PERCEPTION, IMITATION, AND PRODUCTION: EXPLORING A THREE-WAY PERCEPTION-PRODUCTION LINK

Charles Nagle

Iowa State University
cnagle@iastate.edu

ABSTRACT

The Speech Learning Model (SLM) [3] posits that accurate perception facilitates accurate production in second language (L2) speech learning. Although studies indicate that the two domains are correlated, precisely when and how they become aligned over time merits further attention. The present study therefore investigated relationships among the perception, imitation, and production of L2 Spanish stops.

Thirty L1 English university students participated over their first two semesters of Spanish language coursework. In monthly sessions, they completed an oddity task, delayed word repetition, and picture description. Perception was operationalized as A', imitation as voice onset time (VOT) in word-initial stops on the repetition task, and production as VOT in word-initial stops on the picture description task. Separate mixed-effects models were fit to L2 /b/ and /p/ data. Perception and imitation did not predict VOT in L2 /p/, but imitation was a significant predictor of prevoicing in L2 /b/.

Keywords: perception-production link; second language learning; longitudinal; Spanish

1. INTRODUCTION

Current theoretical models of L2 speech learning such as the SLM [3] argue that accurate perception guides accurate production. According to the SLM, if L2 learners detect the difference between native (L1) and target language sounds that are similar but not identical, then they should be able to create a new phonetic category, leading to accurate L2 production. In contrast, if learners equate the two sounds, associating the L2 sound with the L1 category, then over time, they will converge phonetically under the shared category. The likelihood of learners detecting L1-L2 differences is tied to age of onset, which represents the degree to which the perceptual system has become attuned to the L1. Thus, even though phonetic learning remains possible across the lifespan, adult learners may struggle to perceive and produce differences between cross-linguistically similar sounds. Many cross-sectional [5, 6] and training [10, 17] studies have shown a relationship between the two domains, but perception-production correlations often fall in the weak to moderate range, and null results are not uncommon [8]. Given the SLM hypothesis that with time and L2 experience learners may discern differences between similar sounds, which could then lead to more accurate production, more longitudinal work investigating when and how perception and production become aligned [12] is needed.

Relatedly, even though studies have provided insight into how discrimination and identification of L2 sounds potentially shape production, there is a need to examine other skills that could underlie the perception-production link. For example, Simulation Theory (ST) [7] contends that listeners anticipate upcoming speech gestures through covert imitation, recalibrating the perception-production system if predicted and observed input do not match one another. According to this perspective, imitation could be an important bridge between perception and production, especially if better imitators show greater flexibility in articulatory skills and phonetic categories [14]. Given this hypothesis and the fact that at least in some instances accurate perception seems to be necessary but not sufficient in and of itself to promote accurate production [16], it would be advantageous to investigate the role imitation plays in the perception-production link. Addressing the need for more longitudinal research including a variety of perception and imitation tasks that could lead to more accurate speech production, this exploratory study reports on L1 English speakers' perception, imitation, and production of L2 Spanish stops over two semesters of introductory Spanish language coursework.

Following the SLM, English and Spanish stops can be classified as similar sounds. English voiceless stops are aspirated in word-initial position, but Spanish stops are not. VOT values for English /p, t, k/ are in the 30–60 ms range, whereas values of 10–30 ms are common for Spanish. English voiced stops are variably realized with prevoicing or with a short delay in voicing similar to Spanish voiceless stops. In contrast, Spanish voiced stops are prevoiced [15]. Thus, English speakers need to learn to discriminate voiced and voiceless unaspirated stops (e.g., [b] vs. [p]), to produce phonologically voiced stops with

prevoicing, and to produce phonologically voiceless stops with the shorter VOTs typical of Spanish.

2. METHOD

This study took place over one calendar year while students were enrolled in the first two semesters of university Spanish coursework. Thirty-seven participants were recruited from a first-semester course and invited to participate in monthly sessions over the 8-month academic year. The first meeting was a practice session, allowing participants to become familiar with the experimental tasks. Therefore, data from this session was not analyzed. Three participants were excluded because they were not L1 English speakers, and one participant was excluded because he became aware of the purpose of the study during data collection. Additionally, three participants did not return after session 0.

2.1. Participants

Thirty participants (22 females) returned for session 0, the first session after the practice session. All participants were L1 English speakers who had learned Spanish through classroom instruction. The mean age of onset was 13.07 (*SD* = 4.31) years, and participants reported an average of 1.68 (*SD* = 1.71) years of Spanish instruction in secondary school. None of the participants had spent time in a Spanish-speaking country for the purpose of language learning. Sample size decreased over the study (cf. Table 1) because participants decided to withdraw from their Spanish course, declined to enroll in the second-semester course, or withdrew from the study.

2.2. Tasks

Participants completed a battery of perception, production, and individual difference tasks at each hour-long session. This report focuses on the oddity, delayed repetition, and picture description tasks. Participants completed the picture description first using a PowerPoint file, followed by the delayed repetition and oddity tasks, both of which were presented using SuperLab 5.0 software.

Eight versions of each task were created, and versions were randomized across sessions such that each participant received a unique version at each session. Individual data collection sessions took place in a sound-treated room, and a dynamic, head-mounted microphone was used for recording.

2.2.1. Picture Description

Eight PowerPoint image sets were compiled for this task. Each set consisted of five images depicting an action, such as a man fishing at the beach. Images were selected to elicit words beginning with /b/ and /p/ while keeping in mind the basic vocabulary with which participants were familiar (e.g., *bailar*, 'to dance,' *pescar*, 'to fish'). Up to four keywords or phrases were included on each image. Participants had up to 20 seconds to look over each image before describing it, but they were not allowed to script a response. If they were unable to produce a response, they were instructed to read the words appearing on the image aloud. In practice, only a few individuals took advantage of this option at the first session.

2.2.2. Delayed Word Repetition

Delayed word repetition was used to evaluate participants' ability to imitate prevoicing in Spanish /b/ and short-lag VOT in Spanish /p/. On each trial, they heard a target verb in Spanish and repeated it as accurately as possible after a three-second delay. There were 10 target verbs, five each for word-initial /b/ and /p/, and five distractors, all of which were drawn from participants' introductory textbook. Verbs were conjugated in the present tense in the first or second person singular to mimic the textbook exercises that participants completed as part of daily assignments. Stimuli were recorded by a male native speaker (NS) of Argentinian Spanish.

2.2.3. Oddity Task

An oddity task [4] was used to evaluate participants' perception of three target contrasts: [b]-[p], [p]-[p^h], and [b]-[p^h]. On each trial, participants heard a triplet (e.g., [ba]^1-[ba]^2-[pa]) with tokens separated by a 1.3 second interstimulus interval, and had to indicate the position of the odd item by pressing '1,' '2,' or '3' on the keyboard, or the 'N' key for same trials (e.g., [ba]^1-[ba]^2-[ba]^3). Twelve triplets were included per contrast, six odd trials and six same trials. The position of the odd item was counterbalanced, appearing twice in each position. Two male speakers, one a NS of Argentinian Spanish with nativelike proficiency in English (the same speaker who recorded stimuli for the delayed repetition task) and the other, a NS of American English with nativelike proficiency in Spanish, recorded the stimuli for the task. Stimuli from a single speaker were combined to form each trial, but an equal number of trials was compiled for each speaker to prevent participants from becoming attuned to the speech of a single individual. The stimuli included in each triplet were acoustically distinct, such that even on same trials, participants heard three different renditions of the target syllable.

Three additional contrasts ([sa]-[se], [ma]-[na], and [je]-[le]) were included as distractors.

2.3. Coding

A' scores were computed for the target contrasts on the oddity task for each participant at each session. A' is the nonparametric extension of d', a signal detection theory measure of contrast sensitivity that takes into account response bias.

For the target words on the delayed repetition task, VOT was labelled in Praat version 5.4.08 [2] and extracted using a script. Positive VOT was coded from the release burst of the stop to the onset of voicing in the following segment. Negative VOT (prevoicing) was coded from the onset of low frequency periodic energy during stop closure to the release burst.

For the picture description, audio files were transcribed, and a forced aligner was used to generate an initial text grid of the speech. Errors in the text grids were adjusted by hand, and VOT was labelled and extracted for /b/- and /p/-initial words following the conventions outlined above. Word duration was also measured to control for potential relationships between speech rate and VOT.

3. RESULTS

Mixed-effects models were fit to A' (oddity), VOT imitation for /b/ and /p/ (delayed repetition), and VOT production for /b/ and /p/ (picture description). All models were fit in R [13] using the lme4 package [1]. In each case, data was plotted and inspected, and this visualization process guided the selection of unconditional growth models: piecewise growth models estimating separate slopes over the first (sessions 0–2) and second (sessions 3–6) semesters for the oddity and repetition data, and linear models for the picture description data. Fixed effects were backward-tested and random effects forward-tested by performing a chi-square test on the change in deviance statistics of nested models.

3.1. Performance on oddity task over time

Participants' discrimination of [b]-[p^h], the control contrast, was near ceiling for the duration of the study. Overall, they discriminated [p]-[p^h] (*estimate* = .11, *SE* = .02, $p < .001$) more accurately than [b]-[p] (*intercept* = .61, *SE* = .06, $p < .001$), and perception of both contrasts improved significantly over the first semester (*estimate* = .07, *SE* = .03, $p = .02$).

3.2. Performance on delayed repetition over time

At session 0 participants imitated /b/ targets with an average VOT of –32.99 ms (*SE* = 9.18, $p < .001$), and their VOT production decreased significantly over the first semester (*estimate* = –6.65, *SE* = 2.72, $p = .01$). This finding indicates that participants imitated /b/ targets with increasingly Spanish-like VOT over their first semester of intensive Spanish language instruction.

Participants' imitation of VOT in L2 /p/ was stable across the study (*intercept* = 28.37, *SE* = 5.59, $p < .001$). Although modeling demonstrates that there was no group-level change, mean VOT varied substantially among participants, from a minimum of –7 ms to a maximum of 55 ms.

3.3. Performance on picture description over time

Participants' picture description data at each session was analyzed only if they produced at least three tokens for the target phone. As summarized in Table 1, most participants did so for word-initial /p/, but word-initial /b/ production was more variable.

Table 1: Sample size for picture description by target phone at each session.

Session	0	1	2	3	4	5	6
Total *n*	30	30	27	24	23	21	17
/b/ *n*	25	24	24	15	11	9	11
/p/ *n*	29	29	26	23	22	21	17

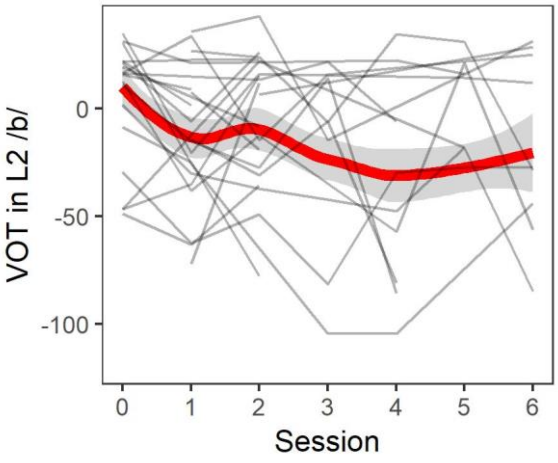
3.3.1. VOT in L2 /b/

Figure 1 plots mean VOT in L2 /b/ for the group (bold red line) against individual trajectories (thin lines) over time. Despite minor fluctuations in rate of change, the overall growth pattern is linear. Thus, a linear growth model was fit to the L2 /b/ data. Session, A' for [b]-[p], and mean VOT imitation for L2 /b/ on the delayed repetition task were included as fixed effects. Age of onset, previous experience, and word duration were grand-mean centered and integrated as control covariates. By-participant and by-word random intercepts were included. By-participant random slopes for session were evaluated but did not improve fit ($\chi^2(2) = 2.04$, $p = .36$).

Participants' average VOT production on /b/-initial picture description words was 12.63 ms (*SE* = 9.72, $p = .20$) at session 0. The fixed effect for session (*estimate* = –4.36, *SE* = .1.54, $p = .005$) was significant, demonstrating that participants produced progressively shorter, more Spanish-like VOT values over time. The mean VOT imitation predictor also reached significance (*estimate* = .25, *SE* = .07, $p < .001$). The positive coefficient shows that

participants who prevoiced /b/ targets on the imitation task (participants who produced /b/ targets with negative VOT) produced a greater amount of prevoicing, or more negative VOT values, on the picture description task. In contrast, A' for [b]-[p] was not a significant predictor of VOT for L2 /b/ (*estimate* = .92, *SE* = 11.94, *p* = .94).

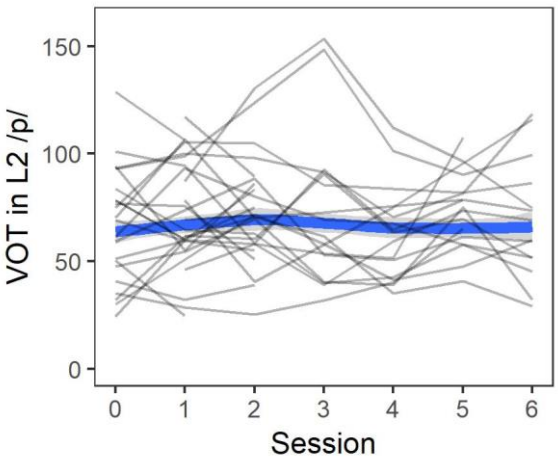
Figure 1: VOT in L2 /b/ over time.



3.3.2. VOT in L2 /p/

Figure 2 plots mean VOT in L2 /p/ on the picture description task. Modeling followed the procedure outlined above. Participants produced an average VOT of 70.84 ms (*SE* = 6.82, *p* < .001) at session 0. As Figure 2 suggests, there was no significant group-level change in VOT over time (*estimate* = –.28, *SE* = .65, *p* = .67). Moreover, neither A' for [p]-[p^h] (*estimate* = 1.23, *SE* = 6.14, *p* = .84) nor mean VOT imitation (*estimate* = –.06, *SE* = .06, *p* = .33) were significantly related to VOT in L2 /p/.

Figure 2: VOT in L2 /p/ over time.



4. DISCUSSION

Contradicting a simple view of the perception-production link, findings demonstrate that sensitivity to the Spanish [b]-[p] contrast was not significantly related to the acquisition of prevoicing in L2 /b/. Research has yet to establish precisely what level of perceptual accuracy is needed before production can improve. Thus, it could be that learners had already established sufficiently robust perceptual representations, laying the groundwork for improvement at a later stage. This would be compatible with a longitudinal view of the perception-production link in which production accuracy lags behind perception [12]. In contrast to the null result for the perception (oddity) measure, imitation was significantly related to prevoicing production in L2 /b/. This finding lends support to Simulation Theory [7] insofar as individual variation in the ability to reproduce L2 articulatory gestures and their timing relations seems to be an additional skill that underlies accurate L2 sound production.

Although sensitivity to [p]-[p^h] improved over the study, imitation and production of Spanish /p/ did not; instead, learners consistently produced longer, English-like VOT. These differing results for L2 /b/ and /p/ may be due to the functional load of the features [11]. On the one hand, acquiring prevoicing for L2 /b/ could be viewed as more communicatively urgent than reducing aspiration in L2 /p/, since producing voiceless unaspirated stops for phonologically voiced stops in Spanish (e.g., realizing /b/ as [p]) could partially neutralize the voicing contrast. This neutralization might then to intelligibility issues. On the other hand, producing aspirated variants for L2 /p/ (e.g., realizing /p/ as [p^h]) would likely not have the same communicative cost, even though it would contribute to foreign accent [18]. In light of these preliminary findings, future work should continue to examine the temporal properties of individual differences [9] and their relationship to L2 speech production over time.

5. ACKNOWLEDGMENTS

This research was supported by a *Language Learning* Early Career Research Grant. I would like to thank Ziwei Zhou, Sonca Vo, Ivana Lucic, Alexandra Urbanski, Laura Valderrama, and Shelby Bruun for their help with data collection, processing, and coding. I also thank Germán Zárate-Sández, Mari Sakai, Pavel Trofimovich, and the anonymous reviews for their insightful comments at various stages of the project.

6. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., Walker, S. 2014. _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1.-7. http://CRAN.R-project.org/package=lme4

[2] Boersma, P., Weenik, D. 2015. Praat: doing phonetics by computer (Version 5.4.08). http://www.praat.org

[3] Flege, J. E. 1995. Second language speech learning: Theory, findings, problems. In: Strange, W. (ed), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Timonium, MD: York Press, 233–277.

[4] Flege, J. E. 2003. A method for assessing the perception of vowels in a second language. In: Fava, E., Mioni, A. (eds), *Issues in Clinical Linguistics*. Padova, Italy: Unipress, 19–43.

[5] Flege, J. E., Bohn, O.-S., Jang, S. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25, 437–470.

[6] Flege, J. E., MacKay, I. R. A., Meador, D. 1999. Native Italian speakers' perception and production of English vowels. *J. Acoust. Soc. Am.* 106(5), 2973–2987.

[7] Gambi, C., Pickering, M. J. 2013. Prediction and imitation in speech. *Frontiers in Psychology* 4, article 340.

[8] Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., Huettig, F. 2012. Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning* 62(S2), 79–109.

[9] Kartushina, N., & Frauenfelder, U. H. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology* 5, article 1246.

[10] 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. *J. Acoust. Soc. Am.* 138(2), 817–832.

[11] Munro, M. J., Derwing, T. M. 2006. The functional load principle in ESL pronunciation instruction: An exploratory study. *System* 34(4), 520–531.

[12] Nagle, C. 2018. Examining the temporal structure of the perception-production link in second language acquisition: A longitudinal study. *Language Learning* 68(1), 234–270.

[13] R Core Team. 2017. R: A language and environment for statistical computing (Version 3.4.2). Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

[14] Reiterer, S. M., Hu, X., Sumathi, T. A., Singh, N. C. 2013. Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. *Frontiers in Psychology* 4, article 782.

[15] Rosner, B. S., López-Bascuas, L. E., García-Albea, J. E., Fahey, R. P. 2000. Letter to the Editor: Voice-onset times for Castilian Spanish initial stops. *Journal of Phonetics* 28, 217–224.

[16] Saito, K., van Poeteren, K. 2017. The perception-production link revisited: The case of Japanese learners' English /ɹ/ performance. *International Journal of Applied Linguistics* 28(1), 3–17.

[17] 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.

[18] Schoonmaker-Gates, E. 2015. On Voice-onset Time as a cue to foreign accent in Spanish: Native and nonnative perceptions. *Hispania* 98(4), 779–791.