Shea Special Issue

Language *and* Speech

Early L2 Spoken Word Recognition Combines Input-Based and Knowledge-Based Processing

Language *and* Speech
1–25

DOI: 10.1177/0023830918761762
journals.sagepub.com/home/las
SAGE

Seth Wiener
Department of Modern Languages, Carnegie Mellon University, USA

Kiwako Ito
Department of Linguistics, The Ohio State University, USA

Shari R Speer
Department of Linguistics, The Ohio State University, USA

Abstract

This study examines the perceptual trade-off between knowledge of a language's statistical regularities and reliance on the acoustic signal during L2 spoken word recognition. We test how early learners track and make use of segmental and suprasegmental cues and their relative frequencies during non-native word recognition. English learners of Mandarin were taught an artificial tonal language in which a tone's informativeness for word identification varied according to neighborhood density. The stimuli mimicked Mandarin's uneven distribution of syllable+tone combinations by varying syllable frequency and the probability of particular tones co-occurring with a particular syllable. Use of statistical regularities was measured by four-alternative forced-choice judgments and by eye fixations to target and competitor symbols. Half of the participants were trained on one speaker, that is, low speaker variability while the other half were trained on four speakers. After four days of learning, the results confirmed that tones are processed according to their informativeness. Eye movements to the newly learned symbols demonstrated that L2 learners use tonal probabilities at an early stage of word recognition, regardless of speaker variability. The amount of variability in the signal, however, influenced the time course of recovery from incorrect anticipatory looks: participants exposed to low speaker variability recovered from incorrect probability-based predictions of tone more rapidly than participants exposed to greater variability. These results motivate two conclusions: early L2 learners track the distribution of segmental and suprasegmental co-occurrences and make predictions accordingly during spoken word recognition; and when the acoustic input is more variable because of multi-speaker input, listeners rely more on their knowledge of tone-syllable co-occurrence frequency distributions and less on the incoming acoustic signal.

Corresponding author:
Seth Wiener, Department of Modern Languages, Carnegie Mellon University, 160 Baker Hall, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.
Email: sethw1@cmu.edu

Keywords
Spoken word recognition, speaker variability, lexical tone, Mandarin Chinese, second language acquisition

Introduction

Listeners recognize spoken words in their first language (L1) with remarkable efficiency. This is primarily accomplished by evaluating the acoustic similarity between the perceived input and stored representations (McQueen & Cutler, 2010; McQueen, Cutler, & Norris, 2006; Samuel, 2011). Because a word's phonetic form can vary across speakers and contexts, native listeners also achieve word recognition by assessing the probability of the signal-to-representation match between new acoustic input and stored exemplars, given the known frequency distribution of cues for a specific sound category or word (e.g., McMurray, Tanenhaus & Aslin, 2002; Norris & McQueen, 2008; Sulpizio & McQueen, 2012; Toscano & McMurray, 2010). A growing body of research suggests that this type of predictive, knowledge-based processing is one way in which listeners overcome acoustic variability and the lack of invariance in the signal (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2015; Kuperberg & Jaeger, 2016). L1 spoken word recognition therefore involves a perceptual trade-off between knowledge of a language's statistical regularities and reliance on the incoming acoustic signal. In this paper, we examine to what degree this perceptual trade-off also takes place in adult second language (L2) spoken word recognition.

Though the cognitive processes involved in recognizing words in an L2 are presumed to be fundamentally the same as those involved in L1 recognition (Weber & Cutler, 2004), listening to speech in one's L2 can be remarkably challenging. L2 input-based processing is often hindered by the interference of L1 phonetic and phonological properties (Best, 1995; Best & Tyler, 2007; Flege, 1995, 1999; So & Best, 2011; Strange, 1995), while L2 knowledge-based processing is limited by a lack of L2 exemplars and uncertainty of L2 frequency distributions (Escudero & Boersma, 2004; Hayes-Harb, 2007).

Previous research on L2 input-based processing has shown that variability in the acoustic signal affects non-native category learning (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lively, Logan, & Pisoni, 1993). For example, Japanese learners struggle to perceive and produce /r/ and /l/ English minimal pairs because Japanese has only one phonemic category for liquid consonants: the (post-)alveolar flap /r/ (Goto, 1971; Logan, Lively, & Pisoni, 1991; Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004). Japanese learners improve both their perception and production of /r/ and /l/ English minimal-pairs by undergoing an extended period of training on /r/-/l/ identification with multi-speaker, high variability stimuli (Bradlow et al., 1997; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). Corroborating evidence from a variety of perceptual learning studies supports the claim that training on stimuli from multiple speakers better facilitates L2 phonemic and word learning than training on low variability stimuli produced by a single speaker (e.g., Barcroft & Sommers, 2005; Flege, 1995; Lively et al., 1991; Sadakata & McQueen, 2013; Sommers & Barcroft, 2007; Wang, Spence, Jongman, & Sereno, 1999).

Acoustic variability, however, is not always beneficial to the listener. While non-contrastive phonetic variation (e.g., in indexical cues) can facilitate learning of contrastive dimensions (Rost & McMurray, 2010), within-category acoustic variability in contrastive dimensions can increase perceptual uncertainty (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008; Nixon, van Rij, Mok, Baayen & Chen, 2016) resulting in a change in cue weighting (Clayards et al., 2008; Holt & Lotto, 2006; Lim & Holt, 2011). For early L2 learners collecting exemplars of non-native sound

categories, acoustically varied input may increase perceptual ambiguity (e.g., Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994), which results in delayed or inaccurate word recognition (Ellis, 2011; Kroll, Michael, Tokowicz, & Dufour, 2002). Therefore, when it comes to learning to discriminate between different sounds in a particular acoustic dimension, between-category variability may help L2 learners form non-native categories, while within-category variability may hinder category formation and word learning.

One possible means to overcoming this increased perceptual uncertainty involves a greater reliance on knowledge-based processing. As L2 learners accrue more exemplars, they may begin to generalize about the likelihood of the input-representation match. Previous research into L2 knowledge-based processing has established that learners are sensitive to the relative repetition frequency of tokens containing non-native acoustic cues (e.g., Escudero, Benders, & Lipski, 2009; Escudero, Benders, & Wanrooij, 2011; Wanrooij, Escudero & Raijmakers, 2013). These distributional learning studies showed increased categorization accuracy when listeners heard greater acoustic distance between tokens instantiating two new speech sound categories in a bimodal distribution. L2 learners, therefore, track the frequency distributions of non-native cues and may increase their reliance on these statistics when faced with highly variable speech and greater perceptual uncertainty.

The challenge in studying the effect of acoustic variability in L2 acquisition, however, stems from the fact that speech recognition requires integrating and weighting various multi-dimensional cues in order to recognize phonemes, syllables, and words (Idemaru & Holt, 2011, 2014; Toscano & McMurray, 2012). For native speakers, the perceptual weighting of co-occurring cues occurs rapidly. Evidence comes from Idemaru and Holt (2011), in which voice onset time values that distinguish minimal pairs (e.g., *beer* vs. *pier*) were crossed with vowel-initial fundamental frequency (F0) values. When the correlation was reversed from the English norm (i.e., higher F0 for *beer* and lower F0 for *pier*), listeners down-weighted their reliance on F0 for voicing categorization after only a few trials of exposure. Idemaru and Holt refer to this as "dimension-based statistical learning," to emphasize that listeners not only track the statistical regularities of larger units like syllables, morphemes, and words (e.g., Saffran, Aslin & Newport, 1996), but also the co-occurring regularities among the dimensions that define these units.

It is unclear whether and how effectively dimension-based statistical learning of non-native cues occurs in L2 speech processing. The present study tests the hypothesis that L2 input with greater speaker variability leads to an increased reliance on the statistical learning of co-occurring cues. To test this hypothesis, learners of an artificial tonal language were trained on distributions of co-occurring segmental cues (i.e., Consonant-Vowel (CV)-syllables) and suprasegmental cues or F0 patterns (i.e., tones). Learners were trained with either single-talker input or multi-talker input across four consecutive days. If greater acoustic variability triggers more active statistical learning, learners who are trained on multi-talker input should show better identification performance based on the signal distribution than those who are trained on single-talker input.

1.1 Input-based and knowledge-based processing of Mandarin Chinese

We draw on empirical evidence from Beijing Mandarin Chinese (hereafter, "Mandarin") to motivate the present study. Mandarin uses four lexical tones, most commonly characterized by their F0 contours, to distinguish otherwise identical (C)V(C) syllables (Ho, 1976; Howie, 1976; Jongman, Wang, Moore, & Sereno, 2006). The F0 contours of the four Mandarin lexical tones can be summarized as: high-level (tone 1), rising (tone 2), low-dipping (tone 3), high-falling (tone 4) (Figure 1). As an example, the syllable /pan/ can mean "class" with a high-level tone (*ban1*) or "to handle" with a high-falling tone (*ban4*).

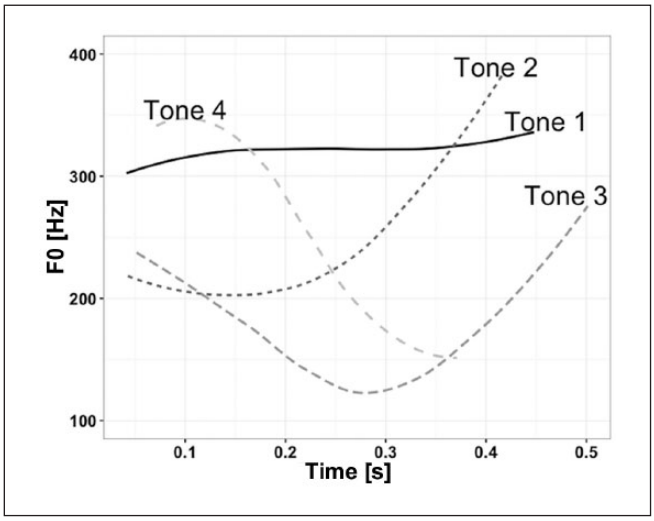

Figure 1. Four Mandarin tones produced in isolation by a native speaker.

While the phonetic features of tone are well defined, naturally produced tones vary given the context (Fox & Qi, 1990; Shen, 1990; Xu, 1994, 1997) and speaker (Leather, 1983; Lee, Tao & Bond, 2009, 2010), and often undergo coarticulation and phonologically induced variation through tone sandhi (Gottfried & Suiter, 1997; Shen & Lin, 1991; Xu, 2001). As a result, even native Mandarin listeners occasionally struggle to accurately categorize tones, for example, demonstrating confusion between tones 2 and 3 (Blicher, Diehl, & Cohen, 1990; Liu & Samuel, 2004, 2007).

Yet, tone recognition is not based on a syllable's perceived tune alone. Recent evidence (Wiener and Ito, 2015, 2016) indicates that native listeners make use of syllable+tone co-occurrence frequency information during the identification process. An often overlooked feature of Mandarin is that only a fifth of the roughly 400 (C)V(C) Mandarin syllables combine with all four tones to form words (Duanmu, 2009). Due to the historical evolution of Mandarin tones, the lexicon contains syllable+tone gaps (Duanmu, 2007; Wang, 1998). For example, the syllable *ban* does not co-occur with the rising second tone, that is, *ban2* is a nonword. Furthermore, syllable+tone combinations rarely have an unambiguous 1:1 mapping of sound to meaning (Tan & Perfetti, 1998). The syllable+tone combination *ban4* can be written with at least nine different orthographic forms (i.e., Chinese characters), all with semantically distinct meanings. The the high rate of homophony in Mandarin combined with non-word gaps results in a distribution of syllable and tone co-occurrences that speakers can track over time and exploit during recognition. Some syllables, like *shi* and *fu*, frequently occur in speech with all tones possible in the language, while other syllables, like *zhou* and *nuan*, occur less frequently overall, and then are much more likely to occur with one or two of the possible tones. As a result, for high frequency syllables with dense tonal neighborhoods, tone is less informative for morpheme identification; syllable+tone combinations like *shi4* and *fu2* result in nearly 30 unique homophonous morphemes, which require additional context for accurate lexical identification (Chen, Vaid, & Wu, 2009; Packard, 1999, 2000). In contrast, for low frequency syllables with sparse tonal neighborhoods, tone can be highly informative. This is the case for the low frequency syllable zhou, which has few homophones and occurs most frequently as zhou1. For some low frequency syllables, tone can be redundant, with acoustic-based processing

of consonant-vowel sequences alone being sufficient for lexical identification. For example, the syllable nuan must carry tone 3, as nuan1, nuan2 and nuan4 are all nonword gaps.

In an eye-tracking study (Wiener & Ito, 2015), mono-dialectal Mandarin speakers heard a target syllable+tone (e.g., *zhou1*) while simultaneously viewing four frequency-controlled characters. Of interest was the level of competition from the tonal-competitor, which shared the same syllable but differed in tone (e.g., *zhou2*). The competitor carried either the more or less probable tone for the target syllable (based on the spoken corpora, see Cai & Brysbaert, 2010). When hearing low frequency syllables with high probability tones, participants looked to the target fastest and mouse-clicked on its character fastest. When hearing low frequency syllables with low probability tones, participants looked to and clicked on the target slowest. This delay was due to participants' initial gazes to the character containing the more probable tone. These anticipatory looks to low frequency syllables with high probability tones began roughly 300 ms after the syllable onset and lasted until 700 ms, at which point listeners recovered from their incorrect predictions and looked to the intended (less probable) target. No effect of tonal probability on click reaction time or eye fixations was found for the high frequency syllables. Wiener and Ito (2015) therefore concluded that tone is less informative for high frequency syllables, which typically require additional context for lexical identification (e.g., Li, Shu, Yip, Zhang, & Tang, 2002; Li & Yip, 1998).

In a follow-up gating study (Wiener & Ito, 2016), native Mandarin speakers demonstrated knowledge-based processing for infrequent syllables even with minimal acoustic input and no visual cues. After hearing a syllable's onset and only 40 ms of the vowel, participants reported the most probable tone for the perceived infrequent syllable, even if the reported tone was acoustically dissimilar to the truncated stimuli. Once again, this effect was only observed for low frequency syllables, that is, syllables appearing with fewer homophones and in sparse tonal neighborhoods, confirming that native listeners draw on stored distributional knowledge of syllable+tone co-occurrences when it is most informative for lexical identification. Thus, the effect of syllable-specific tonal probabilities is robust and observable in both implicit eye movement measures and more metalinguistic gating responses to single-speaker input (see also Shuai and Malins, 2017 for TRACE simulations). Because higher speaker variability is known to evoke greater perceptual uncertainty and thus trigger more frequency-based generalizations (e.g., Kleinschmidt & Jaeger, 2016), multi-speaker stimuli may lead to stronger predictions based on tonal probabilities.

These knowledge-based predictions could improve L2 learners' recognition of speech as non-tonal L1 speakers have difficulty acquiring and categorizing L2 lexical tone (e.g., Hallé, Chang, & Best, 2004; Hao, 2012; Leather, 1983; Wang, Jongman, & Sereno, 2003; Wang et al., 1999; Wayland & Guion, 2004; Wong & Perrachione, 2007). Although some modern Chinese pedagogy curricula call for exposing learners to different instructors throughout a semester/term (e.g., Walker, 1989), studies report that L2 learners struggle with pitch range variability coming from multiple speakers (Lee, Tao & Bond, 2009; Wang, Jongman, & Sereno, 2006), noise in the signal (Gottfried & Suiter, 1997; Lee, Tao, & Bond, 2010, 2013), and varying sentence position (Broselow, Hurtig, & Ringen, 1987). Thus, clarifying the effect of multi-talker input and its role in statistical learning bears a pedagogical implication.

Recent work by Ong, Burnham and Escudero (2015) suggests that non-native listeners can improve their categorization of lexical tones through statistical learning. Ong et al. trained native English listeners on either a unimodal distribution or a bimodal distribution of Thai tones. Learners trained on the bimodal distribution outperformed learners trained on the unimodal distribution in a discrimination test, but only if auditory attention to the speech signal was encouraged in the training phase. The authors conclude that in order to track tone distribution, learners must first be aware of its role in expressing lexical contrast.

Although Ong et al.'s finding suggests that learners can track tonal distributions on a single CV syllable, the offline ABX discrimination task they used did not require lexical knowledge and did not address the time-course of word identification that occurs as listeners combine acoustic input with learned statistical distributions of syllable and tone.

The present study examines the dimension-based statistical learning of syllable+tone co-occurrences for L2 word recognition. We train L2 learners of Mandarin on an artificial tonal language designed to mimic Mandarin's syllable+tone distribution. We use eye tracking as the measure of online signal processing during L2 word recognition, and train and test learners on speech input from either one speaker (low variability) or four speakers (high variability). We look for evidence of knowledge-based tone processing in the form of initial eye fixations to more probable on-screen targets prior to disambiguating acoustic information. We also examine whether learners' looks to more probable on-screen targets increase when hearing speech from multiple speakers. This will serve as evidence that greater speaker variability in L2 input leads to greater initial perceptual uncertainty, which is resolved through more predictive eye movements based on learned syllable+tone co-occurrence distributions.

2 Experiment

To test the interaction between speaker variability and the use of statistical knowledge during word recognition while controlling for participants' experience with the language, the present study uses an artificial tonal language modeled after Mandarin. An artificial language is often used to control certain distributional properties, such as frequency of signal occurrence, and enables an examination of language acquisition in a reduced amount of time, (e.g., Caldwell-Harris, Lancaster, Ladd, Dediu, & Christiansen, 2015; Creel, Aslin & Tanenhaus, 2006; Creel, Tanenhaus & Aslin, 2006; Gomez & Gerken, 2000; Magnuson, Tanenhaus, Aslin, & Dahan, 2003; Saffran, Newport, & Aslin, 1996). Using an artificial language that mimics the segmental and suprasegmental features of Mandarin, we explore how a *Mandarin-like* tonal language is acquired and processed by speakers of a non-tonal L1. Previous behavioral studies have demonstrated that native English speakers are able to effectively learn small, closed sets of real and artificial Mandarin-like words paired with images (Chandrasekaran, Sampath & Wong, 2010; Perrachione, Lee, Ha, & Wong, 2011; Wong & Perrachione, 2007). The present study expands on these previous studies by teaching participants a larger set of syllable+tone combinations, with each combination paired with one or more unique visual nonce symbols. To better facilitate word learning, participants were trained over four consecutive days.

2.1 Methods

2.1.1 Participants. Forty native English speakers learning Mandarin as a second language participated in the study (19 female; 21 male; mean age 21). All participants passed first-year Mandarin at The Ohio State University with a "B" grade or higher and had completed 140 in-class hours, as well as an estimated 140 outside self-study hours. All participants were enrolled in second-year Mandarin classes at the time of testing. Participants also completed a language background questionnaire, which asked learners to self-rate their Mandarin L2 abilities on a four-point Likert scale (1: beginner; 4: fluent): speaking (M = 2.1; SD = 0.5); listening (M = 2.1; SD = 0.6). All participants were paid for their participation.

2.1.2 Materials. Twenty-four CV syllables were built from eight Mandarin onsets: /p, p^h, f, k, k^h, m, ts, ɹ/; three monophthongs: /a, i, ə/; and three diphthongs: /ia, iu, ai/. (See the Supplementary

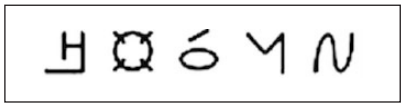

Figure 2. Black and white nonce symbols for *pe2*.

Material for the full stimuli set.) All syllables were considered "phonotactic accidental gaps" (Wang, 1998). These syllables did not violate Mandarin phonotactics, were easily produced by native Mandarin speakers, but were non- existing Mandarin morphemes (i.e., similar to the English nonce word *blick*). Although up to 48 unique syllables could be created from this segmental inventory, only 24 unique syllable types were used to incorporate syllable gaps within the artificial language to match those found in Mandarin (Duanmu, 2007, 2009). The stimuli set included two rare Mandarin syllables: /ka/ and /mə/.[1]

Each of the 24 syllables was paired with one, two, three, or all four tones (tones were directly comparable to those of Mandarin, see Figure 1). Though this maximally yields 96 unique syllable+tone combinations, only 82 combinations were used (i.e., 14 possible combinations were syllable+tone gaps). This ensured roughly the same percentage of syllable+tone gaps as that in the Mandarin lexicon—23% (Duanmu, 2007; Myers, 2002, 2010; Wang, 1998). Additionally, to mimic the proportion of homophony in Mandarin, 16 of the 82 syllable+tone combinations appeared with tonal homophones. For example, the syllable+tone combination *pe2* occurred with four homophones, resulting in five *pe2* nonce words in the artificial language. In contrast, no homophones were created for the syllable+tone combination *pe1*. In total, there were 130 nonce words of which roughly half (66/130) had no homophones (e.g., *pe1*) while the other half (64/130) had at least one homophone (e.g., *pe2*). Each of the 130 nonce words was then given a unique symbol. Thus, as in Mandarin, certain syllable+tone combinations resulted in numerous homophones disambiguated only through the orthography (analogous to English "to," "two" and "too"). Figure 2 shows an example of five nonce symbols, all sharing the same syllable+tone combination *pe2.*

Of these 130 items, 64 served as the test items (16 syllables and their corresponding four syllable+tone combinations, such as *pe1, pe2, pe3*, and *pe4*) while the other 66 served as filler items. Within the test items, two factors—syllable frequency and tonal probability—were crossed to create four within-subject test conditions. Syllable frequency was manipulated by varying the number of tokens of particular CV sequences participants were exposed to. Within the 64 test items, 32 items had high syllable frequency (F+), while 32 items had low syllable frequency (F−). For example, participants were exposed to the F+ syllable *pe* with all four tones as *pe1, pe2, pe3*, and *pe4* (along with the four other *pe2* homophones) for a total of 28 *pe* exemplars per training day. Participants were exposed to F− syllables at just over half that rate. For instance, participants heard *fe1, fe2, fe3*, and *fe4* (along with two other *fe3* homophones) for a total of 16 exemplars per day. This resulted in a F+ to F− syllable frequency ratio of 7:4, roughly approximating Wiener and Ito's (2015) calculations.

Tonal probabilities were assigned by coupling each syllable in the test items with one tone as the most probable (P+) and another tone as the least probable (P−). This resulted in four test conditions: F+P+, F+P−, F−P+, and F−P−. Following Wiener and Ito, the P+ tone occurred in at least half of the spoken tokens for a given syllable. For example, each day, 14 of the 28 *pe* exemplars carried the second tone as *pe2* making that syllable+tone combination a high frequency, high probability (F+P+) target. Only four of the 28 *pe* exemplars carried the fourth tone as *pe4*, making that syllable+tone combination a high frequency low probability (F+P−) target. The other two *pe* items—*pe1* and *pe3*—each appeared five times. The same P+/P− manipulation occurred for low

frequency items: *fe3* served as the F−P+ target (8/16 exemplars heard as *fe3*), *fe1* served as the F−P− target (2/16 exemplars), while *fe2* and *fe4* each appeared three times. To match the occurrence of target symbols across the four conditions, tonal probabilities were manipulated only by increasing or decreasing exposure to filler homophones. High frequency syllables with the high probability tone were presented with five homophones, while low frequency syllables with the high probability tone were presented with three homophones. For example, *pe2* was presented 14 total times as the P+ tone. Of the 14 presentations, 10 occurred using the four right-most nonce symbols in Figure 2, that is, the non-test homophone symbols. By repeatedly showing these four symbols, the probability of the syllable *pe* appearing with tone 2 greatly increased. The left-most nonce symbol in Figure 2 served as the test item and appeared exactly the same number of times as the symbol for the low probability tone (P−) *pe4*. The same homophone manipulation occurred for low frequency syllables: *fe3* appeared six times using the two *fe3* homophone nonce symbols and two times with the test item symbol. The P− symbol for *fe1* appeared two times, as well, resulting in equal P+/P− target symbol occurrence. Thus, the test items' P+ and P− visual symbols appeared the same number of times while the auditory occurrence of certain tones with that syllable varied through manipulation of filler homophone symbols. Importantly, tones 2 and 3 were not paired as a syllable's P+/P− duo because they are perceptually confusable for both native and non-native listeners.[2] As a result of avoiding this pair, the artificial language contained the fourth tone as the most frequent tone (28% of all exemplars), followed by the first tone (25%), the second tone (24%), with the third tone as the least frequent tone (23%)—a pattern nearly identical to that of spoken Mandarin (Duanmu, 2007, 2009). For the 64 test items, however, all four tones occurred 16 times.

All auditory stimuli were digitally recorded by four (two female and two male) native Mandarin speakers at 16 bits/44,100 Hz. Two additional native speakers from Beijing China correctly identified the pronunciation of the syllables and tones with 100% agreement. Acoustic analysis of the tones showed previously demonstrated temporal differences, such that tones 2 and 3 were longer than tones 1 and 4 (Moore & Jongman, 1997; Repp & Lin, 1990; Zee, 1980). Statistically, the mean duration of each tone type did not differ across the four speakers: tone 1, $F(3,96) = 0.166$, $p = 0.69$; tone 2, $F(3,96) = 0.825$, $p = 0.37$; tone 3, $F(3,96) = 3.28$, $p = 0.08$; tone 4, $F(3,96) = 0.114$, $p = 0.73$.

2.1.3 Design and procedure. Participants came to the laboratory for 30-minute sessions on four consecutive days. Each participant was randomly assigned to one of two lists that counterbalanced the tonal probability for each item. For example, in list one, tone 2 was the most probable tone (P+) for *pe*, while in list two it was the least probable tone (P−) for *pe*. Daily training and testing consisted of four tasks in the same order each day.

Participants began with (1) a passive listening task in which a nonce symbol was presented on a computer monitor while its audio label was simultaneously presented over headphones. Participants were instructed to memorize the symbol–audio pair and then mouse click on the symbol at their own pace to advance to the next item (131 trials). Participants next performed (2) a speech shadowing task in which a symbol–audio pair was presented and then participants repeated aloud the CV+tone of the symbol as accurately as possible (131 trials). Next, participants completed (3) a symbol naming task in which a symbol appeared on the screen and participants were asked to name its audio label learned in the previous listening and shadowing tasks. Participants were told explicitly to guess a label, even if they were unsure. After producing a label, participants were told to click the mouse to hear the correct audio label and then click again to advance to the next trial. Only the 64 test items were presented in the naming task, that is, no homophones or other fillers.

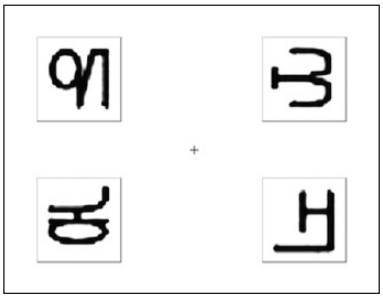

Figure 3. Example of 4AFC display for *pe2*.

After completing the first three tasks in a sound-attenuated booth, participants were seated in front of a computer monitor to complete the final task (4): four-alternative forced-choice (4AFC) symbol identification with eye-tracking (with feedback). In the 4AFC task, which is the focus of the present study, four symbols were presented simultaneously as participants heard the target symbol's audio label over headphones. Participants were told to click on the symbol that matched what they heard. During the 4AFC task, participants' eye movements were continuously recorded at 50 Hz using the Tobii 1750 system. During each trial, participants fixated upon the center of the screen during the carrier phrase "I will say…" ("*wo3 yao4 shuo1*" spoken in Mandarin), followed by the simultaneous presentation of the symbol and the auditory syllable+tone combination. After clicking, a red box appeared around the correct target symbol so that feedback could guide learning. There were 48 total trials (32 filler and 16 target trials with four trials per experimental condition) with a 2 second inter-trial interval. Target trial visual displays showed the target and three other trained test items: a tonal competitor, a rhyme competitor, and a distractor. Figure 3 shows a sample trial slide with *pe2* (P+) as the target (bottom right). The three other visual candidates were a tonal competitor, which shared the same syllable but had the opposite tonal probability (e.g., *pe4*; bottom left), a rhyme competitor, which shared the same vowel and tone but had a different onset (e.g., *fe2*; top left), and a distractor, which had a unique syllable and tone (e.g., *riu1*; top right). The positions of the target and competitors were counterbalanced within each testing session and across all four days of testing. Each day different distractor syllables were used on target trials in addition to wholly new filler trials. This resulted in 192 unique slides across all four days of testing.

To manipulate speaker variability, half of the participants were trained and tested on only one female speaker (low variability condition: V−). In this V− condition, participants heard the same speaker during all four tasks, thus the 4AFC testing consisted of a familiar syllable+tone exemplar spoken by a familiar speaker. The other half of the participants were trained and tested on four different speakers (high variability condition: V+). In this V+ condition, the 4AFC testing consisted of a familiar syllable+tone learned in the first three tasks but spoken for the first time by a particular talker. For example, participants heard *pe2* in the first three tasks spoken by male speaker one, but were then tested on female speaker two's production of *pe2* in the 4AFC task. Thus, each day, participants heard all four speakers producing different syllable+tones across all four tasks, but testing in the 4AFC task involved hearing each speaker produce a particular syllable+tone combination for the first time. Participants in the V− condition ($M = 2.1$, $SD = 0.6$) did not differ in their self-rated Mandarin speaking abilities from those in the V+ condition ($M = 2.2$, $SD = 0.2$); $t(38) = 0.83$, $p = 0.40$. Similarly, V− participants ($M = 2.1$, $SD = 0.6$) did not differ from V+ participants ($M = 2.2$, $SD = 0.5$) in their self-rated Mandarin listening abilities; $t(38) = 1.01$, $p = 0.31$.

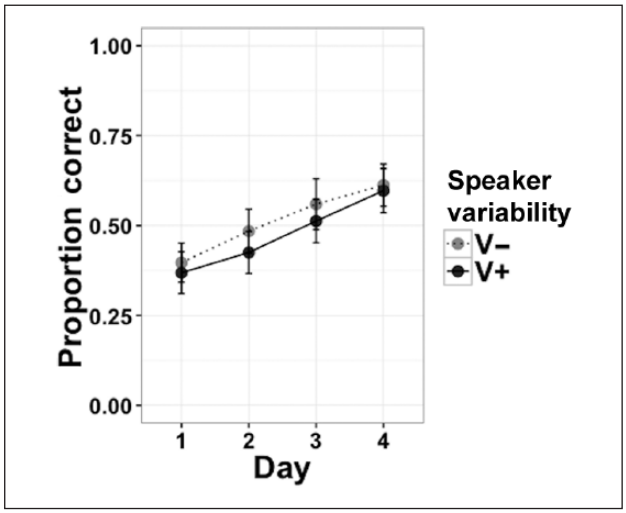

Figure 4. Mean 4AFC mouse click accuracy (with standard error (SE)).

We calculated 4AFC mouse-click accuracy as an offline measure of acoustic-based processing accuracy; learning to correctly perceive tonal contrasts must largely depend upon correct categorization of phonetic cues. Following Wiener and Ito's (2015) findings, knowledge-based processing will primarily be captured through participants' online eye movements within the first 1000 ms. We test whether the two-way interaction between syllable frequency and tonal probability reported in Wiener and Ito (2015, 2016) can be replicated in L2 learners of an artificial tonal language. In particular, we test whether low frequency syllables with high probability tones (F−P+) show the least amount of competition from the tonal competitor, and whether the same low frequency syllables with low probability tones (F−P−) exhibit the greatest amount of competition to the more probable tonal competitor. If speaker variability enhances the usage of tonal probabilities during word recognition, high variability (V+) input should lead to less certainty in acoustic-based processing and a greater reliance on probability-based processing. Therefore we expect participants in the V+ condition to show more erroneous looks to competitors with high probability tones, a greater number of probability-based mistakes on the more probable tonal-competitor, and slower recovery of looks to the intended (but less probable) target. This should be confirmed with a two-way interaction between speaker variability and tonal probability or a three-way interaction between speaker variability, syllable frequency, and tonal probability.

3 Results

3.1 Mouse click accuracy

Three participants (one from low variability; two from high variability) were removed from all subsequent analysis due to at-chance identification on day four. Figure 4 shows the remaining 37 participants' overall mean 4AFC mouse click accuracy for the low variability (V−) and high variability (V+) conditions across the four days of testing. Both groups demonstrated above chance (.25) performance on the first day followed by consistent daily improvement. By the fourth day, participants' accuracy was nearly identical across variability conditions: .61 (V−), .60 (V+). Day

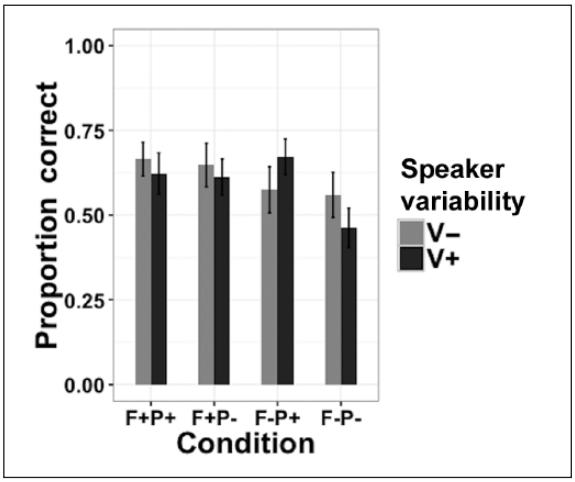

Figure 5. Day 4 mean 4AFC proportion correct (with *SE*).

four incorrect mouse clicks consisted almost entirely of tonal competitor mistakes; distractor and rhyme mistakes were below 7% for both groups.

Figure 5 shows the mean accuracy for each of the four conditions (CV frequency × tonal probability) on Day 4. For high frequency syllables (F+), participants in the V− and V+ conditions identified the intended target with similar accuracy. Likewise, participants made roughly the same proportion of competitor mistakes for high frequency syllables with high probability tones (F+P+) V+: .33; V−: .28; as they did for high frequency syllables with low probability tones (F+P−) V+: .32; V−: .29.

For the low frequency syllables (F−), however, a difference between the variability conditions emerged. Participants exposed to high variability (V+) input identified the low frequency, high probability (F−P+) target with the highest overall accuracy (.68) and the corresponding low probability (F−P−) target with the lowest overall accuracy (.46). Participants who heard only a single voice (V− condition) responded with nearly equal accuracy, irrespective of tonal probability (F−P+: .58; F−P−: .57). This asymmetry was reflected in incorrect mouse clicks on the tonal competitor. V− participants made nearly the same proportion of tonal mistakes (F−P+: .35; F−P−: .38), while V+ participants made far fewer mistakes when the low frequency syllable appeared with the high probability tone (F−P+: .25) than when it appeared with the low probability tone (F−P−: .46).

Day 4 mouse click accuracy results were analyzed using mixed-effects logistic regression models with the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.3.1; R Core Team, 2016). Syllable frequency, tonal probability and speaker variability were treated as sum coded factors. Model fit including selection of predictors (and their interactions), random effects structure, and *p*-values were based on the χ^2-test of the change in deviance between the models with and without the effect of interest. The resulting model contained main effects and two-way interactions alone. The optimal random effects structure contained subject and item intercepts, by-subject random slopes for syllable frequency and tonal probability, and by-item random slopes for speaker variability (Table 1). A marginal two-way interaction between tonal probability and speaker variability was found. Subset analysis indicated that this two-way interaction was restricted to the low frequency syllables in which the high variability group more accurately identified F−P+

Table 1. Model output from the logistic mixed effects model predicting accuracy.

	β	*SE*	*z*	*p*
(Intercept)	0.527	0.15	3.45	0.001
Syllable frequency	0.237	0.22	1.07	0.28
Tonal probability	0.367	0.22	1.63	0.10
Speaker variability	0.089	0.29	0.30	0.75
Frequency: probability	–0.567	0.44	–1.28	0.19
Frequency: variability	–0.104	0.41	–0.25	0.79
Probability: variability	0.861	0.43	1.96	0.05

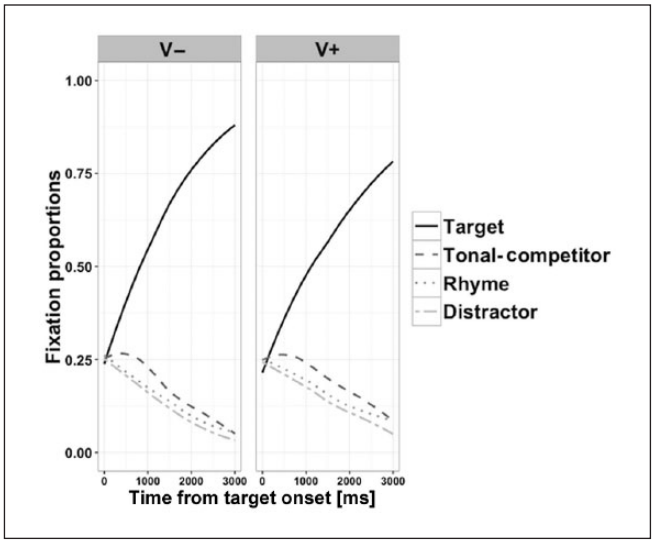

Figure 6. Smoothed grand mean fixation proportions from the onset of the target word across four days.

targets and less accurately identified F−P− targets (β = 0.70, *SE* = 0.28, *Z* = 2.51, $p < 0.05$); the low variability group showed no such difference in accuracy.

Thus, Day 4 mouse click accuracy results indicated that L2-tonal learners exposed to high variability input relied on syllable-conditioned tonal probabilities to choose the symbol for low frequency syllables. Participants in the low variability condition did not demonstrate the same probability-based processing in their mouse clicks; V− participants ostensibly relied less on probability-based processing since acoustic-based processing involved recognition of a familiar speaker with little variability. We next turn to eye movements to explore whether participants demonstrated anticipatory looks to the more probable competitor at an early stage of word recognition.

3.2 Eye fixations

Figure 6 plots the smoothed[3] grand mean fixations across all correct trials on all four days, for each variability group. Both variability groups showed the same overall pattern in which the tonal competitor attracted more looks than the rhyme and distractor.

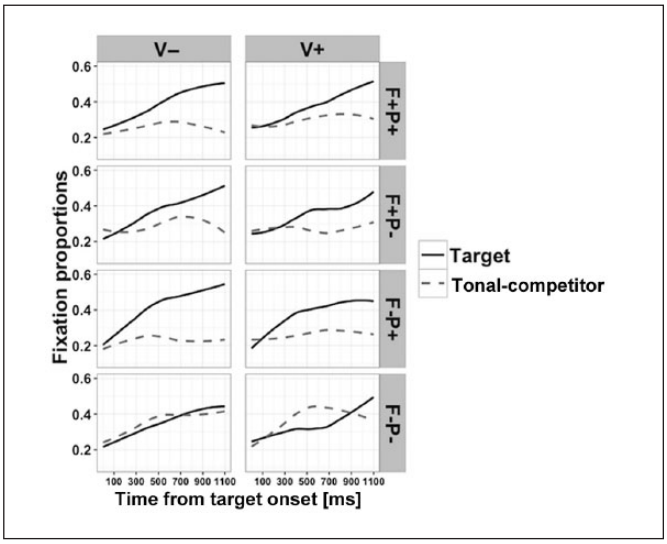

Figure 7. Smoothed day 4 fixations by condition and variability (0–1100 ms).

Figure 7 plots the smoothed Day 4 fixations to the target and tonal competitor for each condition from the target onset until looks to the tonal competitor peaked at approximately 1100 ms. For high frequency syllables (F+), participants in both variability conditions looked to the target and tonal competitor at comparable rates during the first 1100 ms, regardless of tonal probability. For low frequency syllables, both groups showed minimal competition of the tonal competitor (i.e., earlier and greater divergences between the target and the tonal competitor) in the low frequency, high probability (F−P+) condition and greater competition of the tonal competitor in the low frequency, low probability (F−P−) condition. In both groups, fixations to the more probable tonal competitor were greater than fixations to the target from 300 ms until roughly 800 ms for low frequency syllables with low probability tones (F−P−).

To compare looks to the target across conditions, the empirical logit (elogit) and the associated weights were calculated at 200 ms time intervals for each trial (Barr, Gann, & Pierce, 2011). Since roughly 200 ms is required to plan and execute an eye movement (Matin, Shao, & Boff, 1993), and since Figure 6 showed a divergence of segmental cohorts (target and competitor) from the distractors (rhyme and unrelated) at approximately 300 ms from the stimulus onset, the first analysis was set from 300–500 ms with the last analysis set at 900–1100 ms, at which point looks to the tonal competitor reached their peak. Mixed-effects models testing the fixation likelihood for the target were built with the abovementioned sum coding scheme. Inclusion of predictors (and their interactions) in the model, random effects structure, and *p*-values were based on the χ^2–test of the change in deviance between the models with and without the effect of interest. The resulting models contained main effects and two-way interactions only. The optimal random effects structure contained subject and item intercepts, by-subject random slopes for syllable frequency and tonal probability, and by-item random slopes for speaker variability (Table 2).

From 300 ms to 500 ms, a main effect of speaker variability was found, $\chi^2(1) = 7.06$, $p < .01$: participants in the low variability V− condition looked to the target at a higher proportion than participants in the high variability V+ condition. Subset analyses indicated that this effect of variability was an aggregate effect and did not stem from a single condition.

Table 2. Model output from the weighted mixed effects models on target fixations.

Time		β	*SE*	*t*
300–500 ms				
	(Intercept)	–1.832	0.27	–6.74
	Syllable frequency	–0.296	0.54	–0.54
	Tonal probability	–0.376	0.25	–1.48
	Speaker variability	0.608	0.21	2.76
	Frequency: probability	–0.063	0.44	–0.14
	Frequency: variability	0.128	0.46	0.27
	Probability: variability	–0.326	0.50	–0.64
500–700 ms				
	(Intercept)	–1.498	0.20	–7.16
	Syllable frequency	0.382	0.43	–0.87
	Tonal probability	–0.465	0.29	–1.59
	Speaker variability	0.350	0.26	1.30
	Frequency: probability	–1.108	0.51	–2.13
	Frequency: variability	0.318	0.59	0.53
	Probability: variability	–1.033	0.58	–1.78
700–900 ms				
	(Intercept)	–0.447	0.12	–3.61
	Syllable frequency	0.196	0.24	0.79
	Tonal probability	0.019	0.20	0.09
	Speaker variability	–0.004	0.20	–0.02
	Frequency: probability	0.252	0.40	0.61
	Frequency: variability	–0.455	0.40	–1.12
	Probability: variability	0.898	0.41	2.22

From 500 ms to 700 ms, a two-way interaction between syllable frequency and tonal probability was found, $\chi^2(1) = 4.57$, $p < .05$: participants' looks to the low frequency, high probability (F−P+) target were significantly higher than looks to the low frequency, low probability (F−P−) target. Subset analyses for each variability condition confirmed this two-way interaction for low frequency syllables only: V− condition ($\beta = -1.81$, $SE = 0.76$, $t = -2.36$, $p < 0.05$); V+ condition ($\beta = -0.98$, $SE = 0.36$, $t = -2.72$, $p < 0.01$).

From 700 ms to 900 ms, a two-way interaction between tonal probability and speaker variability was found, $\chi^2(1) = 4.55$, $p < .05$: subset analyses indicated that participants in the V− condition looked to low probability (P−) targets at a greater proportion than participants in the V+ condition ($\beta = 0.97$, $SE = 0.42$, $t = 2.29$, $p < 0.05$).

From 900 ms to 1100 ms, no effects of the factors or their interactions were found.

To test whether the reduction in looks to the target was due to participants looking to the more probable tonal competitor, the weighted elogit of the tonal competitor was analyzed (Table 3). From 500 ms to 700 ms, a two-way interaction between syllable frequency and tonal probability was found, $\chi^2(1) = 7.32$, $p < .01$, as well as a marginal main effect of tonal probability. $\chi^2(1) = 3.48$, $p = .05$. Subset analyses of the low frequency syllables confirmed that, for each variability condition, participants looked to the high probability P+ tonal competitor at a greater proportion than the corresponding low probability P− tonal competitor: V− condition ($\beta = -0.98$, $SE = 0.36$, $t = -2.72$, $p < 0.01$); V+ condition ($\beta = 1.74$, $SE = 0.56$, $t = 3.06$, $p < 0.01$). From 700 ms to 900 ms, a main effect of speaker variability was found, $\chi^2(1) = 12.05$, $p < .001$, indicating V+ participants looked

Table 3. Model output from the weighted mixed effects models on competitor fixations.

Time		β	*SE*	*t*
500–700 ms				
	(Intercept)	–2.072	0.15	–13.28
	Syllable frequency	–0.051	0.31	–0.16
	Tonal probability	–0.467	0.24	–1.90
	Speaker variability	–0.287	0.23	–1.24
	Frequency: probability	1.197	0.44	2.68
	Frequency: variability	–0.673	0.47	–1.42
	Probability: variability	–0.071	0.48	–0.14
700–900 ms				
	(Intercept)	–2.209	0.14	–15.06
	Syllable frequency	0.058	0.32	0.17
	Tonal probability	0.202	0.27	0.72
	Speaker variability	–0.456	0.20	–2.22
	Frequency: probability	0.392	0.41	0.93
	Frequency: variability	–0.282	0.50	0.56
	Probability: variability	0.095	0.55	0.17

to the tonal competitor at a higher proportion than V− participants. Subset analyses indicated that this was an aggregate effect and did not stem from a single condition.

To summarize, from 300 ms to 500 ms, participants exposed to high speaker variability initially demonstrated less certainty about the incoming acoustic signal irrespective of syllable frequency or tonal probability. From 500 ms to 700 ms, participants in both variability conditions demonstrated the expected two-way interaction between syllable frequency and tonal probability. Participants looked to the low frequency, high probability F−P+ target at a higher proportion than the corresponding low frequency, low probability F−P− target. The reduction in looks to the F−P− target was due to participants looking first to the more probable F−P+ tonal competitor despite the incongruent acoustic signal. In the following window from 700 ms to 900 ms, after anticipatory eye movements to the more probable competitor proved to be incorrect, participants exposed to a single speaker swiftly recovered and looked to the intended target. Participants exposed to multiple speakers required additional time to switch from the more probable tonal competitor to the less probable target.

4 Discussion

This study used lexical tone as the non-native speech cue of interest to examine L2 learners' perceptual trade-off between acoustic-based processing of F0 information and knowledge-based processing of syllable+tone co-occurrences. We manipulated speaker variability to test whether multi-speaker input increases perceptual uncertainty, causing a greater reliance on statistical learning to achieve word recognition.

On each of the four days of testing, listeners trained and tested on the same familiar speaker in the low variability condition (V−) identified syllable+tone combinations slightly more accurately than those participants trained and tested on different speakers in the high variability condition (V+). This accuracy difference, however, was not statistically significant on any test day (Figure 4). Thus, on the surface the accuracy of offline word identification improved to a similar degree, regardless of speaker variability. A closer examination of response patterns by condition revealed

that, despite this overall similarity, acoustic variability significantly influenced how low frequency syllable+tone combinations were processed. Participants exposed to multi-speaker input relied more on syllable-specific tonal probabilities to identify low frequency syllable words; V+ participants were most accurate at identifying low frequency, high probability F−P+ targets and least accurate at identifying the corresponding low frequency, low probability F−P− targets. Participants trained and tested on low variability, single speaker input V− showed no apparent sensitivity to tonal probabilities in their offline mouse click results. The mouse click results alone seem to suggest that syllable-specific tonal probabilities were traced only when learners had to process words spoken by unfamiliar speakers.

The eye movements to the target and tonal competitor, however, revealed that learners used syllable-specific tonal probabilities at an early stage of word recognition, regardless of speaker variability. The time course of the probability-based anticipatory fixations were similar to those found in the native Mandarin speakers tested by Wiener and Ito (2015). The learners' fixation data, like that of native speakers in Wiener and Ito, exhibited an immediate effect of tonal probability only for low frequency syllables. When the stimulus contained a low frequency syllable with a low probability tone (F−P−), participants in both variability conditions looked first to the more probable competitor and only considered the less probable target after sufficient tonal information became available over time. Importantly, the time course of recovery from these incorrect anticipatory looks was influenced by the degree of variability in the signal: participants exposed to less variability recovered from incorrect probability-based predictions of tone more rapidly than participants exposed to greater variability.

These results shed light on several issues of L2 spoken word recognition. First, the present study's eye fixations to the target and competitor indicate that L2 learners track the distribution of segmental and suprasegmental co-occurrences and make predictions accordingly during spoken word recognition. These results add to previous research that established L2 listeners' sensitivity to non-native phonetic frequencies (e.g., Escudero, et al., 2009, 2011; Escudero & Williams 2014; Hayes-Harb, 2007; Liu & Kager, 2014; Wanrooij et al., 2013). Importantly, these results demonstrate that sensitivity to L2 phonetic distributions extends beyond cues varying along a single dimension to multiple co-occurring cues as a form of dimension-based statistical learning (e.g., Idemaru & Holt, 2011, 2014). Participants in the present study had limited L2 experience with a tonal language—roughly one year of classroom training—and yet they exhibited a degree of anticipatory probability-based processing similar to what has been reported in native Mandarin speakers, albeit with a much smaller artificial language. This suggests that statistical learning in L2 acquisition may not differ in its fundamental mechanism from L1 statistical learning (Gómez & Gerken, 2000; Misyak & Christiansen, 2012; Saffran, 2003).

Second, the present data demonstrate how speaker variability affects the acoustic-based and knowledge-based processing of spoken syllable+tone words. The previously reported finding that high variability input facilitates L2 word learning (e.g., Barcroft & Summers, 2005) was not supported by the offline 4AFC mouse-click accuracy results. No statistical difference was found between speaker variability conditions. We note that the present findings do not necessarily invalidate the hypothesis that high variability input facilitates L2 word learning. This lack of a speaker variability effect on mouse clicks may have been due to the limited number of unfamiliar speakers tested, participants' individual aptitude for tone perception (e.g., Perrachione et al., 2011; Sadakata & McQueen, 2014) or the fact that the participants were somewhat atypical learners given that their university classroom experience involved five different instructors each week. It is unclear to what degree this early exposure to different speakers impacted the offline results.

Yet, despite this similarity in learners' offline mouse clicks, online eye movements revealed an important difference. An increase in variability resulted in greater uncertainty: as the speech signal

unfolded in time, familiarity with the speaker led to faster acoustic-based word recognition of the syllable's onset. From 300 ms to 500 ms, participants exposed to a single speaker showed a steeper increase in the fixations to the target, hinting at a greater certainty about the incoming acoustic signal irrespective of syllable frequency or tonal probability. Those trained on multiple speakers showed slower detection of the target, corroborating previous findings on the effect of speaker variability on segmental identification (e.g., Magnuson & Nusbaum, 2007; Mullennix, Pisoni & Martin, 1989; Nusbaum & Morin, 1992). For listeners exposed to multi-speaker input, the observed reliance on statistical knowledge was more prominent for low frequency syllables. Those participants exposed to multiple voices drew on their knowledge of low frequency syllables co-occurring with probable tones as a way of overcoming poor acoustic-based processing. For these learners, overcoming speaker variability was largely accomplished through knowledge-based processing.

Importantly, these results revealed an inverse relationship between the degree of certainty in acoustic-based signal identification and the degree of incorrect anticipatory looks driven by statistical knowledge. Participants trained and tested on the same speaker demonstrated earlier certainty of the intended syllable and a lower proportion of anticipatory looks to the more probable competitor. Though many participants trained with a single speaker incorrectly predicted the more probable competitor while hearing a low frequency syllable, their ultimate mouse-click choices confirmed that they were able to recover from these limited incorrect predictions. In contrast, participants trained and tested on multiple-speakers demonstrated less certainty of the incoming syllable and relied on syllable-specific tonal probabilities to a much greater degree. This supports the claim that under conditions of uncertainty, participants look for and rely more on other cues (e.g., Nixon et al., 2016). For these L2 learners, increased variability in the acoustic signal resulted in a greater use of probability-based processing. Indeed, for many of the multi-speaker learners, this reliance on tonal probabilities proved too strong to overcome, resulting in a significant decrease in mouse-click accuracy for low frequency syllables with low probability tones (F−P−).

Our results indicate that in addition to bolstering the acoustic–phonetic category learning of novel speech sounds by exposing learners to a full range of acoustic phonetic cues (e.g., Hirata, Whitehurst, & Cullings, 2007; Logan et al., 1991; Sadakata & McQueen, 2013; Wang et al., 1999), high variability stimuli can also facilitate the statistical learning of novel speech cues and their co-occurrences through dimension-based statistical learning (e.g., Idemaru & Holt, 2011, 2014). This variability forces learners to generalize about the likelihood of the input-representation match (given enough appropriate experience with the language). Both this ability to generalize and the ability to successfully recover from incorrect probability-based predictions contribute to the dynamic perceptual trade-off between signal-based and knowledge-based processing. These findings point to a general mechanism of speech processing where a broader range of acoustic input triggers more knowledge-based processing in both L1 and L2.

With respect to Mandarin tone processing, the present data converge with Wiener and Ito's (2016) gating results from native speakers: a truncated low frequency syllable consisting of only the onset and 40 ms of the vowel prompted native Mandarin speakers to report the most probable tone for the perceived syllable, even if the reported tone was acoustically dissimilar to the stimuli. We argue that given tone's highly volatile nature across contexts and speakers (e.g., Huang & Holt, 2009; Xu, 1994, 1997), listeners exert input-based and knowledge-based processing flexibly in order to achieve correct word recognition. As the acoustic signal becomes more variable due to unfamiliar speakers or even truncated speech, listeners draw on their previous experience with the language to assess the likelihood of syllable+tone combinations. The present study demonstrates that the input-driven use of the statistical knowledge is not restricted to native speakers alone (e.g., Fox & Unkefer, 1985; Wiener & Turnbull, 2016) but the same mechanism can operate in non-native learners. L2 learners, like native speakers, draw on this statistical knowledge when it is most

informative for lexical identification, that is, for low frequency syllables. For high frequency Mandarin syllables that may have many more syllable+tone homophones, listeners must tune to acoustic–phonetic cues along with context to achieve lexical identification (Fox & Qi, 1990; Liu & Samuel, 2007; Ye & Connine, 1999).

An interesting future line of inquiry concerns whether truly naïve listeners unaware of tone's lexical role are able to track syllable+tone distributions and make predictions during word recognition. Evidence from Ong et al.'s (2015) ABX discrimination task suggests that distributional learning of tone only occurs if learners' auditory attention is encouraged in the acquisition phase. The L2 learners tested in the present study had completed a year of classroom Mandarin and were therefore well aware of tone's lexical role. This leaves open the possibility that probability-based processing of non-native speech cues emerges only once the listener is aware of the novel cue's importance in word recognition, that is, through explicit instruction of the cue. In a follow-up study we are examining whether naïve listeners who have no previous exposure to a tonal language can use tonal probability information during word recognition and whether explicit instruction on tones is effective for dimension-based statistical learning.

Our results best fit with models of spoken-word recognition that characterize the lexical search and selection as the outcome of a competition between multiple representations and the incoming fine-grained acoustic information (e.g., MacWhinney, 2005; McClelland & Elman, 1986; Norris, 1994; Norris & McQueen, 2008; Shook & Marian, 2013). In particular, the present findings are compatible with models that account for knowledge-based processing influencing the earliest moments of lexical access (e.g., Dahan, Magnuson, & Tanenhaus, 2001). For instance, in their eye-tracking study, Dahan et al. (2001) showed participants four images, three of which shared the same initial phonemes (bench, bed, and bell). As participants heard the initial phonemes of the target word "bench," they were more likely to fixate on the picture with the higher frequency name—"bed"—than the target or the picture with the lower frequency name—"bell." These findings were interpreted as evidence against a late-acting, decision-bias locus for frequency information. Dahan et al. proposed modifications to TRACE (McClelland & Elman, 1986) by adjusting resting–activation levels or connection strengths. The present results, like those reported in Wiener and Ito (2015), demonstrate that syllable frequency and syllable-conditioned tonal probability information affects the earliest moments of syllable+tone lexical access. Modifications to TRACE that account for tone (e.g., Malins & Joanisse, 2010; Shuai & Malins, 2017; Ye & Connine, 1999) in combination with Dahan et al.'s proposed resting–activation level adjustments would serve as an empirically well-motivated starting point for modeling lexical-tone processing.

We acknowledge that the external validity of these findings, like those of any artificial language study, is unknown. Similar to previous artificial language studies (e.g., Caldwell-Harris et al., 2015; Creel et al., 2006a, 2006b; Magnuson et al., 2003; Sulpizio & McQueen, 2012), the present study used a limited number of test items to measure the effects of speaker variability, syllable frequency, and tonal probability. It remains an open question whether the present findings generalize to natural second-language acquisition of Mandarin and other tonal languages. Retrospective power analyses following Gelman and Carlin (2014) and "observed power" simulations using the SIMR package in R (Green & MacLeod, 2016) were run on all our mixed-effects models. For our more robust speaker variability finding, the effect estimate had a power of .8. For our two-way interactions between: (a) tonal probability and speaker variability; and (b) syllable frequency and tonal probability, the effects estimates both had powers of .6.

In conclusion, the present study demonstrated that L2 learners are able to track the statistics of co-occurrences of multiple cues. Learners relied on probability-based processing to a higher degree when faced with high variability in the speech signal due to an increase in perceptual

uncertainty. Although learners drew on tonal probabilities when listening to a familiar talker, the use of statistical knowledge was reduced due to the higher reliability of acoustic cues and greater perceptual certainty. Future studies should clarify how early this perceptual trade-off between knowledge-based and input-based processes emerges in naïve learners, as well as whether explicit attention to novel cues (such as lexical tones) facilitates the statistical learning. Finally, more studies are required to see the long-term effect of multi-speaker exposure in L2 acquisition.

Acknowledgements

We are grateful for the constructive comments from the audience members as well as valuable feedback from Marjorie Chan, Mineharu Nakayama, Chao-Yang Lee, Shravan Vasishth, Christine Shea, and two incredibly helpful anonymous reviewers.

Authors' Note

An earlier version of this work was presented at the Sound to Word in Bilingual and Second Language Speech Perception 2016 conference at the University of Iowa.

Funding

This work was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (grant number BCS-1451677) to the first two authors.

Notes

1. The syllable /ka/ primarily appears in standard Mandarin as a phonetic or onomatopoeic syllable. Based on Cai and Brysbaert (2010)—a 46.8 million character corpus—the syllable /ka/ appears in speech less than one-tenth of a percent. The syllable /mə/ appears only as a suffix in interrogatives, but never carries a tone.
2. Tone 3 undergoes phonologically conditioned tone sandhi in Beijing Mandarin and is produced as a tone 2 variant in certain contexts (Peng, 2000; Shih, 1997). The tone 3–tone 2 contrast is typically hardest for children and L2 learners to acquire (Gottfried & Suiter, 1997; Li & Thompson, 1977; Shen & Lin, 1991). This perceptual confusion has also been shown to affect lexical access and processing in L1 speakers (Nixon, Chen, & Schiller, 2014; Wu, Chen, van Heuven, & Schiller, 2014).
3. Locally weighted scatter-plot smoothing (loess) was used for all fixation plots (Cleveland, 1979)

ORCID iD

Seth Wiener https://orcid.org/0000-0002-7383-3682

References

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, *32*(2), 233–250.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387.

Barr, D. J., Gann, T. M., & Pierce, R. S. (2011). Anticipatory baseline effects and information integration in visual world studies. *Acta Psychologica*, *137*, 201–207.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, (pp. 171–204). Baltimore, MD: York Press.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam, the Netherlands: John Benjamins Publishing Company.

Blicher, D. L., Diehl, R., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin tone 2 tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, *18*, 37–49.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*, 2299–2310.

Broselow, E., Hurtig, R. R., & Ringen, C. (1987). The perception of second language prosody. In G. Loup & S. Weinburger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system* (pp. 350–361). New York, NY: Newbury House Publishers.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *Plos ONE*, *5*(6), e10729.

Caldwell-Harris, C. L., Lancaster, A., Ladd, D. R., Dediu, D., & Christiansen, M. H. (2015). Factors influencing sensitivity to lexical tone in an artificial language. *Studies in Second Language Acquisition*, *37*(2), 335–357.

CC-CEDICT. (2016). Online open source Chinese dictionary. Retrieved from http://cc-cedict.org/

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cueweighting and lexical tone learning. *Journal of the Acoustical Society of America*, *128*(1), 456–465.

Chen, H. C., Vaid, J., & Wu, J. T. (2009). Homophone density and phonological frequency in Chinese word recognition. *Language and Cognitive Processes*, *24*(7–8), 967–982.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association, 74*(368), 829–836.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006a). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language*, *54*, 1–19.

Creel, S. C., Tanenhaus, M. K., & Aslin, R. N. (2006b). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 15–32.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Duanmu, S. (2007). *The phonology of standard Chinese*. New York, NY: Oxford University Press.

Duanmu, S. (2009). *Syllable structure: The limits of variation*. New York, NY: Oxford University Press.

Ellis, N. C. (2011). Frequency-based accounts of SLA. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 193–210). London, UK: Routledge/Taylor & Francis.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–EL212.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition*, *133*(2), 408–413.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research*, (pp. 233–277). Baltimore, MD: York Press.

Flege, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 101–131). Mahwah, NJ: Lawrence Erlbaum Associates.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication*, *16*(1), 1–26.

Fox, R. A., & Qi, Y.-Y. (1990). Context effects in the perception of lexical tones. *Journal of Chinese Linguistics*, *18*, 261–284.

Fox, R. A., & Unkefer, J. (1985). The effect of lexical status on the perception of tone. *Journal of Chinese Linguistics*, *13*, 70–89.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*, 178–186.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, *9*, 317–323.

Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, *25*, 207–231.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Hallé, P. A., Chang, Y-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese and French listeners. *Journal of Phonetics*, *32*(3), 395–421.

Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65–94.

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, *121*(6), 3837–3845.

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*, *33*, 353–367.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Howie, J. (1976). *Acoustical studies of Mandarin vowels and tones*. Cambridge, UK: Cambridge University Press.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America*, *125*(6), 3983–3994.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension–based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *Handbook of East Asian psycholinguistics* (pp. 209–217). Cambridge, UK: Cambridge University Press.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, *18*(2), 137–171.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59.

Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*, *11*, 373–382.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*, 1–15.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2010). Identification of multi-speaker Mandarin tones in noise by native and non-native listeners. *Speech Communication*, *52*, 900–910.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2013). Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners. *Speech, Language and Hearing*, *16*(1), 1–9.

Li, P., Shu, H., Yip, M., Zhang, Y., & Tang, Y. (2002). Lexical ambiguity in sentence processing: Evidence from Chinese. In M. Nakayama (Ed.), *Sentence processing in East Asian languages* (pp. 111–129). Stanford, CA: CSLI Publications.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language*, *4*(2), 185–199.

Li, P., & Yip, M. C. (1998). Context effects and the processing of spoken homophones. *Reading and Writing, 10*(3–5), 223–243.

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405.

Liu, L., & Kager, R. (2014). Perception of tones by infants learning a non-tone language. *Cognition*, *133*(2), 385–394.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47*(2), 109–138.

Liu, S., & Samuel, A. G. (2007). The role of Mandarin lexical tones in lexical access under different contextual conditions. *Language and Cognitive Processes*, *22*, 566–594.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242–1255.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886.

MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). New York, NY: Oxford University Press.

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*, *64*, 407–420.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Attention, Perception, & Psychophysics*, *53*, 372–380.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*(2), 202–227.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In W. J. Hardcastle, J. Lavers, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (pp. 489–520). Wiley-Blackwell, Hoboken, NJ. USA.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302–331.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese Tones. *Journal of the Acoustical Society of America*, *102*, 1864–1877.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378.

Myers, J. (2002). An analogical approach to the Mandarin syllabary. *Journal of Chinese Phonology*, *11*, 163–190.

Myers, J. (2010). Chinese as a natural experiment. *The Mental Lexicon*, *5*(3), 423–437.
Nixon, J. S., Chen, Y., & Schiller, N. O. (2015). Multi-level processing of phonetic variants in speech production and visual word processing: Evidence from Mandarin lexical tones. *Language, Cognition and Neuroscience*, *30*(5), 491–505.
Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.
Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*, 189–234.
Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395.
Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 113–134). Amsterdam, the Netherlands: IOS Press.
Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.
Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*, *60*(3), 355–376.
Ong, J. H., Burnham, D., & Escudero, P. (2015). Distributional learning of lexical tones: A comparison of attended vs. unattended listening. *PloS one*, *10*(7), e0133446.
Packard, J. L. (1999). Lexical access in Chinese speech comprehension and production. *Brain and Language*, *68*, 89–94.
Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge, UK: Cambridge University Press.
Peng, S.-H. (2000). Lexical versus 'phonological' representations of Mandarin Sandhi Tones. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology 5: Acquisition and the lexicon* (pp. 152–167). Cambridge, UK: Cambridge University Press.
Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of The Acoustical Society of America*, *130*(1), 461–472.
R Core Team. (2016). R: A Language and Environment for Statistical Computing. Version 3.3.1. Retrieved from http://www.r-project.org
Repp, B. H., & Lin, H.-B. (1990). Integration of segmental and tonal information in speech perception: A cross-linguistic study. *Journal of Phonetics*, *18*, 481–495.
Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy, 15*(6), 608–635.
Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, *134*(2), 1324–1335.
Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, 5 DOI: 10.3389/fpsyg.2014.01318.
Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.
Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*, 49–72.
Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, *18*, 281–295.
Shen, X. S., & Lin, M. C. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech*, *34*, 145–156.
Shih, C.-L. (1997). Mandarin third tone sandhi and prosodic structure. In J. Wang & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 81–124). Dordrecht, the Netherlands: Foris.

Shook, A., & Marian, V. (2013). The bilingual language interaction network for comprehension of speech. *Bilingualism: Language and Cognition*, *16*, 304–324.

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jTRACE: A simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, 49, 230–241.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech*, *53*(2), 273–293.

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231–249.

Strange, W. (Ed.). (1995). *Speech perception and linguistic experience: Issues in cross-language research*. Baltimore, MD: York Press.

Sulpizio, S., & McQueen, J. M. (2012). Italians use abstract knowledge about lexical stress during spoken-word recognition. *Journal of Memory and Language*, *66*, 177–193.

Tan, L. H., & Perfetti, C. A. (1998). Phonological codes as early sources of constraint in Chinese word identification: A review of current discoveries and theoretical accounts. In C. K. Leong & K. Tamaoka (Eds.), *Cognitive processing of the Chinese and the Japanese languages* (pp. 11–46). Dordrecht, the Netherlands: Springer Netherlands.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

Walker, G. (1989). Designing an intensive Chinese curriculum. In S. McGinnis (Ed.), *Chinese Pedagogy* (pp. 181–227). Columbus: The Ohio State University National Foreign Language Publications.

Wang, H. S. (1998). An experimental study on the phonetic constraints of Mandarin Chinese. In B. K. Tsou (Ed.), Studia Linguistica Serica (pp. 259–268). Hong Kong: City University of Hong Kong Language Information Sciences Research Center..

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, *113*, 1033–1043.

Wang, Y., Jongman, A., & Sereno, J. (2006). Second language acquisition and processing of Mandarin tones. In E. Bates, L. Tan, & O. Tzeng (Eds.), *Handbook of East Asian psycholinguistics* (pp. 250–257). Cambridge, UK: Cambridge University Press.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of The Acoustical Society of America*, *106*(6), 3649–3658.

Wanrooij, K., Escudero, P., & Raijmakers, M. E. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*(5), 307–319.

Wayland, R. P., & Guion, S. G. (2004). Training English and Chinese listeners to perceive Thai tones: A preliminary report. *Language Learning*, *54*(4), 681–712.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25.

Wiener, S., & Ito, K. (2015). Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers. *Language, Cognition and Neuroscience*, *30*(9), 1048–1060.

Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, *56*, 38–51.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*(1), 59–82.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585.

Wu, J., Chen, Y., van Heuven, V. J., & Schiller, N. O. (2014). Tonal variability in lexical access. *Language, Cognition and Neuroscience*, *29*(10), 1317–1324.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*, *95*, 2240–2253.
Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83.
Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics*, *17*, 1–31.
Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes, 14*(5–6), 609–630.
Zee, Y.-Y. (1980). Phonetic Studies of Chinese Tones. Ph.D. thesis, University of California, Los Angeles.