Sensitivity to semantic context and acoustic-phonetic cues plays a crucial role in spoken word recognition. Previous research (Miller et al., 2010; Benichov et al., 2012) has shown that semantics and dynamic fundamental frequency (F0) enhance comprehension of a native language (L1), particularly in degraded listening environments such as background noise. However, non-native (L2) listeners may not benefit from semantic cues to the same extent (Bradlow and Alexander, 2007; Zhang et al., 2016). This study investigates how semantics, F0, and background noise interact to affect spoken word recognition among L1 listeners, experienced and inexperienced L2 learners. In addition, the study examines whether variability in speech recognition can be explained by individual differences in cognitive ability, auditory acuity, and fine-grain language history.

This study was preregistered, and all materials and associated analysis scripts are available on the Open Science Framework. L1 Mandarin listeners (n = 60), experienced (n = 63), and inexperienced L2 Mandarin learners (n = 62) completed an auditory transcription task that manipulated semantic context (meaningful vs. nonsensical sentences), F0 (natural vs. flattened), and background noise (quiet vs. noise). Additionally, participants completed a battery of secondary tasks to measure memory digit span, pitch pattern sensitivity, and familiarity with items presented during the transcription task. Participants also completed a language history questionnaire assessing language proficiency, immersion, dominance, dominance ratio, and multilingual language diversity.

Results revealed clear main effects: L1 listeners demonstrated the highest accuracy, followed by experienced L2 learners, who outperformed inexperienced L2 learners. In addition, accuracy was higher for meaningful compared to nonsensical sentences and accuracy was higher for natural F0 compared to flat F0. However, no significant difference was found between noise and quiet conditions. As for interactions, L1 listeners were not affected by noise, but did show an decrease in accuracy when F0 was flattened. In contrast, L2 learners were not affected by degraded bottom-up cues; that is, accuracy was equivalent in quiet, noise, and flattened F0 conditions. Furthermore, L1 listeners showed a facilitative effect of semantics on recognition when bottom-up acoustic cues were degraded (i.e., when F0 was flattened or background noise was present) but not in ideal listening conditions. In contrast, both learner groups showed a facilitative effect of semantic context in both ideal and adverse listening environments.

Collectively, these results indicate that L1 listeners use both bottom-up F0 cues and top-down semantic information in Mandarin speech recognition, whereas L2 learners primarily rely on semantic cues in the absence of bottom-up F0 use. Unlike L1 listeners, whose use of semantics is modulated by listening conditions and is rarely observed under ideal conditions, L2 learners do not show this difference and instead rely on semantics consistently across both ideal and suboptimal listening environments. As for individual differences, speech recognition is significantly correlated with memory capacity, pitch pattern sensitivity, and all four scores in the language history questionnaire, excluding multilingual language diversity score. The results shed light on theories of L2 speech recognition and have implications for the pedagogical practices of tonal languages such as Mandarin.