syllables 1 and 2 with significant difference (F=45, RR=55% for syllable 1; F=62, RR=46% for syllable 2). In general, the F0 ranges of the static tones H and L were increased less than those of the dynamic tones R and F, except that the F0 range of the word-final tone L was increased to the same extent as that of the dynamic tones. Moreover, the high-offset of the stressed tones H and R increased the onset F0 of the post-stressed syllables, while the low-offset of the stressed tones L and F decreased the onset F0 of the post-stressed syllables. As a result, there is a significant difference between the effects of

the two tonal categories (F=15, RR=18%).

The results illustrate that: (1) the word-final tone L of a stressed syllable shows a unique lengthening effect on the variation of vowel duration, which may be due, at least in part, to the phenomenon that a rising F0 contour is appended to the word-final tone L, while the word-initial tone L keeps falling only; (2) the F0 range of the post-stressed syllable is strongly affected by the high-or-low offset of the tone of the preceding stressed syllable. Contrary to the expectation that the F0 range of the post-stressed syllable would be decreased according to the asymmetric property of the F0 range, it is increased when preceded by the high-offset tones. Xu [14] examined F0 contours of bi-tonal sequences and found a carry-over tonal effect on F0 contours: the starting F0 of a tone is assimilated to the offset of a previous tone. It is considered that the carry-over effect resulted in the increase of the F0 range of the post-stressed syllable.

Table 3: Independent and dependent variables used in the CHAID analysis

Independent variables (I_V)	
T_I	Tone identities of syllable 1 (H, R, L, F)
T_F	Tone identities of syllable 2 (H, R, L, F)
Spk	Subjects who uttered the target word (F-1, F-2, M-1, M-2, M-3)
Dependent variables (D_V)	
$\hat{\alpha}_1$	Standardized value of α_1
$\hat{\alpha}_2$	Standardized value of α_2
$\hat{\beta}_1$	Standardized value of β_1
$\hat{\beta}_2$	Standardized value of β_2

Table 4: Major results of CHAID analysis for vowel duration. * indicates that the subspace represented by the independent variables was further split into a subspace.

Stress pattern Sp01

Depth of split	Independent variables		Dependent variables $\hat{\alpha}_1$	F	RR[%]
1st split	Spk:	F-1, M-1, M-2, M-3* / F-2	0.07 / -0.56	28	28
2nd split	T_F:	H, R, F / L	-0.13 / 0.89	11	15
			$\hat{\alpha}_2$	F	RR[%]
1st split	T_F:	H, R, F / L	0.01 / 2.17	97	57

Stress pattern Sp10

Depth of split	Independent variables		Dependent variables $\hat{\alpha}_1$	F	RR[%]
1st split	T_I:	H, R / L, F	0.67 / 1.07	8	10

Table 5: Major results of CHAID analysis for F0 range. * indicates that the subspace represented by the independent variables was further split into a subspace.

Stress pattern Sp01

Depth of split	Independent variables		Dependent variables $\hat{\beta}_1$	F	RR[%]
1st split	T_I:	H, F / R, L	-0.69 / 0.42	8	10
			$\hat{\beta}_2$	F	RR[%]
1st split	T_F:	H / R, L, F	-0.61 / 0.99	62	46

Stress pattern Sp10

Depth of split	Independent variables		Dependent variables $\hat{\beta}_1$	F	RR[%]
1st split	T_I:	H / L / R, F	-0.36 / 0.68 / 1.10	45	55
			Dependent variables $\hat{\beta}_2$	F	RR[%]
1st split	T_I:	H, R / L, F*	-0.37 / -0.92	15	18
2nd split	T_F:	H / R, L, F	-0.45 / -1.11	10	10

2.3 Results of formant

Figure 1 shows the formant frequencies of vowels in (F1, F2) plane. As can be seen, the F1 of /a/ /e/ and /o/ was considerably increased but there was insignificant change in the F2 of the vowels, respectively. Both F1 and F2 of /i/ showed significant changes. F1 was decreased and F2 was increased systematically. On the whole, there became a tendency that the formants of the vowels plotted in the (F1, F2) plane were stretched outwards.

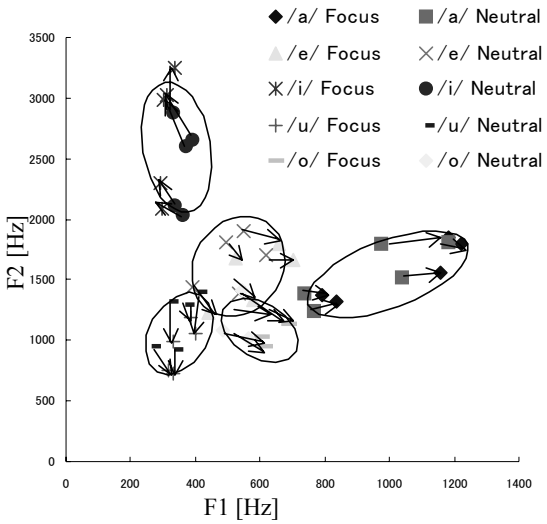

Figure 1: Distribution of the formant frequencies(F1, F2) of the vowels in focused and neutral words from 5 speakers.

3. Generation model and perceptual validation

3.1 F0, duration and formants control

- The F0 range modification involves an additive coefficient in addition to the multiplicative coefficient, which is modeled as following:

$$F0'_{jkt}(n) = \kappa_{jkt}(F0(n) - \underline{F0}_j) + \theta_{jkt} + \underline{F0}_j \quad (1)$$

Where $\underline{F0}_j$ is the minimum of F0 range, κ_{jkt} is a multiplicative coefficient for expanding the F0 range from the bottom up. θ_{jkt} is an additive coefficient for shifting up or down the bottom, j, k and t represent focus positions, stress positions, and four Chinese tones, respectively.

- The duration modification model is defined as:

$$Dur'_{jkt}(n) = \gamma_{jkt} Dur_{jkt}(n) \quad (2)$$

Where γ_{jkt} is a multiplicative coefficient for strengthening the duration.

- The formant frequencies (F1, F2) modification model is defined as:

$$F'_i(n) = F_i(n) \times \lambda_{iv}^{\sin(\pi \times R(n))} \quad (3)$$

Where $R(n)$ $(0 < R(n) < 1)$ represents the relative position of the formant inside a vowel. In order to avoid the production of click sound due to the discontinuity of the formant frequencies series, a sine window smooths the multiplicative coefficient.

All the conversion parameters were estimated statistically from the data recorded in the first experiment.

3.2 Perceptual validation experiment

3.2.1 Materials and procedure

We performed conversions from 10 neutral utterances to perception stimuli in which arbitrarily selected words are focused, using 4 combinations of the control rules of F0, duration and formant frequencies. The combinations are:

All: F0, duration and formant frequencies are modified

F0_D: F0 and duration are modified

F0_F: F0 and formant frequencies are modified

D_F: Duration and formant frequencies are modified

Thus 40 stimuli (10 utterances × 4 combinations of control rules) were produced. We instructed the subjects to judge perceptually which word was focused in a stimulus. The stimuli were randomized and played 10 times to 10 subjects who participated in the perception test. All of them are native speakers of Mandarin Chinese.

3.2.2 Results

The results are summarized in Figure 2, which gives the average identification ratios of focused word as a function of the control rules combinations across the 10 subjects. As can be seen in the Figure, the highest identification ratio is 94% when using all the control rules and the ratio remarkably reduced to 38% when the F0 is not controlled, which demonstrate that the F0 is the dominant cue closely related to the judgment of focused word and the other two cues, duration and formant frequencies do not contribute much to the judgment.

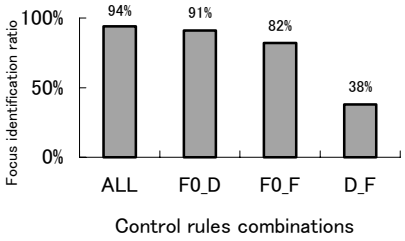

Figure 2: Results of perceptual experiments.

4. Summary

The results of our two experiments reveal that: (1) there is a significant asymmetry of vowel duration as well as F0 range between the pre-stressed and post-stressed syllables, implying that different strategies are employed in the task of focusing disyllabic words, i.e., emphasizing the first syllable as well as weakening the second syllable for the former, but emphasizing the second syllable only for the latter; (2) the tonal combinations significantly affect the variations of both the vowel duration and F0 range; (3) the formant frequencies (F1, F2) are changed systematically in a way that that the formants of the vowels plotted in the (F1, F2) plane were stretched outwards; (4) F0 is the dominant cue closely related to the judgment of focused word, and the other two cues, duration and formant frequencies, contribute less to the judgment.

References

[1] Ladd, D.R., *Intonational phonology* (CUP, Cambridge, 1996), p.153, p.180.

[2] Campbell, N. and Bechman M.E., "Stress, prominence, and spectral tilt", ESCA Tutorial/Research Workshop on Intonation: Theory, Models, Applications, Athens, Greece, 67-70, 1997.

[3] Fant G., Kruckenberg A. and Liljencrants J., "Acoustic-phonetic analysis of prominence in Swedish", *Intonation: Analysis, Modelling and Technology* (Kluwer Academic Publisher, 2000), 55-86.

[4] Maekawa K., "Effects of focus on duration and vowel formant frequency in Japanese", *Computing Prosody: Computational Models for Processing Spontaneous Speech* (Springer Verlag, 1996), 129-153.

[5] Heldner, M., and Strangert, E., "Temporal effects of focus in Swedish", *Journal of Phonetics* 29(3), 329-361.

[6] Gårding, E., "Speech act and tonal pattern in standard Chinese constancy and variation", *Phonetica* 44, 13–29, 1987.

[7] Xu, Y., "Effects of tone and focus on the formation and alignment of f_0 contours", *Journal of Phonetics* 27, 55–105, 1999.

[8] Chao, Y-R., *A grammar of spoken Chinese* (CUP, Cambridge, 1968), p.35.

[9] Zhang, J.L., "Acoustic parameters and phonological rules of a Text-to-Speech system for Chinese", *IEEE ICASSP*, 2023–2026, 1986.

[10] Iwata, R., "Tone and accent across the Chinese Dialects", *Journal of the Phonetic Society of Japan.* 5–1, 18–27, 2001.

[11] Yao, Q. and Pan, W., "Prosodic word: the lowest constituent in the Mandarin prosody processing", *Speech Prosody 2002*.

[12] Otsuka, T. and Kausya, H., "Robust ARX-based speech analysis method taking voicing source pulse train into account", *Journal of the Acoustical Society of Japan* 58:386—397, 2002.

[13] Biggs, D., de Ville, B. and Suen E., "A method of choosing multiway partitions for classification and decision trees", *Journal of Applied Statistic*s, 18, 49–62, 1991.

[14] Xu, Y., "Contextual tonal variations in Mandarin", *Journal of Phonetics* 25, 61–83, 1997.