Pitch and Duration as Cues in Perception of Neutral Tone under Different Contexts in Standard Chinese

Aijun Li [*], Jun Gao [*], Yuan Jia [*] and Yaru Wang [†]
[*] Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China.
E-mail: {liaj,gaojun,jiayuan}@cass.org.cn Tel: +86-1-65237408
[†] School of Computer Sciences, Tianjin University, Tianjin, China.

***Abstract*— This paper conducted a psychoacoustic experiment to investigate the influence of pitch and duration on neutral tone perception and the distribution pattern of the perceptual spaces in three contexts, namely, isolation, on-focus and post-focus. A minimal pair differed in stress pattern was employed in this experiment, i.e., neutral tone word "蘑菇" /mo2ku0/ (mushroom, strong-weak) and its counterpart with normal stress "魔箍" /mo2ku1/ (magic ring, strong-strong). The stimulus continuum was achieved by systematical changes of the F0 and duration on the second syllable of /moku/ in three contexts. ANOVA and multinomial ordinal logistic analysis were conducted on the perceptual results and showed that the distribution of perceptual spaces varies in three contexts. For neutral tone perception, the pitch is always a more reliable cue than the duration. However, the amplitude of the influence of pitch and duration is closely related to the context. In isolation, the pitch is a stronger cue to neutral tone perception than the duration. In on-focus condition, the pitch is still a stronger cue than the duration, but less than that in isolation. Under post-focus condition, even though the pitch plays a slightly more significant role, the duration is quite important as well. The results showed that in a tone language, i.e., Standard Chinese, the way the two acoustic cues influence the perception of neutral tone is not an exact match to that in Indo-Euro languages such as Dutch and English.**

***Index Terms*—neutral tone perception, Standard Chinese, pitch, duration**

I. INTRODUCTION

In Standard Chinese, besides the four lexically distinctive full tones, there exist weak elements in terms of neutral tones [1, 2]. Neutral tone can be considered to be related to both tone and stress system in Standard Chinese [4]. Neutral tone doesn't occur in the initial position of a word and is assumed to be associated with weak syllable that is short and light, such as the second syllable in /ti^{4}ti^0/ (little brother, hereafter "1~4" stands for four lexical tones, tone1 to tone4 respectively, "0" for neutral tone). The neutral tone is a mid-pitch target and its F_0 realization is associated with the tone of the preceding syllable [3]. Morphologically, some neutral tone words have contrastive meanings with their normal stress counterparts. For example, with normal stress (strong-strong), "地道" /ti^{4}tao^4/ means 'tunnel', while with neutral tone (strong-weak) "地道" /ti^{4}tao^0/ means 'purely'. The research on the acoustic correlates of neutral tone and its perceptual space will help understand the characteristics of the spoken Chinese.

Many previous acoustic studies confirmed that the correlative acoustic cues for perceiving stress include pitch, duration, intensity, spectral balance or spectral tilt (timber). In stress languages, such as English and Dutch, pitch is the most salient acoustic cue [5] for pitch accents at utterance level [6, 7, 8]. For lexical stress, duration is the most related cue for Dutch listeners. Spectral balance or spectral tilt is an important cue, but not as reliable as duration. Overall, intensity is the least important cue for stress perception of Dutch and English. In English, vowel reduction is a pervasive phenomenon in unstressed syllables. However, vowel reduction is less pervasive and the poorest acoustic cue in Dutch [7, 9].

In Standard Chinese, the acoustic correlates of word stress or sentence stress are identical as in stress languages. However, in regards with pitch and duration, it's still a disputing issue which cue is mostly related to the weak syllable in neutral tone. Some researches indicated that pitch outranks duration [10, 11], whereas others convinced the opposite result [12]. The unstressed syllable of neutral tone shrinks to 50% [13, 14, 15, 16, 17] or 60% [3] of its stressed value. On average, the unstressed syllable's duration is 60% of the preceding stressed syllable [10]. Regarding intensity, the unstressed syllable is not necessarily lighter than the stressed one. Therefore, intensity is not a reliable cue for neutral tone perception [10, 12, 14, 18]. Spectral tilt may have great influence on neutral tone perception, but it is less important than duration [19].

In the previous studies [11, 12, 13], neutral tone was only investigated in isolated words. And different conclusions were drawn due to different methodologies used in perception experiments. The present study, therefore, conducted a psychoacoustic experiment with the continuum of pitch and duration for the minimal pair /mo^{2}ku^0/~/mo^{2}ku^1/ differed in stress in three conditions, i.e., isolation, on-focus and post-focus, with the aim to explore the perception spaces and the contribution of pitch and duration to neutral tone perception.

II. IDENTIFICATION EXPERIMENT

In previous perceptual experiments [11, 12, 13,], the underlying tone of the neutral tone is not considered in the target word pairs. For example, the pair they adopted was "老师 (/lao^{3}ʂl^1/, teacher)" ~ "老实 (/lao^{3}ʂl^0//, honest)", where the neutral tone syllable "实/ʂl^0/" has Tone 2 as its underlying full tone /ʂl^2/ rather than Tone 1 /ʂl^1/. Thus, the

two words are not real minimal pair. In our perception experiment, however, we used a minimally contrasting stress pair “蘑菇”/mo^{2}ku^0/ (mushroom) and “魔箍”/mo^{2}ku^1/ (magic ring). And we manipulated the stimuli based on the actual production spaces, whose pitch and duration are systematically transformed from normal stress to neutral tone.

A. Stimuli

A male Mandarin speaker, 23 years old, was invited to record the experiment materials. Target words were produced in three conditions:

(1) Isolation: “魔箍”/mo^{2}ku^1/ (magic ring) and “蘑菇”/mo^{2}ku^0/ (mushroom);

(2) On-focus: target words were embedded in a carrier sentence in the focus position:

小赵昨天学了哪个词？小赵昨天学了“**魔箍~蘑菇**”这个词。

(Literally, Xiaozhao yesterday learned which word? Xiaozhao yesterday learned **/mo^{2}ku^1/ ~ /mo^{2}ku^0/** this word.

(Which word did Xiaozhao learn yesterday? Xiaozhao learned the word **/mo^{2}ku^1/ ~ /mo^{2}ku^0/** yesterday.)

(3) Post-focus: target words were embedded in a carrier sentence in the post-focus position:

小赵什么时候学了“魔箍~蘑菇”这个词？小赵**昨天**学了“魔箍~蘑菇”这个词。

(Literally, Xiaozhao when learned /mo^{2}ku^1/ ~ /mo^{2}ku^0/ this word? Xiaozhao **yesterday** learned /mo^{2}ku^1/ ~ /mo^{2}ku^0 /this word.)

(When did Xiaozhao learn the word /mo^{2}ku^1/ ~ /mo^{2}ku^0/? It is **yesterday** that Xiaozhao learned the word /mo^{2}ku^1/ ~ /mo^{2}ku^0/.)

Words and utterances in each condition were recorded five times. The words and utterances that obtained the highest perception score (by the authors) were selected as the original baselines to manipulate the perception stimuli. In the identification experiment, only target words (target words in isolation and target words extracted from utterances in on-focus and post-focus condition) were selected as stimuli. PRAAT was applied to extract F_0 and duration of the target words. The stimuli were achieved through systematical manipulation on the F_0 and duration. (Duration and F_0 data are not presented here.)

Taken the manipulations of the stimuli in isolation condition as an example, firstly, changes along the pitch dimension were made. F_0 values of /ku^1/ and /ku^0/ were extracted and stylized by hand in PRAAT. Then the scale of F_0 was transformed into semitone (St, with 75Hz as reference frequency). Three F_0 curves were equally interpolated between F_0 curve of /ku^1/ and /ku^0/, and one extra curve was added above the original F_0 curve of /ku^1/ and one below /ku^0/ with the same step (see Fig. 1). Therefore, the manipulation of F_0 curve has 7 steps in total, among which step 2 and step 6 were set to the original F_0 of /ku^1/ and /ku^0/, respectively.

The change along the duration dimension was then made. The durations of /ku^1/ and /ku^0/ were measured. Due to the fact that the unstressed syllable in neutral tone shrinks to 50% of its corresponding stressed value, ten step changes were set (see Figure 2 below) with step 3 (1) and step 8 (0.5) being set to the original durations of /ku^1/” and /ku^0/”, respectively.

With PSOLA re-synthesis tool in PRAAT, the second syllable of /mo^{2}ku^1/ was manipulated with step changes of F_0 and duration given in Figure 1 and 2. In sum, 70 stimuli (7 pitch *10 duration steps) were obtained for isolation condition.

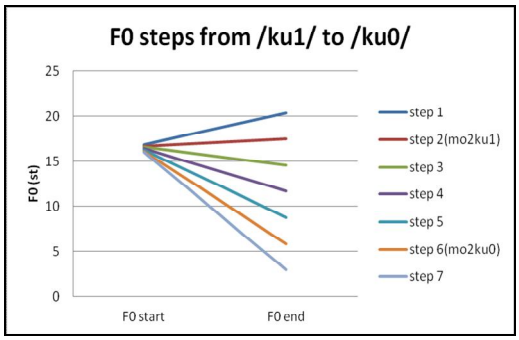

Fig. 1 Seven-step continuum of F_0 between /ku^1/ and /ku^0/ in isolation condition

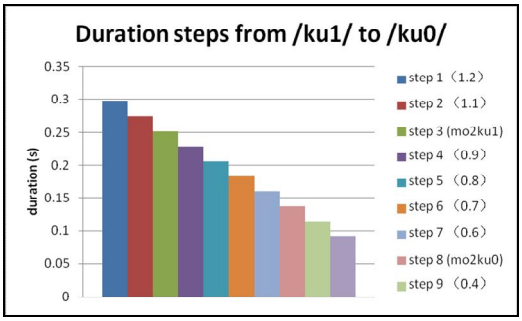

Fig. 2 Ten-step continuum of duration between /ku^1/ and /ku^0/ in isolation condition

The same manipulations in pitch and duration were applied to target words in focused and post-focused condition. We obtained 63 utterance stimuli (7 pitch*9 duration steps; pitch step 2 & 6 and duration step 3 & 7 were the original parameters of /ku^1/ and /ku^0/) in focused condition and 70 utterance stimuli (7 pitch*10 duration steps; pitch step 2 & 6 and duration step 3 & 8 were the original parameters of /ku^1/ and /ku^0/) in unfocused condition. In total, the number of stimuli containing /moku/ was 203. Besides, 82 fillers (including 27 words and 55 utterances containing normal stress and neutral tone words) were added. Therefore, the total number of stimuli was 285.

B. Subjects and identification experiment

Eight male and eight female Standard Chinese-speaking and Beijing-born college students without hearing impairment were enrolled in the experiment, with an average age of 20.44 (Sd=0.46),

The identification experiment (ABX) conducted with E-Prime had two parts, training and test. All the stimuli were randomly presented to the subjects. For each trial, before the audio stimulus was displayed, there was a 2s fixation cross to help the subject concentrate on the experiment. Then the

stimulus was played, afterwards, target words A and B were displayed on the screen. The task was to ask the subjects to judge whether the token(X) they heard was the word “魔箍” /mo^{2}ku^{1}/ (A) or the word “蘑菇” /mo^{2}ku^{0}/ (B). The words A and B were displayed on the screen. The subjects made their judgment by pressing the ‘F’ and ‘J’ keys, specifically, ‘F’ for neutral word “蘑菇” /mo^{2}ku^{0}/,‘J’ for normal stress word “魔箍” /mo^{2}ku^{1}/, and ‘K’ for uncertainty. The experiment lasted about 40 minutes and two short breaks were there during the experiment.

III. ANOVA ANALYSIS

The identification score was set to 3 points for normal stress, 1 point for neutral tone and 2 points for uncertainty. ANOVA was employed to analyze the perception results in three conditions, with the pitch steps and duration steps as independent variable, the identification scores as dependent variable, and subjects as co-variable.

ANOVA results were shown in Table I which demonstrated that: (i) Both pitch and duration significantly affect the perceptual results in all three conditions (all p<0.05). (ii) F value reflects the amplitude of the influence of pitch and duration changes. The larger the F value is, the greater the impact is. Therefore, throughout the three conditions, the impact of pitch outranks duration in a descending order from isolation (F_{pitch} - $F_{duration}$ = 142.6), to on-focus position (F_{pitch} - $F_{duration}$ = 61.2) and to post-focus position (F_{pitch} - $F_{duration}$ = 10.2), i.e. the impact of duration influence is less than that of pitch in all three conditions. (iii) The impact of duration is much constant across the three conditions. In post-focus condition, the impact of duration is closed to that of pitch (F_{pitch} - $F_{duration}$ = 10.2). (iv) The interactive effect between pitch and duration is significant in isolation and focused contexts.

TABLE I
ANOVA ANALYSIS ON PERCEPTUAL RESULTS

Contexts	variances	df	F	Sig.
isolation	Pitch_step	6	153.958	.000
	Duration_step	9	11.318	.000
	Pitchstep * Durationstep	54	1.830	.000
focus	Pitch_step	6	77.533	.000
	Duration_step	8	16.381	.000
	Pitchstep * Durationstep	48	1.996	.000
post-focus	Pitch_step	6	21.702	.000
	Duration_step	9	11.513	.000
	Pitchstep * Durationstep	54	.926	.627

Fig. 3 shows the average results on perception. Within the figure, the first column displays the average perceptual scores as a function of pitch and duration steps. The second and the third column are the average perceptual scores as a function of pitch and duration steps separately. It can be observed from the figures in the first column:

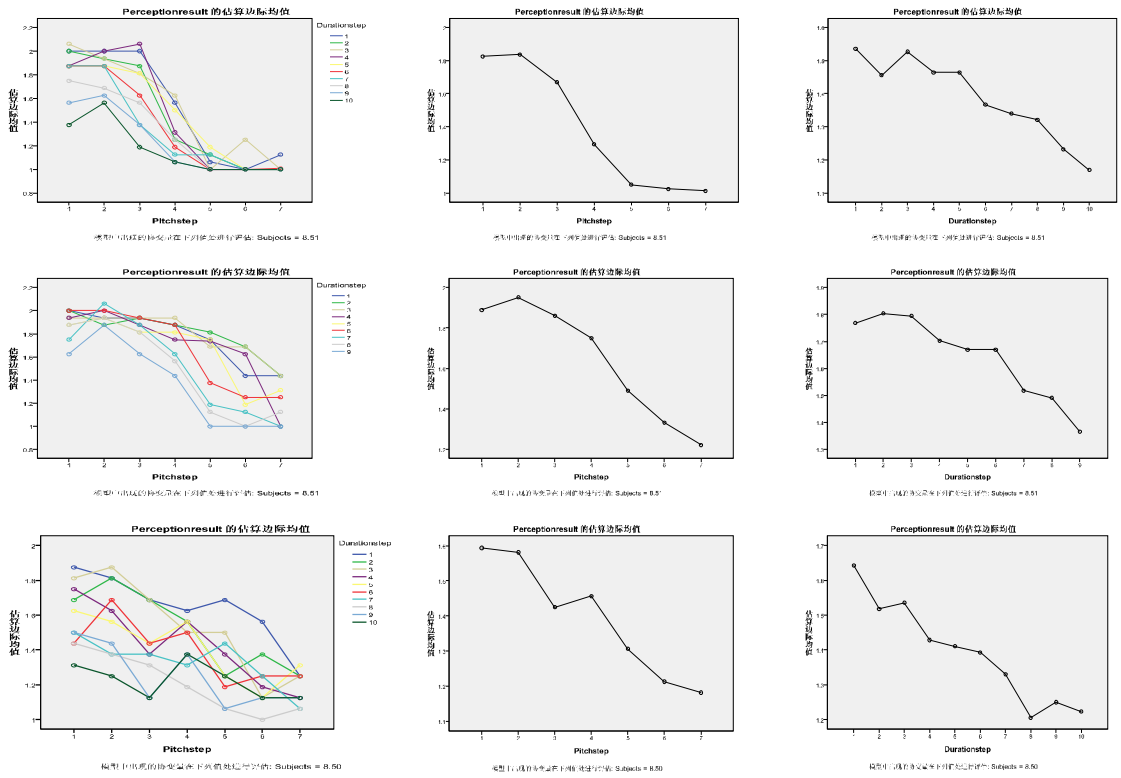

Fig. 3 From top row to bottom row are listed the perceptual results in isolation, on-focused position and post-focused position respectively. The left column shows the average perceptual results as a function of the change of pitch step and duration step. The middle column shows the average perception results with the change of pitch step. The right column shows the average perception results with the change of duration step.

(i) in isolation condition (top-left), when the pitch step 5 is the only one step higher than the actual neutral tone, duration plays an insignificant role in perception. This result further illustrates that duration plays a role only when pitch is not reduced to the neutral target. The pattern implies that duration and pitch have a complementary relationship on the perception of neutral tone to some extent. Under on-focus condition (left in the second row), when the pitch step is lower than step 3 (which is about the original normal stress pitch), duration does not show a significant influence on the perception, and it has the opposite case of the isolation condition. In post-focus condition (bottom-left), pitch and duration work simultaneously throughout all the steps, and the influence of duration is much greater than that in the previous two conditions. There is no significant interactive effect between pitch and duration.

(ii) In the second column, with the increase of pitch step (pitch shifting to the neutral tone /ku^0/), the perception score decreases, which indicates that more subjects identified the stimuli as neutral tone words. We found that the perceptual curve in isolation condition tends to be more similar to the typical function of categorical perception.

(iii) In the third column, across the three conditions, with the increase of duration step (duration shifting to the neutral tone /ku^0/), the perception score also decreases, which also implies that more subjects perceived the stimuli as neutral tone words. However, the perceptual curve does not take on the typical categorical perception pattern.

After all, both pitch and duration contribute to the perception of neutral tone; between the two parameters the pitch has greater magnitude of effect.

IV. MULTINOMIAL ORDINAL LOGISTIC REGRESSION ANALYSIS

In order to compare the distributions of perceptual spaces in three conditions, we normalized the pitch step and duration step according to the difference between the original normal stress (normalized to 0) and the neutral tone (normalized to 1). The steps were higher or lower than the baselines (natural productions of normal stress and neutral tone), after normalization, were smaller than 0 or larger than 1.

Multinomial ordinal logistic regression analysis was conducted to simulate the perceptual results. The perceptual score was set as the dependent variable Y (Y= 3, 2, 1 standing for the perception of normal stress “魔箍” /mo^{2}ku^1/, “uncertainty” and the neutral tone “蘑菇” /mo^{2}ku^0/ respectively). Independent variables were pitch step and duration step. Nine logistic functions were obtained based on the multinomial ordinal logistic regression analysis for three conditions (functions are not listed here for the limited space). P1 and P3 are the probability distributions of Y=1 (for /mo^{2}ku^0/) and Y=3 (for /mo^{2}ku^1/). P2 is the probability distribution of Y=2 (for “uncertainty”).

Fig. 4 shows the three-dimensional perceptual spaces and two-dimensional spaces of P1 and P3 plotted from the regression functions. It can be seen that neutral tone perception relates to both duration and pitch. With the increase of pitch step and duration step (both directing to neutral tone), the perception probability of P1 (neutral tone) increases and the perception probability of P3 (normal stress) decreases.

In Fig. 4, numbers in 2D spaces represent the perception probabilities of neutral tone words. The dark blue area on the right side shows that the probabilities of the perception of neutral tone is higher than that on the left light blue area where the perception probabilities of normal stress is higher. The two areas are demarcated by a boundary line.

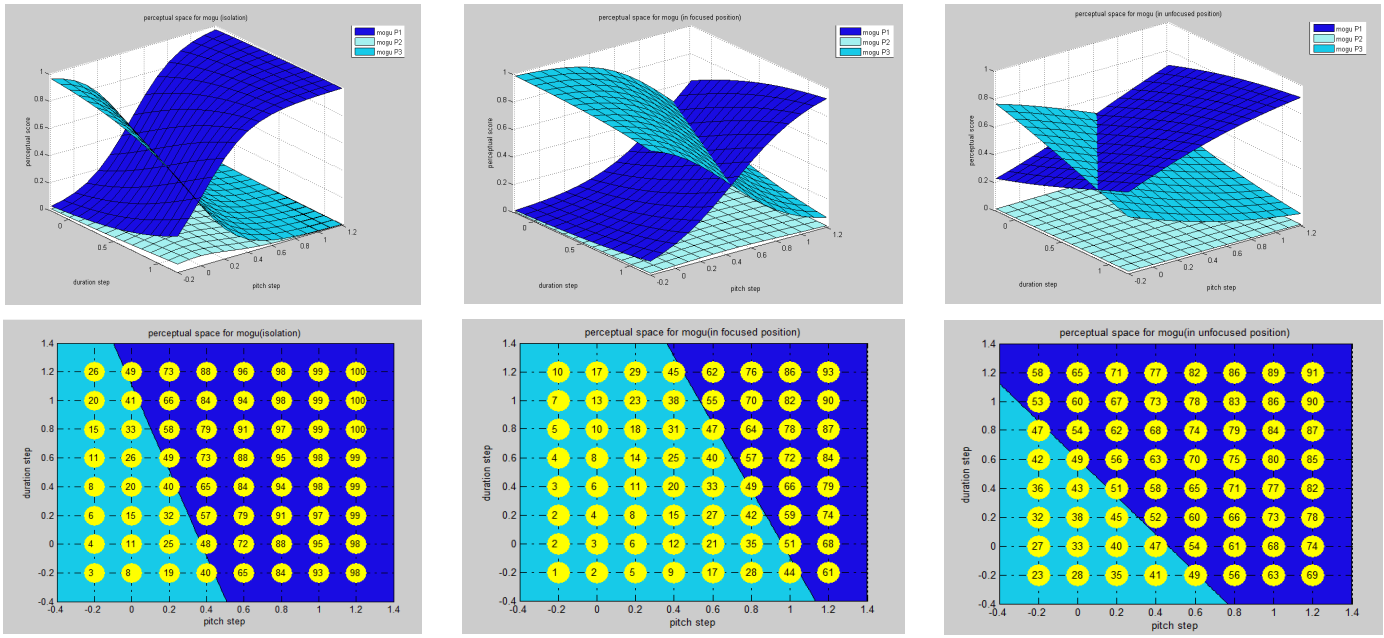

Fig. 4 Upper panel: the 3D simulating space based on multinomial ordinal logistic regression analysis. Lower Panel: 2D simulating space of P1 (dark blue area) and P3 (light blue area). The numbers in the circles are probability of P1 (neutral tone words). From left to right are for isolation, on-focus and post-focus.

The slopes of the boundaries in three conditions reflect that both the pitch and duration influence the perceptual scores (If the boundary is vertical, the perception would be 100% correlated with pitch, and irrelevant to duration at all. If the boundary is horizontal, the perceptual scores would be 100% correlated with duration and irrelevant to pitch), with pitch playing a more significant role than the duration. The absolute value of the slope of the perception boundary is indicative of the extension of correlation. If the absolute value of the slope equals to 1, it means that both cues play an equally important role. Here, the slopes of the boundary in isolation, on-focus and post-focus condition are -2.917；-2.333；-1.309 respectively. It reveals that the impact of pitch on perception decreases from isolation and on-focus condition to post-focus condition. In the post-focus condition here, the absolute value of the boundary slope is close to 1 (-1.309), which indicates that the influence of pitch and duration is quite similar. The simulated space again confirmed the results obtained in previous ANOVA analysis.

Fig. 4 also indicates that the perceptual spaces are different. In order to compare the difference of the simulating functions across the three conditions, we conducted the FDA (Function Discrimination Analysis) for the logistic regression functions. The way to compare the function difference is quite similar to the Least Square Principle. The 'accumulative distance' between two functions can be calculated by integrating the squared difference of them within a certain interval. Here the normalized pitch interval was set between -0.25 and 1.25, and normalized duration interval between -0.5 and 1.5. Within these ranges, difference between the two functions was double integral. The results (not listed here for the limited space) indicate that the perception distribution of 'uncertainty' is similar across the three contexts. The distributions of both P1 and P3 are quite similar, in that difference is smallest between isolation and post-focus, and greatest between isolation and on-focus, and less between on-focus and post-focus.

V. CONCLUSION

The acoustic correlates of neutral tone is always a disputing issue in Standard Chinese, especially the dominate cue between F_0 and duration. Based on the present identification experiment for a continuum of stimuli in pitch and duration spaces for minimally-contrasting stress pair "魔箍" /mo^{2}ku^1/ (magic ring) and "蘑菇" /mo^{2}ku^0/ (mushroom) in three conditions, isolation, on-focus and post-focus, we obtained some novel findings: in the perception of neutral tone, pitch has a greater impact than duration; duration shows a rather constant impact on the perception across the three conditions. The influence amplitude of pitch is closely related to the conditions, specifically, isolation> on-focus > post-focus. In post-focus condition, even though pitch still plays a slightly more significant role, duration is quite important as well. Based on the multinomial ordinal logistic regression analysis we clearly see that perceptual spaces distribute differently across the three conditions, and they are quite similar between isolation and post-focus and more different between isolation vs. on-focus condition and on-focus vs. on-focus condition.

The results found in this paper are totally different from Lin's [12], which confirmed that the impact of duration is much greater than that of pitch in neutral tone perception. Other researches [10, 11] proposed the greater impact of pitch but didn't investigate the situation in utterances.

Beckman &Edwards [20] and Sluijter & van Heuven [7] all proposed that in non-tonal languages, F_0 is a dependent variable of sentence stress instead of word stress; unstressed syllable in word is associated with duration, formants and other features instead of the change of F_0 [21, 22]. Our results stated that in Standard Chinese, as a tone language, F_0 has both functions to express tone and intonation; therefore, the stressing pattern of words, either in isolation as a word stress or in utterance as a nuclear pitch accent, is greatly related to F_0, with pitch as the most important correlate. However, when neutral tone words are in the post-focus position where the F_0 space is compressed, the impact of pitch is reduced to being almost equal to that of duration.

In the present perceptual experiment, only one word pair is tested. And how the underlying tone of the neutral tone affects the perceptual spaces is not explored. In the following studies, more word pairs with the underlying tone of the neutral tone of Tone 2, Tone 3 and Tone 4 will be tested. In the current study, only the weak syllable of the neutral tone word was considered and manipulated. But perception of neutral tone may be more related to the perception of the strong-weak pattern rather than the weak syllable only. Therefore, in the following studies, the relation between the initial strong syllable and the final weak syllable will be explored for the perception of the neutral tone.

ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program (973Program) of China (No. 2013CB329301), NSFC Project with No. 61233009, CASS innovation project 'Key Laboratory of Phonetics and Speech Science' and KNAW China Exchange Program (2010-2014) entitled 'The Early Acquisition of Speech Prosody: a comparative study of Dutch and Chinese".

REFERENCES

[1] Y. R. Chao, *Gwoyeu Romatzyh or the National Romanization.* In Wu, Z. J. (Eds.), Linguistic Essays By Yuenren Chao, Beijing: The Commercial Press. pp. 61-72, 1922.

[2] Y. R. Chao, *Beijing Kouyu Yufa(A Grammar of Spoken Chinese),* Beijing: The Commercial Press, 1979.

[3] Y. Y. Chen, and Y. Xu, "Production of Weak Elements in Speech Evidence from F0 Patterns of Neutral Tone in Standard Chinese," *Phonetica*. 63, 47-75, 2006

[4] J. L. Lu, and J. L. Wang, "Guanyu qingsheng de dingjie (Analysis on the Nature of Neutral Tone)," *Contemporary Linguistics*. 7(2), 107-112, 2005

[5] D. B. Fry, "Experiments in the perception of stress," *Language and Speech*. 1, 126-152, 1958.

[6] A. M. C. Sluijter, and V. J. Heuven, "Effects of Focus Distribution, Pitch Accent and Lexical Stress on the Temporal Organization of Syllables in Dutch," *Phonetica*. 52, 71-89, 1995.
[7] A. M. C. Sluijter, and V. J. Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Acoustical Society of America*. 100 (4), 2471-2485, 1996
[8] V.J. Heuven, and M. de Jonge, "Spectral and Temporal Reduction as Stress Cues in Dutch," *Phonetica*. 68:120–132, 2011.
[9] A. M. C. Sluijter, and V. J. Heuven, "Spectral balance as a cue in the perception of linguistic stress," *Acoustical Society of America*. 101 (1), 503-513, 1997.
[10] J. F. Cao, "The Acoustic Cues of Neutral Tone syllable in Standard Chinese," *J. Applied Acoustics*. 5(4), 1-6, 1986.
[11] Y. J. Wang, "The effect of pitch and duration on the perception of the neutral tone in standard Chinese," *ATTA ACUSTICA* 29 (5), 453-461, 2004
[12] T. Lin, "Experiments on Neutral Tone's qualities in Beijing Mandarin," *In Lin, T. & Wang, L. J. (Eds.), Papers on Experimental Phonetics)*. Beijing: Peking University Press. pp.1-26, 1985
[13] M. Lin, and J. Yan, "Beijinghua qingsheng de shengxue xingzhi," *Dialect* 3: 166–178, 1980.
[14] M. C. Lin, and J. Z. Yan, "Neutral tone and stress in Mandarin Chinese," *Language Teaching and Linguistic Studies*. 3, 88-104, 1990.
[15] T. Lin, and W. Wang, "Tone perception," *J. Chin. Linguist.* 2, 59–69, 1985
[16] S.A. Yang, "The synthetic rules for neutral tone in standard Chinese," *J. Applied Acoustics*. Vol 10(1), 1991.
[17] W.S. Lee, "A Phonetic Study of the Neutral Tone in Beijing Mandarin," *Paper at the 15th International Congress of Phonetic Sciences*, Barcelona, 2003.
[18] T. Lin, and L. Wang, *Course of Phonetics*, Beijing: Peking University Press. 1992.
[19] X. B. Zhong, B. Wang, Y. F. Yang, and S. N. Lv, "Hanyu yunlvci zhijue yanjiu (The perception analyze on Mandarin Chinese prosodic words)," *Xinli Xuebao (Acta Psychologica Sinica)*. 33(6), 481-488, 2001.
[20] M.E. Beckman, and J. Edwards, "Articulatory evidence for differentiating stress categories," *in Keating, Phonological structure and phonetic form. Papers in Laboratory Phonology*, Cambridge University Press, Cambridge. III, pp. 7– 33 1994.
[21] D.B. Fry, "The dependence of stress judgments on vowel formant structure," *in Zwirner, Bethge, Proc. 6th Int. Congr.Phonet. Sci*., pp. 306– 311 (Karger, Basel), 1965.
[22] A.M.C. Sluijter, V.J. Heuven, J.J.A. Pacilly, "Spectral balance as a cue in the perception of linguistic stress," *J.acoust. Soc. Am.* 101: 503– 513, 1997.