Perceptual Confusability of Mandarin Sounds, Tones and Syllables

Bhamini Sharma, Chang Liu and Yao Yao

Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
bhamini.sharma@polyu.connect.hk, c.liu@polyu.edu.hk, ctyaoyao@polyu.edu.hk

ABSTRACT

This paper reports a perceptual identification study for Mandarin sounds, tones and whole syllables, using phonotactically plausible non-word stimuli covered in white noise. The results showed that while the accuracy of whole-syllable identification could be estimated by the independent accuracy of initial and final identification, syllable-level confusability patterns were related to, but not fully predictable from the confusability patterns of initials and finals. Implications of the results on modeling Mandarin phonological neighborhoods are also discussed.

Keywords: speech perception, perceptual confusability, neighborhood activation model, segmental and tonal processing, Mandarin Chinese.

1. INTRODUCTION

Current psycholinguistic models generally describe spoken word recognition as a process of identifying a target word from a set of similar-sounding, competing words [14], [16], [19]. One such model is the Neighborhood Activation Model (NAM; [14]), which provides a quantifiable method for defining perceptual competitors (or "phonological neighbors", using the term of the NAM). The definition of phonological neighbor in the NAM was originally based on perceptual confusability of words but later simplified as the one-phoneme difference rule, i.e. any word that is one phoneme away from the target word by addition, deletion or substitution is a phonological neighbor of the target word [13]. As the NAM gained increasing popularity in recent psycholinguistic research, measures of phonological neighborhood (such as neighborhood density and neighbor frequency) have been widely cited not only in studies of spoken word recognition, but also in studies of word production [23] and phonetic variation [7], [24].

However, previous research on phonological neighborhood has mostly focused on English and a few other European languages (e.g. French, Spanish). Importantly, little is known about the neighborhood structure of tonal languages such as Chinese. The few studies that calculated neighborhood statistics for Chinese languages either followed the one-phoneme difference rule by ignoring tone [22] or treating tone as another feature like phoneme [8], or considered only segmental or tonal neighbor [20]. To resolve the confusion in defining Chinese phonological neighborhoods, we think it is necessary to examine the basis of phonological neighborhood, i.e. perceptual similarity (as defined by confusability) among words, for Chinese lexicons. In this paper, we report a perceptual confusability study for Mandarin Chinese.

1.1. Mandarin syllables

While the concept of "word" may be vague for Chinese languages, it is safe to say that the building blocks of the Mandarin Chinese lexicon are monosyllabic monomorphemic units that correspond to single orthographic units (i.e. Chinese characters). The segmental structure of a Mandarin syllable is relatively simple, and can be described in the general form of (C)(G)V(N), where C=consonant, G=glide (/j w ɥ/), V=vowel, and N=nasal (/n ŋ/). Mandarin's lexical tone system includes four citation tones: a high level tone (Tone 1), a rising tone (Tone 2), a dipping tone (Tone 3) and a falling tone (Tone 4). With pitch contour being the primary cue for tonal distinction, secondary cues such as syllable duration also exist [11].

Linguists have proposed various phonological analyses for Mandarin syllable structure (e.g. [6]), but average Mandarin speakers (in Mainland China) commonly know a Mandarin syllable as consisting of an optional initial (聲母; corresponding to (C)), a final (韻母; corresponding to (G)V(N)) and a tone.

1.2. Mandarin spoken word recognition

Previous literature has illustrated a significant difference between Mandarin and non-tonal languages such as English in the time course of processing segmental and suprasemental information in spoken word recognition. For English, segmental information plays a major role in word recognition, and syllabic and suprasegmental information only enter the play at a later stage [3], [4], [10]; by contrast, in Mandarin word recognition, tone and segments are found to be processed simultaneously, leading some to propose that syllable should be considered as the processing unit in Mandarin spoken word recognition [2], [9], [15], [18], [25], [26], [27]. However, exactly how tones and segments may jointly influence the

perception of a tone-bearing Mandarin syllable remains a question.

1.3. Current study

In the current study, we conducted a perceptual identification experiment that probed into the perception of both syllable components (initial, final, tone) and whole syllables (including tones). The goal of the experiment was twofold: (1) to collect empirical data regarding the confusability of Mandarin initials, finals, tones and whole syllables; (2) to examine the (in)dependence of whole syllable perception on(from) the perception of segments and tones.

The methodology of the current experiment was similar to Tang & Lou's recent study on Mandarin speech perception in noise [21], which in turn followed the methods in Cutler et al.'s study [5]. However, a critical difference between the current study and [21] is that the current study used non-word stimuli while [21] used attested Mandarin syllables. Our main rationale for not using real Mandarin syllables was to avoid influence of lexical frequency, which is known to be an important factor in spoken word recognition. Meanwhile, it was also important to ensure maximal degree of well-formedness of the non-word stimuli so that the identification results could be informative for the perception of real Mandarin syllables. Given these concerns, we decided to use tonal gaps in Mandarin syllabary as experimental stimuli. A tonal gap is defined as a lexical gap only due to its tone (in other words, the same segmental combination exists in the lexicon, but with other tones). For example, /an^2/ is not a real Mandarin word but /an^1/, /an^3/ and /an^4/ all are, which makes /an^2/ a tonal gap.

2. METHOD

2.1. Participants

Fourty native Mandarin speakers (33F, 7M; mean age = 23.65 yr, SD = 3.20) participated in the present study. All participants were born and raised in Mainland China and self-identified as Mandarin native speakers.

2.2. Stimuli

Experimental stimuli consisted of 356 non-word monosyllables (with tones), representing all the tonal gaps in the Mandarin syllabary. The stimuli ranged over 22 different initials (including the null onset 0), 28 finals (V, VN, GV, GVN), and 4 tones. A female native speaker in her 20s recorded the stimuli in a soundproof room with a uni-directional microphone routed to Digi design. Each stimulus was read three times, and the token with the medium length was used in the main experiment. All tokens were normalized for intensity at 70dB. The modified tokens were divided into two equal groups, with white noise added to the two groups at +5dB and -5dB Signal-to-Noise (SNR) levels, respectively. All acoustic processing was done Praat [1].

2.3. Procedure

The experiment was a self-paced auditory identification task, programmed and conducted with Opensesame [17]. Each participant listened to the 356 mixed-SNR stimuli presented in a random order. Before each token was played, instruction was given on the screen about which aspect of the syllable (initial, final, tone, or whole syllable including tone) to pay attention to. After hearing the token, the participant identified the information at question by typing the corresponding *pinyin* symbols in an open-set identification task. Each token was presented to each participant only once, and trial type (initial, final, tone, or whole syllable including tone) was balanced across participants. Each token elicited 40 responses, ten of each trial type. Before the main experiment began, the participant was given four practice trials, one of each trial type.

3. RESULTS

3.1. Overall identification accuracy

Altogether 356*40 = 14240 identification responses were collected, among which 2.4% had to be excluded from analysis because the participants responded with a wrong type of phonetic information (e.g. responding tone when asked to identify the initial) or because the response contained non-*pinyin* symbols. Each trial type had between 3450 and 3550 valid responses. Among the valid responses, the mean overall identification accuracy (across all trial types and participant groups) was 60.0%. No reliable difference was found in either data exclusion rate or identification accuracy between any two participant groups ($p > .1$ in all t-tests), suggesting that the four groups of participants had highly similar overall performance (see Table 1).

Table 1: Data exclusion rates and identification accuracy in each participant group

Group	Data exclusion rates (%)	Identification accuracy (%)
1	1.5 (SD = 1.4)	63.3 (SD = 2.6)
2	2.8 (SD = 2.6)	57.2 (SD = 5.1)
3	3.0 (SD = 1.5)	60.9 (SD = 4.8)
4	2.3 (SD = 1.4)	58.4 (SD = 5.5)

3.2. Confusability patterns

3.2.1. Confusability of initials (in "initial" trials)

The identification of initials (overall accuracy = 44.1%) was less accurate than that of finals and tones, probably because the added white noise was more detrimental to the recognition of consonant features. Some general confusion patterns were observed for initials. As one would expect, consonants with similar phonetic features were likely to be confused with one another. For example, we observed /p/>/t/ (22%)[1], /t/>/p/ (6%), /k/>/x/ (16%), and /m/>/n/ (24%). Contra to [21], frequent confusion between bilabial and velar stops was not observed.

An interesting difference was also noted in the comparison of the perceptual confusability of three series (alveolo-palatal, alveolar, retroflex) of fricatives and affricates. The alveolo-palatal series were most confusable within the series (/tɕ/>/tɕ^h/, 17%; /tɕ^h/>/ɕ/, 18%); but the other two series were also confusable across series (/tʂ/>/ts/, 12%; /tʂ^h/>/ts^h/, 17%; /ʂ/>/tʂ^h/, 20%; /ts/>/s/, 22%; /ts^h/>/tʂ^h/, 7%). While these results disagreed with [21]'s observation of confusion between alveolo-palatals and retroflexes, they did provide support for the claim that the difference between alveolo-palatal and alveolar series was phonemic [12].

3.2.2. Confusability of finals (in "final" trials)

The overall accuracy of final identification was 66.1%. Same as the perception of initials, finals with similar phonetic properties (vowel position, lip rounding, nasality, and presence of glide) were also likely to be confused with one other. In addition, we also observed some strong trends regarding the perception of glides and nasal codas in complex finals. First, misperception of the identity of nasal coda (/n ŋ/) was very common (e.g. /ən/>/əŋ/, 12%; /in/>/iŋ/, 15%; /iŋ/>/in/, 36%), probably due to the fact that some participants were speakers of some southern dialects that neutralized nasal codas. Second, the medial glide was often omitted in the perception of complex finals (e.g. /ja/>/a/, 8%; /jaŋ/>/aŋ/, 20%; /jaʊ/>/aʊ/, 21%; /waɪ/>/aɪ/, 22%; /waŋ/>/aŋ/, 22%; /weɪ/>/eɪ/, 9.3%). Since the participants were instructed to identify only the finals in the "final" trials, these errors could be taken as evidence for an analysis that considered the medial glide as a feature of the consonant and hence part of the initial.

3.2.3. Confusability of lexical tones (in "tone" trials)

The overall accuracy of tone identification was 97.8%. Along the lines of previous studies [11], [21], current results showed that almost all the identification errors happened in Tone 2 and Tone 3 (Tone 2>Tone 3, 1.3%; Tone 3>Tone 2, 5.6%).

3.2.4. Confusability of whole syllables (in "whole syllable" trials)

The overall accuracy of whole syllable identification was 31.7%. In this part of the analysis, we were most interested in verifying whether the perceptual errors of whole syllables could be predicted from the perceptual errors of initials, finals and tones. To this end, we performed two statistical analyses: one on perceptual accuracy (i.e. the probability of correct identification) and the other on perceptual confusability.

In the first analysis, we calculated the probability of correctly identifying each initial, final and tonal type (i.e. p(I), p(F), p(T), based on responses in "initial", "final" and "tone" trials), as well as the probability of correctly identifying each syllable type (i.e. p(Syll), based on responses in "whole syllable" trials). A two-sampled t-test showed that p(Syll) was not significantly different from the product of p(I), p(F) and p(T) (t(709.4) = -0.09, p >.9). Furthermore, when p(T), which was basically 1, was omitted, the product of p(I) and p(F) was still not significantly different from p(Syll) (t(709.8) = 0.11, p >.9). That is to say, the probability of correctly identifying a Mandarin word-like syllable can be estimated by simply multiplying the probabilities of correctly identifying its initial and final (see the formula in (1)), and no consideration of tone or initial-final interaction is necessary.

$$(1)\quad p(Syll) = p(I)p(F)$$

In the second analysis, we investigated whether the probability of perceiving one syllable, e.g. Syll_A (with initial I_A, final F_A, and tone T_A), as Syll_B (with initial I_B, final F_B, and tone T_B; Syll_B may or may not be the same as Syll_A) can be predicted jointly from the probabilities of perceiving I_A as I_B, F_A as F_B and T_A as T_B. In other words, we wanted to test the relationship between syllabic confusability probability $p(\text{Syll}_B|\text{Syll}_A)$, and component confusability probabilities $p(I_B|I_A)$, $p(F_B|F_A)$, $p(T_B|T_A)$. All probabilities were estimated from the identification patterns observed in our dataset. A two-sample t-test showed that the product of $p(I_B|I_A)$, $p(F_B|F_A)$ and $p(T_B|T_A)$ was significantly lower than $p(\text{Syll}_B|\text{Syll}_A)$ (t(2972.6) = -19.7, p<.001). A regression analysis of log transformed $p(\text{Syll}_B|\text{Syll}_A)$ with log transformed

$p(I_B|I_A)$, $p(F_B|F_A)$ and $p(T_B|T_A)$ as predictors revealed significant positive effects from all predictors on $\log(p(Syll_B|Syll_A))$ (see Table 2 for a summary of the regression model), but the adjusted R^2 associated with the model was only 0.31, suggesting that the model predicted less than one third of the variation in $\log(p(Syll_B|Syll_A))$.

Table 2: Summary of the linear regression model on $\log(p(Syll_B|Syll_A))$

Predictor	Coefficient (β)	Std. Error	p
Intercept	-1.10	0.03	<.001
$\log(p(I_B\|I_A))$	0.22	0.01	<.001
$\log(p(F_B\|F_A))$	0.16	0.01	<.001
$\log(p(T_B\|T_A))$	0.11	0.01	<.001

4. DISCUSSION

In this paper, we reported a speech perception study that examined the perceptual confusability of initials, finals, tones and whole syllables in Mandarin. To avoid influence of lexical frequency, we used phonotactically plausible non-word stimuli. Our results of component (initial, final, tone) confusability reflected the role of phonetic similarity in speech (mis)perception, and echoed previous observations of perceptual confusion in Mandarin spoken word recognition. Moreover, we also showed that while syllabic identification accuracy can be estimated by the product of identification accuracy of syllable components (only initials and finals were necessary), confusability probability at the syllable level was hard to predict from confusability probabilities of the components. These findings have at least two important implications for modelling the phonological neighborhoods of Mandarin Chinese.

First of all, the fact that syllabic perceptual confusability does not equal to the product of confusability probabilities of syllable components indicates that the recognition of initial, final and tone cannot be modelled as independent events during the perception of a Mandarin syllable. Our results are compatible with the view that syllable (as opposed to segment) is the basic processing unit in Mandarin speech perception, which would also entail that the Mandarin neighborhood structure should be fundamentally different from what would be mapped by the currently dominant one-phoneme difference rule.

Furthermore, our results confirmed previous observations of robust tonal identification in adverse hearing conditions. When the stimuli are added with white noise or speech-like noise (current study and [21]), tones can still be correctly recognized in >90% of the cases; even when F0 information is partially neutralized, identification accuracy remains over 80% [11]. What does this mean for the neighborhood structure of Mandarin? In particular, how do we model neighbors that share segmental composition but differ in tones? On one hand, we know that tonal neighbors have high phonological similarity, as evinced by previous priming studies (e.g. [25] showed that Mandarin tonal neighbors primed each another, but segmental neighbors did not). On the other hand, as we have observed, tonal neighbors are hardly confusable in identification. How do we represent such "so close, yet so far" perceptual distance in a neighborhood model? We believe that this line of inquiry is important for understanding both Mandarin spoken word recognition and the phonological neighborhood model in general.

Acknowledgment
Research reported in this paper was supported by funding from the Dean's Reserve for Research, Scholarly, and Other Endeavors (Project Account Code: 1-ZVB7), Faculty of Humanities, The Hong Kong Polytechnic University.

5. REFERENCES

[1] Boersma, P., Weenink, D. 2011. Praat: Doing phonetics by computer. Version 5.3. http://www.praat.org

[2] Chen, J. Y., Chen, T. M., Dell, G.S. 2002. Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language* 46, 751–781.

[3] Cholin J., N. O., Schiller, S., Levelt, W. J. M. 2004. The preparation of syllables in speech production. *Journal of Memory and Language* 50, 47–61.

[4] Cutler, A. 1986. Forbear is a homophone: Lexical prosody does not constrain lexical access. *Language and Speech* 29, 201–220.

[5] Cutler, A., Weber, A., Smits, R., Cooper, N. 2004. Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116, 3668.

[6] Duanmu, S. 2007. The Phonology of Standard Chinese. Oxford University Press.

[7] Gahl, S., Yao, Y., Johnson, K. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66(4), 789–806.

[8] Kirby, J. P., Yu, A. C. L. 2007. Lexical and phonotactic effects on wordlikeness judgements in Cantonese. *Proc. 16th ICPhS* Saarbrücken 1389-1392.

[9] Lee, L., Nusbaum, H. C. 1993. Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese. *Perception & Psychophysics* 53, 157-165.

[10] Levelt, W. J. M., Roelofs, A., Meyer, A. S. 1999. A theory of lexical access in speech production.

Behavioral and Brain Sciences 22, 1–75.
[11] Liu, S., Samuel, A. G. 2004. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech* 47, 109-138.
[12] Lu, Y. A. 2011. The psychological reality of phonological representations: the case of Mandarin fricatives. Paper presentation at the *23rd North American Conference on Chinese Linguistics (NACCL)*.
[13] Luce, P.A. 1986. Neighborhods of words in the mental lexicon. Ph.D. Thesis, Indiana University.
[14] Luce, P. A., Pisoni, D. B. 1998. Recognizing spoken words: the neighborhood activation model. *Ear and Hearing* 19, 1–36.
[15] Malins, J. G., Joanisse, M.F. 2010. The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language* 62, 407–420.
[16] Marlsen-Wilson, W. D. 1984. Function and process in spoken word recognition: A tutorial review. In. H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes*. Erlbaum; Hillsdale, NJ.
[17] Mathôt, S., Schreij, D., Theeuwes, J. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44, 314-324.
[18] McBride-Chang, C., Tong, X., Shu, H., Wong, A. M. Y., Leung, K., Tardif, T. 2008. Syllable, phoneme, and tone: Psycholinguistic units in Early Chinese and English word recognition. *Scientific Studies of Reading* 12, 171-194.
[19] Norris, D. 1994. Shortlist:A connectionist model of continuous speech recognition. Cognition 52, 189-234.
[20] Su, L., Chen. J. 2012. The research of phonological neighborhoods in Mandarin Chinese. *Psychological Research* 5, 26–33.
[21] Tang, K., Lou, Y. 2014. Mandarin Chinese speech perception in noise: Phonological implications. Paper presentation at the 11th Old World Conference in Phonology (OCP11), Amsterdam-Leiden, The Netherlands.
[22] Tsai, P. T. 2007. The effects of phonological neighborhoods on spoken word recognition in Mandarin Chinese. Master's thesis, University of Maryland.
[23] Vitevitch, M. S. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning Memory and Cognition* 28 (4), 735–747.
[24] Wright, R. A. 2004. Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation, Papers in Laboratory Phonology* (pp. 75-87). Cambridge, UK: Cambridge University Press.
[25] Zhang, Q. F., Yang, Y. F. 2004. The time course of semantic, orthographic an phonological activation in Chinese word production. *Acta Psychologica Sinica* 36, 1–8.
[26] Zhang, Q. F., Yang, Y. F. 2005. The phonological planning unit in Chinese monosyllabic word production. *Psychological Science* 28, 374–378.
[27] Zhao, J., Guo, J., Zhou, F., Hua, S. 2011. Time course of Chinese monosyllabic spoken word recognition: Evidence from ERP analyses. *Neuropsychologia* 49, 1761–1770.

[1] ">" denotes "is most likely to be confused with" and the rate of confusion is given after.