PERCEPTION OF L2 ENGLISH AFFRICATE ONSETS IN NATIVE MANDARIN LISTENERS

Yizhou Wang[1] & Rikke L. Bundgaard-Nielsen[2]

[1]The University of Melbourne, [2]MARCS Institute (Western Sydney University)
yizwang3@unimelb.edu.au, rikkelou@gmail.com

ABSTRACT

This paper reports on a study investigating native Mandarin listeners' perceptual discrimination and identification of two English phonetic affricate onset contrasts, /tʃ/-/tɹ/ and /dʒ/-/dɹ/, in two vowel contexts, /i/ and /u/. Results suggest that vowel quality influences discrimination accuracy such that /u/ creates a difficult condition for Mandarin listeners, especially for the long voice of onset (VOT) contrast, /tʃ/-/tɹ/. Mouse-tracking data revealed rich information about online processing during the identification procedure, and, consistent with the results from the discrimination task, multiple metrics suggested a robust effect of vowel context. Interestingly, and in contrast to the discrimination results, VOT was not found to be an influencing factor in cursor movements in the identification, indicating that VOT may have a more nuanced role as compared to vowel context.

Keywords: affricate, perception, Mandarin, mouse-tracking.

1. INTRODUCTION

Perceiving unfamiliar nonnative and second language (L2) consonants can be challenging for listeners, especially when certain consonants form phonemic contrasts in the L2 but not in the listener's native language (L1), e.g., the English /l/-/ɹ/ contrast is notoriously difficult for native Japanese listeners to accurately perceive [1]. Many phenomena of this kind stem from differences between L1 and L2 inventories [2]. Another important line of research additionally shows that nonnative category confusions can be influenced by phonological and phonotactic contexts, e.g., L1 English and L1 French listeners are unable to accurately perceive */tl/-/kl/, and */dl/-/gl/ contrasts because both */tl/ and */dl/ are unattested sequences according to English or French phonology [3], [4]. Close analysis of such patterns indicates that */dl/-/gl/ is slightly easier than */tl/-/kl/, suggesting a potential effect of voice onset time (VOT). Consistent with this observation, a recent study [5] showed that L1 Japanese listeners' perception of English /s/-/ʃ/ contrast is conditioned by the nucleus vowel context: Japanese listeners can perform good discrimination in a familiar phonotactic context (*/su/-/ʃu/), but their performance significantly worsens in an unfamiliar phonotactic context (where the nonnative sequence is unattested in the L1, */si/-/ʃi/). These studies indicate the perception of L2 consonant onsets can be influenced by the vowel context, the phonotactic quality of the sequence (unattested or attested), and potentially, the voicing specification of the plosives of the contrast (long- or short-lag).

In the present study, we investigate L1 Mandarin listeners' perception of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ diphone contrasts, which have sometimes been reported to be challenging for Mandarin listeners [6], [7]. It remains unknown, however, how the processing of these two contrasts is influenced by their phonological contexts. Importantly, while English /tɹ/ and /dɹ/ are stop-rhotic sequences phonologically, their realisations are phonetically more similar to affricate-rhotic sequences [8]. Mandarin has long- and short-lag affricates /tʃ, dʒ/ and a rhotic category /ɹ/. For complex onsets, Mandarin phonology permits affricate-/w/ sequences, but affricate-rhotic sequences are not attested [9].

As a result, English /tɹ, dɹ/ may be perceived as *unfamiliar* phonetic affricate categories by Mandarin listeners, while English /tʃ, dʒ/ may be perceived as *familiar* affricate categories. The aim of the present study examines Mandarin listeners' perceptual sensitivity to the cues associated with the /ɹ/ segment in unfamiliar phones, which includes the labial (rounding) and lingual (narrowing) gestures. In case Mandarin listeners do not accurately perceive the rhotic segment in the onset, they would perceive /tɹ, dɹ/ as similar to English /tʃ, dʒ/, which are phonetically similar and phonotactically attested according to Mandarin phonology. Alternatively, Mandarin listeners may exploit some (but not all) the gestural cues in perception, e.g., attending to the labial gesture and thus perceiving English /tɹ/ as similar to the Mandarin sequence /tʃw/. This is supported by evidence in Mandarin loanword adaption patterns, e.g., the English name *Trump* is adapted as '川普' /tʃwan-pu/. If that is the case, English /tʃ/-/tɹ/ will be perceptually mapped to Mandarin /tʃ/-/tʃw/ in perception, and thus even unfamiliar categories may still be discriminated accurately given their acoustic and gestural cues are perceptually salient.

However, if Mandarin listeners heavily rely on the labial gesture (i.e., substituting /ɹ/ with /w/), then the perception of /tʃ/-/tɹ/ will still be difficult when the following segment also has the [+labial] feature, e.g., a rounded vowel /u/, due to anticipatory coarticulation, because then the two affricate categories would be difficult to differentiate based on whether the labial gesture is present. In contrast, an unrounded vowel such as /i/ will create a condition where the labial gesture should be maximally salient (since now only /tɹi/ will be produced with the labial gesture), and good perception would be expected. Our general prediction is thus that Mandarin listeners' discrimination and identification of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ will be influenced by the vowel context: /u/ is predicted to induce poor discrimination and identification but /i/ is predicted to induce more accurate perception.

In the following, we present an AXB discrimination task and a mouse-tracking identification task to investigate Mandarin listeners' perception of the /tʃ/-/tɹ/ and /dʒ/-/dɹ/. The discrimination task tests the pairwise discriminability of English /tʃ/-/tɹ/ and /dʒ/-/dɹ/ by analysing the discrimination accuracy, which reflects the outcome of perception [10]. The mouse-tracking identification task is deployed to investigate the online processing patterns, i.e., how phonetic-phonological information is integrated during the decision-making process, because vision, cognition, and hand motion are tightly coupled, and goal-approaching movement is a valid index of cognitive conflicts [11]–[14].

2. METHODS

2.1. Participants

Twenty right-handed, native Mandarin listeners (18 females; two males) participated in the study. None reported any hearing or speech disorders. All were international students at an Australian university (M_{age} = 24.3), and all spoke English as an L2. In addition to Standard Mandarin, eleven of them also spoke a regional Mandarin dialect, and two spoke Cantonese. None spoke a third language fluently. Their average length of residence in Australia was 2.2 years, and on average, they had learned English in foreign language classroom settings for 13.3 years. Their average age at the onset of acquisition was 6.6 years old, and their mean age of arrival was 22.1 years old. All participants completed a vocabulary size test (VST) [15], and their mean VST score was 8075. Based on these measures, the participants are advanced L2 English speakers.

2.2. Stimuli

The stimuli were eight English CVCV pseudowords, /tʃu-ti, dʒu-ti, tɹu-ti, dɹu-ti, tʃi-ti, dʒi-ti, tɹi-ti, dɹi-ti/, produced by a male native speaker of Australian English, who was phonetically trained. These stimuli were used to create six contrasts, including four critical contrasts /tʃi/-/tɹi, /tʃu/-/tɹu/, /dʒi/-/dɹi/, /dʒu/-/dɹu/, plus two filler contrasts, /dɹu/-/tɹu/, /dʒu/-/dʒi/. The target syllables differ systematically in terms of vowel context (/u/ vs /i/), phonological structure (true affricate vs stop-rhotic sequence), and VOT (short- vs long-lag, or voiced vs voiceless). The second syllable, /ti/, was added to generate a controlled phonological context across all stimuli. The speaker produced each pseudoword three times in a clear speech style in order to maximise the acoustic differences. Stress was on the first syllable of each stimulus word.

2.3. Procedures

The participants completed the two experiments on two subsequent days: First the discrimination task, and then the mouse-tracking task. The tasks were served online and data was collected using *PsyToolkit* [16], [17] and *PsychoPy* [18]. In the discrimination task, participants were presented with 144 trials (six contrasts, four triplets, and six repetitions) testing discrimination of On each trial, the listener was served a sequence of three stimuli (A, X, B) with an interstimulus interval (ISI) at 1.0 s, and the middle stimulus (X) was either phonologically identical to the first or the last stimulus. The long ISI was used to encourage phonological processing [19]. The listener had 3.0 s to decide whether the first two or the last two stimuli were more similar by pressing the "F" key (X = A) or the "J" key (X = B) on their keyboard.

In the mouse-tracking identification task, the participant's computer screen was normalised to a 2 units by 2 units canvas. Following the conventional mouse-tracking paradigm, a "start" box was located in the centre of the screen bottom [0, 0], while the two response labels were printed at the top-left [-1, 2] and top-right [1, 2] corner of the screen. On each trial, the listener clicked the "start" box to play the auditory stimulus, and had to move their mouse and click on the appropriate category label "CH" vs "TR" or "J" vs "DR", representing /tʃ, tɹ, dʒ, dɹ/, respectively. After a trial, the "start" box was printed again, and the listener needed to click on the box to enter the next trial. This procedure ensured that the mouse cursor started from approximately the same location each time. The stimuli were presented in a randomised order, while the directions (left vs right) of correct responses were counterbalanced. The task had 288 trials (four consonants, two vowels, three tokens per combination, two directions, and six repetitions.

The mouse trajectories were recorded during the response procedure. The sample rate of mouse-trajectory recording was 60 frames per second (FPS) for all participants, i.e., two adjacent mouse locations represent cursor displacement within 17 ms. For data analysis, all rightward trajectories were mirrored as leftward trajectories.

When the decision is easy, we expect that the mouse tracking trajectory resembles a straight line connecting the "start" button and the correct response, while cognitive conflicts can lead to more or less curved trajectories. Following previous research [11]–[13], we analysed multiple aspects of mouse movement, including response latency and curvature complexity. In particular, response latency was measured in terms of the identification RT and motor pauses (defined as the idle time after movement initiation, in seconds). Trajectory curvature complexity was measured by total trajectory distance and the maximal deviation from the ideal straight line (both in normalised units).

3. RESULTS

3.1. AXB discrimination

Participants achieved very good performance in discriminating the two filler contrasts (/dɹu/-/tɹu/, 95%, /dʒu/-/dʒi/, 97%), and the filler trials were systematically removed in the statistical analysis. For AXB accuracy (Table 1), we built a generalised linear mixed-effects model (GLMM, binomial link) to evaluate the effect of the onset, vowel, and their interaction whilst controlling participants as a random factor. Tukey-adjusted *post hoc* tests revealed that the contrast /tʃu/-/tɹu/ had significantly lower accuracy as compared to /tʃi/-/tɹi/ (p = .013), /dʒi/-/dɹi/ (p = .017), and /dʒu/-/dɹu/ (p = .027), but these three contrasts did not differ significantly. We further compared mono-dialectal Mandarin speakers with participants who spoke an additional dialect, but no difference was observed in all contrasts (p's > .05, t-tests). The correlation between accuracy data and the speakers' vocabulary sizes was checked, but no significant coefficients were observed (p's > .05, Pearson's r), suggesting that the cohort of participants tested here was homogenous.

3.1. Mouse-tracking identification

The accuracy data of the identification task are summarised in Table 2. In general, the participants achieved very high accuracy (93-97%) in the /i/ context, but the /u/ context led to much poorer performance (57-85%). For statistical analysis, we built a GLMM (binomial link) to model the effect of vowel condition (/u/ vs /i/), phonological structure (affricate vs sequence), and VOT (short vs long). A Wald Chi-squared test revealed a main effect of vowel condition [χ^2 = 479, p < .0001], structure [χ^2 = 149, p < .0001], and VOT [χ^2 = 16, p < .0001]. In addition, there was a significant vowel-structure interaction effect [χ^2 = 9.1, p = .0026], vowel-VOT interaction effect [χ^2 = 13.0, p = .0003], and a vowel-structure-VOT interaction [χ^2 = 7.3, p = .0070]. In addition, a short VOT led to higher identification accuracy in the /u/ context, and the true affricates (/tʃu, dʒu/) had lower accuracy than the corresponding sequence categories (/tɹu, dɹu/).

CV-CV	Condition	Acc. (SD)
/tʃu/-/tɹu/	Long VOT	88 (13)
/dʒu/-/dɹu/	Short VOT	95 (6)
/tʃi/-/tɹi/	Control	98 (4)
/dʒi/-/dɹi/	Control	97 (9)

Table 1: AXB discrimination accuracy.

CV	Acc. (SD)	CV	Acc. (SD)
/tʃu/	57 (31)	/tʃi/	97 (8)
/dʒu/	66 (30)	/dʒi/	93 (20)
/tɹu/	78 (26)	/tɹi/	96 (6)
/dɹu/	85 (18)	/dɹi/	97 (8)

Table 2: Identification accuracy.

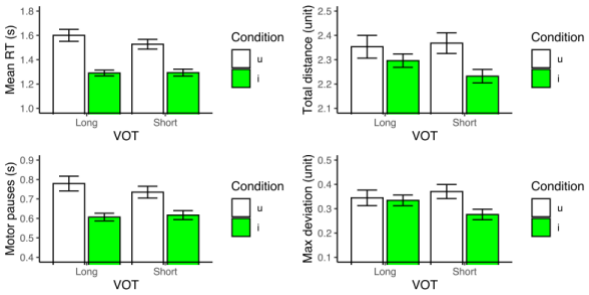

Figure 1: Mouse-tracking metrics.

For the mouse trajectory analysis, we focused on the correct responses where /tʃ/ and /dʒ/ were the target responses, while /tɹ/ and /dɹ/ were the corresponding distractors, as the accuracy data indicated that our listeners had more difficulty in identifying the true affricate categories as compared to the sequence categories. First, we analysed two latency metrics in the responses, including the mean RT and motor pauses, defined as the mouse idle time after the motion initiation, see Figure 1. We applied logarithmic transformations and built two linear mixed-effects models (LMMs) to evaluate the effect of vowel conditions as well as VOT.

For RT, a Wald Chi-squared test revealed a significant effect of the vowel [χ^2 = 69.4, p < .0001], but no significant effect of VOT [χ^2 = 1.9, p = .1686], or vowel-VOT interaction [χ^2 = 1.5, p = .2223], indicating that vowel condition but not VOT

condition affected the response RT. In the /u/ context, listeners spent a significantly longer time making identification choices. For motor pauses, we again found a main effect of the vowel [χ^2 = 32.8, p < .0001], but no significant effect of VOT [χ^2 = 0.6, p = .4533], or vowel-VOT interaction [χ^2 = 1.7, p = .1935], indicating that the listeners had significantly longer pauses in the /u/ context than the /i/ context.

Next, we analysed two curvature metrics of mouse trajectories, including total trajectory distance (length), and maximal deviation from the ideal straight line, see Figure 1. These metrics were similarly analysed using LMMs. For total distance, we found a significant effect of the vowel [χ^2 = 18.7, p < .0001], but no significant effect of VOT [χ^2 = 0.8, p = .3637], or vowel-VOT interaction [χ^2 = 0.9, p = .3474]. These results indicated that vowel condition but not VOT affected the trajectory lengths, and participants had significantly longer mouse trajectories in the /u/ context than in the /i/ context. Similarly, maximal deviation showed a significant effect of the vowel [χ^2 = 15.0, p = .0001], but no significant effect of VOT [χ^2 = 1.6, p = .2025], or vowel-VOT interaction [χ^2 = 1.5, p = .2184], suggesting a similar pattern.

4. GENERAL DISCUSSION

Both the AXB task and the identification task confirmed our prediction that /u/ can create a more challenging condition as compared to /i/ in Mandarin listeners' perception of English (phonetic) affricate onsets, but the two tasks also revealed slightly different patterns. In the AXB task, accuracy data indicated that /u/ created a difficult scenario only in the long VOT condition. However, the identification task revealed that the /u/ context caused perceptual confusion in both short and long VOT conditions, although the accuracy was still higher in the short VOT context as compared to the long VOT context. It is worth noting that Mandarin and English have similar VOT-based contrasts for perceptually distinguishing short- and long-lag obstruents [20]. In general, our finding thus indicates that VOT has a more nuanced interfering effect than the vowel context, and such effect is more apparent when the task complexity increases, because the identification task but not the AXB task requires extra knowledge to draw the correspondence between orthographic and phonological representations. Nonetheless, these findings echo previous research findings that the difficulty level of nonnative consonant perception is influenced by the phonological and phonotactic context [5], and VOT may also play a role in the perceptual easiness of unfamiliar onset categories [3], [4]. One potential explanation is that VOT differences can affect the temporal structure and the phasing relations between the articulatory gestures, and thus the perceptual salience of other gestures (e.g., the lingual and labial gestures for producing the rhotic sound) can be reduced due to an increased salience of aspiration (wide laryngeal). Or perhaps strong aspiration leads to partial devoicing of the following sonorants, and therefore the gesture cues are weakened and become more difficult to attend to, especially for L2 listeners. As for the differences in the results from our two experiments, it is possible that auditory learning precedes the learning of the correspondences between phonology and orthography, or perhaps the explicit metalinguistic knowledge creates another layer of phonological representations beyond perception itself [21], so that the discrimination and identification tasks tap into different subsystems of L2 phonology. This is also shown in the listeners' potential decision bias against the true affricate categories in the /u/ context, see Table 2.

Finally, the four online-processing metrics in mouse-tracking showed a consistent vowel effect, while the VOT effect was non-significant. The findings therefore confirmed our prediction that Mandarin listeners rely on the labial cue, but not the lingual cue, in perceiving English /tɹ, dɹ/ categories. More broadly, this finding indicates that L2 segment perception is sensitive to its immediate phonological context, e.g., the /ɹ/ segment in /tɹ, dɹ/ is likely to be substituted as a /w/ in the /i/ context, while it can be regarded as perceptually 'deleted' in the /u/ context.

To further investigate the nuanced effect of VOT, future research could recruit a group of L2 listeners who are relatively inexperienced with the target language to determine whether and the extent to which English /dʒu/-/dɹu/ can cause perceptual confusion at the beginning of L2 learning. Nonetheless, the mouse-tracking technique provides a wide range of additional metrics for understanding the cognitive processes during decision-making [11]–[14]. We argue that mouse-tracking can complement keystroke paradigms (e.g., AXB/AX tasks or identification by key-pressing) by offering additional perspectives into the online processing of L2 speech input, e.g., the change of mind *en route* as indicated by the curved mouse trajectories. A future study should analyse the types of mouse trajectories and their distribution in different experimental conditions.

5. ACKNOWLEDGEMENT

We would like to thank the twenty participants and Alexander Kilpatrick for his help in generating the experiment stimuli.

5. REFERENCES

[1] A. Sheldon and W. Strange, "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception," *Appl. Psycholinguist.*, vol. 3, no. 3, pp. 243–261, 1982.

[2] C. T. Best, "A direct realist view of cross-language speech perception," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[3] C. T. Best and P. A. Hallé, "Perception of initial obstruent voicing is influenced by gestural organization," *J. Phon.*, vol. 38, no. 1, pp. 109–126, 2010.

[4] P. A. Hallé and C. T. Best, "Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/ clusters," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 2899–2914, 2007.

[5] A. Kilpatrick, R. L. Bundgaard-Nielsen, and B. J. Baker, "Japanese co-occurrence restrictions influence second language perception," *Appl. Psycholinguist.*, vol. 40, no. 2, pp. 585–611, 2019.

[6] Y. Lan, "Perception of English fricatives and affricates by advanced Chinese learners of English," in *Proceedings of INTERSPEECH 2020*, H. Meng, B. Xu, and T. Zheng, Eds. Shanghai, China: International Speech Communication Association, 2020, pp. 4467–4470.

[7] Y. Lan, "Vowel effects on L2 perception of English consonants by advanced learners of English," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, M. Le Nguyen, M. C. Luong, and S. Song, Eds. Hanoi, Vietnam: University of Science, Vietnam National University, 2020, pp. 149–157.

[8] L. Magloughlin, */tɹ/ and /dɹ/ in North American English: phonologization of a coarticulatory effect*. Unpublished doctoral thesis: University of Ottawa, 2018.

[9] S. Duanmu, *The phonology of Standard Chinese*. Oxford, UK: Oxford University Press, 2007.

[10] W. Strange and V. L. Shafer, "Speech perception in second language learners," in *Phonology and second language acquisition*, J. G. Hansen-Edwards and M. L. Zampini, Eds. Amsterdam: Benjamins, 2008, pp. 153–192.

[11] M. J. Spivey, M. Grosjean, and G. Knoblich, "Continuous attraction toward phonological competitors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 29, pp. 10393–10398, 2005.

[12] P. E. Stillman, X. Shen, and M. J. Ferguson, "How mouse-tracking can advance social cognitive theory," *Trends Cogn. Sci.*, vol. 22, no. 6, pp. 531–543, 2018.

[13] D. U. Wulff *et al.*, "Movement tracking of cognitive processes: A tutorial using mousetrap," *PsyArXiv*, 2021.

[14] Y. Wang, R. L. Bundgaard-Nielsen, B. J. Baker, and O. Maxwell, "Native phonotactic interference in L2 vowel processing: Mouse-tracking reveals cognitive conflicts during identification," in *Proceedings of INTERSPEECH 2022*, H. Ko and J. H. L. Hansen, Eds. International Speech Communication Association, 2022, pp. 5223–5227.

[15] P. Nation and D. Beglar, "A vocabulary size test," *Lang. Teach.*, vol. 31, no. 7, pp. 9–13, 2007.

[16] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behav. Res. Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[17] G. Stoet, "PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teach. Psychol.*, vol. 44, no. 1, pp. 24–31, 2017.

[18] J. W. Peirce, "PsychoPy: Psychophysics software in Python," *J. Neurosci. Methods*, vol. 162, no. 1–2, pp. 8–13, 2007.

[19] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, pp. 35–44, 1985.

[20] J. Gong, M. Cooke, and M. L. Garcia-Lecumberri, "Towards a quantitative model of mandarin Chinese perception of English consonants," in *Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech New Sounds 2010*, Poznań, Poland, 2010.

[21] A. Cutler, "Representation of second language phonology," *Appl. Psycholinguist.*, vol. 36, no. 1, pp. 115–128, 2015.

[22] P. Cisek and J. F. Kalaska, "Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action," *Neuron*, vol. 45, no. 3, pp. 801–814, 2005.