

Perceptual consistency of native speech predicts non-native cross-situational word learning

Yuk Sai Nam¹, Quentin Zhen Qin², Yin-To Chui²

¹*Xiamen Institute of Technology*, ²*The Hong Kong University of Science and Technology*

Cross-situational word learning (CSWL) is a powerful mechanism for resolving referential ambiguity [1], yet it is constrained by a phonological bottleneck in second language (L2) acquisition [2], where learners need to distinguish novel phonological forms in addition to statistical tracking. At the same time, there is much individual variability around L2 CSWL success. While previous research has often attributed individual differences in CSWL to broad cognitive traits like working memory [3] or phonological awareness [4], the broader literature suggests that the fine-grained precision of the learner's native (L1) speech processing is a more direct predictor of L2 success. For instance, it was found that the trial-to-trial consistency of native categorization – a measure of internal perceptual stability – predicts the ability to discriminate difficult non-native contrasts [5], as well as higher-level reading ability [6]. However, it remains unknown whether these fine-grained properties of native perceptual stability extend to predict the success of acquiring a new lexicon under referential ambiguity.

To test this, we examined 57 low-proficiency Mandarin-speaking learners of English (Mean age = 19.6; LexTALE score = 51.0). First, to quantify individual differences in native speech perception, participants completed a Visual Analogue Scaling (VAS) task [7] on Mandarin consonant (/pa/-/ta/) and vowel (/ti/-/tu/) continua, created using the WORLD vocoder to ensure high acoustic quality [8]. For each trial (105 per continuum), participants rated the stimulus on a discrete scale of 1-7 based on its similarity to the endpoints. Second, participants completed a five-block CSWL task (280 trials total) utilizing English pseudowords (**Figure 1**), identifying referents in a 2-alternative-forced choice (2AFC) task. Mappings for the target and the distractor created three conditions: consonant minimal pairs (**cMP**; e.g., *bon* vs. *pon*, /b/ vs. /t/), and vowel minimal pairs (**vMP**; e.g., *deet* vs. *dit*, /i/ vs. /ɪ/), and non-minimal pairs (**nonMP**). The trial structures for the VAS task and CSWL task are detailed in **Figure 2**.

For the VAS data, we fitted a logistic curve to each learner's response function and extracted two metrics of perceptual style: categoricity (the slope at the inflection point of the fitted curve) and consistency (negative mean of squared model residuals) [9]. **Figure 3 (Left)** visualizes individual variability in the response patterns, ranging from categorical to more gradient functions. For the CSWL task, generalized linear mixed-effects models revealed a significant effect of Block, confirming overall learning success. A significant effect of Condition confirms a phonological bottleneck: accuracy was highest for nonMPs, significantly lower for cMPs ($\beta = -0.42$, $SE = 0.09$, $p < .001$), and lowest for vMPs ($\beta = -0.78$, $SE = 0.09$, $p < .001$) (**Figure 3; Middle**). Regarding individual differences, a significant Block $\times$ Vowel Consistency interaction ($\beta = 0.17$, $SE = 0.07$, $p < .02$) indicated that participants with higher consistency learned at a faster rate (**Figure 3; Right**). There was no further 3-way interaction with Condition, indicating that the advantage of higher vowel consistency led to comparable benefits to the CSWL of all L2 contrasts (i.e., nonMP, cMP, and vMP alike).

These findings suggest that the ability to resolve referential ambiguity in an L2 is constrained by the internal stability of the learner's native perceptual system. Furthermore, the specific predictive power of vowels aligns with [6]. It is possible that vowel categorization – which requires the integration of complex spectral cues – serves as a more sensitive marker of global processing fidelity than stop consonants. We propose that a "noisier" L1 system (low consistency) degrades the phonological representations of novel words, introducing errors into the statistical tracking process. This indicates that L2 learning models should look beyond structural inventory mappings (e.g., whether L1 and L2 categories overlap) to include a processing dimension: the stability with which learners encode speech.



Fig. 1. The 8 word-referent mappings. Words in the top row form consonant minimal pairs, and words in the bottom row form vowel minimal pairs. Pairing any word from the top row with any word from the bottom row forms a non-minimal pair.

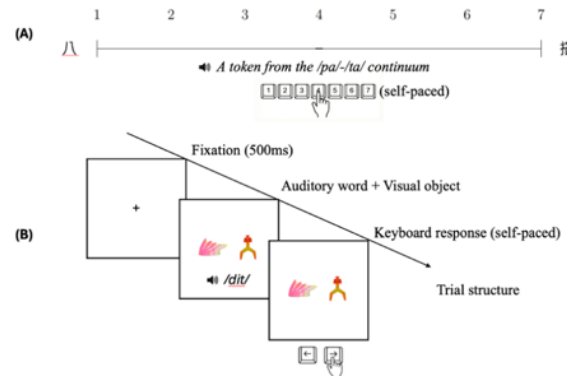


Fig. 2. Schematic representation of typical trials in the (A) VAS task and (B) CSWL task.

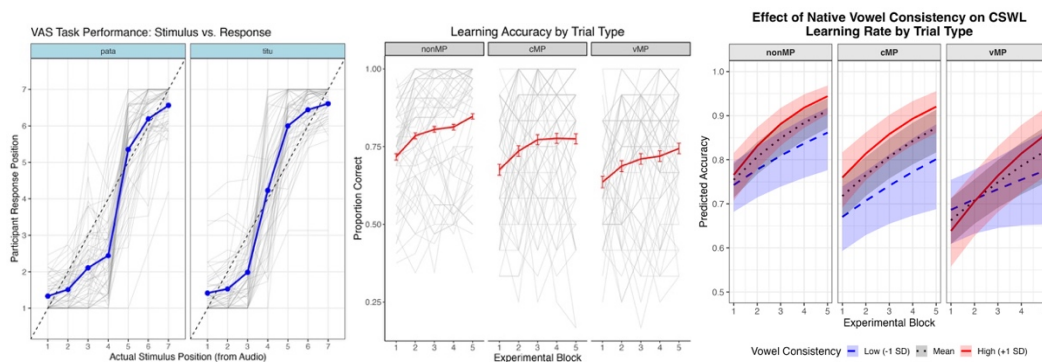


Fig. 3. (Left) The response patterns in the VAS task (blue line = grand mean of all participants). (Middle) The response patterns in the CSWL task (red line = grand mean of all participants). (Right) Model-predicted plots showing the effect of native vowel consistency on CSWL learning rate by Trial Type

References

- [1] Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5), 414–420.
- [2] Escudero, P., Smit, E. A., & Mulak, K. E. (2022). Explaining L2 lexical learning in multiple scenarios: cross-situational word learning in L1 Mandarin L2 English speakers. *Brain Sciences*, 12(12), 1618.
- [3] Li, Y., & Benitez, V. L. (2025). Probing the role of multilingualism and working memory in cross-situational word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 51(5), 804–824.
- [4] Hu, C.-F. (2017). Resolving referential ambiguity across ambiguous situations in young foreign language learners. *Applied Psycholinguistics*, 38(3), 633–656.
- [5] Fuhrmeister, P., Phillips, M. C., McCoach, D. B., & Myers, E. B. (2023). Relationships between native and non-native speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(7), 1161–1175.
- [6] Kim, H., Klein-Packard, J., Sorensen, E., Oleson, J., Tomblin, B., & McMurray, B. (2025). Speech categorization consistency is associated with language and reading abilities in school-age children: Implications for language and reading disorders.
- [7] Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, 59, 40–57.
- [8] Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7), 1877–1884.
- [9] Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, 43(9), 1594–1611.