Effects of Preceding Vocabulary Context on the Perception of Mandarin Vowels

Xunan Huang[1,2], Caicai Zhang[2], Fei Chen[1], Jonathan Sieg[3], Lan Wang[1], Feng Shi[3]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

2Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

3 School of Literature, Nankai University, China

xiaonan0404@126.com

Abstract

This study compares the perceptual performance of Mandarin basic vowels "e" (/ɤ/) and "u" (/u/) in different contexts (independent & contextual). Results indicate that perception of the target vowel is influenced by the adjacent vowel context in a contrastive manner in both identification and discrimination tests. Moreover, in a context of higher F1 and F2, listeners found it more difficult to discriminate stimuli belonging to the /u/ category (which has lower F1 and F2), which may result from the effect of the referential formants of the context. Despite the influence of contextual factors, both /ɤ/ and /u/ in Mandarin showed relatively stable perception categories, and the perceived psychological parameters were consistent with the measured acoustic values Wu Zongji (1964) found for the Mandarin vowels /ɤ/ and /u/.

Index Terms: context, vowel /ɤ/, vowel /u/, categorical perception

1. Introduction

Formant frequency is one of the most important factors for distinguishing different vowels [1]. However, it is highly variable across utterances and talkers [2] [3]. How then do individuals perceive such variations? Earlier studies suggested that vowel perception undergoes a normalization stage during which listeners adapt their perception of vowels according to the talker's vowel space [4] [5]. Generally speaking, the process of vowel normalization is mainly based on two cues, namely word-internal cues and word-external cues. Word-internal cues that are involved during vowel normalization include formant frequencies, pitch, intensity and so on, and the word-external cues mainly involve acoustic cues before or after the target vowels. A word-external cue is, in other words, a contextual cue. All of these acoustic cues contain useful information about the vowel category, among which formant frequency is most important for vowel perception. Consequently, listeners make use of both word-internal and contextual formant frequency for vowel normalization. This study aims to explore the influence of contextual formant frequency on the perception of Mandarin vowels.

The contextual effect of formant frequencies in vowel perception was first examined by Ladefoged *et al.* [4]. They presented listeners with synthesized vowels after the manipulated carrier phrase, and asked them to decide if the target word was bit /bit/, bet /bɛt/, bat /bæt/, or but /bʌt/. Results showed that the same test word was more likely to be perceived as bet /bɛt/ rather than /bit/ when the F1 frequency in the carrier phrase was lowered, and if the F1 and F2 frequency of the adjacent contextual vowel was higher, participants tended to choose the target vowels with lower F1 and F2. These findings were referred to as the "contrastive context effect". In 1989, Ladefoged [5] had replicated this research with natural speech and further confirmed the contrastive effect.

In the same year, Nearey [6] summarized both intrinsic and extrinsic cues in vowel normalization. He indicated that a speaker's general voice character was necessary for speech perception, because it established a frame of reference. Additionally, Watkins and Makin [7] presented the traditionally used carrier phrase "please say what this word is" in a reversed order, and found similar contrastive effects when the carrier phrase was played in reverse. It seems that regardless of whether the carrier phrase is meaningful or not, contrastive effects still exist, even though the effect of a backward phrase is smaller. They further suggested that vowel normalization can be achieved by evaluating each vowel sound on the basis of the long-term average (LTA) spectrum of the preceding carrier phrase.

However, the contextual effect does not function well all the time. Mitterer's study [8] indicated that extrinsic factor effects were dependent on the range of the vowels in the carrier phrase and the target vowel. Extrinsic context factors might play a role in vowel perception only if the listeners have been exposed to vowels with a similar range of formant frequencies in the context.

In order to test whether the vowel context exerts a contrastive effect on target vowels when it is different from the target sound in terms of acoustic space, the perceptual features of both monosyllabic words (independent perception) and disyllabic words (contextual perception) are observed in the present study. We adopted the paradigm of categorical perception (CP) that included both identification and discrimination tests. By comparing boundary positions, discrimination peak positions and some other parameters of independent and contextual perceptions, we endeavored to detect the contextual effects of preceding vocabulary on the perception of Mandarin vowels.

Unlike most of the previous experiments on vowel normalization, which have used sentence-length contexts [4] [7] [9-11], the present contexts consisted of only one syllable. Basic vowels "e" (/ɤ/) and "u" (/u/) in Mandarin were chosen as the target vowels with vowel "a" /a/ serving as the carrier.

2. Method

2.1. Participants

Twenty four Mandarin native listeners (twelve female and twelve male; mean age = 20.73 years old, *SD* =1.09) from Beijing were recruited. They were all right-handed and had not received any formal musical or speech training. None of the participants reported having learning or memory problems, neurological or psychiatric disorders, or speech or hearing difficulties.

2.2. Stimuli

Two types of continua (monosyllabic word and disyllabic word) were constructed for the experiment, and the basic materials for synthesizing are listed in Table 1. A native male speaker from Beijing was instructed to read through the materials five to ten times, and three ideal pronunciations were chosen for each word as the synthesizing samples.

Table1. *Experimental materials used as the basis for stimuli continua.*

Monosyllabic word IPA Gloss	婀-乌 /ɤ 55/-/u 55/ fair-black
Disyllabic word IPA Gloss	大哥-大姑 /ta51//kɤ55/-/ta51//ku55/ eldest brother-aunt

The lowest three formants (i.e., F1, F2, and F3) are the main means to distinguish different vowels, but the most obvious difference between vowels /ɤ/ and /u/ is reflected in F1 and F2. Consequently, in this research, only F1 and F2 were manipulated, while the other formants remained unchanged. Based on the recorded utterance of /ɤ/, nine speech stimuli ranging from /ɤ/ to /u/ with equal distances in F1 and F2 were synthesized using Praat [12]. The formant frequency manipulation is illustrated in Figure 1, with stimulus 1 representing typical /ɤ/ and stimulus 9 representing typical /u/.

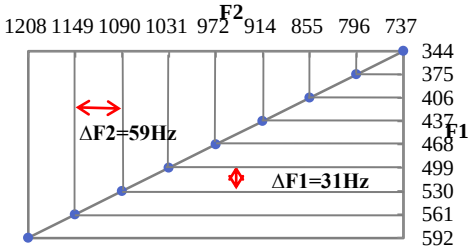

Figure 1: *The schematic diagram of stimulus continuum: Formant frequencies of the stimuli were synthesized from /ɤ/(592 Hz,1208 Hz) to /u/(344 Hz,737 Hz) with ΔF1=31Hz and ΔF2= 59Hz in every step.*

Due to the coarticulation effect of adjacent sounds, the acoustic characteristics of the target vowel in disyllabic words would be likely to exhibit "vowel reduction" [13]. For example, the average value of F1 in the vowel /i/ will be larger when it follows a low vowel like /a/. In order to eliminate the context effect exerted on the acoustic features of target vowels in disyllabic words, the stimuli continuum from /ta51//kɤ55/ to /ta51//ku55/ were synthesized based on the natural productions of /ta51//kɤ55/, and were generated from the parameters in the single word continuum (Figure 1).

2.3. Procedure

Participants were asked to perform two tasks: categorical perception in monosyllabic words (independent perception) and in disyllabic words (contextual perception). Both tasks involved two tests: vowel identification and vowel discrimination. A practice block was presented before the formal test to familiarize participants with the stimuli and the experimental procedure. Subjects were instructed to press 1 on the keyboard to represent /ɤ/ and press 2 to represent /u/ in the identification task. They were instructed to respond by pressing button 1 to indicate the same vowel and 2 to indicate different vowels in the discrimination task. The inter-stimulus-interval (ISI) was set to 500 ms, and the maximum reaction time to 2000 ms. Stimuli were presented in random order and repeated 3 times in each task. There were 27 trials in every vowel identification task and 69 trials in the discrimination task. Each task was divided into three blocks with a 20-second break in between.

2.4. Data Analysis

The identification score was defined as the percentage of responses with which participants identified that stimulus as being either /ɤ/ or /u/. In the following identification curves, only the percentage of /ɤ/ response is presented, and the percentage of /u/ is equal to 100% minus that of /ɤ/. Boundary position and boundary width were assessed by Probit analyses of individual identification curves [14]. The boundary position was defined as the 50% crossover point, and the boundary width was defined as the linear distance between the 25th and 75th percentiles. In the discrimination task, four types of pairwise comparison (AB, BA, AA, and BB, for stimuli A and B separated by two steps) were involved where AB and BA were the "different" pairs, AA and BB were the "same" pairs. The discrimination accuracy of each pair was calculated by using the formula below described by Xu *et al.* [15], where P("S"/S) denotes the percentage of "same" responses to all "same" pairs and P("D" /D) is the percentage of "different" responses to all "different" pairs. P(S) and P (D) are the percentages of "same" and "different" pairs respectively.

$$P = P(" S" /S) \times p(S) + P(" D" /D) \times P(D)$$

3. Hypothesis

The average formant frequencies of contextual and target vowels were calculated using Praat (see Table 2).The F1 and F2 frequencies of the carrier vowel /a/ are higher than those of the target vowels /ɤ/ and /u/, and the F1 and F2 frequencies of /u/ are lower than those of /ɤ/. If the contrastive context effect exists in this experiment, participants are more likely to choose /u/ (i.e., /ta51//ku55/) in the identification task. Therefore, the boundary position in the contextual condition would probably move forward to the end of stimulus 1, and the perceptual space of /u/ may be enlarged. If no context effect exists in this condition, there would not be any significant differences.

Considering the discriminations, the peak positions of the discrimination task would also change according to the boundary positions when the context effect exerts its influence; if not, these would not have any significant differences either.

4. Results

4.1. Vowel Identification

Identification curves and the calculated boundary positions are shown in Figure 2. The boundary positions under the contextual condition are farther forward than those of independent perception, which is consistent with the predicted contrastive context effect.

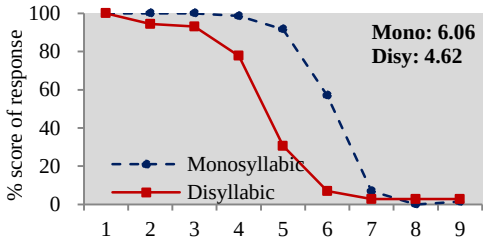

Figure 2: *Mean percentages of /ɤ/ responses in independent and contextual conditions are shown by dashed and solid lines respectively. The average categorical boundary position for each group is shown in the upper right corner*

An independent sample *t*-test was conducted on the perceptual boundary, contrasting independent perception and contextual perception. There was a significant difference between two groups [t = -5.879; p < .001; df =37.825]. This indicated that the perceptual boundary positions were significantly different across different contexts, with the boundary positions in contextual conditions occurring consistently towards a smaller stimulus number.

4.2. Vowel discrimination

The discrimination accuracies of seven comparison cohorts in two groups are shown in Figure 3. The average discrimination accuracies, discrimination peaks and peak positions are shown in Table 2.

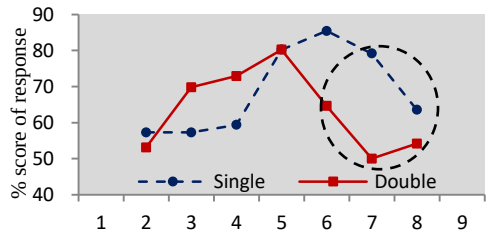

Figure 3: *Discrimination accuracies in isolated and contextual conditions are shown by dashed and solid lines respectively.*

An independent sample *t*-test was employed on the average discrimination accuracies, contrasting independent perception and contextual perception. Results showed that average discrimination accuracies were significantly different between two conditions [t = -2.173, p <.05; df =46], which indicated that the average accuracies of discriminations in the contextual condition were significantly lower than those in the isolated condition, as shown in Table 2.

Additionally, the discrimination accuracy in the isolated condition reached its maximum at pair 5-7, while that in contextual condition arrived at their peaks at pair 4-6 (see Table 2). An independent sample *t*-test was conducted for comparison of the peak discrimination accuracies between two conditions (independent and contextual). No significant difference was found [t = -7.796, p =.430; df =46].

Table 2. *Some parameters in discrimination.*

	Mean accuracy	Peaks	Peak position
Monosyllabic	68.9	85.42	5-7
Disyllabic	63.54	80.21	4-6

It is worth noticing that the average discrimination accuracy was significantly lower in contextual condition in terms of the /u/ category. In order to examine whether or not the difference was significant, their discrimination accuracies in the /u/ category were calculated. The stimuli after the boundary positions were classified in the /u/ category, thus the average discrimination accuracies of the stimuli after the boundary positions were calculated. An independent sample *t*-test was conducted for comparison of the average discrimination accuracies of /u/ category in two conditions (independent and contextual). The discrimination accuracies in the /u/ category were significantly higher in independent perception compared with those in contextual perception [t = -4.076; p <.001; df =38.506].

4.3. Perceptual space and perceptual boundary

An attempt was made to evaluate the performance of categorical perception under both independent and contextual conditions. The perceptual parameters are shown in Table 3.

Table 3. *Some perceptual parameters.*

		Monosyllabic	Disyllabic
Boundary Positions	F1	434.77 Hz	479.47 Hz
	F2	909.98 Hz	994.76 Hz
Boundary Widths	F1	28.56 Hz	32.59 Hz
	F2	54.17 Hz	61.82 Hz
Maximum Identification Rate		100%-100%	100%-97.22%

The perceptual distributions were obtained according to the boundary positions and boundary widths. Specifically, the formant frequencies less than that of the boundary position minus half of the boundary width were mainly perceived as /u/. Similarly, the formant frequencies more than that of the boundary position plus half of the boundary width were mainly perceived as /ɤ/. In the present study, the acoustic spaces where F1<420.49Hz and F2<882.90Hz were mainly perceived as /u/, and the acoustic spaces where F1>449.05Hz and F2>937.07Hz were mainly perceived as /ɤ/ in the independent condition. With regard to the contextual effect, the spaces where F1<463.18Hz and F2<963.85Hz were mainly perceived as /u/, while the spaces where F1>495.77Hz and F2>1025.67Hz were mainly perceived as /ɤ/ in the contextual condition.

Regarding the acoustic parameters of vowels /ɤ/ and /u/, we may refer to a study by Wu Zongji [16], which estimated the acoustic distributions of Mandarin basic vowels. He found that the average F1 and F2 frequencies of males' phonation of e /ɤ/ were around 540Hz and 1040Hz respectively, while those of /u/ were around 380Hz and 440Hz respectively. The

perceptual spaces of vowels /ɤ/ and /u/ in the present study nicely overlap with the acoustic parameters introduced by Wu Zongji.

Furthermore, the maximum identification rates were approaching 100% in both conditions. The context mainly showed its impact on the ambiguous stimuli around the boundary position; the identification rates remained stable in terms of both the /ɤ/ and /u/ categories.

5. Discussion

5.1. Context effects on identification performance

In the identification test, perceptual boundary positions were significantly different in two conditions. Considering boundary positions, since F1 and F2 frequency of the carrier vowel /a/ was higher than those of the target vowels /ɤ/ and /u/, the proportions of choosing /u/ (lower F1 and F2) increased in contextual conditions. Boundary positions were changed in a contrastive manner in the present research, although the formant range of the vowels in the context syllable is not the same with the target vowel. This indicates that the impact of extrinsic contextual formants depends on the relative differences between the target vowels and the vowels in the carrier phrase, rather than the theory proposed by Mitterer [8] that extrinsic factors might play a role in vowel perception only when the listener has been exposed to vowels with a similar formant range in the context.

5.2. Context effects on discrimination performance

Discrimination peak positions were altered in line with the boundary positions, which also originated from the contrastive context effect. As was calculated, discrimination peak positions moved forward in pace with the boundary positions, that is, discriminations showed their peak accuracy near the boundary positions. In the current study, the average boundary position moved from stimulus 6.06 to 4.62 in the contextual condition, and the discrimination peak position moved from stimulus pair 5-7 to 4-6, which is synchronous with the boundary position.

Furthermore, the average accuracies of discriminations in the contextual condition were significantly lower than those in the independent condition. This may result from the effect of the reference character of the context. Specifically, the average F1 and F2 frequency of the preceding vowel /a/ were much higher than those of /ɤ/ and /u/, therefore the differences between /ɤ/ and /u/ were weakened perceptually. Consequently, the discrepancies in stimulus pairs were also difficult to detect.

Discrimination peaks did not differ greatly. However, the average discrimination accuracies in the /u/ category were significantly lower in contextual conditions than in independent perception. This is consistent with the findings of Sjerps *et al.* [17]. They conducted a 4I-oddity discrimination task in which participants were instructed to decide the position of the deviant stimuli. Research suggested that discrimination performance was dependent on speaker context. In the context with higher F1, listeners found it more difficult to discriminate between the target vowel /ɪ/ and the ambiguous sound than between vowel /ɛ/ and the ambiguous sound. In the context of a low-F1 speaker, this pattern was reversed. In the present study, although the paradigm differed, a similar conclusion was reached. Evaluating target vowels on the basis of the preceding carrier phrase, the formant frequencies of vowels in the context played a referring effect. When the formant frequencies of reference character were much higher than those of the target vowels, listeners would find it more difficult to detect the subtle differences between the target stimuli.

5.3. The scope of context effect

In the present study, there is a stable perceptual space for both vowels /ɤ/ and /u/ where it is scarcely affected by the context. The spaces where F1<420.49Hz and F2<882.90Hz were mainly perceived as /u/, while the spaces where F1>495.77Hz and F2>1025.67Hz were mainly perceived as /ɤ/ in the contextual condition. Between the two spaces, there is an ambiguous section in which perception is easily influenced by context.

These findings can be explained by “Quantal Theory” [18] [19]. When a particular articulatory dimension is manipulated through a range of values, there is a nonlinear relation between this dimension and its acoustic consequence. The acoustic parameter is relatively insensitive to the change in the articulatory parameter over one portion of its range and shows a relatively rapid change with articulation over another part of its range. Additionally, regions of insensitivity of acoustic attributes to changes in articulation could provide a quantitative basis for defining distinctive features. In the current perceptual research, those distinctive portions remain stable in categorical perception, while the in-between sensitive proportion was more likely to be influenced by different context conditions.

6. Conclusions

The present study investigated the perceptual features of the Mandarin basic vowels /ɤ/ and /u/ in monosyllabic words (independent perception) and disyllabic words (contextual perception). Identification and discrimination tasks were both conducted. It turned out that target vowel perceptions were influenced by the adjacent vowel context in a contrastive manner in both identification and discrimination tasks. In the identification task, boundary positions were altered in a contrastive manner in the current research. For the discrimination task, the discrimination peak positions changed in line with the boundary positions, which is also due to the contrastive context effect. Furthermore, resulting from the effect of the reference character of the context, average discrimination accuracies are lower in the contextual condition, and the average discrimination accuracies in the /u/ category are significantly lower in contextual conditions.

Despite the influences of the contextual factors, both Mandarin vowels /ɤ/ and /u/ showed stable perceptual categories, and the perceived psychological parameters matched well with the measured values of Wu Zongji [16] for the acoustic parameters of Mandarin vowels /ɤ/ and /u/.

7. Acknowledgements

This work was supported by grants from the National Outstanding Youth Science Fund Project: “1150040716 Production and perception of lexical tone and music in Cantonese-speaking amusics” and the Key Program of National Social Science Fund: “13&ZD134 Acoustic and Perceptive Parameter Database of Mandarin Pronunciation Standard”.

8. References

[1] A. Bladon, and G. Fant, "A two-formant model and the cardinal vowels," *STL-QPSR*, vol.19, no.1, pp. 1-8, 1978.

[2] J. M. Heinz, K. N. Stevens, J. M. Heinz, *et al*. "On the properties of voiceless fricative constants," *Journal of the Acoustical Society of America*, vol.33, no.51, pp. 589-596, 1961.

[3] J. Hillenbrand, L. A. Getty, M. J. Clark, *et al*. "Acoustic characteristics of American English vowels," *Journal of the Acoustical Society of America*, vol.97, no.51, pp. 3099-3111, 1995.

[4] P. Ladefoged, and D. E. Broadbent, "Information conveyed by vowels," *Journal of the Acoustical Society of America*, vol.29, no.11, pp. 98-104, 1957.

[5] P. Ladefoged, "A note on 'Information conveyed by vowels'," Journal of the Acoustical Society of America, vol. 85, no.5, pp. 2223-2224, 1989.

[6] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *Journal of the Acoustical Society of America*, vol. 85, no.5, pp. 2088-2113, 1989.

[7] A. J. Watkins, and S. J. Makin, "Perceptual compensation for speaker differences and for spectral-envelope distortion," *Journal of the Acoustical Society of America*, vol. 96, no.3, pp. 1263–1282, 1994.

[8] H. Mitterer, "Is vowel normalization independent of lexical processing?" *Phonetica*, vol. 63, no.4, pp. 209-229, 2006.

[9] A. J. Watkins, "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," *Journal of the Acoustical Society of America*, vol. 90, no.6, pp. 2942-55, 1991.

[10] A. J. Watkins and S. J. Makin, "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," *Journal of the Acoustical Society of America*, vol. 99, no,6, pp. 3749–3757, 1996.

[11] M. J. Sjerps, H. Mitterer, J. M. Mcqueen, "Constraints on the processes responsible for the extrinsic normalization of vowels," *Attention Perception & Psychophysics*, vol.73, no.4, pp. 1195-215, 2011.

[12] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer," (Computer program). Version 5.3.56, retrieved 15 Sep 2013 from http://www.praat.org/, 2013.

[13] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, vol.35, no.11, pp. 1773-1781, 1963.

[14] D. J. Finney, "Probit Analysis (3rd)," *Cambridge: Cambridge University Press*, 1971.

[15] Y. Xu, J. T. Gandour, A. L. Francis, "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," *Journal of the Acoustical Society of America*, vol.120, no.2, pp. 1063-74, 2006.

[16] Z. J. Wu, "The Spectrographic Analysis of the Vowels and Consonants in Standard Colloquial Chinese," *Acta Acutica*, vol.1, no.1, pp. 32-40, 1964.

[17] M. J. Sjerps, J. M. Mcqueen, H. Mitterer, "Evidence for precategorical extrinsic vowel normalization," *Attention Perception & Psychophysics*, vol.75, no.3, pp. 576-587, 2013.

[18] K.N. Stevens, "On the quantal nature of speech," Journal of Phonetics, vol.17): pp. 3–45, 1989.

[19] K. N. Stevens, and S. J. Keyser, "Quantal theory, enhancement and overlap," *Journal of Speech*, vol.38, no.1, pp. 10-19, 2010.