Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Individual variability in the distributional learning of L2 lexical tone

Seth Wiener [1], *Kiwako Ito* [2], *Shari R. Speer* [2]

[1] Department of Modern Languages, Carnegie Mellon University, USA
[2] Department of Linguistics, The Ohio State University, USA
sethw1@cmu.edu, ito.19@osu.edu, speer.21@osu.edu

Abstract

This study tested whether successful learners of an artificial tone language exhibit sensitivity to varying degrees of tonal informativeness, which has previously been shown to effect spoken word recognition in native Mandarin speakers. Twenty naïve listeners, whose L1 is American English, learned an artificial language in which each visual nonce symbol was arbitrarily associated with a Mandarin-like monosyllable and tone. The stimuli were designed to mimic Mandarin's uneven distribution of syllable-tone combinations; syllable frequency and the likelihood of a syllable co-occurring with a particular tone were manipulated across 4 days of training. The results showed that successful learners (those whose perception and production accuracy were consistently above the daily median) most accurately perceived and produced frequent syllables with probable tones and infrequent syllables with probable tones. Successful learners were least accurate in perceiving and producing infrequent syllables with least probable tones. Learners whose daily accuracy was below the median showed no such sensitivity to syllable-conditioned tonal probability. This finding supports the claim that L2 learners can be sensitive to statistical information available from novel input, and further demonstrates that statistical learning takes place even from an early stage of acquisition in successful L2 learners.

Index Terms: Lexical tone, second language acquisition, distributional learning, individual variability, spoken word recognition

1. Introduction

Mandarin Chinese uses four different F0 contours to distinguish otherwise identical (C)V(C) syllables. 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). For instance, /pan/ can mean 'class' with a high-level tone (*ban1*) or 'to handle' with a high-falling tone (*ban4*). Not all of the roughly 400 (C)V(C) syllables in Mandarin appear with all four tones. Due to the historical evolution of Mandarin tones, the lexicon contains syllable-tone gaps [1]. For example, *ban* is a nonword with a rising second tone as *ban2*. Moreover, Mandarin, like most tonal languages, exhibits a relatively high degree of syllable-tone homophony [2]. Syllable-tone combinations rarely have an unambiguous 1:1 mapping of sound to meaning. For instance, *ban4* can be written as at least nine different orthographic forms, all with semantically unrelated meanings. Together, homophony and syllable-tone gaps create a learnable distribution of syllable-tone combinations that speakers can track as part of their input. Some syllables frequently occur in speech with all tones possible in the language, while other syllables occur less frequently and often with only one or two highly probable tones. Thus, multiple segmental and suprasegmental cues, and their complex relative frequencies, result in a range of tonal informativeness. For example, the syllable *shi* appears with all four tones but most often as *shi4*. Among the over forty unique *shi4* morphemes is the copula verb, which contributes to *shi4* being the most common *shi* plus tone co-occurrence. This is juxtaposed with a syllable like *neng*, which only occurs as *neng2* and can ostensibly be recognized with segmental cues only. Recent experimental evidence has shown that native Mandarin speakers are sensitive to this distribution and draw on their knowledge of syllable-tone probabilities during spoken word recognition [3,4]. In particular, native speakers were found to most often predict probable tones for infrequent syllables, which, unlike frequent syllables, often carry far fewer homophones and therefore have higher tonal informativeness. This allows native listeners to rely less on purely acoustic-based processing and more on probability-based processing for the recognition of highly likely syllable-tone combinations.

Although acquisition of lexical tone in a second language (L2) is challenging for speakers of a non-tonal first language (L1) [5,6], statistical learning of tonal informativeness may be beneficial to early learners. The present study examines whether naïve, monolingual American English speakers exhibit the same sensitivity to tonal informativeness as native speakers and whether an individual's ability to track the distributional information of tone is a strong predictor of successful L2 learning. Previous studies on L2 tone learning have primarily focused on the individual variability in the learning of highly variable acoustic-phonetic cues [e.g., 7,8]. These studies show a wide range of learning ability; some learners are incredibly adept at learning tone and require little training, while other learners struggle to discriminate novel sounds despite ample training. [7,8] argue that the ability to attend to pitch direction and movement are strong predictors of good spoken language learners. While this ability is undoubtedly useful for listeners exposed to a closed set of syllable-tones, natural language processing involves learning the distributional correspondences among the sounds of a language and making predictions based on accruing experience with the language [9]. The degree to which such distributional learning influences L2 tone acquisition remains an open question. Therefore, in contrast with previous L2 tone studies, the present study explores individual variability in the use of tonal informativeness for the learning of L2 syllable-tone distributions.

To achieve these goals, we created a realistic artificial tone language that mimics Mandarin's distribution of homophony, syllable-tone gaps, and controlled syllable frequencies with

 doi: 10.21437/SpeechProsody.2016-110

syllable-conditioned tonal probabilities. As a result, the artificial language shared Mandarin's inherently asymmetric tonal distribution, allowing us to manipulate word-specific syllable-tone frequency as a vital part of the L2 input. This design allows us to address a confound present in previous L2 tone acquisition studies, which tested sets of syllable-tones that were fully symmetric. For example, in studies like [7,8] each syllable was presented with each tone, creating stimuli devoid of any variation in distributional properties. This is especially important given converging evidence that speakers are universally sensitive to relative frequency effects drawn from the distribution of both L1 and L2 input at a very early stage of acquisition [10,11,12].

Furthermore, previous L2 tone learning experiments have primarily examined the acquisition and processing of purely acoustic-phonetic information in the absence of visual stimuli. For example, [13,14] taught and monitored the L2 learning of non-lexical tone stimuli while [5,15] examined L2 participants' ability to discriminate and categorize Mandarin tones but only as acoustic-phonetic information. Recent studies have shown that the use of non-lexical tonal stimuli often results in task-specific perception strategies which may not reflect the processes and mechanisms involved in real tonal word learning [8]. To address this issue, the present study used sound-to-image pairs to explore L2 word learning and tone's role during lexical access [e.g., 7,8]. Thus by incorporating an asymmetric syllable-tone distribution and pairing these combinations with meaningful symbols, the present study aims for a careful examination of realistic L2 tonal acquisition in a reduced period of time.

2. Experiment

2.1. Methods

2.1.1. Participants

Twenty students at a Midwestern U.S. university (13 female; 7 male; mean age: 20) took part in the experiment. All participants spoke English as their L1. Due to language requirements in most U.S. high schools and colleges, all participants had studied an L2, but no participant self-rated as a fluent speaker of their L2 (questionnaire scale: 1=beginner, 5=fluent; mean score: 2.1). Furthermore, no participant had ever studied Mandarin or any other tonal language. All participants received a small payment.

2.1.2. Stimuli

Twenty-four CV syllables were paired with four specific tonal contours (directly comparable to those of Mandarin). All CV syllables made use of Mandarin phonemes and thus were syllable gaps easily produced by native Mandarin speakers (i.e., similar to the English nonce word *blick*). Though this maximally yields 96 unique syllable-tone combinations, only 82 combinations were used. This ensured roughly the same percentage of syllable-tone gaps in the artificial language as in Mandarin. To equally ensure the same proportion of homophony as in Mandarin, 48 homophones were added, resulting in 130 total nonce words. Each nonce word was then given a unique black and white symbol. Thus, like Mandarin, certain syllable-tone combinations resulted in numerous homophones disambiguated only through the orthography. Figure 1 shows an example of five nonce symbols, all sharing the same syllable-tone combination *pe2*.

Figure 1: *Black and white nonce symbols for "pe2."*

Of these 130 items, 64 served as the test items while the other 66 served as filler items and allowed for the distribution of the artificial language to emerge. Within the test items, two factors – syllable frequency and tonal probability – were crossed to create four test conditions. To manipulate syllable frequency, the number of exemplars participants were exposed to was either increased or decreased. Within the test items, 32 had high syllable frequency (F+), while 32 had low syllable frequency (F-). To avoid confounding syllable frequency with the phonemic construction of syllables, each consonant onset appeared in both a F+ and F- syllable. As for tonal probabilities, each syllable in the test items appeared with one tone as the most probable (P+) and one tone as the least probable (P-) while the other two tones had identical middle range probabilities. This resulted in four test conditions: F+P+, F+P-, F-P+, F-P-. To control the occurrence of target symbols, tonal probabilities were manipulated by increasing or decreasing exposure to filler homophones. For example, the syllable-tone combination *pe2* was presented multiple times using the four left-most nonce symbols in Figure 1. By repeatedly showing these four symbols, the probability of tone 2 appearing with the syllable *pe* greatly increased. The right most nonce symbol in Figure 1 served as the test item and appeared exactly the same number of times as the respective test item symbols (i.e., non-fillers) for *pe1*, *pe3*, and *pe4*.

All auditory stimuli were recorded by a monolingual 28 year-old female from Beijing (who spoke no other dialects) at 16 bits/44,100 Hz. Two additional native speakers from China correctly identified 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 in duration than tones 1 and 4 [16,17].

2.1.3. Training and testing procedure

Participants came to the lab for 30-minute sessions on four consecutive days. Participants were randomly assigned to one of two lists, with each list differing in its controlled distribution. For example, in list one tone 2 was the most probable tone for *pe*, while in list two it was the least probable tone for *pe*. Daily training and testing consisted of four tasks in the same order each day: passive listening, shadowing, naming and 4-alternative forced-choice. Participants were seated in front of a computer in a sound booth and wore headphones. For the naming task, a nonce 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. For the final task, participants completed 32 4-alternative forced-choice (4AFC) trials (16 target trials, 16 fillers). Four symbols were presented at a time on a monitor, while participants simultaneously heard one symbol's audio label. Participants were told to click on the symbol that matched the perceived audio. After clicking, a red box appeared around the correct target in order to provide feedback. There were 16 target trials, four in each

experimental condition, showing the target and three other trained test items: a tonal competitor, a rhyme competitor and a distractor. For example in the high frequency, high probability (F+P+) condition, *pe2* served as the target. The three other on-screen items included a tonal competitor, which shared the same syllable but had the opposite tonal probability (e.g., *pe4* as F+P-), a rhyme competitor, which shared the same vowel and tone but had a different onset (e.g., *fe4*), and a distractor, which had a unique syllable and tone (e.g., *riu1*).

2.2. Predictions

Following previous studies on L2 tone acquisition [e.g., 7,8], we predict high individual variability in the learning of the artificial language. If learners are unable to track distributional information, accuracy in the naming and identification tasks should be the same regardless of syllable frequency or tonal probability. If learners are sensitive to syllable frequency of occurrence, they will correctly name and identify high frequency F+ syllables more often than they correctly name and identify low frequency F- syllables. If learners track syllable-conditioned tonal probabilities, they will correctly name and identify syllables with high probability P+ tones more often than they name and identify those with low probability tones P-. If learners, like native speakers, selectively track tonal probability only for infrequent syllables with higher tonal informativeness, then participants will correctly name and identify low frequency syllable targets with high probability tones (F-P+) more often than any other combination [e.g., 3,4]. Following [7], if a difference emerges between good and poor learners based on median accuracy, an interaction will be found in which learners' performance interacts with one (or more) of the distributional variables. In particular, if good learners are more native-like than poor learners, we expect a three-way interaction between learning group, syllable frequency and tonal probability.

2.3. Results

For each task, good and poor learning was assessed by calculating the daily median accuracy for all participants. Following [7], good learners (GL) had a consistent daily accuracy for both naming and 4AFC identification in the top half (n=7), while poor learners (PL) had a consistent daily accuracy in the bottom half (n=13). That is, no participant's accuracy scores placed him/her in the GL group on one day but the PL group on another. For both tasks, day four results were analyzed using mixed effects regression models (linear for the naming task and logistic for the 4AFC task) using the *lme4* package [18] in R (version 3.2.1). Syllable frequency, tonal probability and learning groups were treated as sum coded factors. The random effects structure included subject and item intercepts, by-subject random slopes for syllable frequency and tonal probability and by-item random slopes for learning groups.

2.3.1. Naming task

Naming responses were coded by three trained transcribers (inter-rater reliability: 95%). Utterances in which a participant made no response or uttered an inaudible response were removed from subsequent analysis (< 2%). Figure 2 shows individual naming accuracy across the four days, faceted by the four test conditions. Each participant's mean accuracy is plotted as a smaller gray point, while overall learning group means are plotted as larger black points. The circles indicate good learners (GL), while the triangles indicate poor learners (PL).

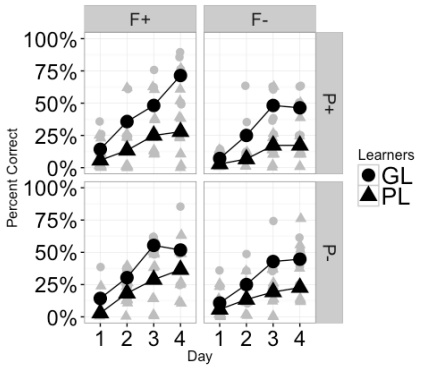

Figure 2: *Participants' daily average on naming task.*

Figure 2 shows a considerable amount of individual variability, with naming accuracy on the first day roughly the same regardless of condition or learning group. As expected, however, GL accuracy was higher in every condition across days two, three and four. For the GL group, day four accuracy was highest for the F+P+ condition (71%) and lowest for the F-P- condition (45%). In contrast, PL day four accuracy was highest for the F+P- condition (36%) and lowest for the F-P+ group (17%). The PL showed neither a sensitivity to syllable frequency nor tonal probability, by naming targets of all four conditions at roughly the same rate. This contrasts sharply with the GL who showed sensitivity to both syllable frequency and tonal probability by naming F+ syllables and P+ tones more accurately than the PL. A linear regression model on the fourth day naming results showed a main effect of syllable frequency (β = 0.16, *SE* = 0.05, *t* = 3.11, $p < .01$), a main effect of tonal probability (β = -0.10, *SE* = 0.05, *t* = -2.08, $p < .05$) and the expected main effect of learning group (β = -0.27, *SE* = 0.07, *t* = -3.78, $p < .01$) in which the GL had a higher overall naming accuracy than the PL. Additionally, a two-way interaction between tonal probability and learning group was found (β = 0.17, *SE* = 0.06, *t* = 2.78, $p < .05$). Further subset analyses revealed that the main effect of frequency was driven by the GL as they correctly named F+ targets more often than F- targets (β = 0.09, *SE* = 0.04, *t* = 2.13, $p < .05$). The main effect of tonal probability and the two-way interaction of tonal probability and learning group were driven by the GL group's overall significantly higher accuracy for P+ targets (β = -0.22, *SE* = 0.04, *t* = -5.10, $p < .001$). Subgroup analyses for the PL group showed no significant difference for either frequency or tonal accuracy main effects.

2.3.2. 4AFC identification task

Only the 16 target trials were analyzed in the identification task. Figure 3 shows individual 4AFC accuracy across the four days, faceted by the four test conditions and plotted by individual and learning group means. The individual means corroborate findings from [7,8] showing a high degree of variability across learners, and larger variability in perception than the naming production task. For some PL, accuracy on the last day was still 0% for F- targets, while some GL reached 100% accuracy for F+ targets. On day 1, GL identified targets slightly better than PL, with GL showing the largest accuracy advantage in the F-P+ condition. The F-P+ condition continued to be the only condition for which a difference

between GL and PL emerged. By day four, only the F-P+ condition showed a clear separation as the GL had an accuracy of 72% while the PL had an accuracy of only 35%. Thus the difference between good and poor learning was almost entirely seen in perception of F-P+ targets. For P- targets, the GL performed only slightly better than the PL across the four days, but neither the F+P- nor F-P- condition showed a large difference between the two learning groups.

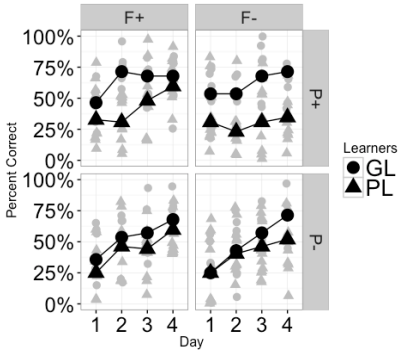

Figure 3: *Participants' daily average on 4AFC task.*

A logistic regression model on day 4's results found the expected main effect of learning group (β = -0.78, *SE* = 0.25, *z* = -3.08, $p < .01$) such that the GL accurately identified the target more often than the PL. A marginal interaction was found between learning group and tonal probability (β = 0.39, *SE* = 0.23, *z* = 1.71, *p* = .07) indicating that GL group performed slightly better at P+ tones than the PL group. A subset analysis of the F- targets indicated that this probability difference between groups was significant in F-P+ condition (β = -1.21, *SE* = 0.36, *z* = -3.32, $p < .001$) but not in the other three conditions. Thus, for infrequent targets, GL were more sensitive to tonal probability than PL.

3. Discussion

This study set out to test learners on a Mandarin-like artificial tonal language with controlled syllable-tone distributional properties, in order to explore whether naïve listeners from a non-tonal L1 exhibit sensitivity to syllable frequency and varying degrees of tonal informativeness. Perception and production were measured across four days of testing, with good and poor learner groups defined as scoring above and below the daily median, respectively. Using a 4AFC perception task, an accuracy difference was found between GL and PL, and a marginal trend was found in which GL tended to identify syllables with more probable tones better than PL did. Further subset analysis indicated that for F-P+ targets, i.e., targets carrying a higher degree of tonal informativeness, GL's identification was significantly more accurate than PL's. This pattern of results is very similar to those found in [3,4] which demonstrated that native mono-dialectal Mandarin speakers make use of tonal probabilities primarily for infrequent syllables that appear in the language with few tones and consequently a higher tonal informativeness. Thus the present study's perception results strongly suggest that naïve L2 learners track tonal informativeness in identification tasks.

Similarly, results from the naming production task showed that GL were more accurate at producing F+ syllables than PL, and more accurate at producing P+ tones than PL. Taken together, these results strongly support previous claims that L2 learners can be sensitive to statistical information from novel input [e.g., 10,11,12] in not only perception but also production. Our results indicate that after the first 30 minutes of training, good learners began to show small perception and production advantages over poor learners. From the second day on, i.e., after an hour of perceptual training, good learners repeatedly demonstrated better perception and production of tone in F+ syllables and for P+ tones. By the fourth day, good learners showed a robust sensitivity to tonal informativeness for F-P+ targets in the perception task and for F+P+ and F-P+ targets in the naming task. It is important to point out that unlike the results found in native Mandarin speakers [3,4], the present results showed a high degree of individual variability [e.g., 7,8], suggesting that only good learners were able to identify when tonal informativeness was highest through distributional learning of the input. To underscore this point, Figure 4 shows day four naming (top plot) and 4AFC (bottom plot) accuracy for the top four GL and bottom five PL participants.

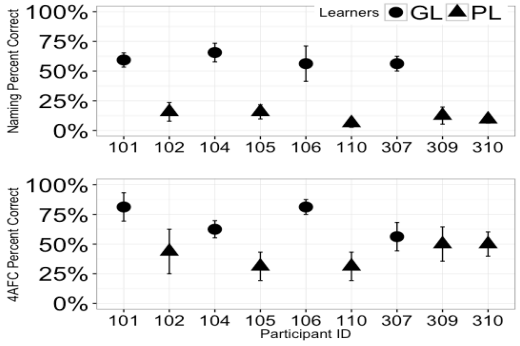

Figure 4: *Day 4 GL and PL individual naming and 4AFC accuracy*

Figure 4 highlights an important point and the direction of our current follow-up study: the difference between GL and PL is largely derived from the naming production task. The regression results support this claim as the GL named F+ and P+ targets more accurately than PL. In the 4AFC task, PL were at or slightly above chance (25%) even on the fourth day, while GL were above 50% accuracy. These results suggest that accurate production may play a larger, critical role in the ability to track statistics of an L2. Accurate perception of novel contrasts may not be sufficient to effectively track the contrasts' distributions. While [7,8] argue that successful learners of tone are better able to attend to acoustic-phonetic details, such as pitch direction, our results suggest that successful learners may first need to learn to produce these pitch direction differences and by doing so, these learners are better able to track distributional input such as syllable frequency and syllable-tone co-occurrences. We are currently collecting native Mandarin speaker ratings of the present study's productions in order to see if GL's tone production are objectively more native-like than PL's and whether phonetic features such as pitch direction account for this difference or whether statistics such as frequency and probability of syllable-tones better explains this difference. This will allow us to further unpack the perception-production link and explore to what degree distributional learning affects accurate productions of novel contrasts.

4. Acknowledgements

This research was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (BCS-1451677) to K.I. and S.W.

5. References

[1] H. S. Wang, "An experimental study on the phonotactic constraints of Mandarin Chinese," *Studia Linguistica Serica,* pp. 259-268, 1998.

[2] M. Yip, *Tone*. Cambridge U. Press, 2002.

[3] S. Wiener and K. Ito, "Do syllable specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers," *Language, Cognition and Neuroscience,* vol. 30, no. 9, pp. 1048-1060, 2015.

[4] S. Wiener and K. Ito, "Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners," *Journal of Phonetics,* vol. 56, pp. 38-51, 2016.

[5] Y. Wang, M. Spence, A. Jongman and J. Sereno, "Training American listeners to perceive Mandarin tones," *Journal of Acoustic Society of America,* vol. 106, no. 6, pp. 3649-3658, 1999.

[6] X. S. Shen, "Toward a register approach in teaching Mandarin tones," *Journal of Chinese Language Teachers Association,* vol. 24, pp. 27-47, 1989.

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

[8] P. C. M. Wong and T. K. Perrachione, "Learning pitch patterns in lexical identification by native English-speaking adults," *Applied Psycholinguistics,* vol. 28, no. 4, pp. 565-585, 2007.

[9] D. Norris and J. M. McQueen, "Shortlist B: A Bayesian model of continuous speech recognition," *Psychological Review,* vol. 115, pp. 357-395, 2008.

[10] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science,* vol. 274, no. 5294, pp. 1926-1928, 1996.

[11] B. Ambridge, E. Kidd, C. F. Rowland, and A. L. Theakston, "The ubiquity of frequency effects in first language acquisition," *Journal of Child Language,* vol. 42, no. 2, pp. 239-273, 2015.

[12] N. C. Ellis, "Frequency effects in language processing," *Studies in Second Language Acquisition,* vol. 24, no. 2, pp. 134-188, 2002.

[13] J. Leather, "*F0* pattern interference in the perceptual acquisition of second language tone," in *Sound Patterns in Second Language Acquisition,* 1987, pp. 59-81.

[14] Y. Wang, D. M. Behne, A. Jongman, and J.A. Sereno, "The role of linguistic experience in the hemispheric processing of lexical tone," *Applied Psycholinguistics,* vol. 25, no. 3, pp. 449-466, 2004.

[15] C.-Y. Lee, L. Tao, and Z. S. Bond, "Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners," *Journal of Phonetics,* vol. 37, pp. 1-15, 2009.

[16] A. T. Ho, "The acoustic variation of Mandarin tones," *Phonetica,* vol. 33, pp. 353-367, 1976.

[17] J. M. Howie, *Acoustical Studies of Mandarin Vowels and Tones*, Cambridge U. Press, 1976.

[18] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear mixed-effect models using s4 classes*, http://cran.r-project.org/package=lme4