Relating production and perception of L2 tone

CHAPTER 9

Relating Production and Perception of L2 Tone

James Kirby and Đinh Lư Giang[*]

9.1 Introduction

The perception and production of second language (L2) speech has been widely studied in a variety of populations with a range of methods. One of the central questions in this line of research has been the degree to which perception guides production of L2 sound categories. According to Flege's Speech Learning Model (SLM; Flege, 1995, 1999), the accuracy with which nonnative segments are perceived will limit how well they can be produced. The SLM posits that L2 ability is not simply a function of age, but rather depends on the nature of L2 exposure and usage as well as the structural similarities between the L1 and L2.[1] The SLM attributes the often observed decrease in L2 production accuracy over the life-span to age-related changes in how the L1 and L2 systems interact: as perception becomes increasingly tuned to the L1, the likelihood of establishing new categories progressively decreases, because L2 sounds are increasingly perceived through the "filter" of L1. Thus, although L2 perceptual ability is predicted to decrease with age, the SLM posits that this is due to perceptual attunement rather than the effects of a critical acquisition period (Flege, 1999). In general, however, the SLM predicts that perception should precede production, and that perception and production abilities will converge over the course of learning. If this is the case,

* This project was funded in part by a Council of American Overseas Research Centers (CAORC) Senior Research Fellowship from the Center for Khmer Studies to J. Kirby. Thanks to Charles Nagle and audiences at the Institute of Phonetics and Speech Processing, LMU Munich; the Phonology Laboratory at the University of Chicago; and LabPhon 16 for thoughtful comments on earlier versions of this work. The authors are solely responsible for any errors of fact or interpretation. We also extend our thanks to the People's Committee of Giồng Riềng province, the clergy of the Cái Đuốc Giữa temple, and to all of the participants, without whom this work would not have been possible.

[1] This basic premise is also shared by other models of L2 perception such as the Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007) and the Second Language Linguistic Perception Model (Escudero, 2005).

production and perception should generally be correlated, at least for novice and advanced speakers, and moderate correlations have been found in studies of both vowels and consonants in a variety of languages (Bettoni-Techio et al., 2007; Elvin et al., 2016; Flege, 1993; Flege et al., 1999; Levy & Law, 2010; Llisterri, 1995; Morrison, 2003).

However, there is also evidence suggesting that learning in production is not always dependent on perception developing first. In a longitudinal study of late L1 English learners of the L2 Spanish onset voicing contrast, Nagle (2018) found that production of the L2 contrast began to improve before learners' ability to discriminate the contrast had reached native-like levels. In fact, there is some evidence that producing sounds during perceptual training may actually impede the formation of perceptual representations. Baese-Berk (2019) studied how L1 English speakers' ability to produce a Spanish-like obstruent contrast was affected by training modality. She manipulated training modality (perception only or interleaved perception and production) while holding testing modality constant (all participants were tested for both production and discrimination). Participants who were trained in both perception and production showed substantial improvement in production accuracy, but their perceptual improvement lagged behind. In other words, they were more accurate at producing the contrast than perceiving it, suggesting that performance in production may be unrelated to performance in perception. Baese-Berk suggests this may be an effect of interleaving production and perception training, while Nagle raises the possibility that the production–perception link may be lagged or asynchronous. Studies like these provide evidence that perceptual ability does not always appear to be a necessary prerequisite for facility in production to improve (see Chapter 1).

There is also some evidence that perceptual difficulties may persist even after production is objectively "mastered" (Strange, 1995). An example is provided by Sheldon and Strange (1982), who tested L1 Japanese learners of L2 English on their ability to perceive and produce the /r/-/l/ contrast. The authors found that native English listeners were more accurate at distinguishing L2 productions of /r/ and /l/ than the Japanese listeners themselves were. This was interpreted as evidence that the production of an L2 contrast can be superior to the perception of that contrast, and thus that production and perception performance may be uncorrelated. The reasons underlying these apparent instances of "perceptuo-productive heteromorphism" (Bohn & Flege, 1997) – that correlations are sometimes observed and sometimes not – has been a source of ongoing investigation, potentially involving age limits on learning new forms of

articulation, the type of contrast being studied, and a diverse range of methodological differences such as the phonetic dimensions being measured to assess production (Flege, 1999), the interstimulus interval used in perception studies (Peperkamp & Bouchon, 2011; Wayland & Guion, 2003), and the tasks used to evaluate performance in each modality (Sakai & Moorman, 2018). All this work makes it clear that the relationship between production and perception is unlikely to be as straightforward as the classical models might suggest.

The topic of production and perception of L2 tone has been studied for East Asian tone languages such as Thai (Gandour, 1983; Wayland & Guion, 2003), Vietnamese (Blodgett et al., 2008; Nguyen & Macken, 2008), and Mandarin Chinese (Wang et al., 2012; Yang, 2015). Much of this literature focuses on how properties of a learner's L1, such as whether or not it is also a tone language, may affect their success at tone production and perception in L2. In general, speakers of a tonal L1 are more accurate at identifying and discriminating tones in a tonal L2 compared to speakers whose L1 is nontonal (Francis et al., 2008; Hallé et al., 2004; Lee et al., 1996; Wayland & Guion, 2004), although even for tone language speakers, the specifics of the tone systems involved may play a nontrivial role (So & Best, 2010). Furthermore, listeners who speak a tonal L1 have been found to be more sensitive to pitch direction when perceiving L2 tones, while listeners with nontonal L1 backgrounds are more apt to attend to pitch height (Francis et al., 2008; Gandour, 1983; Guion & Pederson, 2007; Hallé et al., 2004). In production, L2 learners whose L1 is nontonal often have a compressed pitch range compared to native tone language speakers (Chen, 1974), show interference with certain segments (Nguyen & Macken, 2008; Yang, 2012), and often have difficulty with accurately producing complex contour tones as well as determining the correct starting pitch height (Bauman et al., 2009; Blodgett et al., 2008).

Compared to the literature on segments, however, much less attention has been given to the production–perception relationship for L2 tone. The current, tentative consensus seems to be that, *contra* the predictions of L2 acquisition models, production leads perception, both in the sense of order of acquisition (production is mastered earlier) and facility (production ability is superior to perception). For example, in Yang's (2012) study of American English learners of Mandarin Chinese, learners had considerable difficulty correctly identifying the rising tone /35/. However, this perceptual difficulty was not matched in production: learners' productions of this tone were not any less likely to cause errors

for native-speaker transcribers (but cf. Miracle, 1989; Ding et al., 2011). Yang (2012) suggests this may be because L2 tone production is primarily phonetic in nature, involving imitation and generalization of acoustic targets such as pitch heights, turning points, and perhaps durations. This same sensitivity to phonetic detail, however, works against learners in perception, because they lack robust phonological tone categories in the first place (see also Hallé et al., 2004). The perceptual advantage for L1 speakers of other tone languages would then be explained by their having phonological representations for tone categories that can be carried over from their L1.

As far as we are aware, almost all work explicitly addressing the production/perception relationship in L2 tone has focused on populations acquiring the L2 (usually Mandarin Chinese) in postsecondary instructional environments. This suggests another possible reason production has been found to lead perception, namely, the emphasis on repetition and assessment typical of this setting. In many scenarios, however, learners are receiving little or no formal training in the L2, but instead find themselves in immersion environments where the L2 is the medium of instruction. In these environments, learners are unlikely to be receiving targeted feedback on the phonetic realization of L2 tones (or segments, for that matter). The degree to which the L1 is used relative to the L2 would also presumably play a role (Flege et al., 1997), but as far as we know, this has not been studied for tone.

This study contributes to our understanding of production and perception of L2 tone by investigating how production and perception are realized at the level of individual speakers in a noninstructional setting. We consider how speakers of a nontonal language (Khmer) treat the tones of their L2 (Southern Vietnamese). Because of the social and linguistic dynamics of southern Vietnam, this setting presents an interesting opportunity to study L2 tone acquisition "in the wild," complementing studies of L2 tone acquisition looking at populations who have undertaken formal second language instruction, as well as those who have received explicit training specifically focused on improving tone production and/or perception. In an attempt to mitigate the methodological issue of selecting potentially arbitrary acoustic features, we opt to use global measures of curve similarity to measure the distance between tonal realizations. We consider how well L1 Khmer speakers of L2 Vietnamese distinguish Vietnamese tones in production by measuring their acoustic distances from native Vietnamese productions, but also by considering the extent to which they are acoustically distinctive in a speaker's own

tone space. We also look at both native and nonnative listeners' ability to discriminate these tones. By working with participants who have a broad range of ages and educational backgrounds, we can also gain some insight into how experience shapes the relationship between production and perception of L2 tone.

9.2 Language Background

9.2.1 Khmer Krom

Khmer is an Austroasiatic language spoken primarily in Cambodia, northeastern Thailand, and southern Vietnam.[2] Khmer speakers have probably inhabited the Mekong Delta region from at least the seventh century CE. Today, there are around one million ethnic Khmers in Vietnam (General Statistics Office of Vietnam, 2010). Around 5 percent of speakers (mostly older) are monolingual in Khmer, while around 15 percent (mostly younger and/or of mixed Khmer-Vietnamese ethnicity) are monolingual in Vietnamese (Đinh Lư Giang, 2011).

The Khmer dialects spoken in present-day Vietnam are referred to variably as Southern Khmer or *Khmer Krom* (literally "Khmer from below"). Mutually intelligible with Khmer varieties spoken in central Cambodia, they are often subsumed as part of the Central Khmer construct. That said, Khmer Krom varieties have at least some lexical and phonological features which differentiate them from Standard Khmer (Sochoeun, 2006, pp. 64–66), some of which are probably the result of contact (Đinh Lư Giang, 2011, 2015; Nguyễn Thị Huệ, 2010; Thạch Ngọc Minh, 1999). The Khmer varieties of Vietnam remain underdescribed.

Kiên Giang, one of Vietnam's southernmost provinces, shares its northwestern border with Kampot province in Cambodia. Ethnic Khmer in Kiên Giang make up around 10 percent of the provincial population. The present study was conducted in the district of Giồng Riềng, where Khmers account for about 15 percent of the total population. In the hamlet of Ngọc Chúc, home to most of the participants in our study, nearly one-third of the population is Khmer. Although not a tone language, pitch does play a (very) limited contrastive role in at least some Khmer dialects, including the local variety spoken in Kiên Giang (Kirby, 2014; Kirby & Đinh Lư Giang, 2017; Thạch Ngọc Minh, 1999). Whether

[2] This section is adapted from section 2 of Kirby and Đinh Lư Giang (2017); the reader is directed to that article for more detailed information on Kiên Giang Khmer.

or not this impacts the production and perception of their L2 Vietnamese tones is a question we return to in Section 9.5.

9.2.2 Southern Vietnamese

"Southern Vietnamese" refers to the relatively homogenous language varieties of the Kinh (Vietnamese) people spoken in and south of Khánh Hoà province (Brunelle, 2015). Vietnamese dialects differ considerably in terms of phonetics, phonology, and lexicon, but with the exception of some central dialects, they maintain a high level of mutual intelligibility. The tone systems of the major Vietnamese dialects are well described (Brunelle, 2015; Hoàng Thị Châu, 1989; Phạm, 2003; Vũ Thanh Phương, 1982). Northern Vietnamese (NVN) has six tones that contrast in voice quality as well as pitch (Nguyễn Văn Lợi & Edmondson, 1998), while Southern Vietnamese (SVN) has five tones that are distinguished exclusively by differences in fo height and excursion (see Table 9.1).

9.3 Methods and Materials

9.3.1 Participants

Eighteen adult speakers of Kiên Giang Khmer (18–47, 5 female; hereafter KG) and 10 monolingual native speakers of Southern Vietnamese (19–52, 7 female; hereafter VN) were recruited from the local population. The Khmer speakers also took part in a separate study (Kirby & Đinh Lư Giang, 2017).

All Khmer participants completed a short questionnaire which asked their year of birth (AGE), their highest completed grade (EDUCATION), as

Table 9.1 *Production stimuli*

Item	Tone	Orthography	Gloss
ta:^{33}	*ngang*	*ta*	'1SG (neutral, nonformal)'
ta:^{21}	*huyền*	*tà*	'dusk, twilight'
ta:^{35}	*sắc*	*tá*	'dozen'
ta:^{214}	*hỏi-ngã*[a]	*tả*	'describe'
ta:^{212}	*nặng*	*tạ*	'picul (100 kg)'

Note: Vietnamese names for tones are given for reference.
[a] The *hỏi* and *ngã* tones, which are distinct in Northern Vietnamese, are merged in Southern Vietnamese.

well as a self-reported assessment of what percentage of their daily language usage was Vietnamese as opposed to Khmer (VIETNAMESE USAGE). We did not explicitly ask about age of first exposure to Vietnamese, although we surmise that for most participants it coincided with the onset of formal education (so between ages four and six). Participants' ages ranged from 18 to 52 (mean 35). Education level ranged from no formal schooling of any kind to 12 years (completion of upper secondary education in the Vietnamese system), with the average being completion of grade 7. Self-assessment of percentage of Vietnamese used in daily life ranged from 10 to 80 percent (mean 40 percent). All Khmer participants self-reported as native speakers of Khmer, and our impressions corroborated these self-assessments.

Khmer participants completed the production and perception studies at the Cái Đuốc Giữa temple in Ngọc Bình village, Ngọc Chúc hamlet, Giồng Riềng district, Kiên Giang province. Sessions with the Vietnamese participants took place at the Trung tâm Học tập Cộng UBND xã Ngọc Chúc (Community Learning Center of the Ngọc Chúc People's Committee). All data were collected in August 2011.

9.3.2 Production Study: Methods and Materials

Participants were recorded producing the syllable /taː/ three times with each of the five Southern Vietnamese tones in the carrier phrase *Tôi nói* _____ *cho anh biết* [toj^{33} noj^{35} ____ co^{33} an^{33} biək^{45}] "I say ____ for you." This syllable was selected as it can be combined with all five tones to give commonly occurring lexical items (see Table 9.1). 24 bit, 44.1 kHz recordings were made using an omnidirectional headset condenser microphone and portable solid-state recorder. Recordings were annotated in Praat (Boersma & Weenink, 2015) to indicate the onset and offset of phonation, and a Praat script was used to measure fo at 11 equidistant points in the vowel.

9.3.2.1 Measuring Production Accuracy

Typically, studies of L2 tone production measure accuracy either in terms of acoustic landmarks like pitch range, overall fo change, timing of turning points, and so on, and/or in terms of native-speaker evaluations (e.g., Chen, 1974; Wang et al., 2003; Yang, 2012). In order to facilitate comparison to perception data, however, it can be useful to have a "one-number summary" of similarity, which potentially captures other aspects of the fo contours, such as slope. For this, we considered two global

measures of trajectory comparison: the dynamic time warping (DTW) distance (Müller, 2007) and the Fréchet distance (Chambers et al., 2010).

The DTW distance derives from an algorithm originally developed in the context of speech recognition to find the optimal alignment between two sequences of different lengths. The DTW distance between two sequences X and Y is the minimum of the *sum* of distances:

$$DTW(X,Y) = \min\left\{c_{p^*}(X,Y)\right\}, \text{ where}$$

$$c_p(X,Y) := \sum_{l=1}^{L} c\left(x_{n_l}, y_{m_l}\right), c \text{ a local cost measure.}$$

The Fréchet distance between two curves, sometimes also called the "dog-walking distance," is "the minimum length of a leash required to connect a dog and its owner as they walk without backtracking along their respective curves from one endpoint to the other" (Chambers et al., 2010, p. 295). Because the Fréchet metric takes the shape of the curves into account, it can provide a more accurate similarity measure than alternative measures which first reduce the curves to a small number of points. It can be thought of as the minimum of the *maximum* distance between the curves. The Fréchet distance δ is given as

$$\delta(X,Y) = \min_{\alpha,\beta}\left\{\max_{t\in[0,1]} d\left(X\left(\alpha(t)\right), Y\left(\beta(t)\right)\right)\right\}.$$

This reads as: for every possible function $\alpha(t)$ and $\beta(t)$, find the largest distance between the man and his dog as they walk along their respective path, and keep the smallest distance found among these maximum distances.

9.3.3 Perception Study: Methods and Materials

Following their production session, each participant completed an AX discrimination task. Five syllables (/taː/ with each of the five Southern Vietnamese tones) were synthesized using the KlattSyn implementation in Praat 5.4.08 (Boersma & Weenink, 2015), based on pilot recordings taken from two native speakers of the local Southern Vietnamese dialect who did not otherwise participate in the study. A spectrogram of the stimulus and the synthesized fo contours are shown in Figure 9.1. Stimuli were then arranged to form 30 AX pairs, 10 "same" pairs and 20

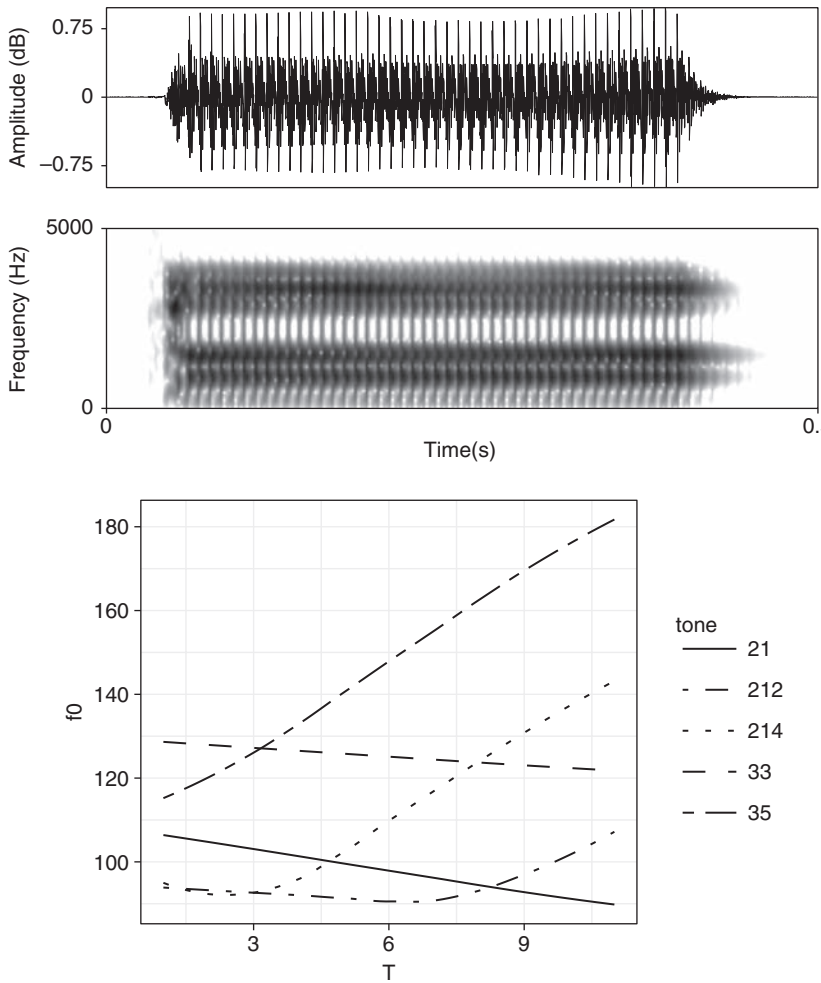

Figure 9.1 (top) Waveform and spectrogram of stimulus /ta:³³/. (bottom) fo contours of synthesized perception stimuli.

"different" pairs, forming all possible permutations of both orders. Within a pair, stimuli were separated by a 300 ms interstimulus interval (ISI). Responses were recorded by pressing keys on a laptop keyboard (g for "same," k for "different," corresponding to the first letter of the corresponding words in Vietnamese). Five hundred milliseconds of silence followed each button press before the next stimulus pair was

presented. A short ISI was selected as nonnative listeners are typically found to have better discrimination in short ISI conditions (Burnham & Francis, 1997; Wayland & Guion, 2003; Werker & Tees, 1984). Participants heard each pair five times, with presentation order randomized within block and participant.

All participants completed a short pretest with 10 pairs (5 same, 5 different) to insure they understood the nature of the experimental task. The entire experiment took most participants about 10–15 minutes to complete.

9.4 Results

For brevity and expositional clarity, and given the small sample size of the study, we focus here primarily on descriptive statistics and informative visual displays. The reader interested in more sophisticated statistical summaries should consult the data and code, available at https://doi.org/10.7488/ds/2635.

9.4.1 Production

Figure 9.2 plots the f0 contours for the five Southern Vietnamese tones averaged over VN (left) and KG (right) speakers. Among the KG speakers we observe pitch range compression, typical of both tonal (Chen, 1974) and nontonal (Mennen, 1998; Zimmerer et al., 2014) L2; deviation from native-speaker targets in terms of the timing of the turning points (Wang et al., 2003); and a possible merger/confusion between the two complex contour tones 212 and 214, perhaps unsurprising given that they are acoustically indistinguishable for at least the first 30 percent of their excursions.

Table 9.2 shows the mean global distances between the KG and VN productions of the Vietnamese tones. As the Fréchet and DTW distances are strongly correlated (ρ = 0.82), the remainder of the chapter will focus on the Fréchet distance.[3] For the KG speakers, mean Fréchet distance correlates most strongly with speaker AGE (0.72), followed by EDUCATION (−0.53) and to a lesser extent VIETNAMESE USAGE (−0.35). AGE and

[3] It is worth noting that the ranking is not perfectly matched, with the Fréchet distance penalizing the shallow slope of the KG realization of the /35/ *sắc* tone more heavily than DTW.

Table 9.2 *Mean global Fréchet and DTW distances between KG and VN tone productions, from most to least similar*

Tone		Fréchet	DTW
33	ngang	1.1	8.0
21	huyền	1.9	10.6
212	nặng	2.2	15.8
214	hỏi-ngã	2.9	14.1
35	sắc	3.1	13.7

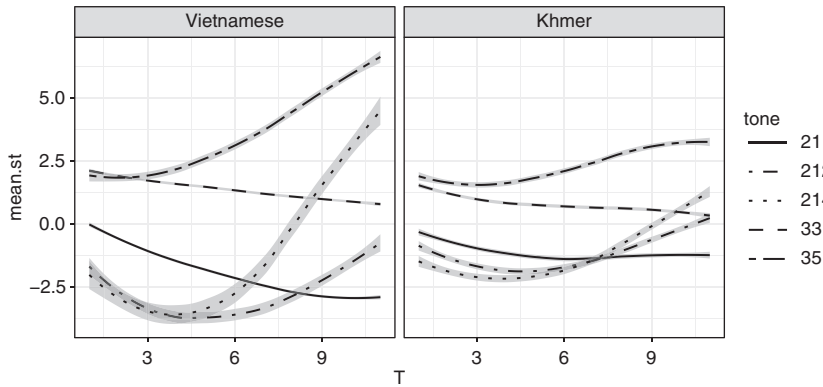

Figure 9.2 Average f0 contours for Southern Vietnamese tones across speakers by L1. Shading ribbon, where present, indicates 95 percent confidence interval.

EDUCATION are negatively correlated (−0.67), as are VIETNAMESE USAGE and AGE (−0.5), while self-reported USAGE increases with EDUCATION (0.62).

Although the averages in Figure 9.2 are broadly representative, there was also considerable individual variation among the Khmer (but not Vietnamese) participants. Figure 9.3 shows the tones produced by 6 of the 18 KG speakers, averaged over utterances (plots for all speakers can be found in the Supplementary Materials). In general, older speakers tended to group tones into two pitch registers, such as high and low (KM7, KF4) or high and rising (KF1). Interestingly, which tones were grouped together was not always consistent: for example, the 33 tone seems to be treated as part of a high register for KM7 and KF1, but as part of a low register by KF4.

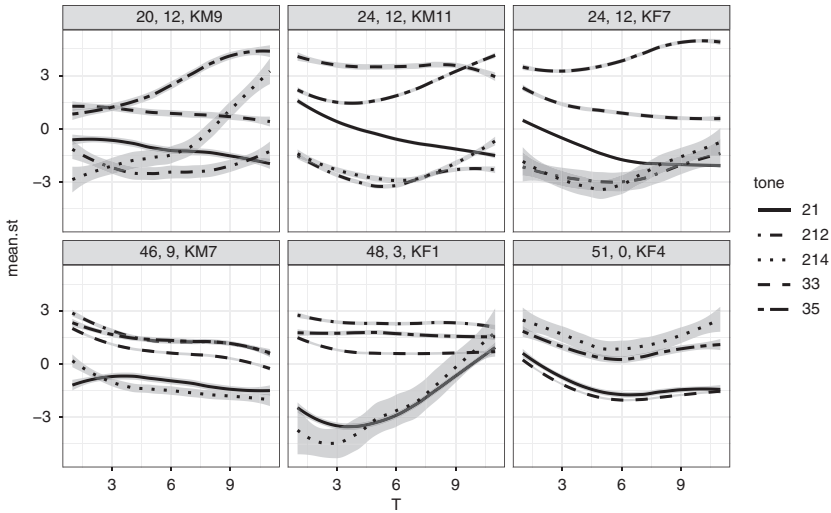

Figure 9.3 Tone productions for six KG participants, averaged over repetitions of each target syllable. The header for each panel shows age, highest grade completed (scale of 0–12), and subject code (KM = male, KF = female). Shading ribbon, where present, indicates 95 percent confidence interval.

9.4.2 Perception

The results of the AX discrimination task were converted into accuracy scores (1 = correct, 0 = incorrect). The results are plotted in Figure 9.4, which shows just the "different" responses; however, including all responses does not meaningfully impact the results (a change of just 1.5 percent in the mean difference in accuracy across all participants). Results are collapsed across presentation order, that is, 33–21 and 21–33 are both treated as a single pair 33/21. Vietnamese participants had an overall mean accuracy of 89 percent, while mean accuracy for Khmer participants was 71 percent. Khmer listeners appeared to have the most difficulty with pairs involving overlapping pitch ranges, especially 21/212 (*huyền/nặng*) and 21/214 (*huyền/hỏi-ngã*). Of note is the fact that the 212/214 (*nặng/hỏi-ngã*) pair was difficult for both groups; this is likely due to the speeded nature of the AX task, combined with the fact that these stimuli are identical for nearly a third of their total excursions. Simple generalized linear mixed-effect logistic regressions predicting the correctness of each trial (correct/incorrect) on the basis of TRIAL, TONE PAIR, and LANGUAGE (with subject-specific intercepts) are consistent with the figure: a model with a predictor LANGUAGE provides a better fit than one with just TRIAL,

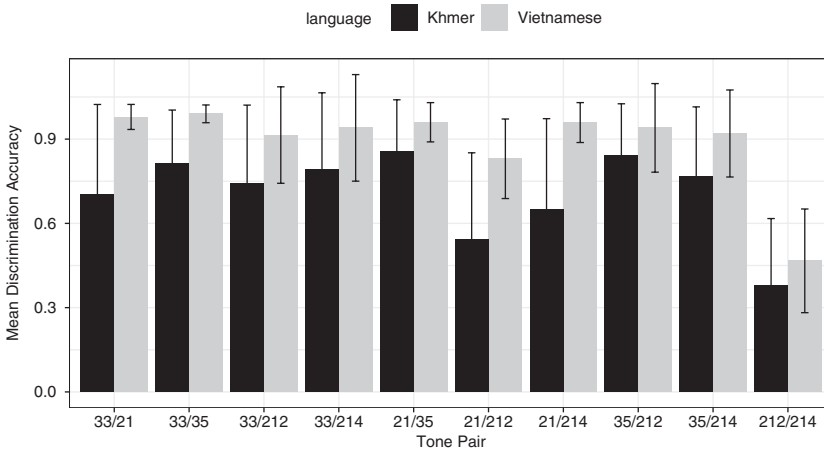

Figure 9.4 Mean discrimination accuracy by tone pair, averaged over speakers and repetitions.

TONE PAIR, and their interaction (χ^2 = 7.26, *df* = 1, *p* = 0.007), and is further improved by the addition of a TONE PAIR: LANGUAGE interaction (χ^2 = 31.24, *df* = 10, *p* < 0.001), which better models the group-level differences in discrimination accuracy of the pairs such as 21/212 and 21/214.

To get a sense of how the demographic variables (AGE, EDUCATION, VIETNAMESE USAGE) correspond to discrimination accuracy, we computed a mean discrimination accuracy for each Khmer listener and correlated this with each variable. Discounting the responses of one clear outlier (KM5, who appeared to have treated this as a *dis*similarity task), mean accuracy was correlated most strongly with EDUCATION (0.65) and to a lesser extent (inversely) with AGE (–0.35). The weakest correlation was with VIETNAMESE USAGE (0.13).

9.4.3 Relating Production and Perception

Figure 9.5 shows the production patterns of two speakers, KM10 (male, age 19, completed seventh grade), and KF1 (female, age 48, completed third grade), with their mean pair-level discrimination accuracies given in Table 9.3. These two speakers illustrate two types of patterns in the data. First, accuracy in distinguishing one tone from another can be quite poor even when production of those tones is objectively native-like. For example, KM10 produces rather native-like tones /33/ and /212/ (Fréchet distances from VN productions of 1.25 and 1.22, respectively), but was at

Table 9.3 *Mean discrimination accuracies for KM10 and KF1 by tone pair*

Pair	KM10	KF1
33/21	0	0.3
33/35	0.9	0.3
33/212	0.56	0.6
33/214	0.6	0.5
21/35	0.75	0.7
21/212	0.38	0.7
21/214	0.57	0.4
35/212	1	0.6
35/214	0.63	0.3
212/214	0.43	0.8

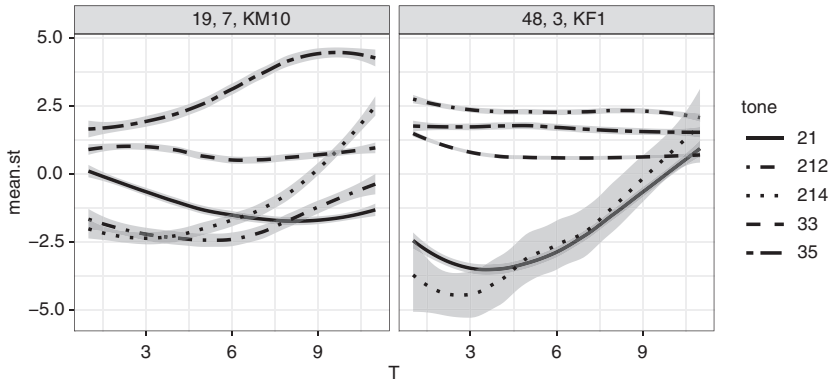

Figure 9.5 Tone productions for KM10 and KF1. Shading ribbon, where present, indicates 95 percent confidence interval.

chance distinguishing them from one another (mean discrimination accuracy of 0.56). Similarly, his native-like tone /21/ production (δ = 1.52) did not seem to help him distinguish it from tone 33, which he failed to do on every trial.

At the same time, these data suggest that listeners can be relatively good at discriminating two tones even when their productions are not native-like, so long as they are acoustically distinct. This is illustrated by

KF1, whose productions of /214/ and /212/ are rather dissimilar to native targets (δ = 2.6 and 5.7 from VN), but who nevertheless is fairly accurate at discriminating these tones, perhaps because she keeps them distinct in her own productions. Conversely, her (non-native-like) production of /21/ (δ = 3.7 from VN) is virtually identical to her (rather more native-like) /214/ tone, and her discrimination accuracy on this tone pair is less than 50 percent. Based on these observations, we explored two possible ways of relating production and perception of L2 tone, based on two different operationalizations of production accuracy: as a deviation from native norms (9.4.3.1), and as a within-speaker difference between tone pairs (9.4.3.2). In both cases, we operationalize perception as discrimination accuracy averaged over all pairs in which a tone occurs.

9.4.3.1 Correlation with Mean Discrimination Accuracy

First, for each tone *T* for each speaker, we compared the Fréchet distance between *T* and its VN exemplar with that speaker's mean discrimination accuracy over all pairs containing *T* (a rough and ready measure of "perception accuracy"). For example, speaker KM10's production of tone /35/ had a (fairly high) mean Fréchet distance from the VN target of 2.25, but a mean discrimination accuracy of (0.9 + 0.75 + 1 + 0.63)/4 = 0.82. The overall correlation was weak (ρ = −0.3), but in the expected direction: smaller Fréchet distances correlate with higher discrimination accuracies. We then fit a linear mixed model predicting DISCRIMINATION ACCURACY from a linear combination of FRÉCHET DISTANCE, AGE, EDUCATION and VIETNAMESE USAGE, with random intercepts for SPEAKER and TONE and by-speaker slopes for DISTANCE. The coefficient estimate for DISTANCE was 0.7, with a standard error of 0.78 and a *t* value of 0.89; thus, even if this effect is robust (and given the small sample size, it is almost certainly anticonservative), this would mean that a fairly large one-unit change in Fréchet distance would on average correspond to less than a 1 percent difference in discrimination accuracy. None of the demographic predictors emerged as statistically significant (*p*-values from 0.06 to 0.33), and coefficient estimates were again very small, ranging from −0.5 to 1.6.

9.4.3.2 Correlation with Pairwise Discrimination Accuracy

Next, on the basis of the within-subject separations observed in Section 9.4.3, we correlated the Fréchet distance between a Khmer speaker's own productions of a particular tone pair – *regardless of their similarity to native-speaker productions* – with their discrimination accuracy for that same tone pair. For example, KF1 has a large Fréchet distance between

her own productions of /21/ and /212/, since she ("incorrectly") produces /21/ as a high level tone, but her discrimination accuracy on this pair is fairly high (0.7). As in Section 9.4.3.1, the overall strength of correlation was weak ($\rho = 0.3$) but in the expected direction: larger Fréchet distance correlates with higher discrimination accuracy. Here, in a linear mixed model predicting DISCRIMINATION ACCURACY from a linear combination of DISTANCE, AGE, EDUCATION and VIETNAMESE USAGE, with random intercepts for SPEAKER and TONE PAIR and by-speaker slopes for DISTANCE, the DISTANCE predictor is statistically significant ($\beta = 2.95$, $SE = 1.34$, $t = 2.20$) but the effect size remains very small.

9.5 Discussion

In general, both Khmer and Vietnamese listeners were able to accurately discriminate most pairs of Vietnamese tones. While the native Vietnamese listeners had overall higher discrimination accuracies, the Khmer listeners were also fairly skilled at this task, and both groups had difficulty with the same pairs of tones. Production, conversely, was much more variable: some KG participants produced Vietnamese tones that were quite close to those of native speakers, while others produced realizations that would potentially confuse a native listener if produced in isolation.

In terms of the Fréchet distance between a given L2 production of a tone and its native-speaker exemplar, we found the largest raw correlation to be with speaker age. All else being equal, younger KG speakers were more likely to produce tones which were more similar to those of native speakers. Discrimination accuracy was best predicted by amount of education, which correlates strongly with age only for the oldest and youngest speakers in our sample. The tonal pairs which presented the most difficulty for KG listeners were those which shared aspects of phonetic realization such as pitch height and contour, although to some extent these proved challenging for the native listeners as well, probably due to the speeded nature of the discrimination task.

We also considered two approaches to relating tone production and discrimination. The first compared KG speakers' tone productions to those of native speakers by measuring the acoustic distance between the f0 contours of KG speakers and VN exemplars. The second compared the acoustic distance between any two tones in a given speaker's own tone productions with that speaker's ability to discriminate between native-speaker productions of those same tones. Modest correlations were observed in both cases, but while the effect of speaker-internal distance

was significant in our second model, the size of the effect was extremely small after parceling out the variation due to individuals and tones.

All of our KG participants demonstrated high, if not completely native-like, perceptual discrimination performance, consistent with the prediction made by models like the SLM that perceptual facility precedes production ability. The productions, compared to native-speaker exemplars, were much more variable. The relative uniformity of perceptual accuracy and the high degree of variability in production mirror the findings of Baese-Berk (2019) and Nagle (2018), and underscore the finding that production accuracy is not necessarily promoted by having achieved a native-like perceptual facility. Although we do not have data on the time course of acquisition, it is clear that strong perceptual skills do not automatically transfer to production, a result which corroborates other L2 studies (e.g., Kartushina et al., 2015). This would appear to hold regardless of whether or not L2 perceptual abilities preceded production for all of our KG participants. In this respect, the present findings do not appear to support the prediction of models like PAM and SLM that perception and production will converge over the course of learning, but it is worth considering the possible reasons why.

One reason may have to do with the interaction of input and usage rates. Bohn & Flege (1997) suggest that experience affects production more than perception. They found that experienced L1 German learners of L2 English (designated as speakers who had lived in the United States for at least five years) were able to produce an /a-æ/ contrast not present in their L1 more accurately than inexperienced German learners of English. However, degree of experience had less of an impact on perception, consistent with the predictions of the SLM. If perception is tuned fairly early in acquisition, the considerable, if passive exposure to Vietnamese tones may explain the relatively good discrimination abilities of our KG participants. Conversely, as shown by Bohn & Flege, improving production at a later stage is possible, but requires a real difference in usage rate. While all of our KG participants grew up in an environment where Vietnamese would be heard, not all of them used it to the same extent, and crucially, these usage rates may have been different at particular time periods over the course of L2 acquisition.

The weak correlation we observe between acoustic separation in a speaker's own L2 production repertoire and his or her ability to distinguish two tones in perception is especially intriguing. This finding seems consistent with work showing that the degree to which a speaker clearly differentiates two L1 categories in production correlates with facility to

discriminate those categories in perception (Byun & Tiede, 2017; Ghosh et al., 2010; Perkell et al., 2004). This type of production–perception correlation is predicted by models of speech production such as DIVA (Guenther & Perkell, 2004), in which planning goals are regions in a multidimensional, acoustic-auditory and somatosensory space. What is interesting in the present case is that this would seem to hold even when the acoustic-auditory input fails to match the production region. What seems more relevant for predicting discrimination accuracy in our study is not whether tones are well separated in the *native* acoustic space, but in the listener's own production repertoire (with the important caveat that the correlation coefficient was rather small). This suggests that the relation between L2 production and perception may be mediated by the L2 acoustic targets, even if these are objectively non-native-like. That is, learners would have categories for each tone class, as abstractions over sets of lexical items, and would learn to associate native Vietnamese pitch contours with those classes. At the same time, they would be developing a separate set of production routines, also associated with those same tone classes/lexemes, but which may not bear any particular resemblance to the pitch targets learned from perception. If the production routines are co-activated when receiving acoustic input, having well-separated production targets for tones A and B would facilitate perception.

This scenario supposes that, even in a setting which is supposed to target low-level, precategorical phonetic information, L2 discrimination is nevertheless mediated through some kind of intermediate representation. This may seem unexpected in the context of the current study, given that the very short (300 ms) ISI used is expected to discourage the use of phonological processing. However, as noted by Wayland & Guion (2003), while a short ISI can facilitate discrimination for inexperienced listeners, this does not necessarily rule out access to phonological information, especially for more experienced learners. We further note sporadic reports of language-specific effects in speeded AX discrimination elsewhere in the literature (e.g., Huang, 2007).

Our findings also lead us to ask how some speakers come to develop tonal production targets that are so divergent from the native-speaker exemplars. One possibility is that L2 Vietnamese tone perception is actually affected by the KG speakers' L1 prosodic system. The tendency of older speakers to group tones into two registers is consistent with findings indicating less proficient listeners are more likely to be sensitive primarily to tone height than contour (Gandour, 1983; Hallé et al., 2004). It might also be related to the fact that KG Khmer has a nascent pitch-based

contrast between level and rising fo (Kirby & Đinh Lư Giang, 2017; Thạch Ngọc Minh, 1999). However, this quasi-tonal use of fo is extremely limited in KG Khmer, distinctive only in items which have lost /r/ in onset position (e.g., Standard Khmer /krɑː/ > KG Khmer [kǎː] "poor," SK /riən/ > KG [hǐən] "to learn") and distinguishing perhaps 20 or 30 minimal pairs. Furthermore, there is no evidence that this use of fo has spread or is spreading to any other contexts. As demonstrated by So & Best (2010), experience with L1 tones (or other prosodic suprasegmentals) does not necessarily facilitate L2 tone perception, but depends heavily on both the phonemic status of the contrast as well as the phonetic features of the tones themselves. For all practical intents and purposes, we view KG Khmer as a nontonal language, and thus are more inclined to attribute the differences between speakers to properties of those individuals such as age, fluency, and degree of usage/exposure.

Finally, it is worth bearing in mind that the statistical evidence of any production–perception link can be impacted by methodological, as well as linguistic factors. As Nagle (2018) and Sakai & Moorman (2018) remind us, the type of task chosen in a given L2 study may considerably impact the results. On the production side, the present study utilized a simple reading task, using aural and orthographic prompts. However, we must recognize the possibility that some participants may simply have been confused about which item they were expected to produce. Despite prompting by a native speaker of Southern Vietnamese (the second author), this procedure did not guarantee imitation; if the participant misheard the cue, they may have been accurately producing the tone they thought they had been asked to produce. The desire to obtain a minimal tone set (where the syllable content did not vary) meant including items that were difficult to depict in a picture-naming task. Similarly, we should be careful not to overinterpret the results of our AX discrimination experiment as a stand-in for "perception." Recall that Yang (2015) determined production abilities tended to be ahead of perception for L1 English late learners of L2 Mandarin. However, Yang's perception study was a 4AFC lexical identification task, in which real lexical items in a meaningful carrier phrase were heard with a range of resynthesized fo contours. This is clearly a very different kind of task from speeded AX discrimination, with the latter tapping primarily into auditory abilities rather than phonological or lexical knowledge. In short, while one can imagine a range of improvements to our experimental procedures, we simply point out that the present findings are likely heavily task-dependent and should be interpreted with appropriate caution.

9.6 Summary

We compared the lexical tone productions by native speakers of Southern Vietnamese with those of speakers of Kiên Giang Khmer with L2 knowledge of Vietnamese, and also considered the discrimination of tones for the same L2 speakers. Production accuracy, as measured by the Fréchet distance between f0 contours, was most strongly predicted by age, while discrimination correlated best with the length of a listener's education. The correlations observed between production and perception – one between discrimination accuracy and the acoustic distance from a native-speaker exemplar, and one between discrimination accuracy and the speaker-specific acoustic separation – were at best modest. Our results are broadly consistent with previous work indicating that L2 production can be independent of perception; however, for the purpose of understanding how production and perception are related, we suggest that the notion of "accuracy" in production may benefit from considering measures in addition to the degree to which a native-speaker target is approximated.

Supplementary Materials

The data and R code necessary to reproduce all figures and statistical results in this chapter, along with additional figures and analyses, is available at https://doi.org/10.7488/ds/2635.

References

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, and Psychophysics*, *81*(4), 981–1005.

Bauman, J., Blodgett, A., Rytting, C. A., & Shamoo, J. (2009). *The ups and downs of Vietnamese tones: A description of native speaker and adult learner tone systems for Northern and Southern Vietnamese* (Technical Report No. E.5.3 TTO 2118). College Park, MD: University of Maryland Center for Advanced Study of Language.

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

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

Bettoni-Techio, M., Rauber, A. S., & Koerich, R. D. (2007). Perception and production of word-final alveolar stops by Brazilian Portuguese learners of English. In *INTERSPEECH 2007* (pp. 2293–2296), Antwerp.

Blodgett, A., Bauman, J., Bowles, A., & Winn, M. B. (2008). A comparison of native speaker and American adult learner Vietnamese lexical tones. In *Proceedings of Acoustics 08* (pp. 688–692), Paris.

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer* (Version 5.4.08).

Bohn, O.-S., & Flege, J. E. (1997). Perception and production of a new vowel category by second-language learners. In A. James & J. Leather (Eds.), *Second-language speech: Structure and process* (pp. 53–74). Berlin: Walter de Gruyter.

Brunelle, M. (2015). Vietnamese (Tiếng Việt). In M. Jenny & P. Sidwell (Eds.), *The handbook of Austroasiatic languages* (Vol. 2, pp. 909–953). Leiden: Brill.

Burnham, D., & Francis, E. (1997). The role of linguistic experience in the perception of Thai tones. In A. S. Abramson (Ed.), *Southeast Asian linguistics studies in honor of Vichin Panupong* (pp. 29–48). Bangkok: Chulalongkorn University Press.

Byun, T. M., & Tiede, M. (2017). Perception-production relations in later development of American English rhotics. *PLoS ONE*, *12*(2), e0172022.

Chambers, E. W., Colin de Verdière, É., Erickson, J., Lazard, S., Lazarus, F., & Thite, S. (2010). Homotopic Fréchet distance between curves or, walking your dog in the woods in polynomial time. *Computational Geometry*, *43*(3), 295–311.

Chen, G. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, *2*(2), 159–171.

Ding, H., Hoffmann, R., & Jokisch, O. (2011). An investigation of tone perception and production in German learners of Mandarin. *Archives of Acoustics*, *36*(3). doi:10.2478/v10168-011-0036-6

Đinh Lư Giang. (2011). Tình hình song ngữ Khmer-Việt tại đồng bằng sông Cửu Long: một số vấn đề lý thuyết và thực tiễn [Khmer-Vietnamese bilingualism in the Mekong Delta: Theoretical and practical issues]. PhD dissertation, Ho Chi Minh City University of Social Sciences and Humanities.

Đinh Lư Giang. (2015). Các đặc điểm chính của song ngữ Khmer-Việt vùng Nam Bộ [The main features of Khmer-Vietnamese bilingualism in the South]. *Ngôn ngữ & Đời sóng*, *4*(234), 81–88.

Elvin, J., Williams, D., & Escudero, P. (2016). The relationship between perception and production of Brazilian Portuguese vowels in European Spanish monolinguals. *Loquens*, *3*(2), e031.

Escudero, P. R. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. Utrecht: LOT.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of the Acoustical Society of America*, *93*(3), 1589–1608.

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

Flege, J. E. (1999). The relation between L2 production and perception. In *Proceedings of the XIVth International Congress of Phonetics Sciences* (pp. 1273–1276), Berkeley.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, *25*(2), 169–186.

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America*, *106*(5), 2973–2987.

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268–294.

Gandour, J. T. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, *11*, 149–175.

General Statistics Office of Vietnam. (2010). *The 2009 Vietnam population and housing census: Major findings*. Hanoi: General Statistics Office of Vietnam. Retrieved from www.gso.gov.vn/default_en.aspx?tabid=515&idmid=5&ItemID=9813

Ghosh, S. S., Matthies, M. L., Maas, E., … Perkell, J. S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *Journal of the Acoustical Society of America*, *128*(5), 3079–3087.

Guenther, F. H., & Perkell, J. S. (2004). A neural model of speech production and supporting experiments. In *From sound to sense: 50+ years of discoveries in speech communication* (pp. B98–B106). Cambridge, MA: MIT Press.

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 57–77). Amsterdam: John Benjamins.

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

Hoàng Thị Châu. (1989). *Tiếng Việt trên các miền đất nước: Phương ngữ học* [Vietnamese in the various areas of the motherland: A dialectological study]. Hà Nội: NXB Khoa học Xã hội.

Huang, T. (2007). Perception of Mandarin tones by Chinese-and English-speaking listeners. *In Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1797–1800), Saarbrücken.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *Journal of the Acoustical Society of America*, *138*(2), 817–832.

Kirby, J. (2014). Incipient tonogenesis in Phnom Penh Khmer: Acoustic and perceptual studies. *Journal of Phonetics*, *43*, 69–85.

Kirby, J., & Đinh Lư Giang. (2017). On the r>h shift in Kiên Giang Khmer. *Journal of the Southeast Asian Linguistics Society*, *10*(2), 66–85.
Lee, Y.-S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, *25*(5), 527–542.
Levy, E. S., & Law, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *Journal of the Acoustical Society of America*, *128*(3), 1290–1305.
Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. In *Proceedings of the 13th International Congress of Phonetic Sciences* (Vol. 4, pp. 92–99), Stockholm.
Mennen, I. (1998). Can language learners ever acquire the intonation of a second language? In *STiLL-1998* (pp. 17–20), Marholmen, Sweden.
Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, *24*, 49–65.
Morrison, G. S. (2003). Perception and production of Spanish vowels by English speakers. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1533–1536), Barcelona.
Müller, M. (2007). Dynamic time warping. In *Information retrieval for music and motion* (pp. 69–84). Berlin: Springer.
Nagle, C. L. (2018). Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning*, *68*(1), 234–270.
Nguyen, H. T., & Macken, M. A. (2008). Factors affecting the production of Vietnamese tones: A study of American learners. *Studies in Second Language Acquisition*, *30*(1), 49–77.
Nguyễn Thị Huệ. (2010). Tiếp xúc ngôn ngữ giữa tiếng Khmer với tiếng Việt (trường hợp tỉnh Trà Vinh) [Language contact between Khmer and and Vietnamese in Tra Vinh province]. PhD dissertation, Ho Chi Minh City University of Social Sciences and Humanities.
Nguyễn Văn Lợi & Edmondson, J. A. (1998). Tone and voice quality in modern northern Vietnamese: Instrumental case studies. *Mon-Khmer Studies*, *28*, 1–18.
Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *INTERSPEECH* (pp. 161–164), Florence.
Perkell, J. S., Guenther, F. H., Lane, H., … Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America*, *116*(4), 2338–2344.
Phạm, A. H. (2003). *Vietnamese tone: A new analysis*. New York: Routledge.
Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261.

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

Sochoeun, C. (2006). Khmer Krom migration and their identity. MA thesis, Royal University of Phnom Penh, Phnom Penh.

Strange, W. (1995). Phonetics of second-language acquisition: Past, present, future. In P. Branderud & K. Elenius (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (pp. 76–83), Stockholm.

Thạch Ngọc Minh. (1999). Monosyllabization in Kiengiang Khmer. *Mon-Khmer Studies*, *29*, 81–95.

Vũ Thanh Phương. (1982). Phonetic properties of Vietnamese tones across dialects. In D. Bradley (Ed.), *Papers in Southeast Asian linguistics: No. 8. Tonation* (pp. 55–76). Canberra: Pacific Linguistics.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, *113*(2), 1033–1043.

Wang, Y., Sereno, J. A., & Jongman, A. (2012). L2 acquisition and processing of Mandarin tones. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *Handbook of East Asian psycholinguistics: Vol. 1. Chinese* (pp. 250–256). Cambridge: Cambridge University Press.

Wayland, R. P., & Guion, S. (2003). Perceptual discrimination of Thai tones by naive and experienced learners of Thai. *Applied Psycholinguistics*, *24*(01), 113–129.

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

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, *75*(6), 1866–1878.

Yang, B. (2012). The gap between the perception and production of tones by American learners of Mandarin – an intralingual perspective. *Chinese as a Second Language Research*, *1*(1), 33–53.

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Berlin: Springer. Retrieved from http://link.springer.com/10.1007/978-3-662-44645-4

Zimmerer, F., Jügler, J., Andreeva, B., Möbius, B., & Trouvain, J. (2014). Too cautious to vary more? A comparison of pitch variation. In *Proceedings of the 7th International Conference on Speech Prosody* (pp. 1037–1041), Dublin.