Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

jTRACE modeling of L2 Mandarin learners' spoken word recognition at two time points in learning

Adam A. Bramlett[1], *Seth Wiener*[2]

[1]University of Kansas, United States
[2]Carnegie Mellon University, United States
adamab@hawaii.edu, sethw1@cmu.edu

Abstract

This study used the TRACE model of spoken word recognition to simulate adult second language (L2) learners' spoken word recognition at two time points in learning. The pre-existing architecture of jTRACE with the TRACE-T phonology was used to simulate spoken Mandarin word recognition by adult L2 learners at week 1 and week 15 of structured elementary classroom learning. A modified lexicon with reduced tonal information was used to capture recognition during week 1 of learning. Partially restored tonal information was used to capture the change observed at week 15. jTRACE simulations were validated by comparing the results to eye fixation data taken at week 1 and week 15. The eye-tracking task consisted of viewing four Mandarin words written in *pinyin* while one of the words was presented auditorily. Roughly half of the trials contained words that were segmentally and tonally contrastive (e.g., *gān, chá, pǐ, xiàn*). The remaining trials contained a target and competitor that were segmentally identical but tonally contrastive (e.g., *mā, má, pěn, gòng*). Proportion of looks to the target were calculated and compared to the jTRACE simulations using multiple linear regression. The results showed evidence of activation and recognition, thereby corroborating our modeling approach.

Index Terms: computational modeling, spoken word recognition, Mandarin Chinese, speech perception, second language acquisition, TRACE

1. Introduction

Mandarin Chinese (hereafter 'Mandarin') is a tonal language. That is, listeners of Mandarin use information from pitch pattern variations to distinguish between different word meanings. Whereas all languages use pitch variations to some degree (e.g., intonation and stress) [1], Mandarin uses both segmental and suprasegmental information in order to convey the meanings of words. To recognize a spoken word, a listener must therefore process strings of consonants and vowels as well as the pitch pattern or tone that the segmental string carries [2]. Standard Mandarin has four tones. Tone 1 is a high-level tone (e.g., *fū*). Tone 2 is a rising tone that starts low and ends high (e.g., *fú*). Tone 3 is a low dipping tone that falls first and rises toward the end (e.g., *fǔ*). Tone 4 is a falling tone, which starts high and falls to low (e.g., *fù*).

To fully understand the process of spoken word recognition across languages, it is necessary to understand how speakers use both segmental and suprasegmental information. Yet, traditional approaches to spoken word recognition initially ignored suprasegmental information (e.g., [3, 4]). One of the most influential models of spoken word recognition is the TRACE model [4]. TRACE is a connectionist model of spoken word recognition, which has been used to successfully model speech perception and spoken word recognition for a number of languages [5]. TRACE uses three independent tiers of information. The first tier involves feature information of the particular phonemes (e.g., consonantal, vocalic). The second tier is the phoneme tier, which consists of all of the phonemes for the language under investigation. The phonemes are then combined as words to create the third tier, which makes up the lexicon. As a connectionist network, TRACE involves three types of connectivity. The first is feedforward in which features connect to phonemes and phonemes connect to words. For example, the consonantal feature feeds forward or excites consonant phonemes like /n/ which in turn feeds forward or excites words containing /n/. The second type of connectivity involves lateral or within tier inhibitory connections. For example, as consonantal feature information is presented and fed forward, other feature information like vocalic information is inhibited. Similarly at the phoneme and word tiers, as /n/ and words containing /n/ are excited, non-nasals and words without nasals are inhibited. The third type of connectivity involves top-down feedback in which lexical information excites connections between words and phonemes. For example, ambiguous input in which the word could be either 'nak' or 'nag' would active the latter because it is an English word whereas the former is not. This activation would excite the phonemes /n/, /æ/, /g/ thereby giving lexical feedback for the ambiguous phoneme /g/ over /k/, i.e., the Ganong effect [6].

Crucial to the present study, TRACE did not originally account for suprasegmental information. [7] first put forth a theoretical solution for modeling suprasegmental information such as tone: a fourth tier of representation for tones (tonemes), which would be processed in parallel with the phoneme tier. The tone tier would have tonal information for each tonal contrast similar to feature level contrasts. Words that are segmentally contrastive like *fù* and *mā* would be distinguished through excitement in the feature tier. However, words that are segmentally identical and tonally contrastive like *fù* and *fǔ* would be distinguished through excitement in the tone tier. [8] used the visual world paradigm to study the activation (and competition) of tonal words. First language (L1) Mandarin speakers' eyes were recorded as they heard a spoken word. Onscreen participants saw images of words that were segmentally identical but tonally contrastive (e.g., *chūang* 'window' *chúang* 'bed'). The authors found evidence that segmental information and tonal information were accessed simultaneously, in line with tone and feature information proposed for TRACE.

More recently, [9] created a computational model of Mandarin spoken word recognition by building TRACE-T. TRACE-T implements a modified jTRACE architecture to

10.21437/SpeechProsody.2022-157

allow for the coding of both tones and segments. TRACE-T splits the feature tier into two groups: 1) segment features, and 2) tonal features. The model provides a foundation for not only modeling Mandarin spoken word recognition but also any other tonal language. In the current study, we extend the framework of the TRACE model and TRACE-T phonology to model L2 Mandarin spoken word recognition. We implement our model in jTRACE and compare our results to preliminary eye-tracking data from adult L2 learners. By reducing access to tonal information in the lexicon, we are able to build simulations that are comparable to real L2 word recognition. Unique to our approach, we modeled L2 learning at two time points roughly 15 weeks apart. We created models of the first week of learning by reducing tonal information in the lexicon and the input words. Further, we modeled the fifteenth-week eye-tracking data by providing partially restored tonal information.

2. Methods

2.1. jTRACE modeling

To model L2 Mandarin word recognition, we implemented the TRACE-T phonology [9]. That is, instead of using the seven features for each phoneme which are used for non-tonal languages (consonantal, vocalic, diffuseness, acuteness, voicing, power, and burst amplitude), the features of the Mandarin phonology were split into features for segments and features for tones. Our approach followed the TRACE-T method created from the PatPho dimensions of Mandarin to use four dimensions of jTRACE for consonants and vowels [10]. Because our goal was to model Mandarin L2 learners' word recognition, we chose the Mandarin phonology over an English phonology. Whereas the features of Mandarin and English are different in many ways, the majority of our test items (see eye-tracking task materials) involved segmental contrasts that have English counterparts (e.g., *fu, ma, ba*). Most importantly, the Mandarin phonology has a place for tonal distinction while an English phonology does not have a place to model these changes for either the beginning state or learning over time.

Syllables were represented with a TRACE-T inspired structure, i.e., segment-tone-segment-tone-segment-tone-segment-tone. All words used the same number of segments and tones to ensure that tonal information remained constant within condition. The syllable structure of all Mandarin words used in the eye-tracking experiment were separated based on initial and final. The initials and finals of all Mandarin words in the experiment were then coded from *pinyin* to IPA using a self-made key and value function built in R, which maps *pinyin* initials to IPA symbols. For most syllables in Mandarin, the coding from *pinyin* to Mandarin is a one-to-one match. However, for words that have *pinyin* vowels that contain "i", "u", "y", or "w", individual changes were made. These changes were necessary due to the overlapping sound and symbol representation in *pinyin*. Next, all IPA symbols were mapped to the Mandarin phonology symbols used in TRACE-T [9].

Because recent studies on tone learning and perception have suggested that naïve tone learners rely on F0 less than L1 speakers of a tone languages during lexical access (e.g., [11]), we decided to build two separate lexicons for week 1 and week 15 L2 learners. The words in the week-1 learner lexicon contained 20% of tonal information that an L1 lexicon would have; the words in the week-15 learner lexicon had 40% of the tonal information compared to an L1 lexicon. We note that the words represented are identical across the two time points. Only the representations of the words in jTRACE have been manipulated. Importantly, the pre-existing architecture of jTRACE does not allow for variant levels of tonal information throughout the word. TRACE-T implements tonal information throughout the word in five different places with features depending on the height and contour of the desired tone. Because of the linear nature of processing tonal features in TRACE-T, we chose to have the reduced tonal information starting in the first position for both week 1 and week 15 simulations. Our approach was driven, in part, by recent studies that suggest that segmental and tonal information is accessed simultaneously [8]. For these reasons, the week-1 learner lexicon starts with one instance of tonal information in the first position. To model learning after 15 weeks, we added partially restored tonal information in the second tonal slot. All later tonal slots were filled with a featureless phoneme represented as "-" that would neither result in activation nor inhibition in the model.

Lastly, for both models, we used the Luce choice rule to convert activations to response probabilities. All simulations were iterated for a minimum of 80 cycles and reduced later in data analysis following [9]. Furthermore, the lexicon is controlled by only comparing the response probabilities of the four words in a particular item shown on screen. That is, the model ran one set of words at a time for a total of 120 simulations for each time point. This choice makes the initial probability more comparable to eye-tracking data by reducing competition from other words not visually shown. Whereas the assumption of equal frequency across lexical items is not true for L1 listeners [4], our L2 learners had extremely limited Mandarin lexicons, particularly at week 1. We, therefore, controlled token frequency by setting frequency to 100 across the entire learner lexicon. This assumption allows jTRACE to treat each set of words as new words. That is, like the L2 learners, the model does not have a fully developed lexicon that would help inform the recognition of a word via top-down feedback.

2.2. Eye-tracking task

2.2.1. Participants

Fifteen L1 English-L2 Mandarin participants were tested. All participants were adults enrolled in a US university first-semester Mandarin language class at the time of testing. No participants were heritage speakers and all participants passed a pitch perception test which required the ability to reliably discriminate two pure tones at 20 Hz or lower.

2.2.2. Materials

Stimuli included 72 segmentally and tonally contrastive items (e.g., *gān, chá, pǐ, xiàn*) and 48 segmentally identical but tonally contrastive (e.g., *mā, má, pěn, gòng*) items using a design modeled on [12]. Figure 1 shows an example of a tonally contrastive item. A target (*fú,* top right) appeared with a competitor that only differed in tone (*fǔ,* top left). Two distractors were also shown (*zāng, sà,* bottom left and right). Location of targets, competitors, and distractors were counterbalanced across the experiment. Each of the 120 targets was recorded by a female L1 Mandarin speaker from Beijing and saved at 16 bits/44,100 Hz using Praat [13] with a normalized amplitude of 70 db.

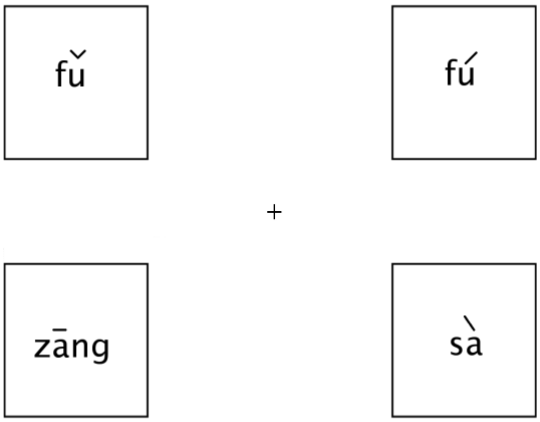

Figure 1: *Example eye-tracking slide containing segmentally identical but tonally contrastive target and competitor.*

2.2.3. Procedure

Participants performed the eye-tracking task twice, once during the first week of a structured, in-person, elementary Mandarin Chinese as a foreign language class, and again during the fifteenth week (i.e., the end of the university semester). The class met three times per week for a total of 210 minutes of class time each week. Participants were tested in a quiet lab and first had their eyes calibrated using a 12-point system. They performed 120 trials in a pseudo-randomized order. On each trial, a fixation cross appeared for 1 second followed by the simultaneous presentation of the 2×2 slide and the target audio. Participants' eyes were continuously recorded at 30 Hz using the Eye-tribe system. A 2-second inter-trial interval preceded each trial, with the eye tracker recalibrated every eighth trial. Participants were told to mouse-click on the word that matched the perceived audio as quickly and accurately as possible. The entire lab visit lasted approximately 20 minutes. Participants were given class credit or a small payment for their time.

2.3. Multiple Linear Regression

To compare the jTRACE data with the L2 eye-tracking data, we first removed trials in which participants incorrectly mouse-clicked on a displayed competitor or distractor word (approximately 2% of all data). The remaining correct trials were coded for fixation to the target, competitor, or distractors at each time point and averaged across subjects and items for the two segment-tone conditions. We then analyzed the fixation proportions from 200 ms post onset to 1,100 ms (i.e., 27 time points) in line with previous visual world paradigm studies given that approximately 200 ms is needed to program a saccade (e.g., [14]). For the jTRACE simulations, we chose to analyze cycles 16-42 in accordance with the 27 time points of eye-tracking data. Multiple linear regression was carried out in R [15] following [16]. The regression model's dependent variable was fixation proportions (or simulated activation probabilities) to the target. The independent variables were data type (eye-tracking/jTRACE), time (week 1/week 15), and segment-tone condition (segmentally and tonally contrastive/ segmentally identical but tonally contrastive). All three variables were treatment coded and tested for two-way and three-way interactions: *lm(fixation-proportaion ~ data type*time*condition)*.

3. Results

The multiple linear regression model yielded a null effect of data type (β = 0.03, SE = .03, t = 1.14, p = .25) indicating that there was no difference between our simulated fixations (activation probabilities) and eye-tracking fixation proportions. Similarly, there was a null effect of time (β = -0.01, SE = .03, t = -0.16, p = .87) indicating that despite the intervening 15 weeks of structured learning, no difference was found between week 1 and week 15 fixation proportions. There was also a null effect of segment-tone condition (β = 0.03, SE = .03, t = 0.95, p = .35) indicating that fixations to segmentally and tonally contrastive targets were no different than those to segmentally identical but tonally contrastive targets. In other words, the presence of a tonal competitor on screen did not significantly affect the participants' fixation proportions across time. All interactions were null at an alpha of .05. Figure 2 summarizes the results by plotting the averaged data (points) with the regression lines.

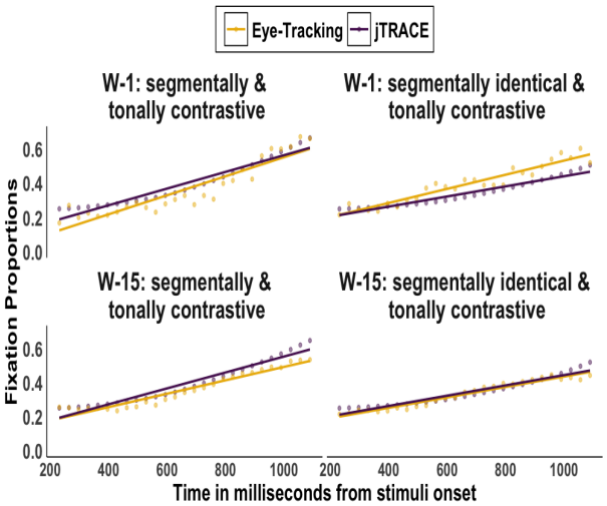

Figure 2: *Eye-tracking fixation proportions to targets over time and jTRACE simulations at week 1 and week 15 (points) with linear regression model fits (lines).*

4. Discussion and conclusion

In this study, we extended previous Mandarin spoken word recognition research beyond the typical L1 group of participants (e.g., [7, 8]) in order to examine adult classroom L2 learners. This served as a proof-of-concept study, in which we augmented previous jTRACE models with the TRACE-T phonology [9, 10] in order to create simulations of spoken word recognition at two points in time during structured L2 acquisition. In line with previous studies that have suggested that Mandarin L2 learners weight F0 movement cues less than native speakers [17, 18], we initially reduced tonal information in our week 1 simulation and then used partially restored tonal information in our week 15 simulation. We compared these simulations to eye-tracking data obtained from L2 learners engaged in classroom learning and found that our jTRACE simulations did not statistically differ from the fixation proportions of real eye-tracking data over time, thus corroborating our modeling approach.

One unexpected finding from our results, however, was that we did not find a significant difference between our data at week 1 and week 15. In other words, we did not observe a

measurable effect of learning. There are many potential explanations for why this may have been the case. We suspect the large individual variability observed at week 1, the small sample size and limited number of data points tested, i.e., low power, and our linear regression modeling approach (cf. growth curve analysis, which may better capture the rise) all contributed to the null effect. It also seems likely that a 15-week, one semester class may not have been enough input and learning to dramatically shift perception and spoken word recognition in our L2 learners [20]. This is in line with a large body of L2 Mandarin acquisition research that has demonstrated learners reach a tone learning plateau in which their abilities do not improve [18] and that any improvement is largely driven by categorical perception of tone [19].

We note that in our modeling approach, tonal information for both week 1 and 15 started at the same time, i.e., the first position. At week 15, we added tonal information at the second position to represent learning. This approach assumes that L2 listeners, like L1 listeners, are able to simultaneously integrate segmental and tonal information as it unfolds in time. Indeed, evidence does suggest that intermediate and advanced L2 listeners are able to do so in an L1-like manner [19]. Whether our beginner L2 listeners were able to do so remains an open question.

In our current follow-up studies, we are improving on our approach in three ways: first, we are exploring a simulation approach which reduces tonal information in the input by increasing stochastic noise in the segments and tones. Because we assumed reduced tonal information for our L2 learners, adding a greater amount of stochastic noise in the jTRACE model may allow for reduced tonal information and increased segment information for early learners compared to simply reducing overall tonal information by having less tonal segments in line with [4] and [8]. Second, we aim to clarify how our jTRACE modeling and eye-fixation data align for L1 Mandarin listeners. We assume our results will be comparable, but we are in the process of exploring to what degree the L1 and L2 groups are similar and different. Third, we are in the process of modeling the tone competitor fixations in the segmentally identical but tonally contrastive condition. Given that the tonal competitors rise and fall were not linear, growth curve analysis or a similar approach may better fit the data. This change may allow us to more fully capture the learning that occurred between week 1 and 15 and eventually model this change in behavior. Different manners of reduction in the lexicon, varying strength of connections, or the addition of stochastic noise may be found to be a superior modeling technique in future work interested in modeling L2 word recognition.

In conclusion, we successfully adapted jTRACE to model adult L2 Mandarin learners' spoken word recognition at week 1 and week 15 of structured classroom learning. Our simulations matched eye-tracking data taken at the start and end of learning and showed that by adjusting tonal information in our TRACE-T phonology, we were able to capture the fixation proportions from our eye-tracking data.

5. Acknowledgements

We thank the participants in the eye-tracking study, the Linguistics Department of the University of Kansas (Dean's Doctoral Fellowship) for supporting the first author, and the undergraduate assistants in the Language Acquisition, Processing, and Pedagogy Lab at Carnegie Mellon University for assisting with data collection.

6. References

[1] C. Gussenhoven, *The Phonology of Tone and Intonation.* Cambridge, UK: Cambridge University Press, 2004.

[2] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkley, CA: University of California Press, 1968.

[3] W. D. Marslen-Wilson and A. Welsh, "Processing interactions and lexical access during word recognition in continuous speech," *Cognitive Psychology*, vol. 10, pp. 29–63, 1978.

[4] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, pp. 1–86, 1986.

[5] T. J. Strauss, J. S. Magnuson, and H. D. Harris, "jTRACE: a reimplementation and extension of the TRACE model of spoken word recognition," *Behavioral Research Methods*, vol. 39, no. 1, pp. 19–30, 2007.

[6] W. F. Ganong, "Phonetic categorization in auditory perception," *Journal of Experimental Psychology*: *Human Perception and Performance*, vol. 6, pp. 110–125, 1980.

[7] Y. Ye and C. M. Connine, "Processing spoken Chinese: the role of tone information," *Language and Cognitive Processes*, vol. 14, no. 5–6, pp. 609–630, 1999.

[8] J. G. Malins and M. F. Joanisse, "The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study," *Journal of Memory and Language*, vol. 62, no. 4, pp. 407–420, 2010.

[9] L. Shuai and J. G. Malins, "Encoding lexical tones in jTRACE: a simulation of monosyllabic spoken word recognition in Mandarin Chinese," *Behavior Research Methods*, vol. 49, no. 1, pp. 230–241, 2017.

[10] X. Zhao and P. Li, "An online database of phonological representations for Mandarin Chinese," *Behavior Research Methods*, vol. 41, no. 2, pp. 575–583, 2009.

[11] K. Connell, A. Tremblay, and J. Zhang, "The timing of acoustic vs. perceptual availability of segmental and suprasegmental information," In Proc. 5th International Symposium on Tonal Aspects of Languages, 2016, pp. 24–27.

[12] J. A. Shaw and M. D. Tyler, "Effects of vowel coproduction on the timecourse of tone recognition," *The Journal of the Acoustical Society of America*, vol. 147, no. 4, pp. 2511-2524, 2020.

[13] P. Boersma and D. Weenick, *Praat: doing phonetics by computer* [Computer program]. Version 6.2.01, 2016.

[14] P. D. Allopenna, J. S. Magnuson, and M. K. Tanenhaus, "Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping model," *Journal of Memory and Language*, vol. 38, no. 4, pp. 419–439, 1998.

[15] R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, 2017.

[16] J. M. Chambers, "Linear models," in *Statistical Models*, J. M. Chambers and T. J. Hastie, Eds. Boca Raton: CRC press, 1992, pp- 95-144.

[17] B. Chandrasekaran, P. D. Sampath, and P. C. Wong, "Individual variability in cue-weighting and lexical tone learning," *The Journal of the Acoustical Society of America*, vol. *128, no.* 1, pp. 456-465, 2010.

[18] S. Wiener, "Changes in early L2 cue-weighting of non-native speech: evidence from learners of Mandarin Chinese," In Proc. INTERSPEECH, 2017, pp. 1765-1769.

[19] W. Ling and T. Grüter, "From sounds to words: the relation between phonological and lexical processing of tone in L2 Mandarin," *Second Language Research*, 2020.

[20] S. Wiener, and E. D. Bradley. "Harnessing the musician advantage: Short-term musical training affects non-native cue weighting of linguistic pitch," *Language Teaching Research*, 2020.