Context effects on second-language learning of tonal contrasts

Charles B. Chang[a)]
Linguistics Program, Boston University, 621 Commonwealth Avenue, Boston, Massachusetts 02215, USA

Anita R. Bowles
Rosetta Stone, 135 West Market Street, Harrisonburg, Virginia 22801, USA

(Received 25 November 2014; revised 17 November 2015; accepted 20 November 2015; published online 21 December 2015)

Studies of lexical tone learning generally focus on monosyllabic contexts, while reports of phonetic learning benefits associated with input variability are based largely on experienced learners. This study trained inexperienced learners on Mandarin tonal contrasts to test two hypotheses regarding the influence of context and variability on tone learning. The first hypothesis was that increased phonetic variability of tones in disyllabic contexts makes initial tone learning more challenging in disyllabic than monosyllabic words. The second hypothesis was that the learnability of a given tone varies across contexts due to differences in tonal variability. Results of a word learning experiment supported both hypotheses: tones were acquired less successfully in disyllables than in monosyllables, and the relative difficulty of disyllables was closely related to contextual tonal variability. These results indicate limited relevance of monosyllable-based data on Mandarin learning for the disyllabic majority of the Mandarin lexicon. Furthermore, in the short term, variability can diminish learning; its effects are not necessarily beneficial but dependent on acquisition stage and other learner characteristics. These findings thus highlight the importance of considering contextual variability and the interaction between variability and type of learner in the design, interpretation, and application of research on phonetic learning.
[http://dx.doi.org/10.1121/1.4937612]

[MSS] Pages: 3703–3716

I. INTRODUCTION

Although phonological contrasts between speech sounds are often conceptualized in terms of discrete invariable phonemes, speech sounds are, in fact, remarkably variable. Among the main sources of variability is phonological context (i.e., the environment in which a sound occurs, including adjacent sounds and prosodic position), which is associated with two kinds of variability: allophonic alternation (which may not have an immediate articulatory motivation) and coarticulatory modification (based in articulatory influence from nearby sounds). For example, the voiceless alveolar plosive (/t/) of American English is, in certain contexts, not actually realized as a voiceless alveolar plosive: in intervocalic contexts it is typically realized as a voiced tap (e.g., later [leɪɾɚ], cf. late [leɪt]), while in a cluster with postalveolar/retroflex /ɹ/ it is realized homorganically as a postalveolar plosive (e.g., trail [t̠ɹeɪl], cf. tail [teɪl]). These kinds of contextual variability—both from alternation and from coarticulation—are one reason for the difficulty of acquiring novel phonological contrasts in a second language (L2). Not only do L2 learners need to overcome fundamental biases from their native language (L1) in the way they process a given kind of variability (e.g., as non-contrastive variation to be abstracted away from), they also need to structure the variability in a different manner—namely, in terms of the contrastive sounds of the L2.

Contextual variability, however, is a characteristic not only of segmental categories such as stop consonants but also of suprasegmental categories such as lexical tones. As such, contextual variability is likely to play an important role in how tones are acquired, yet studies of tone acquisition have largely focused on tones in isolation. This limitation of the literature is problematic for two reasons. On the one hand, the learning of isolated tones can, in principle, provide only a partial picture of tone learning; on the other hand, data from isolated tones may not accurately represent how tones are learned in languages with multi-tone words. In Mandarin Chinese, for example, the majority of the lexicon (over 70%) consists of words containing two or more syllables (Jin, 2011), such that any given tone usually occurs adjacent to another tone within the same word. Thus to the extent that L2 learners of Mandarin encounter tones in disyllabic, rather than monosyllabic, contexts, previous findings on Mandarin tone learning in isolated monosyllables may not provide a realistic picture of how L2 learners build up a vocabulary of tonally contrasting lexical items.

In light of this disparity between the tone learning literature and the typical tone language lexicon, the current study investigated the effects of contextual variability on *ab initio* acquisition of tonally contrasting lexical items, focusing on the case of native English speakers learning Mandarin. The rest of this paper is devoted to describing this study in more detail. In Sec. II, we review the literature on the role of variability in speech learning, the contextual variability of Mandarin tones, and Mandarin tone learning by non-tonal language speakers before motivating specific predictions for

a)Electronic mail: cc@bu.edu

the relative difficulty of acquiring Mandarin tonal contrasts in different contexts. We then present the results of a word learning experiment comparing monosyllabic items and disyllabic items of different types, which support the main hypothesis that contextual variability diminishes the initial learning of tonal contrasts. Finally, we discuss the implications of the study's findings for predicting L2 learning outcomes and future directions for research on L2 speech learning.

II. BACKGROUND

A. Phonetic variability in second-language learning

The phonetic variability characteristic of natural speech has long been known to affect both speech perception and speech learning. Spoken word recognition, for example, is significantly worse when stimuli are made to vary in dimensions such as speech rate and talker identity, and introducing multiple sources of variability simultaneously compounds the effect of each individual source of variability (Sommers *et al.*, 1994; Kirk *et al.*, 1997). These findings suggest that the perceptual normalization processes involved in abstracting from phonetically variable input to phonologically constant categories place non-trivial demands on cognitive resources, making it logical to expect variability in L2 input to diminish the learning of novel L2 categories.

In fact, input variability does make speech learning more difficult at first; however, in the long run, controlled exposure to certain types of variability seems to be helpful because it encourages formation of robust representations that generalize to novel stimuli. For example, introducing talker variability during L1 dialect identification training results in poorer identification for familiar talkers (as the variability interferes with the learning of specific exemplars at early stages of perceptual learning) but, ultimately, better identification for unfamiliar talkers (Clopper and Pisoni, 2004). Along the same lines, the high acoustic variability of L1 speech produced by non-native talkers tends to pose difficulties for perceptual learning by native listeners especially for relatively confusable vowel categories (Wade *et al.*, 2007), while greater talker variability in exposure to unfamiliar L2 vowel contrasts is sometimes found to diminish the initial learning of these contrasts (Kingston, 2003). On the other hand, both L2 identification and discrimination training procedures incorporating talker variability help experienced learners make gains in L2 identification performance that persist over time to a similar degree (Flege, 1995). L2 word learning by novice learners is also better when the input contains talker variability as well as types of within-talker variability that typically decrease L1 word recognition performance (e.g., rate variability); however, linguistically irrelevant types of within-talker variability, such as variability in amplitude and fundamental frequency (f_0) for L1 English/L2 Spanish, seem to have little effect on learning (Barcroft and Sommers, 2005; Sommers and Barcroft, 2007).

As with talker variability, contextual variability can either diminish or enhance L2 learning, and the nature of the effect seems to depend on the stage of learning. At early stages of learning, greater contextual variability fails to result in a learning benefit and, instead, tends to result in poorer outcomes. L1 English speakers with no knowledge of German, for example, are worse at learning certain German vowel contrasts when training stimuli contain greater variation in consonantal context (Kingston, 2003). When learners have prior experience with the L2, however, L2 perceptual training that incorporates contextual variability results in significant gains in performance. Perhaps the most well-known example of this pattern is the case of training L1 Japanese experienced learners of English on the English /l/-/ɹ/ contrast, which shows that a high-variability phonetic training (HVPT) paradigm including variation in context as well as talker can improve both perception and production (Lively *et al.*, 1993; Bradlow *et al.*, 1997; Iverson *et al.*, 2005).

Although HVPT has generally been found to improve the L2 perception of experienced learners, it does not consistently benefit the initial learning of L2 contrasts. In particular, the effects of HVPT differ according to the type of learner with greater variability actually diminishing the learning of individuals with weaker initial perceptual abilities (Perrachione *et al.*, 2011). In short, the findings of the L2 speech literature suggest that while exposure to phonetic variability has the potential to provide learning benefits, its effects are modulated by additional factors such as the stage of learning and individual learner characteristics. Thus in initial L2 learning of Mandarin tonal contrasts, contextual variability may diminish, rather than enhance, learning because the novice learner may not be prepared to benefit from highly variable input.

B. Mandarin tone and variability

Mandarin is standardly analyzed as containing four distinctive tones, in addition to a fifth, "neutral" tone, often analyzed as the absence of one of the four full-fledged tones due to its restriction to weak, unstressed syllables (Duanmu, 2007). The canonical form of each of the four main tones is typically identified with its pitch contour in isolation: a high level contour for tone 1 (T1); a mid-to-high rising contour for tone 2 (T2); a low falling-rising contour for tone 3 (T3); and a high-to-low falling contour for tone 4 (T4). In the five-point tonal representation system of Chao (1930), where "1" indicates the low end of a talker's pitch range, these canonical contours are represented as, respectively, [55], [35], [214], and [51]. Pitch is the primary cue to tone, but duration, phonation, and amplitude properties provide secondary cues allowing native perception to remain surprisingly good when the acoustic correlate of pitch (f_0) is unavailable (Liu and Samuel, 2004; Kong and Zeng, 2006). For example, T3 in isolation is significantly longer than the other three tones, whereas T4 is significantly shorter; furthermore, unlike T1 and T2, both T3 and T4 often dip into creaky phonation during a portion of their duration (Chao, 1933; Chang and Yao, 2007).

Although the Mandarin tones each have a canonical pitch contour, they all show substantial contextual variability contributed by two sources: allotonic alternation and tonal coarticulation. In regard to alternation, T3 in particular is associated with two allotones of interest—a "full" [214]

contour, which occurs before a pause (especially in isolation), and a "half" [21] contour, which occurs before any of the other three tones.[1] In the final syllable of a disyllabic word or phrase, either allotone may occur; however, [21] is more common, as [214] here is associated with emphasis (Duanmu, 2007, pp. 238–239). In regard to coarticulation, the form of a given tone is systematically influenced by the contour of an adjacent tone, which causes modifications to onset and offset f_0 values as well as overall f_0 height that are large enough to be perceptible to native listeners (Shen, 1990; Shen and Lin, 1991). Notably, these effects are bidirectional but asymmetric. Whereas perseverative coarticulation is usually assimilatory, anticipatory coarticulation is usually dissimilatory; furthermore, the magnitude of perseverative effects is larger than that of anticipatory effects (Xu, 1997).

Such coarticulatory perturbations make tone perception significantly more difficult in disyllables compared to monosyllables. Native Mandarin listeners generally compensate for these kinds of perturbations, such that their ability to identify coarticulated tones in context remains high; however, accuracy drops with more profoundly coarticulated tones, especially when the tones are presented out of context (Xu, 1994). In addition, L1 English listeners with no knowledge of a tone language show poorer Mandarin tone discrimination with disyllabic stimuli than with monosyllabic stimuli (Berkowitz, 2010). Furthermore, computers trained to recognize Mandarin tones perform worse with target items of higher syllable counts (Yang *et al.*, 1988). This positive relationship between syllable count and perceptual difficulty suggests that tonal contrasts will be considerably harder to acquire in disyllabic words than in monosyllabic words. Nevertheless, the literature on L2 learning of Mandarin has mostly focused on learning in monosyllables.

C. Tone learning by non-tonal language speakers

Perceptual training has been shown in several studies to improve the perception, processing, and/or production of tones by non-tonal L1 speakers. For example, HVPT using monosyllabic Mandarin stimuli in isolation (albeit with diverse segments and syllable structures) results in significant gains in the tone identification performance of L1 English speakers with one to two semesters of prior Mandarin study; these gains, moreover, generalize to new stimuli and talkers, persevere 6 months after training, and extend to tone production (Wang *et al.*, 1999; Wang *et al.*, 2003). Crossover benefits of perceptual training are also found in L1 Dutch novice learners of Mandarin, who demonstrate acquired knowledge in production even if trained strictly on perception, and vice versa (Leather, 1996). Notably, Leather's results were obtained with a low-variability training paradigm in which the bulk of early stimulus exposure was to one talker uttering one monosyllabic minimal quadruplet. However, when training exposure contains moderate contextual variability introduced by different syllable types, similar benefits of perceptual training are found in semi-novice L2 learners (students at week 6 of an elementary Mandarin course) of both tonal and non-tonal L1 backgrounds (Wang, 2013).

Like Leather (1996), most studies on Mandarin tone learning have focused on monosyllabic stimuli, but some recent studies have included longer stimuli. Hao (2012) shows that tone identification for both L1 English as well as L1 Cantonese experienced learners of Mandarin is better in monosyllables than in either syllable of disyllables and, furthermore, better in the final syllable of a disyllable than in the initial syllable. In addition, analyses of error patterns reveal that for both learner groups the most error-prone tones are T2 and T3, which are frequently confused with each other (in line with previous findings on the high confusability of T3; Gottfried and Suiter, 1997); the next most common confusion types for L1 English learners are T1 being perceived as T2 (and vice versa) and T3 being perceived as T4. While Hao (2012) compares monosyllables to disyllables, Ding (2012) focuses exclusively on disyllables, finding that tone identification accuracy for L1 German experienced learners of Mandarin is similar between initial and final syllables (although most of the items tested were words already familiar to the participants).

Whereas most of the preceding findings are based on metalinguistic tasks such as tone identification, another body of research has examined L2 tone learning through the lens of word learning, a task that is arguably more representative of the L2 acquisition process. This research differs from tone identification studies in focusing on novice learners (L1 English speakers) with no prior exposure to a tone language rather than experienced learners, but the findings are convergent in showing that over the course of a multi-session study, novice learners make significant gains in acquisition of a small tonal lexicon consisting of English-like nonce words combined with Mandarin-like tone contours. However, there is considerable individual variation in learning performance, which is correlated with several experiential, behavioral, and neural variables (Wong and Perrachione, 2007; Chandrasekaran *et al.*, 2010; Wong *et al.*, 2011).

One of the variables affecting tonal word learning is initial perceptual ability, which interacts with type of training to result in different effects of phonetic variability in training exposure. A large amount of talker/token variability is beneficial for learners with relatively strong perceptual abilities (enhancing their ability to generalize to novel stimuli) but is actually detrimental for learners with relatively weak perceptual abilities; furthermore, although talker/token variability ultimately improves "strong" perceivers' generalization ability, it also has the effect of slowing down their learning (Perrachione *et al.*, 2011). This is consistent with the fact that a HVPT procedure for discrimination of Thai tones improves performance for novice learners of tonal L1 backgrounds (who are already familiar with discriminating pitch patterns at the lexical level) but not for novice learners of non-tonal L1 backgrounds (Wayland and Guion, 2004). Given these findings, it is reasonable to suppose that contextual variability might also have the effect of diminishing initial tone learning—especially for non-tonal L1 speakers—and this is the hypothesis tested in the current study.

D. Research questions and predictions

In light of previous findings showing no benefit or even detrimental effects of phonetic variability in perception and learning of L2 phonological contrasts, the current study investigated the effects of contextual phonetic variability on initial L2 perceptual learning of Mandarin tonal contrasts. Initial learning and contextual variability were the empirical focus of the study for two reasons: training studies reporting benefits of input variability have largely focused on experienced learners, while prior work on tone learning has been biased toward isolated target forms, thus removing context as a factor influencing acquisition. The central hypothesis was that acquisition outcomes at this early stage would show an inverse relationship with a tone's contextual variability (i.e., the more variable the tone, the less robust its learning). This hypothesis led to four main predictions regarding the acquisition of tonally contrasting Mandarin words by L1 English novice learners.

First, given the disadvantage observed for multisyllabic words in tone perception and recognition, tonal contrasts were predicted overall to be harder to acquire in disyllabic words than in monosyllabic words. Although the fact that disyllables are longer than monosyllables (and, therefore, impose a higher short-term memory load) was expected to contribute to the higher difficulty of tone learning in disyllables, the difficulty of disyllables was expected to follow primarily from the relatively greater phonetic variability of tonal contours in disyllabic contexts. As such, compared with errors on monosyllables, errors on disyllables were expected to be not only more numerous but also more biased toward tonal errors as opposed to segmental errors.

Second, the relative learnability of a tone was predicted to differ across contexts according to the overall perceptual distinctness of that tone compared with other tones in that context. Thus T3, for example, was expected to be relatively easy to acquire in isolation because in this context, it is distinct from the other tones not only in terms of f_0 contour but also in terms of duration and phonation. Although these secondary cues are not contrastive in English, they provide information that English speakers are able to remember and, moreover, use in speech perception tasks, including discrimination of T3 from T2 (Blicher *et al.*, 1990; Trude and Brown-Schmidt, 2012), such that differences between tones in these phonetic dimensions are likely to enhance perceptual distinctness for L1 English learners. Consequently, where these differences are attenuated (such as in the first syllable of disyllables), T3 was expected to show less of a learning advantage.

Third, the learnability of a given tone in a disyllabic word was predicted to vary across positions according to the degree of divergence of the tone contour from its canonical isolation form. Consequently, T3 and T4 in particular were expected to be learned less successfully in the first (pre-final) syllable of disyllables than in the second (final) syllable, for two reasons: (1) the different allotonic realization of T3 as [21] in pre-final position (i.e., the lack of final rise found in isolation), and (2) the tendency for T3 and T4 to dip in pitch less in pre-final than in final position (and, thus, to be realized with less glottalization). Both these phonetic facts were expected to make pre-final instances of T3 and T4 sound substantially different from their canonical form.

Finally, the learnability of tones in disyllables was also predicted to be lowered by the acoustic consequences of tonal coarticulation, especially the significant tonal perturbations resulting from "tone clash"—that is, a mismatch between the offset and onset f_0 levels of adjacent tone contours. For example, T1 (which ends high) was expected to be less successfully acquired preceding a tone starting lower (e.g., T2) than preceding a tone starting similarly high (e.g., T1) because coarticulation in the former case would result in a falling contour for T1 that could be confused with the high falling tone, T4. For the same reason, T4 was expected to be less successfully acquired preceding a tone starting high (e.g., T1) than preceding a tone starting lower (e.g., T2).

In short, we predicted that context-dependent differences in phonetic variability would systematically affect the learnability of Mandarin tones, resulting in disparities in learning outcomes across different contexts. To test these predictions, native English speakers with no prior tone language experience were recruited to learn a small Mandarin lexicon comprising a variety of word types. Part of a larger correlational study examining predictors of successful tone learning (Bowles *et al.*, 2015), the learning study was designed both to provide a global measure of tone learning and to examine variation in the acquisition of different tonal contrasts. The results we present in the following text focus on context-dependent differences in the acquisition of tonal contrasts.

III. METHODS

A. Participants

Learner participants were recruited from the University of Maryland community and paid for their participation. A total of 166 native speakers of American English completed the study in its entirety; they reported no prior experience with a tone language and no history of hearing, speech, or language difficulties. From this sample, six participants were excluded on the basis of response times or behavior during a session that suggested they were not paying attention during the tasks completed. Thus there were 160 participants included in the current analysis (103 female, 57 male; mean age 21.7 yr, SD 2.5). Most were college-educated and had studied at least one foreign language in high school and/or college (most often Spanish or French).

B. Stimuli

1. Precursor tone perception tasks

Prior to beginning the focal task of Mandarin word learning, participants completed several other tasks designed to measure constructs related to pitch processing, language learning aptitude, and general cognitive ability. Two of these tasks involved tone perception (and are thus relevant to the interpretation of performance in the word learning task): the first was a tone identification task, while the second was a tone discrimination task.

The stimuli in both tone perception tasks were monosyllabic and recorded in the same manner as the stimuli for the word learning task (see Sec. III C). The talker for the identification stimuli (as well as for one set of tokens of the discrimination stimuli) was a female native speaker, while the second talker for the discrimination stimuli was a male native speaker; both talkers were in their 20s, born and raised in mainland China with Mandarin as the primary language spoken at home, and had moved to the U.S. the preceding year. Tone identification stimuli consisted of 80 items in the form of 20 tonally minimal quadruplets (e.g., /xɑʊ˥/ "wormwood," /xɑʊ˧/ "bold and unconstrained," /xɑʊ˨˩/ "good," /xɑʊ˩˧/ "number").[2] Tone discrimination stimuli consisted of 48 items in the form of 24 tonally minimal pairs (pronounced by different talkers), which were evenly distributed over all possible pairs of tones and comprised syllables different from those in the tone identification stimuli.

2. Word learning task

The stimuli for the word learning task were recorded by six talkers recruited from the Mandarin-speaking population in the U.S. to match the background of the talkers in previous studies (Shen, 1990; Xu, 1997). These six talkers (three female, three male; mean age 23.2 yr, SD 2.3) were native Mandarin speakers born and raised until at least the age of 18 in northern China with Mandarin as the primary language spoken at home. They reported no history of hearing, speech, or language difficulties and were paid for their participation. Most were international students who had been residing in the U.S. for a limited amount of time (mean length of residence, 1.9 yr, SD 1.7); however, all had extensive experience with English in formal educational contexts (mean length of formal study, 12.3 yr, SD 2.3). One talker had also studied an additional foreign language (Japanese), although none had ever lived outside of China before moving to the U.S.

The target lexicon in the word learning task consisted of 24 Mandarin pseudowords in the form of six tonally minimal quadruplets—two monosyllabic (MS) and four disyllabic (DS). Table I lists all 24 items (sound-meaning correspondences). The DS quadruplets were evenly split between having the contrastive tone on the pre-final (penultimate) syllable (DSP items) or on the final syllable (DSF items). Target items represented Mandarin segmental sequences that comply with English phonotactic constraints and depart modestly from the segmental inventory of American English; this allowed for a study of tone learning that used ecologically valid segments while minimizing the effect of learning unfamiliar segments on the learning of tones. To limit the lexicon to 24 items, only a subset of the 16 possible two-tone combinations were included in each group of eight DS items: T2-{T1/T2/T3/T4} and T4-{T1/T2/T3/T4} for DSF items, and {T1/T2/T3/T4}-T1 and {T1/T2/T3/T4}-T2 for DSP items. The unalternating tones in the DS items were selected so as to contrast high vs low/mid pitch on both sides of the tone juncture, avoid tone sandhi contexts, and produce contrasts attested in common words. Thus the selected tone combinations represent phonotactically legal final and pre-final tonal contrasts that occur in real Mandarin words (cf. the final contrast in /tɑ˨˩xə˥/ "drink a lot" vs /tɑ˨˩xə˧/ "big river" vs /tɑ˨˩xə˨˩/ "big congratulations," and the pre-final contrasts in /lɑʊ˧kʊŋ˥/ "labor" vs /lɑʊ˩˧kʊŋ˥/ "husband" and /ʂɨ˥jɛn˧/ "accidentally say the wrong thing" vs /ʂɨ˨˩jɛn˧/ "promise").

In the interest of consistency, both MS and DS items were made to represent sound-meaning pairs that do not constitute actual words in Mandarin. Whereas the phonological forms of the DS items do not occur in Mandarin to begin with (the segmental forms occur, but not with the given tone combination), the phonological forms of the MS items do occur.[3] Consequently, care was taken to pair the phonological forms of the MS items with meanings (English translation equivalents) that were different from their actual meanings in Mandarin. For example, the forms /mɑ˥/ and /mɑ˩˧/ (the actual meanings in Mandarin of which are "mother" and "horse") were paired with the meanings "banana" and "cake" in the target lexicon.

The final meanings of the words in the target lexicon were controlled with respect to several psycholinguistic dimensions. Because word meanings were to be represented in the learning task with pictures, possible meanings were drawn from a standardized set of pictures normed for name agreement, image agreement, familiarity, and visual complexity (Snodgrass and Vanderwart, 1980).[4] This set of candidate meanings was narrowed down by eliminating meanings with low imageability, concreteness, frequency, and/or familiarity as reported in the MRC Psycholinguistic Database (Wilson, 1988). Once the set of meanings was narrowed down to 24 in this way, they were randomly assigned

TABLE I. Target lexicon in the learning study. The 24 items comprise monosyllabic items (MS), disyllabic items with final contrast (DSF), and disyllabic items with pre-final contrast (DSP).

Item type	Tone 1 item	Tone 2 item	Tone 3 item	Tone 4 item
MS	/loʊ˥/ 'balloon'	/loʊ˧/ 'window'	/loʊ˩˧/ 'apple'	/loʊ˨˩/ 'horse'
	/mɑ˥/ 'banana'	/mɑ˧/ 'pencil'	/mɑ˩˧/ 'cake'	/mɑ˨˩/ 'ring'
DSF	/tɕi˧nɑn˥/ 'chair'	/tɕi˧nɑn˧/ 'ear'	/tɕi˧nɑn˩˧/ 'door'	/tɕi˧nɑn˨˩/ 'box'
	/ti˨˩wɑ˥/ 'book'	/ti˨˩wɑ˧/ 'pear'	/ti˨˩wɑ˩˧/ 'car'	/ti˨˩wɑ˨˩/ 'hat'
DSP	/pɑʊ˥mi˥/ 'hand'	/pɑʊ˧mi˥/ 'button'	/pɑʊ˩˧mi˥/ 'desk'	/pɑʊ˨˩mi˥/ 'fork'
	/tɑ˥li˧/ 'key'	/tɑ˧li˧/ 'foot'	/tɑ˩˧li˧/ 'eye'	/tɑ˨˩li˧/ 'dog'

to phonological forms with some additional rearrangement to distribute meanings in the same semantic field (e.g., "apple" and "pear"; "dog" and "horse") across different tonally minimal quadruplets.

The 24 audio-visual stimuli corresponding to the lexical items in Table I paired the audio recordings produced by the talkers with color images depicting their associated meanings. The audio recordings selected for use comprised the final tokens (of three total) that talkers uttered of each item except in the few cases where there was an audible error or hesitation (in which case one of the tokens uttered earlier was selected instead). The pictures used were modified versions of the black-and-white line drawings in Snodgrass and Vanderwart (1980) that were developed by Rossion and Pourtois (2004), who enhanced the original images with both texture and color, significantly facilitating recognition of the objects depicted therein.

C. Procedure

Recording of all auditory stimuli was done in a sound-attenuated booth using an Audix HT5 head-mounted microphone and a Zoom H4N recorder at 44.1 kHz and 24-bit resolution. Items were presented in random order on individual index cards showing their respective orthographic forms in simplified Chinese characters and pinyin romanization, although talkers were told to focus on the pinyin because many characters were phonologically ambiguous.[5] Talkers were instructed to speak at a comfortable volume and pace and given the opportunity to take breaks whenever necessary. In addition, to encourage the production of natural tonal coarticulation, talkers were specifically instructed to utter the disyllabic nonce items normally as if they were real words (i.e., without pausing between syllables). With a little practice, all talkers were able to accomplish this. Their pronunciation was monitored during the recording session by the experimenter (in every case, a Mandarin speaker), who asked the talker to repeat any item that was produced in an unnatural manner.

The learning study was part of a larger correlational study comprising a wide range of tasks, which participants came into the laboratory five times over the course of 2 wk to complete. All tasks were completed at individual computer stations in groups of up to 14 participants. Among the first tasks participants completed were a tone identification task and a tone discrimination task; these are not the focus of the current study but are mentioned here because they provided some exposure to Mandarin before the word learning task. The tone identification task (a four-alternative forced-choice task that began with a brief familiarization phase) consisted of 80 test trials during which participants heard a Mandarin monosyllable and had to select, from among four stylized line drawings depicting pitch contours, which tone they thought they had heard. The tone discrimination task (a categorial AX task) consisted of 96 test trials during which participants heard two talkers each utter a Mandarin monosyllable containing the same segments and had to indicate whether the talkers had said the same word or different words. Thus although participants had not been exposed to a tone language prior to entering the study, they were exposed to a total of 272 monosyllabic tokens of Mandarin (over the course of approximately 20 min) in the precursor tasks they completed before the word learning regimen.

Modeled after the learning regimen used in Chandrasekaran *et al.* (2010) and Wong *et al.* (2011), our learning regimen consisted of similarly structured sessions during which auditory forms were presented along with images depicting their meanings in three phases. In an initial familiarization phase, learners were exposed to the sound-meaning correspondences via simultaneous presentation of the auditory and visual stimuli. The items in each quadruplet were presented a total of four times, uttered by two male and two female talkers and blocked (i.e., grouped into experimental blocks) by quadruplet with MS quadruplets presented first (in random order) followed by DS quadruplets (in random order). In the following practice phase (blocked by quadruplet in the same manner), learners were tested on their knowledge of the sound-meaning correspondences and given feedback on their answers by the computer. On each practice trial, learners heard an auditory form, saw pictures of the items in the relevant quadruplet presented in a 2×2 grid on screen in random order, and clicked on the picture they thought was the correct answer. If correct, the screen read "CORRECT"; if incorrect, the screen read "INCORRECT" and showed the correct answer. As in the familiarization phase, each item in the practice phase was presented a total of four times. In the final test phase, learners were tested on their knowledge of the sound-meaning correspondences without feedback. Unlike the first two phases, the test phase was not blocked by quadruplet. On each of the 96 test trials, learners heard any one of the 24 auditory forms, saw all 24 pictures presented in a 6×4 grid in random order, and clicked on the picture they thought was the correct answer (but received no feedback on their accuracy).

The learning regimen was completed during participants' last three visits to the laboratory and consisted of a total of six sessions (two sessions per visit). This condensed completion schedule was the main difference between our regimen and that used in Chandrasekaran *et al.* (2010) and was adopted for two reasons. First, because it was apparent in the results of Chandrasekaran *et al.* that the most successful learners distinguished themselves from less successful learners well before the end of their regimen, the number of sessions in our regimen was reduced to six, the number of sessions it took for the most successful learners of Chandrasekaran *et al.* to reach ceiling performance. Second, to reduce attrition from the study, the six sessions in our regimen were consolidated into three visits to the laboratory, the first session during a visit being completed at the beginning of the list of tasks for that visit and the second session being completed at the end. These differences in design, as well as the inclusion of disyllabic items among the stimuli, were likely to increase the difficulty of our regimen in comparison to that of Chandrasekaran *et al.* (2010). Nevertheless by the end of our regimen, the most successful learners were still able to achieve ceiling performance (so it was not the case that the regimen was unreasonably difficult).

The overall structure of each session of the learning regimen was identical, but the test phase in the sixth and final session used the speech of two talkers (one male, one female) who had not been heard up to that point. Thus whereas the audio stimuli in all phases of sessions 1–5 and in the familiarization and practice phases of session 6 were physically identical, those in the test phase of session 6 were physically different because they were from novel talkers. The purpose of these latter stimuli was to examine the generalization of learners' knowledge to unfamiliar voices. By preventing learners from using perceptual strategies specific to the audio samples they had heard during the preceding sessions, the final test stimuli provided a truer measure of learners' knowledge of the lexicon under study. For this reason, performance in the final test phase was taken as our measure of acquisition.

D. Analysis

Because of the problems with using analysis of variance (ANOVA) on accuracy data represented in terms of percentages (Jaeger, 2008), the data from the final test phase were analyzed in a series of logistic mixed-effects regression models, with participant and meaning as random effects and item type (MS, DSF, DSP; reference level = MS), tone (T1–T4; reference level = T1), and their interaction as fixed effects. With respect to incorrect responses, we distinguished between errors in general and errors that were specific to tone (i.e., responses that were incorrect in terms of tone only). Thus if the trial audio was /loʊ˥/ "balloon," the response "apple" (= /loʊ˩/) would be a specifically tonal error, whereas the response "cake" (= /maɪ˩/) would not; the latter response would be instead a segmental (as well as tonal) error because the segments are incorrect. All of the model results presented in the following text are from the final (sixth) test phase in the learning regimen.

IV. RESULTS

A. Acoustic variability of tones across contexts

Before proceeding to the learning results, we first present a summary of acoustic analyses that were conducted on our word learning stimuli to confirm that these stimuli showed the same patterns of tonal variation described in previous studies (Xu, 1994, 1997). The pitch contour of a tone was measured in terms of f_0 at each of 10 evenly spaced points in the tone's time span, ranging from 5% to 95%. The time span of each tone contour (i.e., the voiced interval of the relevant syllable) was demarcated via joint inspection of the waveform and a wide-band spectrogram. The beginning of a first-syllable contour was marked at the onset of visible periodicity; the beginning of a second-syllable contour (=the end of a first-syllable contour in a disyllable) was marked at the drop in amplitude and/or onset of antiresonances corresponding to the second-syllable onset consonant (/m n l w/); and the end of an utterance-final contour was marked at the end of visible periodicity. Measurements of f_0 were taken in Hz (using the cross-correlation method in PRAAT) and then standardized against each talker's f_0 mean and range. The final data set thus consisted of 2400 standardized f_0 measures [6 talkers × 40 (8 MS + 32 DS) contours × 10 time points].

The results of the acoustic analyses revealed that tonal variability (measured in terms of the standard deviation of f_0 at each of the ten time points in the tone interval) was, overall, greater across contexts (within a talker) than across talkers (within the same context—namely, the MS context showing the canonical tone contour). In contrast to a mean standard deviation of f_0 across talkers (averaging across time points and tones) of 0.353, the mean standard deviation of f_0 across contexts (averaging across time points and tones) ranged from 0.447 to 0.736 in the set of six talkers. All differences between cross-context and cross-talker variability were significant [$|t|(39) > 2.176$, $p < 0.05$].[6]

TABLE II. Summary of results from acoustic analyses of tonal variability. Tone "onset," "offset," and "contour" refer to mean standard f_0 at the 5% point, at the 95% point, and over all time points, respectively. Tone "rise" and "fall" refer to the absolute value of the difference between the lowest and the following highest mean standard f_0 values (T3) and between the highest and the following lowest mean standard f_0 values (T4), respectively.

		Context								
Tone	Property	Isolation	T1_	T2_	T3_	T4_	_T1	_T2	_T3	_T4
T1	Onset	0.403	0.922	0.433	−1.285	−0.787	0.970	1.285	–	–
	Offset	0.912	1.304	1.178	0.816	0.650	1.050	0.950	–	–
	Contour	0.710	1.033	0.867	0.044	0.109	0.957	1.113	–	–
T2	Onset	−0.379	0.453	0.683	−1.582	−0.993	−0.413	−0.365	−0.122	−0.395
	Offset	0.871	0.316	0.781	0.363	0.167	0.357	0.759	0.898	0.144
	Contour	−0.362	−0.238	0.037	−0.810	−0.798	−0.317	−0.101	−0.007	−0.347
T3	Onset	−0.225	–	1.045	–	−0.808	−0.365	−0.175	–	–
	Offset	−1.036	–	−1.225	–	−2.005	−2.099	−1.645	–	–
	Contour	−1.233	–	−0.679	–	−1.723	−1.150	−1.051	–	–
	Rise	1.484	–	1.078	–	0.680	none	0.007	–	–
T4	Onset	1.223	–	0.334	–	−0.446	1.358	1.549	1.747	1.468
	Offset	−2.099	–	−1.637	–	−1.718	−0.634	−0.726	−0.464	−0.311
	Contour	0.337	–	0.414	–	−0.135	0.563	0.623	0.917	0.794
	Fall	3.489	–	3.030	–	2.680	1.996	2.276	2.216	1.868

The acoustic analyses also replicated several of Xu's (1997) findings on tonal coarticulation (see Table II for specific f_0 values and the online supplementary material for mean tone contours by context). With respect to tone onset, first, T1 started lower after T2-T4 than after T1; in addition, compared with its isolation form, T1 started higher after T1 but lower after T3-T4. Second, T2 started higher after a tone with a high offset (T1, T2) than after a tone with a low offset (T3, T4); this was also the case for T3. In addition, T2 and T3 each started lower after T3/T4 and higher after T1/T2 compared to their respective isolation forms. Third, T4 started lower after a low-offset tone (T4) than after a high-offset tone (T2) and lower after T2/T4 compared with its isolation form. With respect to tone level, the overall T2 contour was higher before low-onset tones (T2, T3) than before high-onset tones (T1, T4) and was considerably higher before T2/T3 compared with its isolation form. In addition, the T1 contour was slightly lower before another T1 than before a different tone (T2); in both cases, the initial T1 contour was also higher than its isolation form.

In addition to modifications of tone onset and level, the analyses also showed coarticulatory consequences for tone offset. For example, T1 ended higher after high-offset tones (T1, T2) than after low-offset tones (T3, T4) or in isolation. Similarly, T2 ended higher after T2 than after low-offset tones and also ended higher after T1 than after T4; in addition, T2 ended lower after any other tone than in isolation. As for T3, in DSF contexts this tone sometimes, but not always, showed the final rise characteristic of the canonical (MS context) contour; when this rise occurred, however, it was shallower than that of the canonical contour, resulting in lower T3 offset values after T2 and T4 compared to the MS context. The other falling tone, T4, was affected by a preceding tone in a similar manner: T4 ended higher after another tone compared to the MS context, resulting in a smaller f_0 fall in DSF contexts. These patterns, too, were largely in line with the findings of Xu (1997).

In short, acoustic analysis of the word learning stimuli supported the predictions in Sec. II D. It was found that the magnitude of contextual variability in the stimuli was substantial—in fact, larger than that of talker variability—providing further motivation for a study of context effects on tone learning. Moreover, the stimuli evinced patterns of tonal variation that were very similar to those documented in previous work, making it reasonable to expect these patterns to influence the acquisition of tonal contrasts in the current study.

B. Learning across item types

As predicted, DS items were learned at significantly lower rates than MS items. Model results showed that the odds of a MS item being identified correctly in the final test phase were better than 50–50 [$\beta = 1.277$, $z = 5.448$, $p < 0.0001$];[7] however, both DSF items [$\beta = -1.487$, $z = -5.246$, $p < 0.0001$] and DSP items [$\beta = -1.825$, $z = -6.435$, $p < 0.0001$] were significantly less likely to be identified correctly. Whereas MS items were identified correctly at a rate of 71%, DSF and DSP items were identified correctly at rates of 46% and 40%, respectively. An additional model showed that the small decrease in accuracy from DSF to DSP items was not significant [$\beta = -0.324$, $z = -1.277$, $p = 0.201$]. That the difficulty of DS items was due to their tones rather than their segments was clear from learners' errors, the majority of which were specifically tonal errors. This bias toward tonal errors was evident for all item types, but slightly stronger for DS items (68% of errors) than for MS items (64% of errors) and, in fact, most pronounced (76% of errors) for the item type that proved to be the most difficult to learn (namely, DSP items), suggesting further that DS items were learned less successfully at least in part because of their tonal variability.

Further inspection of the data by contrastive tone revealed that the four tones differed in their relative learnability across item types as shown in Fig. 1. In the case of MS items, T1 items were identified correctly with better than 50–50 odds [$\beta = 0.949$, $z = 3.574$, $p < 0.001$]—at a rate of 66%—and the rates of correct identification for T2 items and T4 items were not significantly different [$|\beta| < 0.152$, $|z| < 0.452$, $p > 0.651$]. T3 items, on the other hand, showed by far the highest rate of correct identification (85%), which was significantly higher than that for T1 items [$\beta = 1.386$, $z = 4.089$, $p < 0.0001$]. The pattern of relative learnability in DS items was markedly different, however, especially for T3. In comparison with the pattern in MS items, T3 DS items showed significantly lower rates of correct identification for both DSF items [$\beta = -1.013$, $z = -2.131$, $p < 0.05$] and DSP items [$\beta = -1.894$, $z = -3.979$, $p < 0.0001$]. In other words, as expected, T3 in DS contexts did not show the advantage in learning apparent in MS contexts.

In addition to the disparities between MS and DS items, there were further disparities between DSF and DSP items. In the case of DSF items, an additional model showed that T1 items were identified correctly with worse than 50–50 odds [$\beta = -0.539$, $z = -2.877$, $p < 0.01$], and the rate of correct identification for T2 items was not significantly different [$\beta = -0.143$, $z = -0.683$, $p = 0.495$]. The rate of correct identification for T3 items was higher than that for T1 items but only marginally so [$\beta = 0.362$, $z = 1.729$, $p = 0.084$]. T4 items, by contrast, showed a significantly higher rate of correct identification (59%) than T1 items [$\beta = 0.989$, $z = 4.710$, $p < 0.0001$]. In the case of DSP items, an

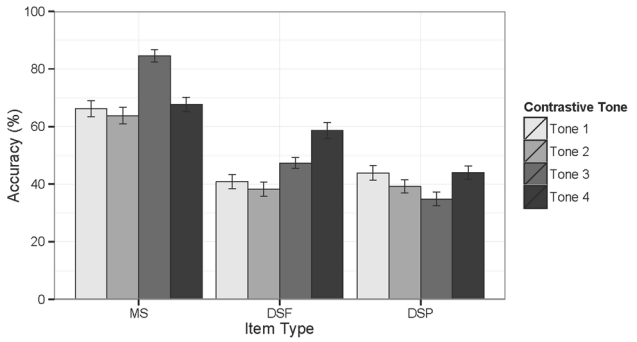

FIG. 1. Accuracy at final test by item type and contrastive tone. Item types are monosyllabic (MS), disyllabic with final contrast (DSF), and disyllabic with pre-final contrast (DSP). Error bars indicate ±1 standard error of the mean over participants.

additional model showed that the odds of T1 items being identified correctly were not significantly different from 50–50 [$\beta = -0.299$, $z = -0.817$, $p = 0.414$]. Among the other three tonal sets, T3 items showed the lowest rate of correct identification (35%); however, none of the other three tonal sets differed from the T1 set with respect to likelihood of correct identification [$|\beta| < 0.479$, $|z| < 0.967$, $p > 0.333$]. Notably, models built by tone showed that whereas neither T1 nor T2 differed significantly between DSP and DSF items [$|\beta| < 1.600$, $|z| < 0.232$, $p > 0.816$], both T3 [$\beta = -0.657$, $z = -7.335$, $p < 0.0001$] and T4 [$\beta = -0.867$, $z = -2.655$, $p < 0.01$] were learned less successfully in DSP items than in DSF items.

Examination of DS items by quadruplet revealed additional differences among the various coarticulatory contexts in the learnability of the same tone. In general, tones were learned less successfully in contexts of tone clash (where a disparity between the onset and offset f_0 levels of the adjacent tones results in significant coarticulatory perturbation of one or both tone contours). As shown in Fig. 2, T1 was learned much less successfully in contexts of tone clash (following T4 in /ti˨˩wa_/; preceding T2 in /ta_li˧˥/) than no tone clash (preceding T1 in /pɑʊ_mi˥˩/). This was also the case for T4, which was learned less successfully preceding T1 (in /pɑʊ_mi˥˩/) than following T2 (in /tɕi˧˥nan_/) or preceding T2 (in /ta_li˧˥/). However, following T4 (in /ti˨˩wa_/), T4 was learned unexpectedly well, perhaps benefiting from final creaky phonation as a secondary cue (or from a possible default response bias toward T4; see Sec. IV C). Effects of tone clash were also apparent in the learning of T2 (which was less successful following T2 in /tɕi˧˥nan_/ and preceding T2 in /ta_li˧˥/) but not in the learning of T3, which instead closely followed position. DSP items showed less successful learning of T3 regardless of the presence or absence of tone clash, suggesting that T3 acquisition was heavily influenced by the allotonic divergence of the pre-final form of T3 from its canonical form.

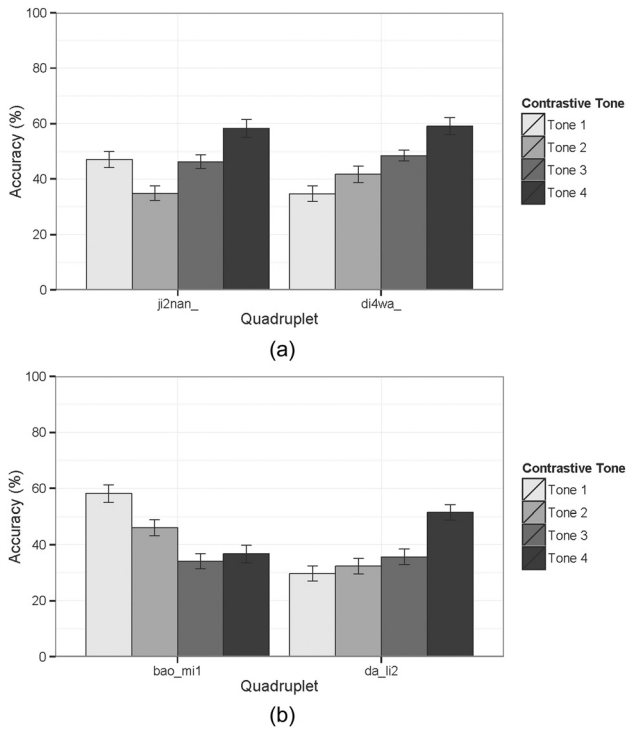

FIG. 2. Accuracy at final test for disyllabic items with (a) final tonal contrast (DSF) and (b) pre-final tonal contrast (DSP) by minimal quadruplet and contrastive tone. The DSF and DSP quadruplets are, respectively, *ji2nan_* and *di4wa_*, and *bao_mi1* and *da_li2* (in pinyin romanization; underscores mark the locus of tonal contrast). Error bars indicate ±1 standard error of the mean over participants.

An anonymous reviewer wondered how learning, as reflected in the likelihood of accuracy at test, related to reaction time. In particular, might participants have responded correctly in the test phase only because they took a long time to do so? This possibility was investigated by comparing the reaction times (i.e., the intervals between the end of audio stimulus presentation to the registering of a mouse click response) of correct vs incorrect responses via the nonparametric Mann–Whitney test. This analysis showed that reaction times were significantly faster for correct responses than for incorrect responses ($W = 33550232$, $n_1 = 8145$, $n_2 = 7215$, $p < 0.0001$) by approximately 388 ms on average. Thus although our research questions did not concern response speed specifically, reaction time data were consistent with the accuracy data: responses in the final test phase that were correct (thus indicating successful acquisition of the target contrasts) were also faster than responses that were incorrect.

C. Error patterns

To further address our last prediction (that learnability of tonal contrasts would be lowered by coarticulatory perturbations in disyllabic contexts), we examined whether the types of errors that learners made were those that would follow from specific coarticulatory effects. Because this analysis pertained to tonal confusions specifically, it focused on specifically tonal errors as opposed to segmental errors, which represented 32% of all errors. Tonal errors represented 64%, 60%, and 76% of all errors on MS, DSF, and DSP items, respectively, and were the majority error type for every item except for T3 MS items (see Fig. 3).

Detailed analyses of specifically tonal errors revealed systematic patterns of tonal confusion consistent with effects of tonal coarticulation. As shown in Fig. 3, confusion patterns were similar between the two MS quadruplets: T1 tended to be confused with T4 (and vice versa), T2 with T1 or T4, and T3 with T2. The dominant confusions in DSF items generally resembled those in MS items and were also similar between the two DSF quadruplets, although a notable exception was T3, which in DSF items was more commonly confused with T4 than with T2 (consistent with the lack of final rise or smaller final rise in this context; see Secs. II B and IV A). However, the distribution of confusion types in DSF items showed differences vis-a-vis MS items that followed from consequences of perseverative coarticulation discussed in Sec. IV A—in particular, lower T1 onset after T4, higher T1 offset after T2, higher T2 and T3 onset after T2, and lower T4 onset after T2 and T4. These perturbations were reflected in more frequent confusion (vis-a-vis confusions in MS items) of T1 with T2 in /ti˨˩wa_/, more frequent confusion of T2 and T3 with T1 and T4, respectively,

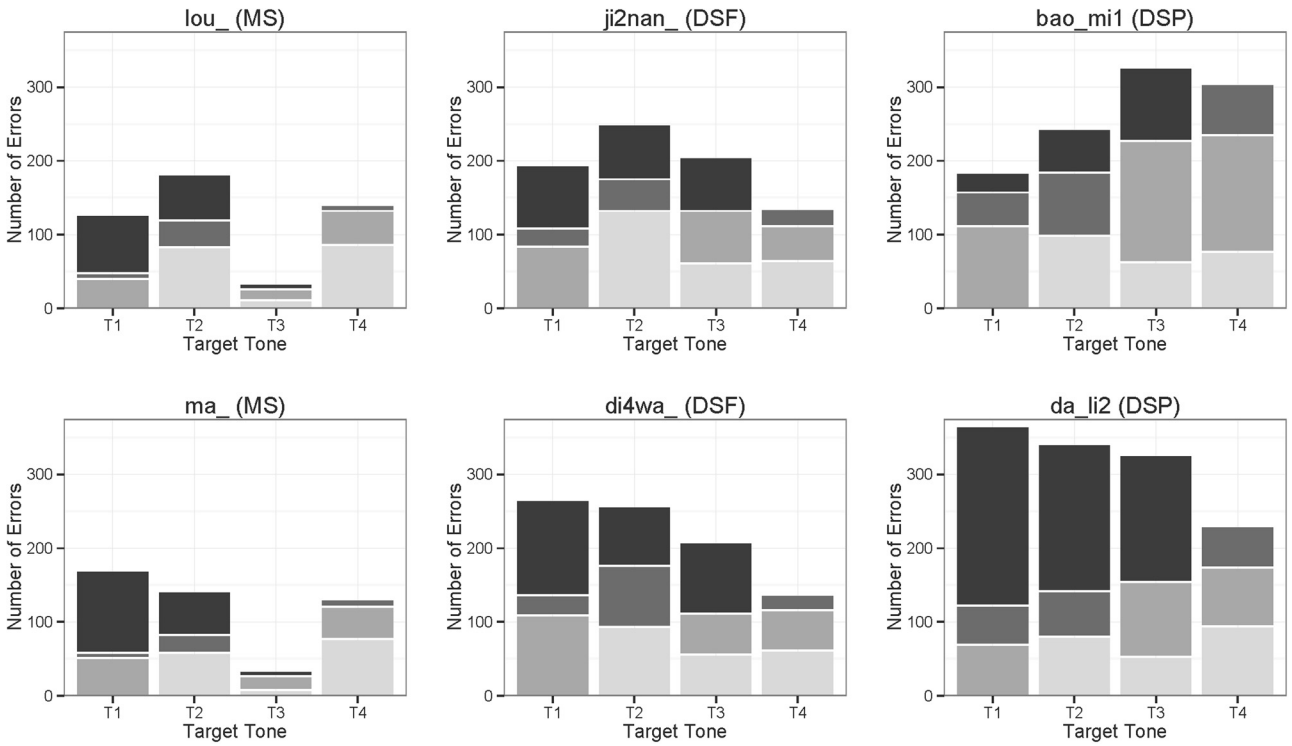

FIG. 3. Tonal error counts, by quadruplet, target tone, and response tone. Phonological forms are given in pinyin romanization. For each target tone, error types are shaded progressively darker according to response tone, with T1 in the lightest gray and T4 in black. By item, the percentage of all errors on that item represented by these tonal errors was (in left-to-right, top-to-bottom order): 66%, 73%, 36%, 68%; 57%, 60%, 60%, 52%; 69%, 71%, 78%, 75%; 71%, 67%, 34%, 67%; 64%, 69%, 63%, 53%; 82%, 80%, 79%, 75%.

in /tɕiˊnan_/, and more frequent confusion of T4 with T3 in both DSF contexts. In addition, the relatively low T1 onset after T2 (resulting from the rise of the initial T2 extending into the time domain of the second syllable) was reflected in more frequent confusion (vis-a-vis confusions in MS items) of T1 with T2 in /tɕiˊnan_/, similar to /tiˋwa_/.

Of the three item types, DSP items showed the highest number of errors, which was also the most biased toward tonal errors, as shown in Fig. 3. In comparison to the MS and DSF quadruplets, the two DSP quadruplets showed a more marked disparity in their patterns of tonal confusions, and the differences between the quadruplets as well as their differences with respect to MS contexts again followed from effects of tonal coarticulation. As discussed in Sec. IV A, the T2 contour was higher before tones with a low onset (e.g., T2) than before tones with a high onset (e.g., T1), resulting in a relatively steep transitional pitch fall from T2 offset to T2 onset. This was reflected in /ta_liˊ/ showing much more confusion of T2 with T4 than seen in /paʊ_miˇ/ or MS contexts. In addition, when the first tone in /ta_liˊ/ was T1, the medial pitch fall coming from the mismatch between T1 offset and T2 onset resulted in more confusion of T1 with T3 and T4 than seen in MS contexts. Compared to MS contexts, /ta_liˊ/ also showed relatively more confusion of T3 and T4 with each other, which followed from the ambiguity of the long interval of low pitch followed by rise characteristic of both T3-T2 and T4-T2 sequences. As for /paʊ_miˇ/, this context, in contrast to both /ta_liˊ/ and MS contexts, showed more confusion of T1 with T2 than with T4; this was likely due in part to the lower contour of T1 before another T1 (see Sec. IV A and Xu, 1997, p. 69, Fig. 6a), which results in a slight rise from the first T1 to the second T1 that may be perceived as the rising T2, and/or to a final uptick in pitch found in several of the T1-T1 stimuli. Also unlike both /ta_liˊ/ and MS contexts, /paʊ_miˇ/ showed more confusion of T4 with T2 than with T1, attributable to the medial pitch rise coming from the mismatch between T4 offset and T1 onset.

Although our explanation of these tonal errors is based on variation in tonal implementation, an anonymous reviewer pointed out that the tonal errors may instead be due to default perceptual biases. Perhaps, for example, L1 English learners are, *a priori*, biased to identify an ambiguous tone contour as T4 (e.g., because T4 resembles the declarative intonation contour of English). Such default biases, in and of themselves, do not provide a convincing explanation of the observed errors because the errors did not favor one tone in particular (e.g., T4) but every one of the four tones depending on context. In other words, there would have to be several different tone- and/or context-specific biases to account for the diversity of tonal confusions seen in this study. Consequently, while acknowledging that learners may be influenced by perceptual biases independent of context effects, we attribute the patterns of tonal confusions seen here primarily to patterns of contextual tonal variation because these provide a principled, as well as plausible, explanation of these confusions without the need to invoke the notion of preexisting biases.

V. DISCUSSION

In summary, phonological context was found to have pervasive effects on L2 tone learning, with contextual variation in tone contour consistently diminishing novice learners' ability to acquire tones in a Mandarin word learning task. Learning patterns observed in this task were consistent with all four of our predictions regarding learnability. First, disyllabic (DS) words—characterized by greater variation in individual tone contours—were indeed significantly harder to acquire than monosyllabic (MS) words overall, at least in part due to this increased tonal variability. Second, acquisition was further influenced by context-dependent differences in perceptual distinctness of the contrastive tone. Third, acquisition was also influenced by divergence of the contrastive tone contour from its canonical isolation form, especially that due to the allotonic alternation of T3. Fourth, coarticulatory perturbations affecting all tones in DS contexts exerted predictable—namely, negative—effects on tone learning. These findings suggest that learners acquiring lexical tones for the first time are not generally aided by the introduction of contextual variation in tone contour; on the contrary, such variability seems to interfere with their initial learning of tonal contrasts, although it remains an open question whether this difficulty introduced at the initial stage of learning might lead to more robust representations of the tones in the long term (if, for example, learners were given additional training time).[8] At least with 3 hr of training, it is clear that the difficulty does not help.

Given that DS words in the target lexicon contained not only more contextual tonal variation but also more tones than MS words, it is reasonable to think that DS words were learned less successfully than MS words simply because there were more tones to learn in DS words. However, there are two reasons why the observed MS-DS disparity in word learning is unlikely to be due to differences in tone count *per se*. First, if it was specifically the higher tone count of DS words that made them overall more difficult to acquire, we would expect DS item types to show more tonal errors than their MS counterparts across the board, but this is not the case. As shown in Fig. 3, DSF T4 items and MS T4 items showed virtually identical numbers of tonal errors; this is consistent with the "contextual variation" explanation of the MS-DS disparity (as T4 in the DSF context shows a contour that is similar to T4 in the MS context) but inconsistent with the "tone count" explanation. Crucially, the tone count explanation also fails to predict learning disparities within the set of DS items. If tone count was the primary factor affecting the acquisition of MS and DS words, we would expect DS words to show a similar decrement in learning relative to MS words because they all had the same number of tones (namely, two). However, as discussed in the preceding text, DS items showed marked learning disparities (Figs. 1–2), which were correlated with context effects.

Although context effects clearly had an influence on performance in the word learning experiment, it remains an open question how much of this influence was due to specific difficulty with learning the tones of a word (i.e., encoding the tonal information into the word's mental representation) as opposed to general difficulty with perceiving tones. Being able to abstract away from phonetic variability introduced by context (as well as other factors) to identify tonal categories would seem to be a prerequisite for learning tones; in other words, learners can only acquire tones to the extent that they can perceive them. This is consistent with the fact that accuracy in the word learning experiment was highly correlated with accuracy in the precursor tone identification task ($r = 0.75$; see Fig. 4): the more successfully learners were able to explicitly identify tones, the more successfully they acquired tonal word forms. To be precise, however, a correct response in the word learning experiment required both veridical acquisition of the target word forms (i.e., correct sound-meaning pairings) as well as accurate perception of the test stimuli. Consequently, some tonal errors could have arisen not due to faulty lexical representations but rather due to faulty perception of the tone(s) in a test stimulus. Although we cannot say for sure how many of learners' errors fall into this category, we regard it as most likely for perceptual deficits with tone to have caused problems with all of the stimuli, not just the test stimuli. That is, we have no reason to believe that the tone perception abilities of learners who perceived the familiarization/practice stimuli well enough to construct accurate mental representations of the target words failed suddenly on the test stimuli. Nevertheless, it would be interesting to examine learners' performance in other tasks (e.g., elicited production) to try to tease apart their acquired mental representations from their general tone perception abilities.

As for why we used a perceptual task to probe learners' acquired lexical knowledge, recall that the learning regimen in this study was designed with the goal of making the results more comparable with those reported in previous work on L2 tone learning that also used perceptual tasks. Such a comparison underscores a recurrent disparity between novice learners and more advanced learners of an L2 with

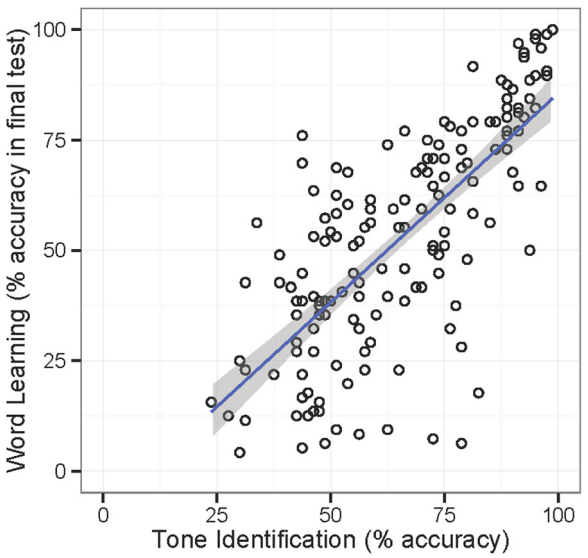

FIG. 4. (Color online) Performance in word learning vs the precursor tone identification task. The scale of both axes is identical; each data point represents percent accuracy for one participant. The shaded area represents the 95% confidence region around the regression line.

respect to how they cope with phonetic variability. Whereas studies of high-variability phonetic training (HVPT)—generally conducted with learners who are already familiar with the target L2 rather than total beginners—largely show benefits of exposure to phonetic variability, the current findings converge with those of Perrachione *et al.* (2011) in suggesting that variability is not always beneficial. In particular, there seem to be fundamental differences in how learners cope with input variability in a familiar vs unfamiliar L2. Learners are better able to benefit from exposure to variability once they have acquired a modicum of experience with the L2; with little in the way of a mental framework for the L2, novice learners tend instead to have difficulty with variability. In light of the findings of Perrachione *et al.* (2011) that show that novice learners benefit from input variability only if they have relatively strong perceptual abilities, this points to a larger conclusion that variability is a double-edged sword: it can either help or hurt L2 learning depending on a range of factors, including the stage of learning.

This duality in the effects of variability on learning begs the question of how and why later stages of learning come to diverge from *ab initio* learning. To our knowledge, there is no published research that systematically compares gains from HVPT at different stages of L2 learning. This gap in the literature highlights the need for controlled training studies designed to investigate how benefits of exposure to variability change over the course of L2 learning and, more generally, what factors prepare learners to benefit from variable input. The current state of the science suggests that two such factors are prior knowledge of the target language (including a sizable lexicon) and acute perception of relevant phonetic dimensions (arising from inherent aptitude and/or relevant experience), which may in fact be related to each other. For example, it would not be surprising if the non-tonal L1 speakers with L2 Mandarin experience examined by Wang (2013) were better at pitch perception than comparable individuals without this L2 experience; that is to say, linguistic experience with a phonetic property as a lexically contrastive feature is probably not orthogonal to perception of that property but instead influential in shaping perception. Such an effect would be consistent with the findings of Wayland and Guion (2004) that show that only tonal L1 speakers benefit from HVPT on unfamiliar L2 tones as well as recent evidence suggesting that tonal L1 experience increases sensitivity to melodic properties that have linguistic analogues in the L1 (Bradley, 2012).

When novice learners lack the perceptual abilities that are crucial for learning an L2 phonological contrast, phonetic variability in the input tends to be problematic because it makes the task of storing a mental representation of the input inherently more difficult; it remains to be seen, however, whether this is the case for all types of variability. In particular, it is reasonable to believe that there may be a difference between language-specific sources of variability (e.g., allophonic alternation) and language-universal ones (e.g., certain kinds of sex-based talker variability) in terms of their learning benefits at a given stage of acquisition. Work done in the HVPT paradigm often combines multiple types of variability in training exposure, given that listeners will eventually need to handle all types of variability in natural speech. This conflation can make it difficult to assess where observed training benefits are coming from or even whether the types of variability included in the training stimuli are all in fact improving outcomes for the given learner group. Although the final test phase discussed in the preceding section incorporated talker variability along with contextual variability, comparisons of performance in this test phase with that in a different test phase—namely, the preceding (fifth) test phase completed the same day (during which only familiar talkers were heard)—show that on average there was less than a 1% decrement in percent accuracy associated with the introduction of new talkers in the final test phase. Given that the average decrement in percent accuracy on tones between MS and DS contexts was much larger than this (27%; see Fig. 2), this suggests that context effects had a stronger influence on learning than did talker effects, consistent with our acoustic analyses showing greater tonal variability in the stimuli across contexts than across talkers.

Although the magnitude of different kinds of context effects is not something this study was designed to compare, in light of the overall sizable influence of context, it is worth pointing out that the allotonic realization of T3 as "half" T3 ([21]) in pre-final position was associated with the single greatest decrement in percent accuracy from MS to DS items (50%). This fact is consistent with the view that language-specific sources of variability are particularly problematic for novice learners as these patterns are not predictable on general phonetic grounds. For example, there is no articulatory reason why T3 has to be pronounced as [21] in pre-final contexts and, thus, little reason for a novice learner to posit, without morphophonemic evidence, that [21] in pre-final contexts corresponds to T3. By contrast, many of the coarticulatory perturbations learners had to cope with in this study are inevitable. When T4 occurs in the context of a following T1, for instance, there is no way to remove—short of a full stop—the transitional rise between T4's fall and T1's high onset. Certainly it is not the case that all tonal coarticulation patterns are predictable; nevertheless, the fact that a significant portion of coarticulatory variability can be understood in terms of language-universal tendencies may make this less problematic for novice learners. In short, although the line between allophonic variation and coarticulatory variation is not always clear, the current findings highlight the relevance of this distinction for future research on how novice learners cope with phonetic variability in L2 input.

VI. CONCLUSION

The findings reported in this article are intrinsically important to the study of L2 speech learning because they demonstrate why the construct of "phonetic variability" is a matter of concern in the design and interpretation of research on initial L2 learning. It is clear from the literature on speech learning—both of segmental and of suprasegmental categories—that, contrary to conventional wisdom, the difficulty introduced by variability in L2 input does not necessarily benefit learners; instead, effects of variability are correlated with learners' perceptual abilities as well as the amount of

prior experience they have with the L2. Moreover, phonetic variability is a complex construct consisting of multiple types of variability, and our findings suggest that contextual variability—especially language-specific types of contextual variability—may pose special problems for novice learners' acquisition of L2 contrasts.

A nuanced understanding of the role of variability in learning—one in which variability is viewed in relation to the type of learner and stage of learning rather than as a general booster of acquisition outcomes—is critical for the field to develop further for two reasons: building a better theory of speech learning and improving methods of phonetic training. For example, our findings, taken together with the other findings of the L2 speech learning literature, suggest that training learners by throwing as much and as many different kinds of variability at them may not be the most effective training method for all learners. Instead, tailored and adaptive methods (e.g., adjusting the amount of variability according to learners' interim performance) are more likely to result in the greatest gains across learners. Such tailoring may be further improved by taking into account additional relevant factors, such as the different timescales of acoustic cues to segmental vs suprasegmental categories.

In addition to the implications for research on speech learning in general, our findings also have implications for research on L2 Mandarin specifically. As discussed at the beginning of this paper, previous studies on L2 learning of Mandarin tonal contrasts have largely investigated tone learning in isolated monosyllabic items even though the nature of the Mandarin lexicon virtually requires learners to acquire tones in multisyllabic contexts. The current study attempted to address this disparity by investigating the effects of phonological context on tone learning. In short, our results show a profound influence of context on the learnability of tones, which suggests that findings limited to monosyllabic contexts are inadequate for generating predictions about L2 acquisition of Mandarin tones in natural speech. Precise and broadly applicable predictions may instead require systematic examination of a wide variety of phonological contexts.

In closing, our findings speak to the need for future research on speech learning to take into account the kinds of aptitude/attribute-by-treatment interactions (ATI) that have long been observed in other branches of L2 research, in education, and in psychology (e.g., Snow, 1989; Vatz *et al.*, 2013). Taken together, our results and those of previous studies evince a clear interaction between type of learner and type of input variability. Consequently, careful consideration of how specific combinations of learner profile and variability type may lead to different outcomes is likely to provide new insight into the development of L2 speech and the role of phonetic variability in influencing learning outcomes.

ACKNOWLEDGMENTS

The findings reported in this paper are based upon work supported, in whole or in part, by funding from the United States government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the authors' institutions or any agency of the United States government. The authors are grateful to Price Bingham, Janet Cook, Eli Cooper, Ryan Corbett, Joseph Dien, Dimitrios Donavos, Meg Eden, Christian Frable, Lora Grasso, Henk Haarmann, Valerie Karuzis, Yao Yao, and audiences at the University of Maryland and the Annual Meeting of the Linguistic Society of America for assistance, discussion, and feedback at various stages of this research.

[1]A third allotone of T3 occurs before T3 due to a tone sandhi rule that applies to T3-T3 sequences. In addition, T1 and T4 are associated with allotonic alternations specific to a limited set of lexical items. These alternations are irrelevant for the current study because the stimuli contained neither T3-T3 sequences nor any of the lexical items subject to T1 or T4 alternation.

[2]All tones are transcribed using the Chao system of tone letters (Chao, 1930) adopted by the International Phonetic Association.

[3]Note that this did not create an inconsistency in the stimuli because learners had no knowledge of which phonological forms are and are not attested in Mandarin. The stimuli were not limited to attested forms because real tonally minimal DS quadruplets are difficult to find such that each member of the quadruplet is a common word. The benefit of using real quadruplets in this case would thus have been marginal (due to not all of the words being known to talkers); consequently, all nonce words were used for the DS items to avoid unintended production disparities between real and nonce words. On the other hand, the stimuli were not limited to unattested forms instead because MS gaps in the Mandarin lexicon are rare.

[4]Name agreement refers to the degree to which subjects asked to name the object in a given picture agree on the same name for that object. Image agreement refers to the degree to which subjects' mental image of an object agrees with the depiction of that object in the given picture. Familiarity refers to the degree to which subjects "come in contact with or think about the concept" in the picture, while visual complexity refers to "the amount of detail or intricacy of line in the picture" as opposed to the complexity of the object depicted (Snodgrass and Vanderwart, 1980, p. 183).

[5]For the nonce DS items, which had no standard orthographic form, common characters were selected to represent the target pronunciation.

[6]See supplementary material at http://dx.doi.org/10.1121/1.4937612 for plots of variability over time across talkers vs across contexts.

[7]The output of a binomial logistic regression is relativized to 50-50 odds because evenly split probability for a binomial outcome (e.g., "correct" vs "not correct") corresponds to 50-50 odds. Note, however, that "chance" performance in the learning task is actually 1 of 24 (=1–23 odds or 4.2% accuracy) for completely random guessing or 1 of 4 (=1–3 odds or 25% accuracy) for random guessing within the correct tonally minimal quadruplet.

[8]Although this remains a possibility, one should be skeptical of the idea that learning is unconditionally enhanced by exposure to high variability; this leads to the prediction that after a sufficient amount of training time, learners exposed to more variable input will always show an advantage over those exposed to less variable input regardless of other factors such as the learners' initial perceptual abilities. Do "weak" perceivers (who tend to be overwhelmed, not buoyed, by high variability) experience a sudden reversal at a later point where the initial exposure to overwhelming variability "kicks in" and helps them catch up to—and even surpass—learners exposed to less variability? There are no training studies of *ab initio* learners over the long term that can speak directly to this question; however, in light of the findings of Perrachione *et al.* (2011) that show marked differences in the learning trajectories of "weak" and "strong" perceivers (see, e.g., their Figs. 3 and 6), this seems unlikely.

Barcroft, J., and Sommers, M. S. (**2005**). "Effects of acoustic variability on second language vocabulary learning," Stud. Second Lang. Acquis. **27**, 387–414.

Berkowitz, S. S. (**2010**). "Discrimination of tone contrasts in Mandarin disyllables by naive American English listeners," Ph.D. thesis, City University of New York, New York.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (**1990**). "Effects of syllable duration on the perception of the Mandarin tone 2/tone 3 distinction: Evidence of auditory enhancement," J. Phon. **18**, 37–49.
Bowles, A. R., Chang, C. B., and Karuzis, V. P. (**2015**). "Pitch ability as an aptitude for tone learning," Lang. Learn. (in press).
Bradley, E. D. (**2012**). "Crosslinguistic perception of pitch in language and music," Ph.D. thesis, University of Delaware, Newark, DE.
Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/. IV: Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**, 2299–2310.
Chandrasekaran, B., Sampath, P. D., and Wong, P. C. M. (**2010**). "Individual variability in cue-weighting and lexical tone learning," J. Acoust. Soc. Am. **128**, 456–465.
Chang, C., and Yao, Y. (**2007**). "Tone production in whispered Mandarin," in *Proceedings of the 16th International Congress of Phonetic Sciences*, edited by J. Trouvain and W. J. Barry (Pirrot, Dudweiler, Germany), pp. 1085–1088.
Chao, Y.-R. (**1930**). "A system of tone-letters," Le Maître Phon. **45**, 24–27.
Chao, Y.-R. (**1933**). "Zhong guo zi diao gen yu diao [Tone and intonation in Mandarin Chinese]," Guo Li Zhong Yang Yan Jiu Yuan Li Shi Yu Yan Yan Jiu Suo Ji Kan [J. Inst. History Philos., Acad. Sinica] **4**, 121–135.
Clopper, C. G., and Pisoni, D. B. (**2004**). "Effects of talker variability on perceptual learning of dialects," Lang. Speech **47**, 207–239.
Ding, H. (**2012**). "Perception and production of Mandarin disyllabic tones by German learners," in *Proceedings of the 6th International Conference on Speech Prosody*, edited by Q. Ma, H. Ding, and D. Hirst (Tongji University Press, Shanghai, China) pp. 378–381.
Duanmu, S. (**2007**). *The Phonology of Standard Chinese*, 2nd ed. (Oxford University Press, Oxford, UK), 352 pp.
Flege, J. E. (**1995**). "Two procedures for training a novel second language phonetic contrast," Appl. Psycholing. **16**, 425–442.
Gottfried, T. L., and Suiter, T. L. (**1997**). "Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones," J. Phon. **25**, 207–231.
Hao, Y.-C. (**2012**). "Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers," J. Phon. **40**, 269–279.
Iverson, P., Hazan, V., and Bannister, K. (**2005**). "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults," J. Acoust. Soc. Am. **118**, 3267–3278.
Jaeger, T. F. (**2008**). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. Mem. Lang. **59**, 434–446.
Jin, W. (**2011**). "A statistical argument for the homophony avoidance approach to the disyllabification of Chinese," in *Proceedings of the 23rd North American Conference on Chinese Linguistics*, edited by Z. Jing-Schmidt (University of Oregon, Eugene, OR), Vol. 1, pp. 35–50.
Kingston, J. (**2003**). "Learning foreign vowels," Lang. Speech **46**, 295–349.
Kirk, K. I., Pisoni, D. B., and Miyamoto, R. C. (**1997**). "Effects of stimulus variability on speech perception in listeners with hearing impairment," J. Speech Lang. Hear. Res. **40**, 1395–1405.
Kong, Y.-Y., and Zeng, F.-G. (**2006**). "Temporal and spectral cues in Mandarin tone recognition," J. Acoust. Soc. Am. **120**, 2830–2840.
Leather, J. (**1996**). "Interrelation of perceptual and productive learning in the initial acquisition of second-language tone," in *Second-Language Speech: Structure and Process*, edited by A. James and J. Leather (Mouton de Gruyter, Berlin, Germany), pp. 75–101.
Liu, S., and Samuel, A. G. (**2004**). "Perception of Mandarin lexical tones when F0 information is neutralized," Lang. Speech **47**, 109–138.
Lively, S. E., Logan, J. S., and 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," J. Acoust. Soc. Am. **94**, 1242–1255.
Perrachione, T. K., Lee, J., Ha, L. Y. Y., and Wong, P. C. M. (**2011**). "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," J. Acoust. Soc. Am. **130**, 461–472.
Rossion, B., and Pourtois, G. (**2004**). "Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition," Perception **33**, 217–236.
Shen, X. S. (**1990**). "Tonal coarticulation in Mandarin," J. Phon. **18**, 281–295.
Shen, X. S., and Lin, M. (**1991**). "Concept of tone in Mandarin revisited: A perceptual study on tonal coarticulation," Lang. Sci. **13**, 421–432.
Snodgrass, J. G., and Vanderwart, M. (**1980**). "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," J. Exp. Psychol. Hum. Learn. Mem. **6**, 174–215.
Snow, R. E. (**1989**). "Aptitude-treatment interaction as a framework for research on individual differences in learning," in *Learning and Individual Differences: Advances in Theory and Research*, edited by P. L. Ackerman, R. J. Sternberg, and R. Glaser (Freeman, New York), pp. 13–59.
Sommers, M. S., and 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," Appl. Psycholing. **28**, 231–249.
Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1994**). "Stimulus variability and spoken word recognition, I: Effects of variability in speaking rate and overall amplitude," J. Acoust. Soc. Am. **96**, 1314–1324.
Trude, A. M., and Brown-Schmidt, S. (**2012**). "Can listeners use creaky voice to constrain lexical interpretation?," in *the 53rd Annual Meeting of the Psychonomic Society*, Minneapolis, MN.
Vatz, K., Tare, M., Jackson, S. R., and Doughty, C. J. (**2013**). "Aptitude-treatment interaction studies in second language acquisition," in *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment*, edited by G. Granena and M. Long (John Benjamins, Amsterdam, The Netherlands), pp. 273–292.
Wade, T., Jongman, A., and Sereno, J. (**2007**). "Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds," Phonetica **64**, 122–144.
Wang, X. (**2013**). "Perception of Mandarin tones: The effect of L1 background and training," Modern Lang. J. **97**, 144–160.
Wang, Y., Jongman, A., and Sereno, J. A. (**2003**). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. **113**, 1033–1043.
Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (**1999**). "Training American listeners to perceive Mandarin tones," J. Acoust. Soc. Am. **106**, 3649–3658.
Wayland, R. P., and Guion, S. G. (**2004**). "Training English and Chinese listeners to perceive Thai tones: A preliminary report," Lang. Learn. **54**, 681–712.
Wilson, M. (**1988**). "MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00," Behav. Res. Methods Instrum. Comput. **20**, 6–10.
Wong, F. C. K., Chandrasekaran, B., Garibaldi, K., and Wong, P. C. M. (**2011**). "White matter anisotropy in the ventral language pathway predicts sound-to-word learning success," J. Neurosci. **31**, 8780–8785.
Wong, P. C. M., and Perrachione, T. K. (**2007**). "Learning pitch patterns in lexical identification by native English-speaking adults," Appl. Psycholing. **28**, 565–585.
Xu, Y. (**1994**). "Production and perception of coarticulated tones," J. Acoust. Soc. Am. **95**, 2240–2253.
Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phon. **25**, 61–83.
Yang, W.-J., Lee, J.-C., Chang, Y.-C., and Wang, H.-C. (**1988**). "Recognition of lexical tones for isolated syllables and disyllables in Mandarin speech," Int. J. Pattern Recogn. Artificial Intell. **2**, 49–69.