2021

# Maximizing the effect of visual feedback for pronunciation instruction: A comparative analysis of three approaches

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

Heather M. Offerman
*Davidson College*, heofferman@davidson.edu

## Recommended Citation

Olson, D. J., & Offerman, H. M. (2021). Maximizing the effect of visual feedback for pronunciation instruction: A comparative analysis of three approaches. *Journal of Second Language Pronunciation,* 7(1), 89-115.

# Maximizing the effect of visual feedback for pronunciation instruction: A comparative analysis of three approaches

**Daniel J. Olson[a] and Heather M. Offerman[b]**

[a]Purdue University
640 Oval Drive
West Lafayette, IN, 47907-2039
danielolson@purdue.edu

[b]Davidson College
Box 7140
Davidson, NC 28035
heofferman@davidson.edu

## Abstract

Visual feedback, in which learners visually analyze acoustic speech characteristics, has been shown to significantly improve pronunciation, but extant research has varied widely with respect to the target feature, length of the intervention, and type of intervention. This study presents a comparative analysis of three methods of visual feedback for L2 segmental pronunciation instruction. These methods, all focused on training voice onset time for English-speaking learners of Spanish, differed in duration of instruction (i.e., short and long) and the nature of each intervention (i.e., phonemes presented simultaneously or sequentially). Results show that while all forms of visual feedback significantly improve L2 Spanish pronunciation, evidenced by a reduction in voice onset time, the greatest improvement was found following both longer treatments and a sequential approach. Theoretical and pedagogical implications are discussed.

**Keywords:** visual feedback; voice onset time; Spanish; classroom; technology; English; pronunciation

## 1. Literature Review

### *1.1. Trends in Pronunciation Instruction*

As a growing body of research has begun to establish the benefits of pronunciation instruction on second language (L2) learner production, several trends have emerged. First, as illustrated in both a recent meta-analysis (Lee et al., 2015) and a narrative review (Thomson & Derwing, 2015), technology has begun to play a significant role in pronunciation instruction. A number of authors have noted that technology-aided pronunciation instruction is unique in its ability to “promote learner autonomy” and “afford the possibility of individualized instruction” (Thomson & Derwing, 2015, p. 336). While feedback has long been shown to improve learner outcomes in lexical and morphosyntactic acquisition (for a review, see Plonsky & Brown, 2015), feedback has also been shown to significantly improve outcomes in pronunciation (Lee et al., 2015). As such, technology-aided pronunciation instruction, particularly with an individualized feedback component, represents a positive avenue for further research. One such approach that has been the source of a small, but growing body of research, is the use of visual feedback for pronunciation instruction (e.g., Olson, 2014a). Finally, there has been a growing call for research that takes a comparative approach to different methods of pronunciation instruction (e.g., Derwing & Munro, 2015). Responding to these trends, the current study presents a comparative analysis of the effectiveness of three visual feedback methodologies for L2 pronunciation instruction. These three methodologies can be broadly characterized as a short visual feedback paradigm (i.e., single intervention), a long simultaneous paradigm (i.e., multiple interventions that simultaneously address several phonemes during each intervention), and a long sequential paradigm (i.e., multiple interventions that address a single phoneme during each intervention).

### *1.2. Visual Feedback in Pronunciation Instruction and Learning*

Broadly, *visual feedback* refers to any paradigm in which L2 learners receive feedback on their productions through a visual modality. Kartushina et al. (2015) drew a distinction between direct and indirect visual feedback. Direct visual feedback describes those paradigms in which learners are given a direct image of the position of the articulators during speech production, such as through ultrasound or electropalatography. Indirect visual feedback describes paradigms in which learners are given a visual representation of the speech sound, usually their own productions and those of a native speaker (NS), corresponding to some facet of the acoustic signal. In indirect visual feedback, learners are required to extrapolate, either implicitly or explicitly, from the acoustic signal to the motor movements required to produce that acoustic signal. A further distinction should be made between real-time visual feedback (e.g., Garcia et al., 2018), in which the visual representation is produced concurrent with speech production, and delayed visual feedback, in which the visual representation is delivered (sometimes immediately) following speech production (e.g., de Bot, 1980). A typical indirect visual feedback paradigm consists of: (a) learner production of the target stimuli, (b) visual display of the relevant acoustic features, (c) visual display of NS productions for comparison, with or without auditory presentation, and (d) subsequent production in which the learner attempts to “match” the NS production.[1] Indirect feedback often takes the form of an intonation contour (e.g., Chun, 1998), waveforms (e.g., Akahane-Yamada et al., 1998; Motohashi-Saigo & Hardison, 2009), and/or spectrograms (e.g., Olson, 2014b). The type of visual feedback is generally tailored to the feature under examination.

Within the field of L2 pronunciation instruction, widespread use of visual feedback is a relatively recent development allowed by easy access to computing technology. Many of the earliest studies examined intonation contours and other suprasegmentals (e.g., de Bot, 1980, de Bot, 1983; Chun, 1989). This line of research has shown that visual feedback, largely through the display of both native-speaker and learner-produced intonation contours, results in more native-like L2 productions. This finding holds for lexical tone (for L1 English–L2 Mandarin see Fischer, 1986, as cited in Chun, 1998), utterances (e.g., Hardison, 2004), and discourse (Chun, 1998, 2002; Levis & Pickering, 2004). Other research has shown that gains made on target stimuli generalize to non-trained stimuli (Hardison, 2004).

More recently, visual feedback has been used for the instruction of segmental features in L2 speech (see Chun, 2007). While early research on the use of visual representation for L2 instruction intended to provide a roadmap for classroom integration (e.g., Lambacher, 1999; Moholt, 1988), more recent work has sought to quantify its potential benefits. For vowels, overall results have been somewhat mixed, with some studies showing a positive impact of visual feedback on vowel production (e.g., Saito, 2007) and others showing no effect (e.g., Carey, 2004; Ruellot, 2011).

Results, however, have been more consistent for duration-based contrasts. Motohashi-Siago and Hardison (2009) found significant improvement in the accuracy of the singleton–geminate distinction for English-speaking learners of L2 Japanese with visual feedback, beyond gains seen in an audio-only condition. Similar results were found for vowel duration by Okuno (2013), in which the addition of visual feedback significantly improved the accuracy of vowel duration contrasts. Finally, a series of studies (Offerman & Olson, 2016; Olson, 2019; Olson, under review) has shown a significant impact of visual feedback on the production of voice onset time by English-speaking learners of L2 Spanish relative to a control group.

Visual feedback relies on noticing, as described in the 'noticing hypothesis' originally proposed by Schmidt (1990). Within this framework, conscious awareness of L2 forms, or differences between L1 and L2 forms, may be necessary for some L2 development. Schmidt (1995) argues that what learners are able to notice in the input becomes 'intake', which leads to acquisition. Noticing may be particularly important for phonetic acquisition (Derwing & Munro, 2005), as models of L2 phonetic acquisition suggest that phonemes in the L2 that are similar to existing L1 categories may be perceptually assimilated to the existing category at the initial stages of L2 acquisition (e.g., Perceptual Assimilation Model- L2 (PAM-L2): Best & Tyler, 2007; Speech Learning Model (SLM): Flege, 1995). Increasing metalinguistic awareness through training that raised consciousness of L2 phonetic features has been shown to significantly improve L2 contrast perception (e.g., Flege & Wang, 1989). As such, visual feedback offers a distinct (visual) modality to aid L2 learners in the task of noticing features in the L2 and differences between the L1 and L2. Moreover, given the links between L2 phonetic perception and production (e.g., Bradlow et al., 1997), the benefits of visual feedback for production may be tied to a learner's ability to notice L2 phonetic contrasts.

Olson (2014b) argued that as visual feedback may function to increase learners' ability to notice specific contrasts, certain features may be inherently suited to visual feedback. Stylized intonation contours (Spaii & Hermes, 1993), for example, may be relatively intuitive for learners to interpret (Léon & Martin, 1972; Anderson-Hseih, 1992). At the segmental level, Olson (2014b) noted that differing features may be easier or harder to interpret. Durational contrasts, for example, may be both easy to perceive visually and relatively straightforward in their correlation to articulatory gestures. Other features, such as rhoticity, nasality, or even vowel articulation, may be either harder to perceive visually (e.g., anti-formants in nasal vowels) or be more abstract in their relation to articulatory gestures (e.g., F1 is inversely correlated to tongue height in vowel production).

Within this overarching approach to visual feedback, a number of different methodologies have been implemented, including a visual model of NS production (e.g., de Bot, 1980), a combined audio-visual model (de Bot, 1983), audio and visual feedback (e.g., Ruellot, 2011), and visual feedback as part of a larger, multi-faceted course on pronunciation (e.g., Lord, 2005). Studies have differed in terms of length, from single interventions with accuracy measured immediately before and after training (e.g., Olson, 2019) to multiple interventions delivered over the course of a semester (Okuno, 2013). In addition, visual feedback has been implemented in both simultaneous and sequential approaches. Here, *simultaneous* is defined as addressing multiple phonemes in the same intervention (e.g., Olson, under review) and *sequential,* more closely following the Processing Instruction approach (Lee & VanPatten, 1995), as addressing a single unique phoneme in each intervention (e.g., Offerman & Olson, 2016). Despite the varied methodologies, there has been no comparison of different visual feedback methodologies, and more careful cross-methodological comparison is clearly warranted.[2]

### *1.3. Voice Onset Time in English and Spanish*

The three methods examined in the current study incorporated visual feedback for the instruction of L2 voiceless stops, exploiting cross-linguistic differences in voice onset time between Spanish and English. Voice onset time (VOT), which has been shown to be a reliable cue to stop consonant voicing (Lisker & Abramson, 1964), is defined as the temporal difference between the release of the oral closure and the onset of vocal fold vibrations. Traditional descriptions often divide VOT into three categories: prevoicing, in which vocal fold vibration occurs prior to the release (VOT < 0 milliseconds); short-lag, with vocal fold vibration occurring within 0–30ms after the release (0 milliseconds < VOT < 30 milliseconds); and long-lag, in which the onset of voicing occurs 30ms or more after the release (30 milliseconds < VOT). For word-initial voiceless stops, English generally employs long-lag VOT and Spanish employs short-lag VOT (e.g., Lisker & Abramson, 1964: Spanish means = 4–29ms, English means 58–80ms).

English-speaking learners of Spanish, particularly at the beginning and intermediate levels, have been shown to employ English-like, long-lag VOTs in word-initial position in their L2 (Lord, 2005). Within the theoretical framework of intelligibility, comprehensibility, and accentedness (e.g., Munro, 2016), this cross-linguistic difference has been shown to impact judgments of accentedness (e.g., Lord, 2005) and foreignness (McBride, 2015), although it is unlikely to significantly impact intelligibility (Lord, 2005). While many have argued for a focus on intelligibility in L2 pronunciation training, VOT was chosen for the current study for both its pedagogical and theoretical value, as its gradient nature allows for a nuanced comparison between differing visual feedback methodologies. Furthermore, prior research has shown that visual feedback is effective for durational contrasts, making VOT a strong test case for visual feedback paradigm comparisons.

### *1.4. Research Questions*

Responding to a call in the literature for a carefully implemented comparative analysis (Derwing & Munro, 2015), this study provides an empirical comparison of several different approaches to visual feedback for pronunciation instruction. Three specific research questions are addressed:

(1) Does visual feedback, in the form of waveforms and spectrograms, contribute to a more target-like production of the L2 segmental feature VOT?

(2) How does the duration of the visual feedback paradigm, operationalized as the number of iterations or interventions, impact the degree of change in L2 VOT?

(3) How does the specific nature of the visual feedback paradigm, specifically whether all related stop consonants are addressed at the same time (i.e., simultaneous) or individually (i.e., sequential), impact the degree of change in L2 VOT?

## 2. Methodology

To address the research questions, this study provides a detailed quantitative comparison of three previous studies, each employing different methods of visual feedback. While much of the methodology is identical or similar across the three studies, two key differences should be noted: the duration of the treatment (i.e., number of visual feedback activities) and the nature of these treatments (i.e., simultaneous or sequential). Table 1 illustrates the distribution of these key characteristics in the previous studies.

Table 1. Key Methodological Characteristics

| Method | Author(s), Year | Phonetic Feature | Number of Interventions | Nature of Interventions |
|---|---|---|---|---|
| Short | Olson, 2019 | VOT | 1 | – |
| Long Simultaneous | Olson, under review | VOT | 3 | Simultaneous |
| Long Sequential | Offerman & Olson, 2016 | VOT | 3 | Sequential |

While full details of these studies are available in the respective publications, relevant details are provided here to highlight the comparability of the participants, stimuli, and procedure across the three different studies. In addition, in each original study, the results of the experimental group(s) were compared to a control group, which received either unrelated cultural lessons for an equal amount of time (Offerman & Olson, 2016) or training on an unrelated phonetic feature (Olson, 2019; Olson, under review). As the experimental group outperformed the control group in each of the three studies, details of the control groups are not included here. For each of the studies, the interventions were included in the course curriculum and participants received a grade for their completion of each part of the study. No additional feedback was given. Participation in the studies was voluntary.

### *2.1. Method 1: Short Visual Feedback*

Method 1, focused on training and improving L2 voiceless stop consonants (i.e., evidence by a reduction in VOT), consisted of a single instantiation of a visual feedback paradigm (i.e., short intervention) that addressed one of the three voiceless stop consonants. The full details of Method 1 are available in Olson (2019) (Study 1).

#### *2.1.1. Participants*

Twenty-five English-speaking learners of Spanish participated in Study 1. Participants were NSs of English, having learned English from birth, and L2 learners of Spanish, having begun acquisition of Spanish after the age of 5 (*M* =12.9; *SD* = 3.8). All participants reported speaking only English in the home and not having spent any significant time (> six weeks) in a non-English-speaking region. Participants were recruited from several intact fourth-semester Spanish-language courses at a large, public Midwestern university. All participants were considered to be intermediate-level language learners, having been placed at the intermediate-level via a standardized institutional placement exam or successful completion of a previous language course.

### *2.1.2. Stimuli*

Two sets of stimuli were included in Study 1: tokens embedded in novel utterances and tokens in isolation. All stimuli consisted of target words with word-initial voiceless stops /p, t, k/. For the words in utterances, 90 unique two-syllable words were embedded in utterance-medial position, with thirty tokens for each voiceless stop. All targets were non-cognate paroxytonic words that were balanced for the vowel immediately following the initial voiceless stop (/a/ or /o/). Each of the resulting CV syllables was represented in 15 different words (3 initial stops × 2 vowels × 15 words = 90 target tokens). Thirty distinct tokens were recorded as part of the pretest, posttest, and delayed posttest. Tokens were also balanced for word familiarity via a word familiarity norming study in which a unique group of L2 learners of Spanish (*N* = 19) evaluated each of the tokens for familiarity on a 7-point Likert scale (Auer et al., 2000). Results demonstrate that, although familiarity was variable, ranging from highly unfamiliar to highly familiar, there was no significant difference in the familiarity of the tokens between the pretest, posttest, and delayed posttest. All tokens were placed in medial position of utterances that were novel to participants. Example (1) shows sample stimuli for /p,t,k/.

(1) a. *Miles de patos nadan en los lagos grandes.* '
‘Thousands of ducks swim in the big lakes.’ "
b. *Si llego tarde mañana, mi madre me gritará.*
‘If I arrive late tomorrow, my mother will yell at me.’
c. *Mi abuelo vende camas en su tienda.* '
‘My grandfather sells beds in his store.’ "

For words in isolation, a unique set of 30 words balanced for initial voiceless stops, were recorded. As with the words in utterances, the words in isolation were all non-cognate, paroxytonic, two-syllable words. Unlike the words in utterances, the words in isolation were repeated in each of the recording sessions (pretest, posttest, delayed posttest) and produced in isolation, without any surrounding utterance.

### *2.1.3. Procedure*

Broadly, the visual feedback paradigm employed here followed recommendations by Olson (2014a) and consisted of pre-recording, self and NS visual analysis, NS and non-native speaker (NNS) comparison, and re-recording. During the pre-recording phase, participants recorded their own productions of the target stimuli. Using Praat (Boersma & Weenink, 2018), participants produced and printed the waveforms and spectrograms of five words in isolation and attempted to segment them into their individual component phonemes. This pre-recording phase formed the basis of the pretest and was conducted prior to the in-class analysis.

For the self and NS analysis phase, participants answered a series of guiding questions about the visual images of their own productions, drawing attention to the relevant phonemes. Subsequently, participants were provided with the waveform and spectrogram of the same tokens produced by a NS of Spanish (male, peninsular variety) and asked to answer a similar set of guiding questions. Example 2 illustrates the guiding questions for self-analysis, originally provided in the target language.

(2) a. How did you decide where the boundaries of each sound were?
b. What are the visual characteristics of your *p*?
c. Is your *p* longer or shorter than the *e*?

During the NNS and NS comparison phase, participants were asked to describe the visual differences between NNS and NS productions, as well as to create hypotheses about the auditory differences between the two. To confirm their hypotheses, they were given three pairs of spectrograms/waveforms for novel words and asked to identify which word in the pair was produced by a NS of Spanish. Both the analysis and comparison phases were conducted as part of a single 50-minute class period.

Finally, during the re-recording stage, participants were given three days to record the second set of stimuli, including a new set of tokens in novel utterances and the repeated set of tokens in isolation. The delayed posttest, in which participants again recorded a new set of tokens in utterances and the same set of words in isolation, was conducted approximately four weeks after the re-recording phase. The timing of the delayed posttest was chosen following the mean delayed posttest timing reported by Norris and Ortega (2000). All recordings were conducted as homework.

Because the goal of Study 1 was to test whether changes in VOT for a single (trained) voiceless stop (e.g., /p/) generalize to other (non-trained) voiceless stops (e.g., /t/ and /k/), participants received training on only one of the three voiceless stop consonants. Of the 25 participants in Study 1, eight received visual feedback training on the production of /p/, 13 on /t/, and four on /k/. Unequal group sizes are accounted for in the analysis by using a normalized VOT value (see Data Analysis).[3]

### *2.2. Method 2: Long Simultaneous Visual Feedback*

Parallel to Method 1, Method 2 also focused on the training of voiceless stop consonants by English-speaking L2 learners of Spanish. In contrast to Method 1, which included a single intervention and separated the pretest and posttest by a maximum of 5 days, Method 2 consisted of three separate iterations of the visual feedback paradigm, separated by approximately 3–4 weeks. The full details of Method 2 are available in Study 2 (Olson, under review), which included an analysis of both voiced and voiceless stop consonants. Only the voiceless subset is included in the current comparison.

#### *2.2.1. Participants*

Participants (*N* = 20) in Study 2 were drawn from a similar population as those in Study 1. Specifically, they were all native English-speaking learners of Spanish (age of acquisition *M* = 13.2, *SD* = 2.6), enrolled in intermediate-level Spanish language courses, with no long-term stays in Spanish-speaking regions.

#### *2.2.2. Stimuli*

Again, two sets of stimuli were included in Study 2: tokens in novel utterances and tokens in isolation. All tokens consisted of Spanish words with word-initial voiceless stops /p,t,k/.

For the words in utterances, 54 tokens were balanced for each of the three word-initial stops (18 tokens per voiceless consonants). All tokens were two-syllable, non-cognate, paroxytonic words, and controlled for word familiarity. The vowel following the consonant of interest was also controlled, balanced between /a,e,u/. Each resulting CV combination was represented by six unique stimuli (3 initial stops × 3 vowels × 6 words = 54 target tokens). Stimuli were divided equally (*n* = 18) across the three sessions (pretest, posttest, delayed posttest). All tokens were embedded in utterance-initial position of novel utterances.[4]

For words in isolation, a single set of 18 tokens, balanced for initial voiceless stops, were also recorded. Again, all tokens were non-cognate, two-syllable, paroxytonic Spanish words. As in Study 1, the same words in isolation were repeated in each of the recording sessions. The words in isolation in Study 2 were recorded only at the pretest and delayed posttest.

#### *2.2.3. Procedure*

Study 2 employed a visual feedback paradigm, with a pretest, posttest, and delayed posttest. The intervention in Study 2 consisted of three separate iterations of the visual feedback paradigm, conducted over the course of 4 weeks. Each iteration of the visual feedback paradigm included a pre-recording, self and NS analysis, NS and NNS comparison, and a re-recording phase. Within each of these individual phases, the methods were identical to those detailed in Study 1. The pretest was defined as the pre-recording phase of the first visual feedback iteration, and the posttest was defined as the re-recording phase of the third visual feedback iteration.

The three training sessions of the visual feedback paradigm in Study 2 employed a simultaneous approach, in which all three voiceless phonemes were addressed during each intervention. These trainings were scaffolded, such that each intervention addressed tokens in increasingly complex and naturalistic structures. Participants analyzed words produced in isolation in training 1, words in utterances in training 2, and words in paragraph-length connected discourse in training 3. A delayed posttest was conducted approximately four weeks after the posttest.

### ***2.3. Method 3: Long Sequential Visual Feedback***

Method 3 also focused on the training of voiceless stop consonants by English-speaking L2 learners of Spanish, and consisted of three separate iterations of the visual feedback paradigm, separated by 2–3 weeks. In contrast to Method 2, each intervention in Method 3 provided training on one single voiceless stop (i.e., sequential). The full details of Method 3 are available in Offerman and Olson (2016) (Study 3).

#### *2.3.1. Participants*

Participants (*N* = 17) in Study 3 were drawn from a similar population as those in Studies 1 and 2. All participants were native English-speaking learners of Spanish, having begun acquisition after the age of 12, enrolled in fourth semester Spanish language courses, and reported no significant immersion experience in Spanish-language regions.

*2.3.2. Stimuli*

Two sets of stimuli from Study 3 are relevant for the current comparison: tokens in novel utterances and tokens in carrier phrases. All tokens were Spanish words with word-initial voiceless stop /p,t,k/.

For the words in utterances, there were a total of 15 target tokens, balanced for the three word-initial voiceless stops. Tokens were also balanced for the vowel immediately following the voiceless stop, /a,e,i,o,u/. Each of the CV combinations as represented by a single stimulus (3 initial stops × 5 vowels × 1 word = 15 target tokens). All words were non-cognate. Most tokens were two-syllable, paroxytonic words (*n* =13) and were embedded in utterance-medial position of novel utterances. In contrast to Studies 1 and 2, the tokens in novel utterances were repeated at the pretest and delayed posttest. Words in utterances at the posttest were different from those at the pretest and delayed posttest, but all tokens followed the same phonetic CV pattern.

Words in carrier phrases consisted of a single set of 30 stimuli, balanced for word-initial stops. All tokens were non-cognate, and most were two-syllable, paroxytonic words (*n* = 26). The target tokens embedded in novel utterances were a subset of the tokens used in carrier phrases. In addition, while Studies 1 and 2 included words in isolation, Study 3 included words in carrier phrases. Participants produced the carrier phrase *Di* ______ *de nuevo* (‘Say ______ again’) with the different target tokens embedded within the utterance. As with the words in utterances, the same tokens were repeated at the pretest and delayed posttest. A different set of words in carrier phrases were employed at the posttest but followed the same CV pattern and were non-cognate. While the carrier phrase methodology is not identical to the words in isolation methodology used in the other studies, both methodologies represent a structure with presumably lower cognitive load than the words in novel utterances. As a result, the words in carrier phrases represented the most relevant point of comparison with the words in isolation in the previous studies.

*2.3.3. Procedure*

Again, a visual feedback paradigm was employed with a pretest, posttest, and delayed posttest design and included three visual feedback interventions. Each iteration of the visual feedback paradigm included a pre-recording, self and NS analysis, NS and NNS comparison, and a re-recording phase. Methodology for each of the individual phases was parallel to that described in Study 1.[5] The pretest was defined as the pre-recording phase of the first visual feedback intervention, and the posttest was defined as the re-recording phase of the third intervention.

The three interventions in Study 3 employed a sequential approach, in which only one single voiceless phoneme was addressed during each of the training sessions. Participants analyzed words, produced within carrier phrases, with word-initial /p/ in training session 1, /t/ in training session 2, and /k/ in training session 3. The delayed posttest was conducted approximately two weeks after the posttest.

### *2.4. Summary of Key Methodological Differences*

While the three methodologies above were similar in many respects (participant type, target phonemes, intervention design), they had two key differences. First, the studies differed with respect to the duration of the treatment (Table 1). Specifically, the Short Visual Feedback study (Olson, 2019) consisted of a single intervention while the Long Simultaneous (Olson, under review) and Long Sequential (Offerman & Olson, 2016) studies each consisted of three different interventions. Second, these studies differed with respect to the nature of the intervention. In the Long Simultaneous paradigm, participants received training on all three voiceless stop consonants (/p,t,k/) during each intervention. In contrast, in the Long Sequential paradigm, participants received training on a single stop consonant during each intervention (/p/, /t/, *or* /k/).

### *2.5. Data Analysis*

#### *2.5.1 Words in Novel Utterances*

For the three-study comparison of words in novel utterances, a total of 4,095 tokens were included in the initial analysis. Study 1 accounted for 2,250 tokens (25 participants × 30 tokens × 3 sessions), Study 2 accounted for 1,080 tokens (20 participants × 18 tokens × 3 sessions)[6], and Study 3 accounted for 765 tokens (17 participants × 15 tokens × 3 sessions). Of the initial set of tokens, approximately 9% of tokens were eliminated for various errors, including missing data and poor recording quality, leaving 3,734 total tokens. Finally, responses ± 3 SD from the total mean (*n* = 20) were eliminated. After accounting for the participants that missed a session, the percentage of tokens was eliminated from the analysis of each study was roughly 6-8% (Study 1 = 6.3%; Study 2 = 8.6%; Study 3 = 6.4%). A total of 3,714 tokens were included in the final analysis.

#### *2.5.2. Words in Isolation*

The initial comparison of words in isolation included a total of 4,230 tokens. Study 1 contributed 2,160 tokens (24 participants × 30 tokens × 3 sessions)[7], Study 2 accounted for 540 tokens (15 participants × 18 tokens × 2 sessions), and Study 3 accounted for 1530 tokens (17 participants × 30 tokens × 3 sessions). Following elimination of errors (*n* = 116, approximately 3%) and outliers (± 3 SD, *n* = 21), a total of 4093 tokens were included in the final analysis. Again, a modest percentage of tokens was eliminated from the analysis in each study (Study 1 = 4.3%; Study 2 = 4.1%; Study 3 = 1.3%).

#### *2.5.3. Voice Onset Time*

VOT was measured using Praat (Boersma & Weenink, 2018), with particular attention given to the waveform. VOT was defined as the temporal difference between the release of the oral closure, identified by the burst in the waveform, and the onset of vocal fold vibration, identified as the onset of periodicity in the waveform. VOT differs across place of articulation, with bilabials produced with the shortest VOT and velars produced with the longest VOT (see Cho & Ladefoged, 1999), and the difference between mean English and Spanish voiceless stop VOT is dependent on place of articulation. For example, using data from Lisker and Abramson (1964), English and Spanish word-initial /t/ differ by an average of 61ms, but English and Spanish /k/ differ by only 51ms. As such, a similar shift in raw VOT (ms) represents a different percentage for these two phonemes. A such, a normalized VOT measure, which allowed for all phonemes to be analyzed on a similar scale, was calculated to facilitate comparison (see Olson, 2019).

To normalize VOT, the raw VOT (ms) for each token was converted into a ratio based on the average Spanish and English VOT reported in Lisker & Abramson (1964). In this ratio, a value of 0 represents a token with the average Spanish VOT value (4 milliseconds for /p/, 9 milliseconds for /t/, 29 milliseconds for /k/). A value of 1 represents a token with the average English VOT value (58 milliseconds for /p/, 70 milliseconds for /t/, 80 milliseconds for /k/).

## 4. Results

### *4.1. Words in Utterances*

Initial analysis was conducted using a linear mixed effects model, with normalized VOT as the dependent variable and method (short, long simultaneous, long sequential) and session (pretest, posttest, delayed posttest) as fixed effects. Participant and phoneme were included as random effects. Phoneme, rather than token, was selected as a random effect given the lack of consistency in the tokens across studies. Following recommendations by Barr et al. (2013), the maximal random effects structure that permitted convergence was employed: random intercepts and random slopes by method for participant and random intercepts for phoneme. The significance criterion was set at $|t| \geq 2.00$. Statistical analysis was conducted using R software (R Core Team, 2013) and the lme4 package (Bates et al., 2014). Effect size confidence intervals were conducted using the psych package (Revelle, 2018).

To justify the inclusion of each of the main effects (method and session), two subsequent models were conducted, each dropping one of the main effects. Model comparison showed that the original model (loglikelihood = -1128.9) was a significantly better fit than both the model without method (loglikelihood = - 1263.7, $\chi^2(11) = 269.56$, $p < .001$) and the model without session (loglikelihood = -1249.5, $\chi^2(6) = 241.16$, $p < .001$).

Results from the fixed effects for the full model (Table 2) showed no difference in the normalized VOT produced at the pretest between the intercept (short, pretest) (*M* = 0.70, *SD* = 0.43) and either the long simultaneous (*M* = 0.61, *SD* = 0.42, *b* = -0.081, *t* = -1.131) or long sequential (*M* = 0.79, *SD* = 0.47, *b* = .114, *t* = 1.392) studies, suggesting that the participants in the different studies were adequately matched. There was a significant effect of session, with significant differences between the intercept and both the posttest (*M* = 0.61, *SD* = 0.40, *b* = -0.093, *t* = -5.499) and delayed posttest (*M* = 0.59, *SD* = 0.37, *b* = -0.117, *t* = -6.880), highlighting the overall decrease in VOT following training for the short group.

Finally, the interactions between method and session illustrate that the improvement following treatment was not the same for all groups. First, there was no significant interaction between method and group for the long simultaneous group (posttest: *M* = 0.47, *SD* = 0.35, *b* = 0.002, *t* = 0.050; delayed posttest: *M* = 0.50, *SD* = 0.37, *b* = -0.033, *t* = -1.056). This lack of an interaction suggests that the impact of the training was the same for the short and long simultaneous groups. In contrast, for the long sequential group, there was a significant interaction between method and session at both the posttest (*M* = 0.52, *SD* = 0.42, *b* = -0.182, *t* = -5.315) and delayed posttest (*M* = 0.44, *SD* = 0.41, *b* = -0.235, *t* = -7.085). This interaction shows that the effect of training was different for the short and long sequential groups. An analysis of the mean normalized VOT for participant-produced Spanish voiceless stops, as well as the boxplot in Figure 1, highlight the nature of this interaction. Specifically, the long sequential group demonstrated more improvement (i.e., short VOTs) following training than the short group. This difference was present at both the posttest and delayed posttest.

Table 2. Novel Utterances Main Model Fixed Effects

| | Estimate | Std. Error | 95% CI | *t* value | Cohen's *d* | 95% CI |
|---|---|---|---|---|---|---|
| Intercept (Short, Pretest) | 0.689 | 0.065 | [0.599, 0.818] | 10.667 | | |
| Long Simultaneous | -0.081 | 0.072 | [-0.225, 0.063] | -1.131 | 0.21 | [0.067, 0.361] |
| Long Sequential | 0.114 | 0.816 | [-1.518, 1.745] | 1.392 | -0.21 | [-0.362, -0.067] |
| Posttest | -0.093 | 0.017 | [-0.127, -0.059] | -5.499 | 0.20 | [0.049, 0.343] |
| Delayed Posttest | -0.117 | 0.017 | [-0.151, -0.083] | -6.880 | 0.27 | [0.126, 0.421] |
| Long Simultaneous: Posttest | -0.033 | 0.031 | [-0.096, 0.030] | -1.056 | 0.58 | [0.431, 0.731] |
| Long Sequential: Posttest | -0.182 | 0.034 | [-0.250, -0.113] | -5.315 | 0.41 | [0.266, 0.563] |
| Long Simultaneous: Delayed Posttest | 0.002 | 0.032 | [-0.061, 0.065] | 0.050 | 0.49 | [0.341, 0.639] |
| Long Sequential: Delayed Posttest | -0.235 | 0.033 | [-0.302, -0.169] | -7.085 | 0.60 | [0.448, 0.748] |

A subsequent reordering of the model, with the long sequential pretest as the intercept (not shown in Table 2) further supports these findings. Specifically, there was a significant interaction between method and session for both the long simultaneous posttest (*b* = 0.148, *t* = 3.754) and long simultaneous delayed posttest (*b* = 0.237, *t* = 6.071), suggesting that the effect of the training was more pronounced for the long sequential group. Taken as a whole, while all groups displayed a significant decrease in the normalized VOT following training, the long sequential group outperformed both the short and the long simultaneous groups.

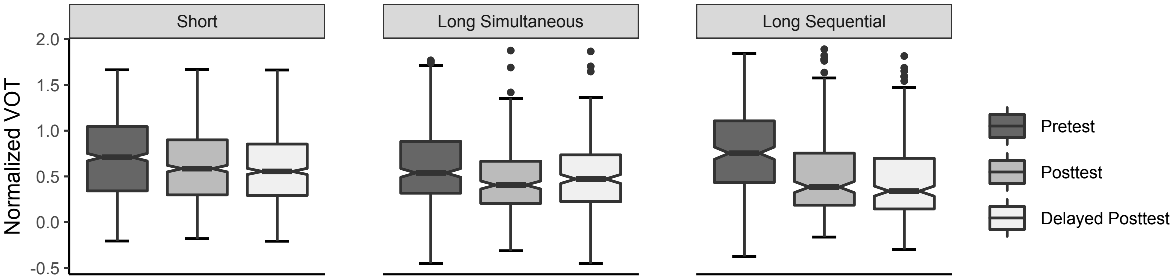

Figure 1. Normalized VOT ratios for words in novel utterances by session (pretest, posttest, delayed posttest) for each of the relevant methods (short, long simultaneous, long sequential).

Finally, to permit comparison with other studies, Table 3 presents the raw means and standard deviations in milliseconds by method, session, and phoneme. Across all studies and sessions, VOTs follow the expected pattern of shortest for bilabial and longest for velar stop consonants. Two trends should be noted. First, all of the interventions resulted in a decrease in L2 VOT, and this pattern was consistent across all places of articulation. Second, the long sequential method resulted in the shortest VOT at the delayed posttest across all places of articulation.

Table 3. Novel Utterances Mean VOT (ms)

| Study | Phoneme | Pretest | Posttest | Delayed Posttest |
|---|---|---|---|---|
| Short | /p/ | 44.3 (23.7) | 38.2 (22.6) | 38.2 (19.5) |
| | /t/ | 52.1 (24.9) | 49.0 (24.0) | 47.0 (22.6) |
| | /k/ | 61.3 (22.3) | 57.3 (19.7) | 54.5 (18.6) |
| Long Simultaneous | /p/ | 42.8 (21.1) | 34.6 (18.0) | 34.1 (17.9) |

| | | | | |
|---|---|---|---|---|
| | /t/ | 44.3 (22.1) | 39.1 (21.1) | 41.5 (21.7) |
| | /k/ | 55.5 (24.0) | 47.1 (17.3) | 49.8 (20.8) |
| Long Sequential | /p/ | 45.1 (21.6) | 28.9 (21.5) | 25.7 (19.7) |
| | /t/ | 57.9 (28.7) | 44.3 (28.2) | 36.5 (23.6) |
| | /k/ | 64.1 (28.5) | 50.1 (25.6) | 45.0 (25.2) |

### *4.2. Words in Isolation*

Statistical analysis for words in isolation paralleled the analysis for the words in utterances: method and session as fixed effects, and participant and phoneme as random effects. The random effects structure differed somewhat, as model convergence was possible only with random intercepts, but not slopes. Again, model comparison demonstrated that the original model (loglikelihood = -1499.7) represented a significantly better fit than either a model dropping the fixed effect of method (loglikelihood = -1677.2, $\chi^2(5) = 354.91$, $p < .001$) or session (loglikelihood = -1893.3, $\chi^2(5) = 787.09$, $p < .001$).

Results for the fixed effects in the full model (Table 4) for words in isolation largely parallel those from the tokens in novel utterances. Again, there was a significant difference between the Intercept (short, pretest, *M* = 0.69, *SD* = 0.40) and both the posttest (*M* = 0.62, *SD* = 0.38, *b* = -0.081, *t* = -4.317) and delayed posttest (*M* = 0.62, *SD* = 0.39, *b* -0.074, *t* = -4.075), showing a decrease in the normalized VOT following the short visual feedback paradigm. There was a significant difference between the initial VOTs for each group at the pretest, with the long sequential group (*M* = 0.95, *SD* = 0.42, *b* = 0.387, *t* = 16.368) producing greater VOTs than the short group (*M* = 0.69, *SD* = 0.40).

There were significant interactions between method and session, suggesting that the impact of session (i.e., training) was not the same for all groups. Specifically, there was a significant interaction between method and session for the long simultaneous group at the delayed posttest (*M* = 0.46, *SD* = 0.38, *b* = -0.153, *t* = -4.325). These results, coupled with an analysis of the means (i.e., Table 5) and Figure 2, show that the long simultaneous group showed greater improvement than the short group at the delayed posttest. (The long simultaneous group did not record words in isolation at the posttest.) Similarly, there was a significant interaction between method and session for the long sequential group at both the posttest (*M* = 0.47, *SD* = 0.42, *b* = -0.393, *t* = -13.741) and delayed posttest (*M* = 0.38, *SD* = 0.41, *b* = -0.492, *t* = -17.409). This interaction, coupled with an analysis of the means, shows that the long sequential group demonstrated a greater reduction in VOT than the short group at both the posttest and delayed posttest. Results from this set of interactions show that both long treatments, simultaneous and sequential, resulted in a greater reduction in VOT than the short treatment.

To compare the long simultaneous and long sequential groups, the model was reordered to set the long simultaneous group (pretest) as the intercept. There was a significant interaction between method and session at the delayed posttest (*b* = 0.339, *t* = 9.130). As seen in Figure 2, while all treatments resulted in a reduction of normalized VOT, the long sequential group outperformed the other two groups.

Table 4. Words in Isolation Main Model Fixed Effects

| | Estimate | Std.<br>Error | 95% CI | *t* value | Cohen's<br>*d* | 95% CI |
|---|---|---|---|---|---|---|
| Intercept (Short, Pretest) | 0.641 | 0.043 | [0.555, 0.728] | 12.694 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Long Simultaneous | 0.047 | 0.074 | [-0.100, 0.194] | -0.162 | 0.01 | [-0.134, 0.161] |
| Long Sequential | 0.387 | 0.024 | [0.339, 0.434] | 16.368 | -0.62 | [-0.774, -0.472] |
| Posttest | -0.081 | 0.018 | [-0.117, -0.045] | -4.317 | 0.19 | [0.046, 0.342] |
| Delayed Posttest | -0.074 | 0.018 | [-0.110, -0.038] | -4.075 | 0.18 | [0.177, 0.324] |
| Long Sequential: Posttest | -0.393 | 0.029 | [-0.451, -0.336] | -13.741 | 0.53 | [0.385, 0.685] |
| Long Simultaneous: Delayed Posttest | -0.153 | 0.035 | [-0.223, -0.082] | -4.325 | 0.60 | [0.448, 0.750] |
| Long Sequential: Delayed Posttest | -0.492 | 0.028 | [-0.548, -0.435] | -17.409 | 0.77 | [0.616, 0.922] |

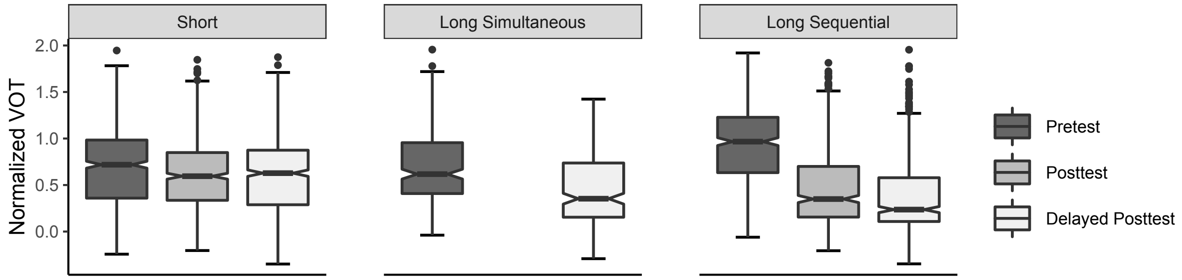

Figure 2. Normalized VOT ratios for words in isolation by session (pretest, posttest, delayed posttest), for each of the relevant methods (short, long simultaneous, long sequential).

Table 5 provides means and standard deviations for the VOTs, in milliseconds, by method, session, and phoneme. As with the words in utterances, the long sequential method produced the shortest mean VOTs for each of the voiceless stop consonants.

Table 5. Words in Isolation Mean VOT (ms)

| | | | | Delayed |
|---|---|---|---|---|
| Method | Phoneme | Pretest | Posttest | Posttest |
| Short | /p/ | 42.0 (22.0) | 36.4 (19.7) | 38.1 (21.1) |
| | /t/ | 52.5 (22.2) | 47.7 (19.8) | 46.1 (21.4) |
| | /k/ | 63.0 (21.3) | 60.6 (22.1) | 61.3 (21.9) |
| Long Simultaneous | /p/ | 39.5 (19.6) | | 29.6 (19.8) |
| | /t/ | 54.9 (22.4) | | 41.3 (22.5) |
| | /k/ | 62.5 (21.1) | | 48.2 (20.2) |
| Long Sequential | /p/ | 53.9 (21.1) | 25.8 (20.1) | 23.9 (19.6) |
| | /t/ | 69.4 (25.8) | 38.1 (23.8) | 36.3 (23.8) |
| | /k/ | 76.4 (22.1) | 56.7 (24.9) | 45.9 (23.9) |

## 5. Discussion

This study presented a comparative analysis of three different approaches to providing visual feedback. The three approaches differed across two key parameters: duration of the training and nature of each visual feedback paradigm (i.e., simultaneous or sequential). The short paradigm consisted of a single visual feedback intervention, while the long paradigms included three visual feedback interventions conducted over the course of multiple weeks. The long simultaneous paradigm addressed all three phonemes during each visual feedback intervention, and the long sequential paradigm trained only one phoneme per visual feedback intervention.

Regarding the overall effectiveness of the visual feedback paradigm (Research Question 1), the results for words in novel utterances revealed that all forms of visual feedback served to effectively improve pronunciation of L2 Spanish voiceless stops, as evidenced by a significant decrease in VOT at the posttest. These gains were maintained through the delayed posttest. The results for the words in isolation also highlight the overall benefits of the visual feedback paradigm, with reductions in VOT found following all three intervention types.

Contextualizing these results within existing theoretical frameworks, a number of authors (e.g., Tyler, 2019) have noted that some prominent models, including the PAM-L2 (Best & Tyler, 2007) and SLM (Flege, 1995) focus on the effect of the L1 on the L2 at the initial stages of acquisition and how phonetic contrasts in the L2 may (or may not) be assimilated into existing L1 phonetic categories. This may certainly be the case for VOT differences in English and Spanish, with participants producing English-like VOTs for Spanish tokens. Tyler (2019) focuses on how the principles of such models may be applied to the L2 classroom. Among several suggestions, Tyler (2019) notes that learners need "experiences that provide opportunities for them to discover phonetic differences that signal phonological contrast in the L2" (p. 607), notably at an early stage of L2 acquisition. Within the framework of the noticing hypothesis (Schmidt, 1990), it is possible that visual feedback helps to overcome limitations in L2 phonetic perception, particularly for acoustically similar phonemes (Best & Tyler, 2007), allowing learners to 'discover' previously obscured phonetic distinctions. In short, while participants may be unable to auditorily perceive differences in word-initial voiceless stops between the L1 and L2, the visual modality, particularly with an intuitively picturable feature, may facilitate the noticing and intake processes (Schmidt, 1995), which in turn improves L2 perception and production.

While effect sizes for the short treatment were small, ranging from *d* = 0.18–0.27, effect sizes for the longer treatments fell within the medium to medium-to-large range (*d* = 0.50–0.77). These effect sizes are similar to, or somewhat larger than, the average effect sizes for studies involving a technological-component, as reported by Lee et al. (2015; *mean d* = 0.53). A number of factors may have mitigated the size of the effects found here, including the focus on consonants (vs. vowels), the relatively short length of the treatment (total time < 3 hours) relative to the median treatment length in previous research (4.25 hours, Lee et al., 2015), and the largely inductive approach that involved little direct instructor input (see Lee et al., 2015; for discussion of explicit and implicit technology-enhanced pronunciation instruction, see Offerman, 2020), potentially resulting in smaller effects for visual feedback relative to other types of pronunciation instruction.

Considering the second research question, namely whether greater benefits of visual feedback may be found following longer treatments, the results showed a significant impact of treatment duration. For words in novel utterances, both the short and long simultaneous approaches resulted in similar reductions in VOT, but the participants in the long sequential approach significantly outperformed the other two groups. The words in isolation illustrated that the long simultaneous treatment led to greater improvement than the short treatment. The long sequential treatment resulted in the greatest improvement in VOT. Taken as a whole, these results are in line with previous research on both L2 instruction more broadly (Plonsky & Oswold, 2014), as well as with findings on L2 pronunciation instruction (Lee et al., 2015), that longer interventions result in greater degrees of improvement.

It should be noted, however, that it is unlikely that such a relationship is linear, with equal gains in pronunciation following each successive intervention. For example, Offerman and Olson

(2016) suggest that learners may quickly reach a stable point, not necessarily target-like, after which additional training may not result in further improvement. Within a noticing framework (Schmidt, 1990), it is possible that as learner productions approach more target-like norms, the resulting smaller differences between learner and target productions become both harder to notice and less relevant for L2 intelligibility, comprehensibility, and accentedness (Derwing, Munro & Wiebe, 1998). As such, researchers and instructors should weigh the potential benefits of additional interventions against the inherent opportunity costs, namely the opportunity to focus time, efforts, and pedagogical resources on other features of the L2 (Plonsky & Oswald, 2014).

Finally, considering the third research question, whether a simultaneous or sequential approach to training multiple phonemes from the same natural class impacts learner outcomes, the results suggest an advantage for a sequential approach. The long sequential approach led to a greater reduction in VOT for both words in utterances and words in isolation than the long simultaneous approach. The words in isolation analysis serves to complement the findings for the words in novel utterances. This parallel is relevant given the differences in procedure in the three approaches for words in utterances, namely that the same utterance stimuli were recorded at the pretest and delayed posttest in the long sequential approach. As such, the confirmatory results from the words in isolation stimuli, in which all three participant groups recorded the same stimuli in each session (pretest and delayed posttest), suggest that the added benefit from the sequential approach in the utterance condition is unlikely to be the result of stimuli repetition.

There are several possible explanations for why a sequential approach results in greater improvement than a simultaneous approach, relying on both general language learning theories, as well as phonetic theory. With regard to L2 learning, Lee and VanPatten's (1995) approach to Processing Instruction (PI), which suggests that instruction is important for overcoming inefficient or insufficient input processing, proposes a learning benefit that results from "present[ing] one thing at a time" (p. 173). This strategy lessens the processing burden for learners, allowing them to make greater gains on each individual concept. In the current study, a sequential approach may serve to lessen the processing burden and show greater pronunciation gains for each phoneme.

From a phonetic perspective, recent proposals have suggested that acquisition of L2 phonetic contrasts may not occur on a phoneme-by-phoneme basis, as is tacitly accepted by several theoretical models (PAM-L2: Best & Tyler, 2007; SLM: Flege, 1995), but rather at the subphonemic level of the feature (e.g., VOT) (e.g., de Jong et al., 2009). As such, training on a single phoneme that employs a given subphonemic feature generalizes to other phonemes with the same featural characteristics. In Olson (2019), training on a single voiceless stop consonant (e.g., /p/) resulted in significant gains for the non-trained phonemes (e.g., /t/ and /k/) that share the relevant characteristic (i.e., VOT). As such, the sequential strategy may allow focus on a single phoneme, leveraging the "present one at a time" strategy, while the featural connections lead to gains across all phonemes in the natural class. While the results here suggest an advantage for the sequential approach, such results should be confirmed through future research, and the possible underlying cognitive mechanisms for these results should be explored further.

## 6. Conclusions

The current comparative analysis demonstrates a significant and lasting impact of visual feedback on L2 pronunciation. Specifically, providing L2 learners with visual images of their own productions and those of NSs of the target language, and promoting comparison, resulted in a significant reduction in VOT. The degree of improvement was dependent on both treatment duration, with longer interventions resulting in greater reductions in VOT, and the nature of each individual treatment, with the sequential approach outperforming the simultaneous approach. Moreover, all three methodologies were conducted in an intact classroom, with no special access to laboratory equipment, and much of the pre- and post-analysis work was given to students as homework. Answering the call for ecological validity in pronunciation instruction research (Thomson & Derwing, 2015), this visual feedback approach can be easily translated into a variety of instructional contexts.

While the study begins to answer the previous call in the literature for comparative studies in instructed L2 pronunciation (Derwing & Munro, 2015; Lee et al., 2015), it has focused narrowly on different implementations of visual feedback on a single durational contrast. Future comparative work, carefully controlling for treatment length, population, and phonetic feature, should directly address different types of phonetic training and different phonetic features.

In adapting the visual feedback paradigm to other classroom settings, a few considerations ought to be made. Researchers and instructors should carefully consider whether visual feedback is appropriate for a given pronunciation feature in the L2, and what type of visual feedback (e.g., direct vs. indirect, raw vs. stylized/simplified) might be most suitable. VOT, while an ideal test case for these particular research questions, may not contribute to issues of intelligibility for this language pairing. Balancing the intelligibility principle, learner and institutional goals, and the inherent temporal constraints on the curricular design, instructors and researchers should be deliberate in choosing pronunciation features. Finally, the current studies were all conducted in a guided inductive approach, with little input from instructors. Future research may consider the role that instructors may play in maximizing the potential of the visual feedback paradigm.

## 7. References

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. S. (1998). Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores. In *Fifth International Conference on Spoken Language Processing* (pp. 1–4). http://www.isca-speech.org/archive

Anderson-Hseih, J. (1992). Using electronic visual feedback to teach suprasegmentals. *System, 20,* 51–62. https://doi.org/10.1016/0346-251X(92)90007-P

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

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/package1⁄4lme4.

Barr, D., Levy, R., Scheepers, C., & 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, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. Munro & O. S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (version 6.0.42) [computer software]. Available from www.praat.org.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310. https://doi.org/10.1121/1.418276

Carey, M. (2004). CALL Visual feedback for pronunciation of vowels: Kay Sona-Match. *CALICO Journal, 21*(3), 571–601. https://www.jstor.org/stable/24149798

Cho, T., & 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

Chun, D. (1989). Teaching tone with microcomputers. *CALICO Journal, 7*(1), 21–47. https://www.jstor.org/stable/pdf/24147465

Chun, D. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology, 2*(1), 61–77. http://llt.msu.edu/vol2num1/article4/index.html

Chun, D. (2002). *Discourse intonation in L2: From theory and research to practice*. Amsterdam: John Benjamins.

Chun, D. (2007). Come ride the wave: But where is it taking us? *CALICO Journal, 24*(2), 239–252. https://www.jstor.org/stable/pdf/24147910

de Bot, K. (1980). Evaluation of intonation acquisition: A comparison of methods. *International Journal of Psycholinguistics, 7*, 81–92.

de Bot, K. (1983). Visual feedback of intonation: Effectiveness and induced practice behavior. *Language and Speech, 26*, 331–350. https://doi.org/10.1177/002383098302600402

de Jong, K. J., Hao, Y., & 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

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: a research-based approach. TESOL Quarterly, 39, 379–397. https://doi.org/10.2307/3588486

Derwing, T. M. & Munro, M. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* Amsterdam: John Benjamins.

Derwing, T., Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Fischer, L. B. (1986). The use of audio/visual aids in the teaching and learning of French. Pine Brook, NJ: Kay Elemetrics Corporation.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-276). Baltimore: York Press.

Flege, J. E., & Wang, C. (1989). Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final /t/–/d/ contrast. *Journal of Phonetics, 17,* 299–315. https://doi.org/10.1016/S0095-4470(19)30446-2

Garcia, C., Kolat, M. & Morgan, T. (2018). Self-correction of second language pronunciation via online, real-time, visual feedback. In J. Levis (Ed.), *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching Conference* (pp. 54–65). Ames, IA: Iowa State University.

Hardison, D. (2004). Generalization of computer assisted prosody training: Quantitative and qualitative findings. *Language Learning and Technology, 8*(1), 34–52.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *Journal of the Acoustical Society of America, 138,* 817–832. https://doi.org/10.1121/1.4926561

Lambacher, S. (1999). A CALL tool for improving second language acquisition of English consonants by Japanese learners. *Computer Assisted Language Learning, 12*(2), 137–156. https://doi.org/10.1076/call.12.2.137.5722

Lee, J., Jang, J., & 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

Lee, J. F., & VanPatten, B. (1995). *Making communicative language teaching happen. Volume 1: Directions for language learning and teaching*. New York: McGraw-Hill.

Léon, P., & Martin, P. (1972). Applied linguistics and the teaching of intonation. *The Modern Language Journal, 56*(3), 139–144. https://doi.org/10.2307/324034

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. https://doi.org/10.2307/3588485

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics 27,* 184–202. https://doi.org/10.1017/S0267190508070098

Levis, J., & 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., & Abramson, A. S. (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

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

McBride, K. (2015). Which features of Spanish learner's pronunciation most impact listener evaluations? *Hispania 98*(1), 14–30. https://www.jstor.org/stable/24368849

Moholt, G. (1988). Computer assisted instruction in pronunciation for Chinese speakers of American English. *TESOL Quarterly 22*(1), 91–111. https://doi.org/10.2307/3587063

Motohashi-Saigo, M., & Hardison, D. (2009). Acquisition of L2 Japanese geminates training with waveform displays. *Language Learning & Technology, 13*(2), 29–47. http://llt.msu.edu/vol13num2/motohashisaigohardison.pdf

Munro, M. J. (2016). Pronunciation learning and teaching: What can phonetics research tell us. *In the Proceedings of the International Symposium on Applied Phonetics* (p. 26–29). DOI: 10.21437/ISAPh.2016

Munro, M., & Derwing, T. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*, 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Norris, J. M. & 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.* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Offerman, H. M. (2020). (Unpublished doctoral dissertation). Effects of pronunciation instruction on L2 learner production and perception in Spanish: A comparative analysis. Purdue University, West Lafayette, IN.

Offerman, H. M., & Olson, D. J. (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, D. J. (2014a). Phonetics and technology in the classroom: A practical approach to using speech analysis software in second-language pronunciation instruction. *Hispania*, *97*(1), 47–68. https://www.jstor.org/stable/24368745

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

Olson, D. J. (2019). Feature acquisition in second language phonetic development: Evidence from phonetic training. *Language Learning, 69*(2), 366–404. https://doi-org/101111/lang.12336

Olson, D. J. (Under review). Considering the scope of phonetic features in second language acquisition.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267-278.

Plonsky, L. & Oswald, F. (2014). How big is “big” Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi-org/10.1111/lang.12079

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

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

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

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.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–159. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R.W. Schmidt (Ed.), Attention and awareness in foreign language learning (pp. 1–63). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Spaai, G. W., & Hermes, D. J. (1993). A visual display for the teaching of intonation. *Calico Journal 10*(3), 19–30. https://www.jstor.org/stable/24147786

Sturm, J., Miyamoto, M., & Suzuki, N. (2019). Pronunciation in the L2 French classroom: Student and teacher attitudes. *The French Review, 92*(3), 60–78.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Tyler, M. (2019). PAM-L2 and phonological category acquisition in the foreign language classroom. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. Hjortshøj Sørensen (Eds.) *A sound approach to language matters – In honor of Ocke-Schwen Bohn* (pp. 607-630). Department of English, School of Communication & Culture, Aarhus University.

[1] Although some research has shown that learners aspire to native-like productions (Sturm et al., 2019), many researchers have explicitly argued against the goal of a native-like accent in favor of intelligibility (e.g., Levis, 2005, 2007; Munro & Derwing, 1995). As such, while visual feedback generally employs a NS comparison, the ultimate objective may not be native-like production.

[2] This lack of empirical methodological comparison is not limited to the subfield of visual feedback, but has remained a trend across the larger field of pronunciation instruction (e.g., Derwing & Munro, 2015).

[3] While participants received training on only a single stop consonant, results from Olson (2019) showed no significant difference in the improvement for trained and nontrained phonemes. Moreover, by-participant analysis demonstrated a strong, positive relationship between the degree of change for trained and nontrained phonemes ($R^2 = .42$, $F(1,23) = 16.63$, $p < .001$, $b = 0.941$). As such, productions for all phonemes, both trained and nontrained, are included in the current comparison.

[4] The utterance-initial position of the target tokens in Study 2, which differed from the utterance-medial position in Study 1, owes to the original study's goals. Specifically, Study 2 examined both voiceless /p, t, k/ and voiced /b, d, g/ stop consonants. The occlusive realizations of the voiced stops [b, d, g] are restricted in their distribution, largely occurring in phrase-initial position. Despite the differing positions within the utterance, there was no significant difference between the normalized VOT values at the pretest between Study 1 and Study 2 (see Results).

[5] While the overall intent of the guiding questions for self and native speaker analysis in Study 3 was similar to those used in Studies 1 and 2, the wording of some of the questions varied slightly. For example, while Studies 1 and 2 ask 'What are the visual characteristics of the native speaker's *p*?', Study 3 asks 'What is the *p* in the photo like?'. It is not anticipated that these minor differences in the guiding questions significantly impacted outcomes.

[6] Several participants failed to complete the posttest ($n = 3$) or delayed posttest ($n = 4$). As participants completed two of the three testing sessions, their data were retained.

[7] Data from the words in isolation were missing from one participant in Study 1.