2022

Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops

Daniel J. Olson
Purdue University, danielolson@purdue.edu

Recommended Citation

Olson, D. J. (2022). Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops. *Second Language Research, 38*(4), 913–940.

Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops

Daniel J. Olson

Purdue University
640 Oval Dr., West Lafayette, IN, USA, 47907
danielolson@purdue.edu

Abstract
Featural approaches to second language phonetic acquisition posit that the development of new phonetic norms relies on sub-phonemic features, expressed through a constellation of articulatory gestures and their corresponding acoustic cues, which may be shared across multiple phonemes. Within featural approaches, largely supported by research in speech perception, debate remains as to the fundamental scope or "size" of featural units. The current study examines potential featural relationships between voiceless and voiced stop consonants, as expressed through the voice onset time cue. Native English-speaking learners of Spanish received targeted training on Spanish voiceless stop consonant production through a visual feedback paradigm. Analysis focused on the change in voice onset time, for both voiceless (i.e., trained) and voiced (i.e., non-trained) phonemes, across the pretest, posttest, and delayed posttest. The results demonstrated a significant improvement (i.e., reduction) in voice onset time for voiceless stops, which were subject to the training paradigm. In contrast, there was no significant change in the non-trained voiced stop consonants. These results suggest a limited featural relationship, with independent VOT cues for voiceless and voices phonemes. Possible underlying mechanisms that limit feature generalization in L2 phonetic production, including gestural considerations and acoustic similarity, are discussed.

Keywords:
phonetics, features, visual feedback, voice onset time, Spanish

Acknowledgements:
I am grateful for the efforts of John Nielsen who provided technical help for this project. In addition, I want to thank Lourdes Barranco Cortes, Bruno Staszkiewicz Garcia, and Aída García Tejada for their assistance in the classroom.

Introduction

As learners develop competency in their second language, the development of second language (L2) phonetic categories plays a key role in determining the intelligibility, comprehensibility, and accentedness of non-native speech (e.g., Munro and Derwing, 1995). As noted by several theoretical models (e.g., Flege, 1987), the ability of a learner to establish a new phonetic category depends in part on the relationship of the new L2 sound to the L1 phonetic system. While this relationship has been well-documented, the nature of the underlying "units" that learners acquire remains subject to ongoing debate. While some studies have adopted a segmental approach (e.g., Flege et al., 1992), in which acquisition occurs phoneme-by-phoneme, recent research has suggested that acquisition takes place at the level of the feature (De Jong, Hao, and Park, 2009). Featural approaches posit that the development of new phonetic norms relies on features which may be shared across multiple phonemes (De Jong, Hao, and Park, 2009; Olson, 2019). Within featural approaches, debate remains as to the fundamental scope or "size" of featural units. Although features were originally described as binary (or monovalent) (Clements, 1985), more recent work considers phonological features in terms of a constellation of articulatory gestures and acoustic cues (Dmitrieva et al., 2015). Thus, during acquisition, L2 learners must acquire the relevant cues, and may apply these new L2-specific cues to the set of phonemes that share a given "feature".

The current study investigates the nature of the phonetic feature by examining the potential featural relationships between voiceless and voiced stop consonants. Here, voice onset time (VOT) is taken as a fundamental acoustic cue for both voiceless and voiced stop consonants. While both classes of stop consonants make use of VOT, it remains unclear whether they employ a single VOT cue that is shared between both categories (i.e., voiceless and voiced), or if each category employs an independent VOT cue. To address this question, this study examines whether a shift in the VOT cue for voiceless stops results in a shift in the VOT cue for voiced stops.

For this study, English-speaking learners of Spanish received a pedagogical treatment (Olson, 2014b) addressing Spanish voiceless stop consonant production. While English and Spanish both employ a voiced–voiceless distinction for stop consonants, they differ in their VOT (Lisker and Abramson, 1964). Analysis focuses on the change in VOT for both voiceless (i.e., trained) and voiced (i.e., non-trained) phonemes, across the pretest, posttest, and delayed posttest. Results are discussed with respect to feature acquisition, relying on gestural considerations (De Jong, Hao, and Park, 2009) and acoustic similarity (Goldinger and Azuma, 2003).

Review of the Literature

Theoretical Models of L2 Phonetic Acquisition

While many theories of L2 phonetic acquisition have largely considered outcomes and developmental trajectories on a segment-by-segment basis, a feature oriented approach, in which acquisition occurs at the subsegmental level of the feature, may provide complementary underlying mechanisms to motivate findings at the segmental level. Notably, as described by Clements (1985), individual phonemes may be conceptualized as a bundle of smaller properties (i.e., features) that can be shared across multiple phonemes. While originally described as binary (or monovalent) distinctive features, more recent interpretations, most notably from experimentally-oriented paradigms, conceptualize of phonological features as being "realized phonetically in terms of a constellation of articulatory gestures" (Dmitrieva et al., 2015). This

"constellation" of gestures corresponds to a variety of acoustic cues that work in conjunction to build relevant contrasts. A given contrast may be signaled by a variety of articulatory gestures and their related acoustic cues, and the relevance of those cues may be language specific. Thus, when acquiring a new set of L2 phonetic norms, it is possible that L2 learners acquire the relevant cues of a given feature and apply them across a set of phonemes that share this feature, rather than acquiring the norms for each phoneme individually.

Among several models that have addressed the acquisition of L2 phonetic norms, the Speech Learning Model (SLM) has been among the most influential (Flege, 1987, 1988, 1991, 1995, 1998). With respect to the interaction of the L1 and nascent L2 phonetic systems, the SLM posits that the relationship between segments in the L1 and L2 is crucial for determining the development of the new L2 sound. Depending on the similarity to the extant L1 sounds, learners may incorporate new L2 sounds into an existing L1 category (Flege, 1987), or they may establish a new, separate category specific to the L2 (Flege and Eefting, 1987). L2 phonemes that are more similar to existing L1 categories are most likely classified as "equivalent", preventing the establishment of native-like (i.e., "authentic", Flege, 1987: 62) production in the L2. In contrast, L2 phones that are more distinct are more likely to be established as their own, unique L2 category. Important for the current discussion, the SLM is broadly concerned with acquisition on a segment-by-segment basis. Highlighting this segment-by-segment approach, De Jong, Hao, and Park (2009) noted that studies within an SLM framework have often considered the acquisition of phonemic contrasts in isolation (e.g., /t/ vs /d/) rather than a set of phonemes that share a given class or feature (e.g., Flege et al., 1992). A focus on the segmental level can also be seen in other perception-oriented models, including the Native Language Magnet theory (Kuhl, 1992, 1993a, 1993b; Kuhl and Iverson, 1995) and the Perceptual Assimilation Model–L2 (Best and Tyler, 2007; Tyler, 2019). These models both suggest that the ability to perceive a new sound in the L2 is conditioned by the existing L1 phonetic system. Although these models differ with regard to the underlying mechanisms, they are similar in that they both approach L2 phonetic acquisition on a sound-by-sound basis.

While each of these models considers the role of existing L1 categories, Brown (2000) argued that many models generally do not provide a mechanism by which sounds are equated with existing L1 categories. One mechanism by which such pairing may occur is through the use of subphonemic components or features. This approach reflects theoretical accounts of feature geometry (Clements, 1985; Sagey, 1986), which note that every segment (i.e., phoneme) can be described via a series of relevant subcomponents or features. Importantly, a given feature or set of features, may be shared across a number of phonemes. For example, the English occlusives /b, d, g/ display common values for the features of continuity (i.e., [–continuant]) and voicing (i.e., [+voice]), and differ with respect to the feature of place of articulation (i.e., /b/ is [labial], /d/ is [alveolar], /g/ is [velar]). Thus, when a new sound in the L2 does not exist in the L1, learners may substitute a segment that is "minimally phonetically distinct" (Hancin-Bhatt, 1994: 244), defined as the L1 phoneme that differs by the fewest number of features from its L2 counterpart. Within this feature-oriented approach, outcomes have been shown to correlate between phonemes that share a similar feature in perception (De Jong, Silbert, and Park, 2009) and production (De Jong, Hao, and Park, 2009; Olson, 2019).

Feature Acquisition in L2 Perception and Production

A number of studies have noted that L2 perceptual abilities are constrained by the native phonemic inventory, and particularly highlighting the role of the extant phonemic contrasts. For

example, Brown (1997, 1998) examined perceptual abilities in L1 Japanese speakers, and found that their ability to correctly perceive contrasts present in the L2 (English) was dependent not only on the L1 phonetic inventory, but the presence of a given featural contrast in the L1 (see also, Brown, 2000). To illustrate, when testing perceptual contrasts of two pairs of phonemes, participants were able to successfully discriminate /b–v/, but not /r–l/. For both phoneme pairs, one phoneme was present in the L1 (i.e., /b/ and /r/ respectively), while the other was absent (i.e., /v/ and /l/). Brown suggests that the difference in the ability to perceive the contrast between the two phoneme pairs relies on the fact that the stop–fricative contrast is present in Japanese (i.e., /t–s/). These results were taken to suggest that L1 influence occurs not at the level of the phoneme, but at the subphonemic component of the phonological feature, which effectively accounts for "a learner's ability to acquire various non-native contrasts… and the varying abilities of learners with different L1s to acquire the same non-native contrasts" (Brown, 1998: 187).

More recently, De Jong, Silbert, and Park (2009) examined the potential for correlations in perceptual discrimination across various pairs of phonemes for L1 Korean speaking learners of English. Participants had varied L2 experience and proficiency, and as such, there were inter-participant differences in perceptual abilities. However, results showed that the discrimination of consonant pairs that differed by a single manner contrast was highly correlated across various points of articulation for a given speaker. For example, participants that were accurately able to discriminate between /f/ and /p/, a pair of consonants that differ with respect to manner (i.e., /f/ is a fricative and /p/ is a stop) were also able to accurately discriminate /θ/ and /t/, which differ by the same manner contrast. Importantly, performance was not correlated between different feature contrasts, such that participants who successfully perceived the stop–fricative contrast (i.e., manner contrast) were not necessarily successful at discriminating across a voiced–voiceless contrast (i.e., voicing contrast). Again, such results suggest that participants may acquire a single feature or feature contrast, which generalizes across multiple phonemes or phoneme pairs, rather than acquiring each individual phoneme in isolation.

Further evidence for a featural approach to phonetic acquisition comes from perceptual paradigms carried out with monolingual listeners. In a process known as selective adaptation, perception of a given stimulus along a contrast continuum (i.e., voiced–voiceless) may be shifted following repeated input of one of the continuum endpoints (for review see Kleinschmidt and Jaeger, 2016). For example, listeners are more likely to categorize an acoustically ambiguous stimulus as /p/ following repeated presentation of /b/ (Eimas and Corbit, 1973; Samuel, 1986; Vroomen et al., 2004; Vroomen et al., 2007). Importantly, selective adaptation has been found to generalize to phonemes that share the relevant feature, but were not involved in the adaptation process. For example, repeated presentations of /b/ may result in a shifted boundary for both the /b–p/ and /d–t/ continua (for a review see Klienschmidt and Jager, 2016; for retuning see Norris et al., 2003; Kraljic and Samuel, 2006). Kleinschmidt and Jaeger (2016) propose that the effect of selective adaptation may be explained not by the "fatigue of feedforward feature detectors" (p. 687), but as a form of distributional learning. In this case, listeners are sensitive to distributions that may be organized "as a sort of feature" (p. 687).

While perception-oriented research has provided significant support for the notion of features, and particularly acquisition of features in L2 phonetics, more recent work has begun to explore the potential role of featural representations in production. Production-oriented research has also demonstrated correlations in production accuracies between phonemes that share a key feature. For example, an L2 learner's ability to successfully produce a distinction between stops

and fricatives has been shown to be correlated between voiced and voiceless phonemes (De Jong, Hao and Park, 2009). That is, learners who successfully produce the stop–fricative contrast for voiceless phonemes (e.g., /f/–/p/) also produce the stop–fricative contrast for voiced phonemes (/v/–/b/). Again, such a correlation suggests that L2 phonetic acquisition may rely on subphonemic components. More recently, Olson (2019) employed a phonetic pedagogical treatment to investigate the potential for feature acquisition resulting from explicit selective training. In this study, L1 English speakers were selectively trained on the production of one of three voiceless stop consonants in L2 Spanish (i.e., /p/, /t/, or /k/, for additional details on English and Spanish VOT see below). Following a pedagogical treatment, participants significantly improved their production of the trained phoneme, as evidenced by a reduction in VOT. Importantly, participants also showed a significant reduction in VOT for non-trained phonemes. Moreover, the degree of improvement for each individual participant for trained and non-trained phonemes was shown to be highly positively correlated: participants who received instruction on /p/ showed similar levels of improvement for the non-trained phonemes (i.e., /t/ and /k/), which share the same manner (i.e., stop) and voicing (i.e., voiceless) features. As the acquisition or shift in VOT for a trained phoneme generalized to other non-trained phonemes that share the same feature, these results support the notion that acquisition may be underpinned by these subphonemic components.

Worth noting, De Jong, Hao, and Park (2009), suggest that expressions of the featural links between phonemes that share a given feature may be influenced by gestural considerations. In some cases, the same contrast (e.g., stop–fricative) may require different articulatory gestures for different places of articulation (e.g., labial vs. coronal). As such, outcomes in production may differ somewhat from those in perception.

While there is considerable evidence to suggest that features play a role in L2 phonetic acquisition, there remains ongoing debate regarding the "size" or scope of the relevant features. Implications of this debate extend beyond L2 acquisition to other areas such as (L1 or L2) word recognition (e.g., Mitterer et al., 2018). Although the traditional notion of the feature relied on an abstract phonological contrast (e.g., Clements, 1985), others have suggested that features may be conceptualized as a constellation of articulatory gestures (e.g., Dmitrieva et al., 2015) and patterns of acoustic similarity (e.g., Goldinger and Azuma, 2003; Goldinger et al., 1989). These inherent connections are described as emergent properties that are determined by a sustained level of acoustic coherence or similarity. Such connections between units may be hierarchical in nature, with stronger connections found between more related phonemes. For example, as shown in an English monolingual perceptual confusion matrix, the phoneme /p/ is most likely to be misidentified as /t/ or /k/, and almost never misidentified as /b/ (Luce, 1986). Others have suggested that word recognition is driven not by phonemes or segments, but by their context-dependent interpretations, like allophones (Mitterer et al., 2018; Reinisch et al., 2014). Taken as a whole, this line of research suggests that L2 acquisition relies in part on subphonemic featural components, but that the "grain-size" or scope of such features, particularly in L2 phonetic acquisition, remains undetermined (Mitterer et al., 2018, p. 77).

The Voicing Feature in English and Spanish

From a theoretical perspective, there remains debate about the underlying nature of laryngeal contrasts. Kager et al. (2007) provide a relevant distinction in previous work between the Single Feature Hypothesis (e.g., Lisker and Ambramson, 1964; Wetzels and Mascaró, 2001), in which voicing relies on a single binary [+/- voice] feature or a monovalent [voice] feature, and

the Multiple Feature Hypothesis (e.g., Jessen and Ringen, 2002), in which laryngeal contrasts may employ both [voice] and [spread glottis] features (i.e., aspiration) (for review see Simon, 2009). Relying on the acoustically-oriented notion of contrasts as a constellation of gestures, this distinction can be reframed as a set of language-specific cues that collectively signal the voiced–voiceless distinction. Among this set of cues, voice onset time (VOT), defined as the temporal difference between the release of the oral closure (i.e., burst) and the onset of vocal fold vibration (i.e., periodic waves), has been shown to be a reliable cue to stop consonant voicing (Lisker and Abramson, 1964). A negative VOT measurement occurs when vocal fold vibration precedes the release of the oral closure. A positive VOT measurement occurs when vocal fold vibration follows the release of the oral closure. With respect to the voicing contrast, Dmitrieva and Dutta (2020) note that VOT "has served as a cornerstone of acoustical investigations of laryngeal contrasts" and provides a unified basis for distinguishing stop voicing contrasts across a wide range of languages (p. 210). While VOT may serve as a primary or unifying cue, additional secondary cues may also be relevant, including F0 onset (Dmitrieva et al., 2015) and aspiration (Cho and Ladefoged, 1999).[1] Within this perspective, adopted for the current study, VOT serves as one of multiple cues to the feature of voice. Furthermore, given that both voiced and voiceless stop consonants make use of VOT, the question arises as to whether VOT operates as a single, shared cue between voiced and voiceless phonemes, or whether VOT operates as a separate, independent cues for each class.

Although both English and Spanish employ a bipartite voicing distinction (voiced – voiceless) across three places of articulation (i.e., labial, alveolar/dental, velar), English and Spanish differ in their production of word-initial stops. Word-initial voiceless stops are produced with long-lag VOT in English (30 – 100ms; Lisker and Abramson, 1964) and short-lag VOT in Spanish (0 – 30ms; Lisker and Abramson, 1964; Rosner et al., 2000; Williams, 1977), although VOT has been shown to vary by place of articulation (English mean: /p/ = 58ms, /t/ = 70ms, /k/ = 80ms; Spanish mean /p/ = 10ms, /t/ = 15ms, /k/ = 25ms; Lisker and Abramson, 1964). English voiced stops are usually produced with short-lag VOT (range: 0 – 30ms; mean English short-lag VOT /b/ = 1ms, /d/ = 5ms, /g/ = 21ms), although pre-voiced stops (VOT: -100 – -20) are also possible. As Lisker and Abramson (1964) note, short-lag voiced stops are most common in English, and implementation of short-lag or pre-voiced voiced stops is largely speaker dependent but internally consistent. Spanish word-initial voiced stops are pre-voiced, with the onset of voicing preceding the release of the oral closure (range: -150 – -50ms; mean Spanish VOT /b/ = -138ms, /d/ = -110ms, /g/ -108ms; Lisker and Abramson, 1964). Voiced stops in Spanish in intervocalic position are produced as approximants, a pattern that holds in word-initial position (Hualde et al., 2011). All tokens in the current study occurred in utterance initial position, and thus the voiced stops are expected to be realized with a full oral closure. Table 1 illustrates the cross-linguistic difference in VOT for English and Spanish. A number of authors have noted that beginning and intermediate-level English-speaking learners of Spanish often produce Spanish

[1] As suggested by the Multiple Feature Hypothesis (see Kager et al., 2007), English also employs aspiration as a cue to stop voicing. VOT and aspiration have been shown to be correlated cross-linguistically (e.g., Lisker and Abramson, 1964; Liu et al., 2007), and Dmitrieva et al., (2015) note that long-lag VOT entails a "relatively long period of aspiration-filled near-silence occur[ing] between the stop release and the onset of vocalic voicing," (p 79). As such, voice onset time and aspiration are considered additive cues from the underlying constellation of gestures.

stop consonants with English-like VOTs (e.g., Hammond, 2001). While this is unlikely to cause issues of intelligibility for voiceless stops (Munro and Derwing, 1995; Lord, 2005), Spanish voiced stops produced with English-like VOT may be perceived as voiceless by native Spanish listeners.

Table 1. Comparison of voicing in English and Spanish

	Pre-voiced < 0ms	Short-lag 0 – 30ms	Long-lag > 30ms
English		[+voice]	[-voice]
Spanish	[+voice]	[-voice]	

Applying the notion of features as a "constellation of gestures" to the expression of voicing in English and Spanish, both languages employ VOT as a cue to stop voicing. In both languages, VOTs are longer or more positive for voiceless stops and shorter or more negative for voiced stops. Yet, it remains to be seen whether this single cue, relevant both voiced and voiceless stops, is inherently linked across the two classes, or whether the VOT cue functions independently for voiced and voiceless stops.

Overview of the Study

Building on recent research on feature acquisition in L2 phonetics and conceptualizing features as acoustic cues, the current study investigates the nature of featural "units." As VOT serves as a fundamental cue to both voiceless and voiced phonemes, it is possible that this single acoustic cue is inherently linked for the two categories. Alternatively, and following more traditional formal phonological approaches, two independent, category-specific VOT cues may be employed for voiceless and voiced stops. While Olson (2019) demonstrated that a reduction in VOT for a single voiceless stop may generalize to other voiceless stops, suggesting that a single VOT cue is shared across difference places of articulation, the question remains as to whether such improvement would also generalize to voiced stops. This question speaks directly to the nature of the featural relationships between voiced and voiceless stop consonants.

To address this question, a pedagogical treatment was administered to a group of English-speaking learners of Spanish. The current study consists of a pretest, three phonetic treatments, a posttest, and a delayed posttest. The treatment employed was a set of three visual feedback paradigms, in which participants produced stimuli containing word-initial voiceless Spanish stops and used visual analysis of waveforms and spectrograms to compare their productions with those of native Spanish speakers. The use of a targeted phonetic treatment allowed for the relative isolation of voiceless and voiced stops. Treatment focused only on the set of voiceless phonemes (/p, t, k/), but outcomes were assessed for both voiceless and voiced (/b, d, g/) stops.

Exploiting the cross-linguistic difference in VOT between English and Spanish, the current study addresses featural relationships between voiceless and voiced stops, through an examination of the VOT cue. Two specific research questions are addressed:

(1) Does improvement in the production of voiceless L2 stops, defined as a reduction in VOT produced by English-speaking learners of Spanish, generalize to the set of voiced L2 stops?
(2) Is there a relationship between the degree of change for trained (i.e., voiceless) and non-trained phonemes (i.e., voiced)?

In the absence of clear hypotheses, three possible outcomes are considered, each indicative of a different scope of the VOT cue and highlighting different potential featural relationships

between voiced and voiceless stops. First, it is possible that both voiceless and voiced stops will shift in the same direction on the VOT continuum, with a similar magnitude of shift. That is, a change in VOT for voiceless consonants (i.e., becoming less positive) will be matched by an equal change in VOT the same direction for voiced consonants (i.e., becoming either less positive or more negative). This outcome would suggest a wider scope of the VOT cue, which is shared not only for voiceless stop consonants across different places of articulation (i.e., Olson, 2019), but operates as a single cue for both voiced and voiceless stops. Second, it is possible that a significant change in the VOT of voiceless stop consonants will result in a change in the VOT for voiced consonants in a similar direction, *only if* such changes create an overlap in the voiced and voiceless categories that could lead to perceptual confusion. In this case, a shift in the VOT for voiced consonants may be smaller in magnitude than the shift in VOT for voiceless consonants, and only for participants whose voiceless VOT values after the treatment approach or overlap their initial voiced VOT values. As such, changes in the VOT for voiced phonemes would be attributable not to inherent links between the VOT cues for voiced and voiceless phonemes, but to a speaker's need to avoid category overlap and perceptual confusion. Finally, it is possible that a significant reduction in VOT for voiceless stop consonants will not result in any significant changes in voiced consonants. This result would suggest a narrower scope of the VOT cue, in which although the VOT cue may be shared across different places of articulation, it is not shared across voicing contrasts. In other words, the VOT parameter is manipulated independently for voiced and voiceless phonemes, suggesting a limited featural relationship.

Methodology

Participants

Twenty-four participants were recruited from several fourth-semester Spanish courses at a large, public Midwestern university. This intermediate-level course focuses on the four basic language skills and relevant cultural topics. Participants were placed into this course via standardized placement exam or successful completion of the previous course in the sequence. An abbreviated version of the Bilingual Language Profile (Birdsong et al., 2012), incorporating the subcomponents of language history, proficiency, use, and attitudes, was administered following the completion of all other aspects of the experiment. While the goal of this language background questionnaire was the application of the inclusionary criteria, results are available in Appendix A.

All participants are considered to be native speakers of English, having learned English from birth and speaking only English in the home. All are L2 learners of Spanish, with an average age of onset of acquisition (AoA) of 13.2 years old (SD = 2.6). An initial inclusionary criteria, with an onset of acquisition of Spanish of greater than five years old, was established to exclude potential heritage speakers of Spanish (Valdés, 2001). Participants who reported exposure to a language other than English from birth were excluded (n = 3). One participant was removed from the analysis, having failed to complete two of the three required sessions. A total of 20 participants were retained for the final analysis.

Stimuli

Stimuli consisted of 54 tokens with word-initial voiceless stops (/p, t, k/) and 54 tokens with word-initial voiced stops (/b, d, g/). As the phonetic environment following the stop consonant has been shown to impact VOT (e.g., Flege, 1991; Klatt, 1975), stimuli were balanced with respect to the following vowel: /a/, /e/, and /u/. Each resulting consonant–vowel (CV)

pairing was represented by two unique tokens in each of the three recording sessions (pretest, posttest, delayed posttest), resulting in a total of 108 individual tokens produced by each participant (6 stops × 3 vowels × 2 tokens × 3 sessions = 108 tokens). All tokens were two-syllable, paroxytonic words, and controlled for learner familiarity (see below). All tokens were embedded in utterance-initial position of novel utterances. Table 2 provides sample stimuli.

Table 2. Sample target stimuli

Phoneme	Voicing	Sample stimuli
/p/	[-voice]	*Paro el coche si veo un semáforo en rojo.*
		'I stop the car if I see a red stoplight.'
/t/	[-voice]	*Tacha mi nombre de la lista.*
		'Cross my name off the list.'
/k/	[-voice]	*Cada persona tiene que traer un plato.*
		'Each person has to bring a plate.'
/b/	[+voice]	*Barre el suelo.*
		'Sweep the floor.'
/d/	[+voice]	*Damos regalos a mi suegro por su compleaños.*
		'We give presents to my father-in-law for his birthday.'
/g/	[+voice]	*Gasto dinero en caramelos.*
		'I spend money on candy.'

Complementary to the main analysis of tokens in novel utterances, a single set of 36 tokens in isolation (18 voiceless, 18 voiced) were also recorded at the pretest and delayed posttest. Tokens were balanced for initial phoneme (/p, t, k/ and /b, d, g/) and following vowel. Unlike the targets in novel utterances, the same tokens were recorded at the pretest and delayed posttest. Tokens in isolation serve to confirm the effect of the pedagogical treatment and support the findings from tokens in isolation. As tokens produced in isolation are less naturalistic than productions in continuous speech, they are considered to be secondary to the main analysis.

Word Familiarity Norming

As the above constraints limited the available pool of lexical items, the tokens selected for the current study varied in their relative frequency. As frequency has been shown to impact phonetic production (e.g., Jurafsky et al., 2001), it was necessary to control for token frequency across the three experimental sessions. As word frequency measurements generally rely on native speaker corpora, word familiarity was considered to be more appropriate for L2 learners.

Participants for the familiarity norming task ($N = 21$), different from those recruited for the full experiment, were drawn from a similar population (Spanish AoA: $M = 13.5$, $SD = 3.7$). Participants were given a randomized list of target tokens and rated each token on a seven-point familiarity scale (Auer et al., 2000): 1 = highly unfamiliar; 7 = highly familiar. Verbs were presented in the infinitive form and nouns in the singular form. All tokens were rated for familiarity, and subsequently divided between the three sessions to ensure equal levels of familiarity.

Average familiarity ratings for individual tokens ranged from highly familiar (e.g., *baile* 'dance', $M = 7.0$, $SD = 0$) to highly unfamiliar (e.g., *tuerca* 'nut', $M = 1.2$, $SD = 0.7$), with an overall moderate average familiarity ($M = 4.5$, $SD = 2.1$). Assessing the familiarity of the stimuli across each of the three sessions, as well as between the voiced and voiceless tokens, a two-way ANOVA was conducted on the average familiarity rating, with session (pretest, posttest, delayed posttest) and voicing (voiced, voiceless) as independent variables. Results demonstrated no

significant difference in familiarity across sessions, $F(2, 102) = 0.255$, $p = 0.775$, $\eta_p^2 = 0.005$, or voicing type, $F(1, 102) = 0.276$, $p = 0.600$, $\eta_p^2 = 0.003$. Thus, while tokens varied in their familiarity, there were no significant differences in familiarity between the three sessions or the differing voicing types.

Procedure and Pedagogical Treatment

The current study consists of a pretest, three phonetic treatments, a posttest, and a delayed posttest. The phonetic treatment employed here is a scaffolded variation of the visual feedback paradigm. Broadly, visual feedback involves participant self-recording and visual analysis of their own productions, as well as visual and auditory comparison with native speaker productions. A number of different types of visual representations have been successfully used for improving phonetic production, including: intonation contours (e.g., Levis and Pickering, 2004), spectrograms (e.g., Ruellot, 2011; Saito, 2007), waveforms (Motohashi-Saigo and Hardison, 2009), and both waveforms and spectrograms (Olson, 2014a). Visual feedback was chosen for the current study as it has been shown to successfully improve performance across a number of duration-based segmental features (e.g., for vowel length see Okuno, 2013; for singleton–geminate see Motohashi-Saigo and Hardison, 2009), including VOT for voiceless stop consonants (Offerman and Olson, 2016; Olson, 2019).

In order to effectively address the research questions, while analysis focuses on the change in VOT for voiceless and voiced stop consonants, the treatment focused exclusively on voiceless consonants. Building on previous visual feedback paradigms, the current study employed a scaffolded approach, in which participants recorded tokens within increasingly complex and more naturalistic structures: words in isolation (treatment 1), words in novel utterances (treatment 2), and words in a continuous discourse (treatment 3). Each treatment consisted of three phases: (a) prerecording, (b) self and native-speaker analysis, and (c) rerecording.

Tokens directly analyzed in the treatment ($N = 24$, 8 per session) differed from those used in the analysis. These training tokens were all Spanish lexical items, with word-initial voiceless stops (/p, t, k/). Due to the lack of real-word lexical items fitting the strict phonetic and familiarity criteria and the need to select semantically appropriate tokens for the more complex tasks (e.g., words in discourse), the training tokens were not balanced for following phonetic environment or syllable structure.

Prior to the first activity, participants were given a brief (< 10 minute) introduction to the assignment and Praat software. The instructors, all native Spanish speakers, answered any questions during this time. For the prerecording phase, participants were given a handout with detailed activity instructions. Using Praat software (Boersma and Weenink, 2017), the handout provided instructions on how to record the target stimuli and print the "visual representation" (i.e., spectrogram and waveform) of the first six training words, two per word-initial voiceless stop. In addition, following these written instructions, participants segmented each of their words into individual "sounds" (i.e., phonemes), generally through repeated listening. All instructions were provided in the target language. Students reported no specific difficulties understanding the task. All recording was conducted on participants' personal computers with a microphone.

During the in-class analysis phase (approximately 25 minutes), participants were given a second handout, which included guiding questions, native speaker spectrograms, and spectrogram pairs for discrimination practice. The questions guided participants to analyze their own productions, with a focus on the target segment. During each treatment, participants

analyzed tokens beginning with each of the voiceless stops (/p, t, k/). Example 1 shows (translated) sample guiding questions.

(1) Look at the word *pido* that you recorded and answer the following questions:
   a. How did you decide to mark the boundaries of each sound or letter?
   b. What are the visual characteristics of your "p"?
   c. Is your "p" longer or shorter than the following "i"?

To promote comparison with native-speaker productions, participants examined a spectrogram and waveform of the same word produced by a native Spanish speaker (peninsular dialect) and were asked to describe the visual characteristics and compare with their own productions. Next, participants were asked to hypothesize about the auditory differences between native and non-native speaker productions (Example 2). Finally, participants listened to productions by native and non-native speakers to confirm their hypotheses. To further emphasize the visual difference, participants were given several pairs of spectrograms and waveforms for novel words, and asked to identify which image in the pair was produced by a native Spanish speaker.

(2) Now compare your word *pido* with the image of the word *pido* produced by a native Spanish speaker.
   a. Describe the visual difference between your "p" and the "p" produced by the native speaker.
   b. What do you think the *auditory* difference is between your "p" and the "p" produced by the native speaker?

Following each in-class analysis phase, participants were given three days to re-record the training tokens and provide them to their instructor. The progression of record, in-class visual analysis, and re-record, was repeated for each of the three treatments.

Pretest stimuli were recorded at the same time as the initial recording phase of the first pedagogical treatment. Posttest stimuli were recorded at the same time as the re-recording phase of the third treatment. The delayed posttest stimuli were recorded approximately 4 weeks after the completion of the third treatment (for timing see Norris and Ortega, 2000). There was no additional discussion of the target segments in intervening class sessions. Table 3 illustrates the timeline for treatments, phases, and target stimuli.

Table 3. Experimental timeline

Week	Treatment Description	Treatment Phase	Target Stimuli
0	Words in isolation	Initial recording	Pretest
		Visual analysis	
		Re-recording	
2	Words in utterances	Initial recording	
		Visual analysis	
		Re-recording	
4	Words in discourse	Initial recording	
		Visual analysis	
		Re-recording	Posttest
8			Delayed Posttest

The peadgogical activities were included in the course curriculum, and each phase was graded on a complete/incomplete basis. Participation in the research study was voluntary. Participants were compensated for their participation.

Data Analysis

A total of 2160 tokens were expected in the initial analysis (20 speakers × 6 stops × 3 vowels × 2 tokens × 3 sessions = 2160 tokens). Data was maintained from any participant who completed both the pretest and either the posttest or delayed posttest. A total of seven participants were retained who failed to complete either the posttest (n = 3) or delayed posttest (n =4). As such, 252 tokens were classified as missing data. The remaining tokens were coded for production or recording errors (e.g., noisy recordings, false starts at target word, etc.), and 5.5% of the remaining data was eliminated. Lastly, outliers (n = 18, approximately 1%), defined as tokens with normalized VOT measures greater than three standard deviations above and below the mean for voiceless and voiced phonemes, were eliminated. A total of 1785 tokens were retained for the final analysis.

VOT, defined as the temporal difference between the release of the oral closure and the onset of vocal fold vibration, was measured by a single trained coder, who was not the researcher, using Praat (Boersma and Weenink, 2017). To assess the intrarater reliability, a subset of approximately 10% (n = 192) of tokens were recoded by the same coder. During the recoding process, the coder did not have access to their original measurements. Data elimination procedures mirrored those described above, with a total of 150 tokens included in the reliability testing. Results of a Pearson correlation comparing the original values with the recoded measurements demonstrated a high degree of intrarater reliability ($r(148) = .964$, $p < .001$).

Crosslinguistically, VOT varies across place of articulation, with labials evidencing shorter VOTs and velars longer VOTs (for review, see Cho and Ladefoged, 1999). In addition, it is worth noting that the difference between expected English and Spanish values also differs by place of articulation. VOT averages for English and Spanish /p/ differ by approximately 50ms, while averages for English and Spanish /b/ are separated by nearly 140ms (Lisker and Abramson, 1964). As such, to allow accurate comparison both within and between voicing categories, a normalized VOT value was calculated (following Olson, 2019) for each token based on the average Spanish and English VOT reported by Lisker and Abramson (1964).[2] The normalization formula employed was: (Token VOT – Spanish mean)/(English mean – Spanish mean). In this ratio, a value of 1.0 represents a token with VOT equal to that of the average English VOT (e.g., 58ms for /p/) and 0.0 represents a token with VOT equal to that of the average Spanish VOT (e.g., 4ms for /p/).[3] Values above 1.0 represents a token with VOT above the mean English VOT; any value below 0 represents a token with a VOT below the mean Spanish VOT. As an example of the normalization procedure for voiceless stops, the token *paro* 'I stop' produced with a raw VOT of 40ms (English /p/ mean = 54ms; Spanish /p/ mean = 4ms) corresponds to a normalized value of 0.76.

---

[2] As noted by an anonymous reviewer, bilingual or monolingual speakers could serve as the basis for normalization. Following this suggestion, a secondary analysis was conducted with VOT values produced by early bilinguals obtained from Flege and Eefting (1987). Comparison of the two sets of results showed no differences in the patterns of significance. Considering that this normalization procedure followed Olson (2019) and that the same pattern of results was found with monolingual and bilingual values, the results presented here are normalized based on monolingual values (Lisker and Abramson, 1964).

[3] Lisker and Abramson (1964) report two patterns for English voiced stops: pre-voiced and short-lag voiceless. As nearly all of the pre-voiced tokens were produced by a single English speaker, the short-lag values were used for normalizing VOT value in this study.

The initial linear mixed-effects model was conducted using R software for statistical computing (R Core Team, 2013) and the lme4 package (Bates et al., 2015). Fixed effects included session (pretest, posttest, and delayed posttest) and voicing (voiceless, voiced). Participant and initial phoneme (/p, t, k, b, d, g/) were included as random effects with random slopes and intercepts for each of the main effects. Initial phoneme was used as opposed to item, as unique items were used in each session. More complex models, with crossed random effects, did not permit model convergence (see Barr et al., 2013). The significance criterion was set at $|t| = 2.00$. Effect size confidence intervals were computed using the psych package (Revelle, 2018).

Results

Words in Utterances

To assess the contribution of each of the main effects to the model, two submodels were conducted by dropping one of the fixed effects but maintaining the random effects structure parallel to that of the initial model. Comparing the initial model (log likelihood = -466.45) to each of the submodels revealed that the inclusion of both session (log likelihood = -470.77, $\chi^2(4) = 8.64$, $p = .071$) and voicing (log likelihood = -472.74, $\chi^2(4) = 12.593$, $p = .006$) improved model fit. Similarly, the random effects structure was assessed, and model comparison showed that the inclusion of both participant (log likelihood = -711.38, $\chi^2(10) = 489.79$, $p < .001$) and initial phoneme (log likelihood = -494.82, $\chi^2(10) = 54.66$, $p < .001$) significantly improved model fit.

[Insert Figure 1]

Results for the fixed effects from the full model (Table 4) demonstrate a significant effect of voicing on the production of normalized VOT, with a significant difference between voiced ($M = 0.88$, $SD = 0.38$) and voiceless phonemes ($M = 0.60$, $SD = 0.41$) at the pretest. For random effects see Appendix B. There was no significant difference between the normalized VOT for voiced phonemes at the pretest and either the posttest ($M = 0.92$, $SD = 0.36$) or the delayed posttest ($M = -0.96$, $SD = 0.30$). However, there was a significant interaction between session and voicing (posttest: $\beta = -0.173$, $t = -3.172$; delayed posttest: $\beta = -0.158$, $t = -3.063$). This interaction suggests that the impact of session was different for voiced and voiceless phonemes. As can be seen in Figure 1, for voiceless phonemes, productions were more native-like at the post-test and delayed post-test than the pre-test. For voiced phonemes, there was no significant difference between productions at the pretest and either the posttest or delayed posttest.

Table 4. Main model fixed effects

	Estimate	Std. Error	*t*-Value	Lower 95%	Upper 95%	*d*	95% CI
Intercept (Pretest, Voiced)	0.879	0.058	15.097	0.763	0.995		
Posttest	0.042	0.048	0.875	-0.054	0.138	-0.11	[-0.32, 0.10]
Delayed Posttest	0.049	0.064	0.799	-0.079	0.177	-0.24	[-0.46,-0.03]
Voiceless	-0.274	0.086	-3.172	-0.446	-0.102	0.69	[0.48, 0.91]
Posttest: Voiceless	-0.173	0.056	-3.119	-0.285	-0.061	1.15	[0.92, 1.37]
Delayed Posttest: Voiceless	-0.158	0.052	-3.063	-0.262	-0.054	1.03	[0.80, 1.25]

To further assess the differential effect of session on voiceless and voiced phonemes that may have been obscured in the initial main model, two separate approaches were considered:

separate mixed effects models and two one-sided equivalence testing. First, separate mixed effects models were conducted for voiceless and voiced phonemes. Session was included as a fixed effect and participant and initial phoneme as random effects with random slopes and intercepts by session. Results from the model with voiceless phoneme VOT as the dependent variable (Table 5, for Random Effects see Appendix C), showed a significant difference between the pretest and posttest (β = -0.132, t = -2.746), although this difference was not significant at the delayed posttest (β = -0.119, t = -1.870). Therefore, VOT production was significantly more native-like following treatment. In contrast, no significant differences were found between the pretest and either the posttest (β = 0.43, t = 0.786) or delayed posttest (β = 0.062, t = 0.862) for voiced phoneme VOT (Table 6, for Random Effects see Appendix D). To ensure that the lack of a significant effect of session on the VOT production of voiced phonemes was not the result of an underpowered study, a power analysis was conducted using a simulation-based approach (simr package: Green and MacLeod, 2016). Results of the power simulation, with the effect size determined by the findings for the voiceless phonemes (d = 0.28) and 500 simulations, showed that the current study design surpassed the 80% power threshold, suggesting that the study was not underpowered.

Table 5. Voiceless phoneme model fixed effects

	Estimate	Std. Error	*t*-Value	Lower 95%	Upper 95%	*d*	95% CI
Intercept (Pretest)	0.605	0.080	7.515	0.444	0.766		
Posttest	-0.132	0.048	-2.746	-0.228	-0.036	0.37	[0.15, 0.58]
Delayed Posttest	-0.119	0.064	-1.870	-0.246	0.008	0.28	[0.07, 0.49]

Table 6. Voiced phoneme model fixed effects

	Estimate	Std. Error	*t*-Value	Lower 95%	Upper 95%	*d*	95% CI
Intercept (Pretest)	0.879	0.064	13.695	0.750	1.007		
Posttest	0.043	0.054	0.786	-0.066	0.151	-0.11	[-0.32, 0.10]
Delayed Posttest	0.062	0.072	0.862	-0.082	0.207	-0.24	[-0.46,-0.03]

Second, equivalence testing was performed separately for the voiceless and voiced phonemes using the two one-sided t-test procedure, with α = .05, and equivalence bounds (Cohen's d) of Δ_L = -.5, Δ_H = .5, comparing the normalized VOT at the pretest and posttest. The posttest was initially chosen as it was most likely to elicit the largest shifts in voiceless VOT. Upper and lower boundaries were determined via benchmark and employed a medium effect size (Cohen, 1988; Lakens, Scheel, and Isager, 2018). For voiceless phonemes, results of the null hypothesis t-test showed a significant difference between pretest and posttest productions (t(616) = 4.604, p < .001), while results of the equivalence test were non-significant (t(616) = -1.644, p = .050), confirming the expected difference. In contrast, for voiced phonemes, results of the null hypothesis t-test showed no significant difference between pretest and posttest productions (t(618) = -1.379, p = .168), while results of the equivalence test were significant (t(618) = 4.880, p < .001), demonstrating that the observed effect is equivalent to zero. Taken as a whole, the two one-sided tests confirm that while production of the voiceless phonemes improved significantly following training, there was no effect of training on the voiced phonemes.

The effect sizes found here are in line with the expected shift based on previous studies using a similar approach (e.g., Olson, 2019). While not reaching the field-specific threshold for "small" within-group effects (d = .6) in L2 acquisition research suggested by Plonsky and Oswald (2014), this effect size is not unexpected. A meta-analysis of L2 phonetic instruction by Lee et al., (2014) notes that a short treatment duration (< 4.5 hours), intermediate-level participant proficiency, and a focus on segmental properties, all lead to smaller effect sizes (see Lee et al., 2014). Again, rather than the size of the effect, the key here is that the change in the trained, voiceless phonemes did not lead to any notable change in voiced phonemes.

Subsequent analysis, related directly to the second research question, was performed to address the potential link between the degree of change for voiceless and voiced phonemes. A Δ normalized VOT value was computed for each subject by subtracting their mean normalized VOT value at the posttest from that at the pretest, with positive values indicating a shift in the direction of native-like norms. Separate Δ normalized VOT values were conducted for voiceless and voiced phonemes. For participants who did not complete the posttest (n = 3), the mean normalized VOT value at the delayed posttest was substituted. A linear regression was conducted comparing the Δ normalized VOT values for voiceless and voiced phonemes (Figure 2). Results demonstrate that there was no significant relationship between the change in voiceless and voiced phonemes (*adj.* R^2 = .122, F(1,18) = 3.65, p = 0.072, b = 0.47). In order to ensure that the results were not influenced by the choice to substitute delayed post-test values, an additional linear regression was conducted excluding the three participants who were missing posttest values. Results closely mirror those presented in the main analysis, *adj.* R^2 = .140, F(1,15) = 3.60, p = 0.077, b = 0.57.

The outcome of the linear regression was influenced by the presence of one participant who showed marked improvement in both the voiceless and voiced phonemes. Parallel analysis excluding this particular subject as an outlier illustrates the impact of this one subject (*adj.* R^2 = -.028, F(1,17) = 0.510, p = 0.485). This result, potentially signaling an avoidance of category overlap between voiceless and voiced phonemes, is addressed further in the discussion.

[Insert Figure 2]

Finally, as many previous studies report raw VOT values (ms), Figures 3 and 4 are provided for comparison. In Figure 3, the pattern of improvement (i.e., becoming more native-like) from pretest to posttest and delayed posttest is consistent across all three voiceless phonemes. VOT values for voiceless phonemes improved an average of 7.4ms across all places of articulation (/p/ = 8.5ms, /t/ = 5.2ms, and /k/ = 8.6ms). While this shift was significant, it did not reach the expected range of native Spanish speakers. In contrast, the pattern is less consistent for voiced phonemes (Figure 4) at the posttest, and all phonemes show some degree of becoming *less native-like* by the delayed posttest, although model results show this change to be non-significant.

[Insert Figure 3]

[Insert Figure 4]

Tokens in Isolation

While the main analysis focuses on the production of words embedded within utterances, an examination of the production of words in isolation provides complementary analysis and serves to validate the findings for words in utterances. Unlike the tokens in utterances participants produced the same tokens in isolation at the pretest and delayed posttest. Fifteen participants provided isolated word data at both the pretest and delayed posttest. As with the words in utterances, the initial 1080 tokens (15 speakers × 6 stops × 3 vowels × 2 tokens × 2 sessions = 1080 tokens) were coded for a variety of production and recording errors. A total of 7% of tokens were removed from the analysis. A total of 987 tokens were included in the final analysis.

The initial statistical analysis was similar to that employed for the tokens within utterances, with a mixed effect model approach with session (pretest, delayed posttest) and voicing (voiced, voiceless) as fixed effects, and participant and initial phonemes as random effects with both random intercepts and slopes by each of the main effects. Results of the main model were similar to those found in the analysis of words in utterances. Namely, while there was no difference between the intercept (pretest, voiced) and the delayed posttest for the voiced phonemes ($\beta = 0.004$, $t = 0.056$), there was a significant interaction between session and voicing ($\beta = -0.213$, $t = -5.271$). As with the words in utterances, the differential effect of session on voiceless and voiced phonemes was examined using separate mixed effects models and two one-sided equivalence testing. The secondary analysis was conducted on the voiceless and voiced phonemes separately, with normalized VOT as the dependent variable and session as the fixed effect. Random effects included subject and initial phoneme, with both random slopes and intercepts by session. Results confirmed the above pattern, namely while there was a significant effect of session on voiceless phonemes ($\beta = 0.206$, $t = 2.387$), this effect was absent for voiced phonemes ($\beta = -0.008$, $t = -0.111$). Equivalence testing was performed using the two one-sided t-test procedure ($\alpha = .05$, $\Delta_L = -.5$, $\Delta_H = .5$) comparing the normalized VOT at the pretest and delayed posttest. For voiceless phonemes, results of the null hypothesis t-test showed a significant difference between pretest and delayed posttest productions ($t(491) = 6.412$, $p < .001$), while results of the equivalence test were non-significant ($t(491) = 0.827$, $p = .796$), confirming the expected difference. In contrast, for voiced phonemes, results of the null hypothesis t-test showed no significant difference between pretest and posttest productions ($t(464) = -0.790$, $p = .430$), while results of the equivalence test were significant ($t(464) = 4.723$, $p < .001$), demonstrating that the difference between the two sessions was equivalent to zero.

Taken as a whole, results for the repeated tokens in isolation serve to confirm the findings for tokens embedded within utterances. The pedagogical treatment produced a significant reduction in normalized VOT for voiceless phonemes. However, the effect of this treatment did not extend to the untrained voiced phonemes, and no significant shift in normalized VOT was found for the voiced phonemes.

Discussion

In this study, native English-speaking learners of Spanish received visual feedback on the production of voiceless stop consonants in Spanish. Analysis focused on VOT for both voiceless and voiced stops. With respect to the production of voiceless stops, results showed a significant reduction in VOT for the voiceless phonemes, all of which were subject to visual feedback treatment. This same pattern was found for both the main analysis of tokens in utterances, as well as the complementary analysis of words in isolation. This finding, expected based on previous results regarding the reduction of VOT for voiceless stops following visual feedback (e.g.,

Offerman and Olson, 2016), thus allowed for an examination of the comparison of interest between voiceless and voiced stops. Related directly to the first research question, which asked whether improvement in L2 voiceless stop production, defined as a reduction in VOT, generalizes to the set of L2 voiced stops, the results showed that there was no significant generalization from voiceless to voiced stops. Again, this result was found for both words in utterances and words in isolation. Assessing the second research question, namely the relationship between the degree of change for voiceless (i.e., trained) and voiced (i.e., non-trained) phonemes following the visual feedback paradigm, results showed that there was no significant correlation between the two. As a whole, although previous research has shown that a shift in VOT for one voiceless stop generalizes to all voiceless stops (Olson, 2019), the current results suggest that a shift in VOT for voiceless stops does not generalize to voiced stops. Therefore, while VOT serves as a cue to voiceless and voiced stop consonants, this cue appears to be independent for the two categories, suggesting a limit on their featural relations. Returning to the original possible outcomes, the results here suggest a dual VOT cue approach, in which voiced and voiceless phonemes employ two unique VOT cues.

Although the current results suggest that the VOT cue in production is not shared between voiceless and voiced phonemes, it is possible that some underlying mechanisms effectively serve to limit the scope of the VOT cue, including both articulatory or gestural considerations and acoustic similarity. First, voice onset time is comprised of two separate, time-locked articulatory gestures – the release of the oral closure and the onset of vocal fold vibration. The voiceless phonemes in English and Spanish both require the same order of these two gestures, namely release of the oral closure followed-by the onset of vocal fold vibration. Therefore as noted by Olson (2019), L2 learners of Spanish (L1 English) are tasked with an adjustment of the timing of these two gestures when acquiring the new L2 voiceless stops, with the onset of vocal fold vibration in Spanish occurring with less temporal distance from the point of release than in English. In contrast, VOT for voiced stops in Spanish employs the same gestures, but the release occurs after the onset of vocal fold vibration. As such, acquiring the voiced stops requires a full reordering of the articulatory gestures. It is possible that such gestural considerations play a role in determining the scope of the voicing feature. A similar pattern can be seen in the production-oriented results from De Jong, Hao, and Park (2009), in which cross-segment correlations for production accuracies were found for some contrasts (i.e., manner contrasts were correlated between voiced and voiceless phonemes), but not others (i.e., manner contrasts were not correlated between different places of articulation). The authors couch such results in terms of a featural acquisition approach, but with production-specific gestural constraints. That is, the stop–fricative manner contrast employs similar gestures for both voiced and voiceless phonemes (i.e., /s/–/t/ vs. /z/–/d/). In contrast, the stop–fricative manner contrast employs different gestures, and different articulators, for different places of articulation (i.e., /s/–/t/ vs. /f/–/p/). Applied to the current study, this approach may suggest although there may be an overarching VOT cue relevant for both voiceless and voiced phonemes, thus accounting for results in perception (e.g., De Jong, Silbert, and Park, 2009), the production of voiceless and voiced phonemes requires a gestural component that is sufficiently different between the two voicing categories to impede accuracy correlations. In this case, while a shift in the timing of the two gestures seems to be "similar enough" (De Jong, Hao, and Park, 2009, p. 369) to generalize across places of articulation for voiceless phonemes (Olson, 2019), a reordering of time course of these gestures is sufficiently different to block generalization between voiceless and voiced phonemes. These gestural considerations parallel previous theoretical approaches which propose

that voiceless stops are unmarked, and voiced stops may be aerodynamically more difficult to produce (see Westbury and Keating, 1986). Again, gestural constraints, framed here as the ease of production or "naturalness", may effectively limit the expression of featural relationships between voiceless and voiced stops. In short, the current results, and particularly the divergence from the generalizability of the VOT cue across voiceless stops found in Olson (2019), may be accounted for by considering gestural constraints.

A second possible, or complementary, explanation for the lack of generalization from voiceless to voiced phonemes relies on the notion of acoustic similarity (for review see Goldinger and Azuma, 2003). A schema of acoustic similarity can be found in perceptual confusion matrices, which are generated by presenting phonemes (or CV syllables) in noise to listeners and recording patterns of incorrect responses. Results from such confusion matrices for English largely demonstrate a strong acoustic similarity within voiceless and voiced phonemes, but not across these classes. For example, syllable-initial /p/ is most likely to be misperceived as /t/ and /k/, but highly unlikely to be misperceived as any of the voiced stops (i.e., /b, d, g/). Similarly, syllable initial /b/ is likely to be misperceived as /d/ or /g/, but less likely to be perceived as a voiceless stop (see Luce, 1986). A such, it is possible that the acoustic properties are sufficiently different to organize into separate units for voiceless and voiced phonemes.

Closely linked to the idea of acoustic similarity, the lack of change in the VOT of voiced stops may be attributed to the cross-linguistic differences in the nature or weights of the multiple acoustic cues to stop consonant voicing. As feature contrasts can be considered as a group of cues, and the weight of such cues may vary cross-linguistically (e.g., Dmitrieva et al., 2015), it is possible that English and Spanish rely on different cue weights. While English voiceless and voiced stops clearly differ in VOT, they also differ in aspiration. In contrast, Spanish voiceless and voiced stops differ primarily in VOT, and aspiration, while present, is reduced in temporal nature and less acoustically salient. As such, L1 English speakers may attend to, and modify, aspiration rather than VOT. While the end result of this change in aspiration includes a change in VOT, the relative weight of the VOT cue may be limited and prevent generalization to the voiced stops.

Additionally, it is worth noting the potential for lexical confusion and a lack of comprehensibility that may arise from independent shifts in voiceless and voiced stops. Again, Spanish voiceless stop VOTs produced by native Spanish speakers have been shown to be short-lag, with a VOT of between 0–30ms (Lisker and Abramson, 1964; Rosner et al., 2000; Williams, 1977) and English voiced stop VOTs are produced within a similar range, 0–30ms (Lisker and Abramson, 1964). As the voiceless and voiced phonemes appear to shift independently, in the case of significant improvement in voiceless stop VOT and no improvement in voiced stop VOT, it would be theoretically possible for a learner to produce Spanish-like voiceless VOTs and English-like voiced VOTs, resulting in an overlap in the production of the two classes of phonemes. This potential overlap in VOT for voiceless and voiced phonemes could result in lexical confusion, in which Spanish words like /peso/ 'weight' and /beso/ 'kiss' would be produced with a similar acoustic signal. In the current study, this was broadly not the case, as the shifts in VOT for the voiceless phonemes were significant, but did not enter the typical range of native Spanish speakers. Thus, while voiceless and voiced stops became more similar, they did not overlap. However, the subject that produced the most significant shift in VOT for the voiceless phonemes (mean delayed posttest VOT = 31ms) also showed a sizeable shift in the voiced phonemes. Considering the voiced phonemes, while only 20% of her voiced phonemes were produced with negative VOTs at the pretest, 69% of voiced tokens were produced with

negative VOTs at the delayed posttest. While caution should be used in interpreting the results from a single participant, it is possible that a shift in voiced tokens occurred to avoid an overlap in the voiceless and voiced phonemes. Moreover, as pre-voiced stops are attested in English (e.g., Lisker and Abramson, 1964), a given participant's L1 phonetic realizations may also be worth examining. While the broader results suggest that VOT cue is acquired independently for voiceless and voiced phonemes, learners may seek to avoid a merger of two distinct L2 phonetic categories.

Lastly, while the current study adopted an experimental approach to phonological features, in which phonological features are conceptualized as bundles of articulatory gestures and corresponding acoustic cues, it is worth considering more formal approaches to the phonological voicing feature. Again, two differing approaches can be highlighted from the previous literature: the Single Feature Hypothesis and the Multiple Feature Hypothesis (see Kager et al., 2007). In the Single Feature Hypothesis (e.g., Lisker and Ambramson, 1964; Wetzels and Mascaró, 2001), voicing relies on a single binary [+/- voice] feature. In the Multiple Feature Hypothesis (e.g., Jessen and Ringen, 2002), laryngeal contrasts may employ both [voice] and [spread glottis] features (i.e., aspiration) (for review see Simon, 2009). Applied to the current context, both approaches posit a separation of the voiced and voiceless stop consonant classes, such that changes in one class (e.g., voiced stops) are unlikely to generalize to the other (e.g., voiceless stops), but the underlying mechanisms may differ. Within the Single Feature Hypothesis (e.g., Lisker and Ambramson, 1964; Wetzels and Mascaró, 2001), the current results would suggest that the acoustic specification of [+ voice] and [- voice] operate independently, and that changes in the [- voice] category do not impact the [+ voice] category, except potentially in cases of acoustic overlap between categories. Within the Multiple Feature Hypothesis (e.g., Jessen and Ringen, 2002), English may be considered an aspiration language, in which the crucial distinction between voiceless and voiced stops is specification of [+ spread glottis] or [- spead glottis], respectively. In contrast, Spanish would be considering a voicing language, in which the crucial distinction is [+/- voice]. If English speaking learners of Spanish are able to effectively modify the spread glottis feature, the result may be a shift in voiceless stops, but not voiced stops. Accurate production of voiced stops would require the acquisition of a new [voice] contrast not employed in English stops. The current results, reinterpreted within such a theory, would suggest that learners may have modified the specification for voiceless stops, from [+ spread glottis] to [- spread glottis], but have failed to acquire the novel phonological feature of [voice]. There are several reasons why such a pattern may emerge, including the difficulty of learning a new phonological feature vs. modifying an existing feature, as well as the prominent acoustic salience of aspiration (for the salience of aspiration see Simon, 2009).

These two hypotheses have clear parallels with the above discussion of phonological features as a "constellation" of articulatory gestures and acoustic cues (Dmitrieva et al., 2015). Relying on notions of gestural considerations, the current results may suggest unique VOT cues for voiced and voiceless stops or a single cue that is impacted by gestural considerations, akin to the single feature hypothesis. In contrast, relying on multiple cue weighting and acoustic similarity, the current results may be attributed to the fact that the underlying shift may be more closely associated with one (i.e., aspiration) of the multiple cues to stop consonant voicing, recalling the multiple feature hypothesis. While differentiating between the single or multiple cue approaches is outside the scope of the current study, future research employing a gestural or cue-oriented approach may further add to this debate.

Conclusion

As learners develop proficiency in an L2, they are tasked with acquiring L2 phonetic norms. While many theories have tacitly adopted a segment-by-segment approach to L2 phonetic acquisition (e.g., Flege, 1987), others have suggested that acquisition occurs at the subphonemic level of the feature (De Jong, Hao, and Park, 2009), or that features may provide the mechanisms by which L1 and L2 sounds are equated (Brown, 2000). As much of the debate regarding the nature or scope of the feature has been driven by work in perception (for L2 perception see De Jong, Silbert and Park, 2009; for word recognition see Mitterer et al., 2018), the current study adds to this body of research by providing evidence for the scope of L2 features from a production-oriented paradigm. Specifically, while previous work has shown that improvement in VOT for a single voiceless stop consonant generalizes across all voiceless stop consonants (Olson, 2019), results from the current study showed that such improvement does not extend to voiced consonants. These results suggest that voice onset time, which serves as a one of several cues to stop consonant voicing, may operate independently for voiceless and voiced stop consonants. These results suggest limits on the featural connections between voiceless and voiced stop consonants. While this paper, drawing on previous research (e.g., De Jong, Hao, and Park, 2009; Dmitrieva et al., 2015), proposed that such connections are limited by a lack of gestural or acoustic similarity, it is also possible that other external factors may account for this finding. For example, the uncontrolled nature of the tokens employed in the training paradigm, a limitation of the current study, may have introduced unforeseen factors that differentially facilitated generalization to voiceless stops, but not voiced stops. Future research, including replication of the current study with carefully controlled training tokens, may serve to clarify the underlying mechanisms, such as gestural considerations or acoustic similarity, that impact the nature or size of each feature constellation. Moreover, as a growing body of work highlights the malleability of L1 phonetic systems (for review see de Leeuw and Celata, 2019), future work may seek to explore the potential role of features in L1 phonetic changes (e.g., drift, attrition, bilingual mode).

Acknowledgements: I would like to thank John Nielsen, Lourdes Barranco Cortes, Bruno Staszkiewicz Garcia, and Aída García Tejada for their support in various aspects of this project. In addition, I am grateful for the insightful comments provided by the anonymous reviewers and editor.

Funding Acknowledgements: The author received no financial support for the research, authorship and/or publication of this article.

References

Auer ET, Bernstein LE and Tucker PE (2000) Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory and Cognition* 28(5): 789–797. https://doi.org/10.3758/BF03198414

Bates D, Maechler M, Bolker B and Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7 http://CRAN.R-project. org/package1⁄4lme4. $

Barr D, Levy R, Scheepers C and Tily H (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68: 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Best CT and Tyler MD (2007) Nonnative and second-language speech perception: Commonalities and complementarities. In: M Munro and OS Bohn (eds) *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production.* Amsterdam: John Benjamins, pp. 13–34.

Birdsong D, Gertken LM and Amengual M (2012) Bilingual language profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin*, retrieved from https://sites.la.utexas.edu/bilingual/

Boersma P and Weenink D (2017) Praat: doing phonetics by computer [Computer program]. Version 6.0.33, retrieved from http://www.praat.org/

Brown CA (1997) *Acquisition of segmental structure: Consequences for speech perception and second language acquisition.* Doctoral dissertation, McGill University, Montreal.

Brown CA (1998) The role of L1 grammar in the L2 acquisition of segmental structure. *Second Language Research* 14(2): 136–193.

Brown CA (2000) The interrelation between speech perception and phonological acquisition from infant to adult. In: J Archibald (ed), *Second Language Acquisition and Linguistic Theory.* Hoboken, NJ: Wiley-Blackwell, pp. 4–64.

Cho T and Ladefoged P (1999) Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics* 27: 207–229. https://doi.org/10.1006/jpho.1999.0094

Clements GN (1985) The geometry of phonological features. *Phonology* 2(1): 225–252. https://doi.org/10.1017/S0952675700000440

De Jong KJ, Hao Y and Park H (2009) Evidence for featural units in the acquisition of speech production skills: Linguistic structure in foreign accent. *Journal of Phonetics* 37: 357–373. https://doi.org/10.1016/j.wocn.2009.06.001

De Jong KJ, Silbert NH and Park H (2009) Generalization across segments in second language consonant identification. *Language Learning* 59(1): 1–31. https://doi.org/10.1111/j.1467-9922.2009.00499.x

de Leeuw E and Celata C (2019) Plasticity of native phonetic and phonological domains in the context of bilingualism. *Journal of Phonetics* 75, 88–93. https://doi.org/10.1016/j.wocn.2019.05.003

Dmitrieva O and Dutta I (2020) Acoustic correlates of the four-way laryngeal contrast in Marathi. *Phonetica* 77(3): 209-237. https://doi.org/10.1159/000501673

Dmitrieva O, Llanos F, Shultz A and Francis A (2015) Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics* 49: 77–95. https://doi.org/10.1016/j.wocn.2014.12.005

Eimas PD and Corbit JD (1973) Selective adaptation of linguistic feature detectors. *Cognitive Psychology* 4(1): 99–109. https://doi.org/10.1016/0010-0285(73)90006-6

Flege JE (1987) The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics* 15(1): 47–65.

Flege JE (1988) Factors affecting degree of perceived foreign accent in English sentences. *The Journal of the Acoustical Society of America* 84(1): 70–79. http://dx.doi.org/10.1121/1.396876

Flege JE (1991) Perception and production: The relevance of phonetic input to L2 phonological learning. In: T Heubner and C Ferguson (eds) *Crosscurrents in Second Language Acquisition and Linguistic Theory.* Philadelphia: John Benjamins, pp. 249–289.

Flege JE (1995) Second language speech learning: Theory, findings, and problems. In: W Strange (ed) *Speech Perception and Linguistic Rxperience: Issues in Cross-language*

Research. Baltimore: York Press, pp. 233–73.
Flege JE (1998) Age of learning and second-language speech. In D. Birdsong (Ed.), *New Perspectives on the Critical Period Hypothesis for Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum, pp. 101–131.
Flege JE and Eefting W (1987) Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication* 6(3): 185–202. https://doi.org/10.1016/0167-6393(87)90025-2
Flege JE Munro MJ and Skelton L (1992) Production of the word-final English/t/–/d/ contrast by native speakers of English, Mandarin, and Spanish. *The Journal of the Acoustical Society of America* 92(1): 128–143. http://dx.doi.org/10.1121/1.404278
Green P and MacLeod C (2016) SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* 7: 493–498. https://doi.org/101111/20410210X.12504
Goldinger SD and Azuma T (2003) Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics* 31(3-4): 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5
Goldinger SD, Luce PA and Pisoni DB (1989) Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language* 28(5): 501–518. https://doi.org/10.1016/0749-596X(89)90009-0
Hancin-Bhatt B (1994) Segment transfer: A consequence of a dynamic system. *Second Language Research* 10(3): 241–269. https://doi.org/10.1177/026765839401000304
Hammond, R. (2001) *The Sounds of Spanish: Analysis and Application*. Somerville, MA: Cascadilla.
Hualde JI, Simonet M, and Nadeu M (2011) Consonant lenition and phonological recategorization. *Journal of Laboratory Phonology,* 2(2): 301–329. https://doi.org/10.1515/labphon.2011.011
Jessen M and Ringen, C (2002) Laryngeal features in German. *Phonology* 19: 189–218. https://www.jstor.org/stable/4420223
Jurafsky D, Bell A, Gregory M and Raymond WD (2001) Probabilistic relations between words: Evidence from reduction in lexical production. In: J Bybee and P Hopper (eds) *Frequency and the Emergence of Linguistic Structure. Typological Studies in Language*, 45. Amsterdam: John Benjamins, pp. 229–254
Kager R, Van Der Feest S, Fikkert P, Kerkhoff A and Zamuner T (2007) Representations of [voice]: Evidence from acquisition. In: Van De Weijer JM and Van Der Torre EJ (eds) *Voicing in Dutch: (De)voicing Phonology, Phonetics, and Psycholinguistics*, pp. 41–80.
Klatt DH (1975) Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research* 18(4), 686–706.
Kleinschmidt DF and Jaeger TF (2016) Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning. *Psychonomic Bulletin and Review* 23(3): 678–691. https://doi.org/10.3758/s13423-015-0943-z
Kraljic T and Samuel AG (2006) Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review* 13(2): 262–268. https://doi.org/10.3758/BF03193841
Kuhl PK (1992) Infants' perception and representation of speech: Development of a new theory. In: J Ohala, T Neary, B Derwing, M Hodge and G. Wiebe (eds) *Proceedings of the International Conference on Spoken Language Processing*. Edmonton, Alberta: University of Alberta, pp. 3–10.

Kuhl PK (1993a) Infant speech perception: A window on psycholinguistic development. *International Journal of Psycholinguistics* 9: 33–56.
Kuhl PK (1993b) Innate predispositions and the effects of experience in Speech Perception. The Native Language Magnet Theory. In: B de Boysson-Bardies, S de Schonen, P Jusczyk, P MacNeilage and J Morton (eds) *Developmental Neurocognition: Speech and Face processing in the First Year of Life*. Dordrecht, Netherlands: Kluwer Academic Publishers, pp. 259–274.
Kuhl PK and Iverson P (1995) Linguistic experience and the "perceptual magnet effect". In: W Strange (ed) *Speech Perception and Linguistic Experience: Issues in Cross-language Research* Baltimore: York Press, pp. 121–154.
Lakens D (2017) Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4): 355–362. https://doi.org/10.1177/1948550617697177.
Lakens D, Scheel AM and Isager PM 2018. Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2): 259–269. https://doi.org/10.1177/2515245918770963.
Lee J, Jang J and Plonsky L (2015) The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, 36(3): 345–366. https://doi.org/10.1093/applin/amu040
Levis J and Pickering L (2004) Teaching intonation in discourse using speech visualization technology. *System* 32: 505–524. https://doi.org/10.1016/j.system.2004.09.009
Lisker L and Abramson AS (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3): 384–422. http://dx.doi.org/10.1080/00437956.1964.11659830
Liu H, Ng ML, Wan M, Wang S and Zhang Y (2007) Effects of place of articulation and aspiration on voice onset time in mandarin esophageal speech. *Folia Phoniatrica et Logopaedica* 59: 147–154. https://doi.org/10.1159/000101773
Lord G (2005) (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania* 88(3): 557–567. https://doi.org/10.2307/20063159
Luce PA (1986) *Neighborhoods of words in the mental lexicon*. Doctoral dissertation, Indiana University, Bloomington
Mitterer H, Reinisch E and McQueen J (2018) Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language 98, 77–92.* https://doi.org/10.1016/j.jml.2017.09.005
Motohashi-Saigo M and Hardison D (2009) Acquisition of L2 Japanese geminates training with waveform displays. *Language Learning and Technology* 13(2): 29–47. Retrieved from http://llt.msu.edu/vol13num2/motohashisaigohardison.pdf
Munro MJ and Derwing TM (1995) Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45(1): 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x
Norris D, McQueen JM and Cutler A (2003) Perceptual learning in speech. *Cognitive Psychology* 47(2): 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9
Norris J and Ortega L (2000) Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50(3): 417–528.
Okuno T (2013) *Acquisition of L2 vowel duration in Japanese by native English speakers.* Doctoral dissertation, Michigan State University, East Lansing, MI.
Offerman HM and Olson DJ (2016) Visual feedback and second language segmental production:

The generalizability of pronunciation gains. *System* 59: 45–60. https://doi.org/10.1016/j.system.2016.03.003

Olson DJ (2014a) Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning and Technology,* 18(3): 173–192. Retrieved from http://llt.msu.edu/issues/october2014/olson.pdf

Olson DJ (2014b) Phonetics and technology in the classroom: A practical approach to using speech analysis software in second-language pronunciation instruction. *Hispania*, 97(1) 47–68.

Olson DJ (2019) Feature acquisition in second language phonetic development: Evidence from phonetic training. *Language Learning* 69(2), 366–404.

Plonsky L and Oswald FL (2014) How big is "big"? Interpreting effect sizes in L2 research. *Language Learning* 64(4): 878–912.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reinisch E, Wozny DR, Mitterer H and Holt LL (2014) Phonetic category recalibration: What are the categories? *Journal of Phonetics* 45: 91–105. https://doi.org/10.1016/j.wocn.2014.04.002

Revelle W (2018) psych: Procedures for Personality and Psychological Research. R package version 1.8.4. https://CRAN.R-project.org/package=psych.

Rosner B, López-Bascuas, LE, García-Albea, JE and Fahey, R (2000) Voice-onset times for Castilian Spanish initial stops. *Journal of Phonetics* 28: 217–224.

Ruellot, V (2011) Computer-assisted pronunciation learning of French /u/ and /y/ at the intermediate level. In: J Levis and K LeVelle (eds) *Proceedings of the 2nd Pronunciation in Second Language Learning and Teaching Conference.* Ames, IA: Iowa State University, pp. 199–213.

Sagey EC (1986) *The representation of features and relations in non-linear phonology*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Saito K (2007) The influence of explicit phonetic instruction on pronunciation teaching in EFL settings: the case of English vowels and Japanese learners of English. *The Linguistics Journal* 3(3): 16–40.

Samuel AG (1986) Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology* 18(4), 452–499. https://doi.org/10.1016/0010-0285(86)90007-1

Simon, E (2009) Acquiring a new second language contrast: An analysis of the English laryngeal system of native speakers of Dutch. *Second Language Research* 25(3), 377–408. https://doi.org/10.1177/0267658309104580

Tyler MD (2019) PAM-L2 and phonological category acquisition in the foreign language classroom. In: Nyvad AM, Hejná M, Højen A, Bothe Jespersen A and Hjortshøj Sørensen (eds) *A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn,* pp. 607–630.

Valdés G (2001) Heritage language students: Profiles and possibilities. In J K Peyton, D Ranard and S McGinnis (eds.) *Heritage languages in America: Blueprint for the future* Washington, DC: Center for Applied Linguistics and Delta Systems, pp. 37–78.

Vroomen J, van Linden S, Keetels M, de Gelder B and Bertelson P. (2004) Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication* 44(1): 55–61.

Vroomen J, van Linden S, de Gelder B and Bertelson P (2007) Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(3): 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031
Westbury J and Keating P (1986) On the naturalness of stop consonant voicing. *Journal of Linguistics* 22(1): 145–166. https://10.1017/S0022226700010598
Wetzels WL and Mascaró J (2001) The typology of voicing and devoicing. *Language* 77(2): 207–244.
Williams L (1977) The voicing contrast in Spanish. *Journal of Phonetics* 5: 169–184.

Appendix A. Language Background Questionnaire Results)

Subcomponent	Scale	English	Spanish
Language Acquisition	Age of Acquisition (years)	0.0 (0.0)	13.2 (2.6)
Language Use	Percentage of Daily Use (0–100)	94.9 (4.7)	5.4 (4.4)
Language Proficiency	Likert Scale (1–7)[a]	6.8 (0.4)	3.4 (0.9)
Language Attitudes	Likert Scale (1–7)[b]	6.6 (1.0)	1.5 (0.6)

[a]1 = do not speak language at all; 7 = speak language very well

[b]1 = do not identify with culture at all; 7 = identify with culture very strongly

Appendix B. Main Model Random Effects

Participant	Variance	Std. Dev.	Corr.		
Intercept	0.051	0.225			
Posttest	0.016	0.125	-0.66		
Delayed Posttest	0.052	0.229	-0.63	0.98	
Voiceless	0.051	0.227	-0.60	0.12	0.26

Initial Phoneme	Variance	Std. Dev.	Corr.		
Intercept	0.002	0.042			
Posttest	0.003	0.053	-0.99		
Delayed Posttest	0.002	0.046	-0.92	0.85	
Voiceless	0.009	0.098	0.00	0.15	-0.40

Appendix C. Voiceless Phoneme Model Random Effects)

Participant	Variance	Std. Dev.	Corr.	
Intercept	0.051	0.227		
Posttest	0.028	0.168	-0.73	
Delayed Posttest	0.058	0.241	-0.51	0.96

Initial Phoneme	Variance	Std. Dev.	Corr.	
Intercept	0.011	0.104		
Posttest	0.001	0.027	0.10	
Delayed Posttest	0.001	0.032	-0.87	0.41

Appendix D. Voiced Phoneme Model Random Effects)

Participant	Variance	Std. Dev.	Corr.	
Intercept	0.059	0.242		
Posttest	0.016	0.127	-0.77	
Delayed Posttest	0.069	0.263	-0.72	1.00

Initial Phoneme	Variance	Std. Dev.	Corr.	
Intercept	0.011	0.104		
Posttest	0.001	0.027	-1.00	
Delayed Posttest	0.001	0.032	-0.98	0.97