# Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish

Wendy Herd,[a] Allard Jongman, and Joan Sereno
*Linguistics Department, University of Kansas, 1541 Lilac Lane, Lawrence, Kansas 66044*

(Received 3 January 2012; revised 11 February 2013; accepted 5 April 2013)

This study investigates the effectiveness of three high variability training paradigms in training 42 speakers of American English to correctly perceive and produce Spanish intervocalic /d, ɾ, r/. Since Spanish spirantization and English flapping both affect /d/ intervocalically, the acquisition of the /d/-/ɾ/ contrast proves difficult for English learners of Spanish. The acquisition of the trill /r/ is also problematic because it is a new phoneme for English learners and is articulatorily difficult to produce. Past research reported that high-variability perceptual training improves both perception and production [Bradlow *et al.*, J. Acoust. Soc. Am. **101**, 2299–2310 (1997); Wang *et al.*, J. Acoust. Soc. Am. **113**, 1033–1043 (2003)] and that production training improves both as well [Hirata, Comp. Assisted Lang. Learning **17**, 357–376 (2004)]. However, trainees were able to listen to stimuli during production training, making it unclear whether production training alone transfers to perception. This study systematically controls both training modalities so they can be directly compared and introduces a third training methodology that includes both perception and production. All three training paradigms proved effective. While perception and production trainees primarily made gains in perception, combination trainees made gains in production. The effectiveness of each training modality depended on the nature of the contrast being trained and the modality of the test.
 [http://dx.doi.org/10.1121/1.4802902]

PACS number(s): 43.71.Hw, 43.70.Kv [BRM] Pages: 4247–4255

## I. INTRODUCTION

When L1 speakers of American English acquire Spanish, they must reanalyze two sounds (i.e., /d/ and /ɾ/) in their native language and learn a new sound (i.e., /r/) to acquire a three-way /d, ɾ, r/ contrast. Although the trill /r/ does not exist as an allophone or phoneme in English, the interdental voiced fricative /ð/, alveolar voiced stop /d/, and alveolar flap /ɾ/ are familiar sounds for speakers of American English. Unlike Spanish, where the dental voiced fricative [ð] or approximant [ð̞] only exists as an allophone of /d/, /ð/ is a phoneme in English which contrasts with /d/ as seen in the minimal pair *though* [ðoʊ] - *dough* [doʊ]. On the other hand, the [ɾ] only surfaces as an allophone of /d/ (and /t/) in American English. For example, the addition of the morphological ending *-er* changes the pronunciation of *ride* [ɹaɪd] to *rider* [ɹaɪɾɚ].

Flapping, a highly productive phonological rule in English, causes /d/ (and /t/) to surface as [ɾ] in post-tonic intervocalic position. In a recent study, Boomershine *et al.* (2008) found that monolingual English speakers rated /d, ɾ/ minimal pairs as more similar than native Spanish speakers and were slower at discriminating the pairs than Spanish speakers. This suggests that American English–speaking learners of Spanish will experience perception difficulties also. Studies have reported that flapping occurs between 94% and 99% of the time in the post-tonic intervocalic position (Patterson and Connine, 2001; Connine, 2004; Zue and Laferriere, 1979; Byrd, 1994; Herd *et al.*, 2010). Since flapping occurs so frequently in an environment where Spanish spirantization (e.g., intervocalic /d/ is produced as a voiced dental fricative [ð] or approximant [ð̞] as in *codo* [koð̞o]) also occurs, American English learners may produce intervocalic /d/ as a /ɾ/ in Spanish also, both failing to spirantize /d/ correctly and producing a form that can be confused with another phoneme in Spanish.

In Spanish, /d, ɾ, r/ are separate phonemes; however, there are no minimal triplets that distinguish the three because /d/ is spirantized to [ð] or [ð̞] intervocalically, /ɾ/ does not occur word-initially, and /ɾ, r/ are in free variation word-finally. However, the contrast can still be illustrated by looking at a minimal triplet and a minimal pair. The minimal triplet *codo* [koðo] "elbow" - *coro* [koɾo] "choir" - *corro* [koro] "I run" illustrates that /ɾ, r/ contrast with each other and [ð], the allophonic variant of /d/. Likewise, the minimal pair *dato* [dato] "fact" - *rato* [rato] "time" shows the /d, r/ distinction.

As with flapping in English, Spanish spirantization, where voiced stops /b, d, g/ are spirantized to [β, ð, ɣ], is a highly productive phonological rule in Spanish, with intervocalic spirantization of /d/ occurring 99% of the time (Waltmunson, 2005). Since /ð/ contrasts with /d/ in English and since Spanish spirantization occurs in the same environment as English flapping, this difference in how /ð, d, ɾ/ are categorized in the two languages may cause difficulties for English learners of Spanish.

While little research has investigated the difficulty with which American English learners of Spanish perceive /d, ɾ, r/, production difficulties are well documented. With respect to spirantizing intervocalic /d/, students in their third or fourth semester of university-level Spanish only spirantize intervocalic /d/ 6%–25% in an environment where native

[a]Author to whom correspondence should be addressed. Present address: English Department, Mississippi State University, 100 Howell Hall, P.O. Box E, Mississippi State, MS 39762. Electronic mail: wherd@english.msstate.edu

Spanish speakers spirantize near ceiling levels (Zampini, 1993, 1994; Waltmunson, 2005). Likewise, Spanish learners at these levels produce intervocalic /ɾ/ correctly only 25%–49%, incorrectly producing it as /ɹ/ in 92% of errors (Face, 2006; Waltmunson, 2005; Rose, 2010). Finally, intermediate American English learners of Spanish produce intervocalic /ɾ/ in only about 5% of cases (Face, 2006; Rose, 2010). Although production of the /d, ɾ, ɾ/ contrast improves with increased experience, even advanced American English learners of Spanish enrolled in Spanish doctoral programs produce the intervocalic /ɾ/ correctly in only about 80% of cases, significantly less often than native Spanish speakers (Johnson, 2008; Rose, 2010).

Previous research in high variability perceptual training, which involves using a large amount of variability in speakers, contexts, and/or words in the training stimuli, has led to the development of a systematic training methodology to improve second language learners' ability to distinguish novel contrasts in the target language. Perceptual training has been shown to improve the perception of /ɹ/ and /l/ by Japanese learners of English (Logan *et al.*, 1991; Bradlow *et al.*, 1997) and the perception of tone by English learners of Chinese (Wang *et al.*, 2003). In addition to demonstrating an improved ability to perceptually distinguish the contrasts as a result of training, trainees also exhibit an improved ability to produce these distinctions immediately following training (Bradlow *et al.*, 1997; Wang *et al.*, 2003) and 3 months after training (Bradlow *et al.*, 1999). Also using high variability training, Hirata (2004) found that using visual images of pitch contours effectively trained English learners of Japanese to both produce and perceive pitch and duration contrasts. However, since participants listened to the stimuli during training, it is unclear whether their improvements were due to the production training alone or due to inadvertent perceptual training.

While these studies clearly establish the effectiveness of high variability training to train second language contrasts, it is unclear how effective they will be for training the /d/-/ɾ/ contrast in Spanish. Instead of the /ɾ/ only existing as an allophone of /d/ in English that must be teased apart to perceive the /d, ɾ/ contrast in Spanish, the /ɾ/ also occurs as the result of American English flapping in the same environment where Spanish spirantization should occur. Thus the acquisition of the Spanish /d/-/ɾ/ contrast may be further complicated by the existence of competing phonological rules in English and Spanish. The present study aims to investigate the effectiveness of training Spanish intervocalic /d/, /ɾ/, /ɾ/. Furthermore, previous research has not manipulated perceptual and production training such that the effects of the two can be compared, a gap which is addressed in the current study. The present study aims to discover the directionality of the link between perception and production and to test the effectiveness of the two training modalities in combination.

## II. METHODS

### A. Participants

The productions of nine native Spanish speakers (five male, four female) with an average age of 26, eight from Peru and one from Spain, were recorded to create pretest, training, posttest, and generalization stimuli. Forty-two native speakers of American English (9 male, 33 female) with a mean age of 20 and enrolled in an intermediate Spanish course at the University of Kansas also participated as trainees and controls. These students had completed 3 to 4 years of high school Spanish and were enrolled in their second or third semester of college Spanish. They were randomly assigned to four groups: perception trainees, production trainees, combination trainees, and controls. An additional eight native Spanish speakers from Chile (three male, five female) with an average age of 23, none of whom had traveled in an English-speaking country, participated as judges for the native speaker identification task. It is important to note that many varieties of Spanish exist; however, the phonemes /ɾ, ɾ/ and the intervocalic allophone [ð/ð̞] investigated in this study are present in all varieties of Spanish. Additionally, every attempt was made to recruit speaking participants who use the same variety of Spanish such that all pretest, training, and posttest stimuli were produced by native speakers from Peru, with the exception of one speaker from Spain whose productions were only used to test for generalization to new speakers and new words.

All participants completed a human consent form and a dialect questionnaire before completing any sessions. All participants were paid $10 per hour for their participation, and the learners of Spanish, who were required to visit the lab from 2 to 12 times depending on group, were paid an additional $20 completion bonus upon the completion of all sessions. None of the participants reported any speech or hearing disorders.

### B. Stimuli

Nine native Spanish-speaking participants [eight from Peru (four male, three female) and one (male) from Spain] read 210 minimal pairs: 70 contrasting /ɾ/ and /ɾ/ (e.g., *coro* "choir" and *corro* "I run"), 70 contrasting /d/ and /ɾ/ (e.g., *moda* "fashion" and *morra* "crown [of the head]"), and 70 contrasting /ɾ/ and /d/ (e.g., *loro* "parrot" and *lodo* "mud"). In order to develop a word list large enough to accommodate unique pretest, training, and generalization stimuli, both words and nonwords were used. For each of these contrasts, half of the minimal pairs were word–word pairs while the other half were word–nonword or nonword–word pairs.

Of the 210 minimal pairs, 60 minimal word–word pairs and 60 minimal word–nonword pairs were used to create pretests, posttests, and generalization tests, which are detailed in Sec. II C 1. A paired samples $t$ test verified that the word–word pairs used on the pretests, posttests, and generalization tests did not significantly differ from each other in word frequency as determined by the Corpus del Español (Davies, 2002) [$t(59) = 0.343$, $p = 0.733$]. Furthermore, paired samples $t$ tests confirmed that the /ɾ/-/ɾ/ pairs [$t(19) = 0.994$, $p = 0.333$], /ɾ/-/d/ pairs [$t(19) = 1.284$, $p = 0.214$], and /d/-/ɾ/ pairs [$t(19) = 0.568$, $p = 0.577$] did not differ significantly from one another in word frequency.

Thirty of the minimal pairs described above (15 word–word and 15 word–nonword) and produced by F1 (a female speaker from Peru) were used to create the perception pretest

and posttest. These 30 minimal pairs read by M1 (a male speaker from Peru) and M5 (a male speaker from Spain) were also used for the generalization to new speakers test. The same list of 30 minimal pairs was then used as the pretest and posttest production stimuli.

An additional 90 minimal pairs read in equal parts by speakers F1, M1, and M5 were used to create the generalization to new words test. A subset of these 90 minimal pairs taken equally from the lists read by F1, M1, and M5 was used as the production generalization stimuli. The remaining 90 minimal pairs read by three female (F2, F3, and F4) and three male speakers (M2, M3, and M4) were used during training sessions. All recordings took place in an anechoic chamber, using a solid-state recorder (Marantz PMD671) and Electro-Voice 767a microphone. Due to the large number of tests, minimal pairs, and speakers, the role of each speaker and the number of unique minimal pairs used on each task are summarized in Table I.

## C. Procedure

### 1. Pretest, training, and posttest designs

English-speaking participants completed perception and production tasks to evaluate their acquisition of the /d/, /ɾ/, and /r/ contrasts in Spanish. Trainees completed these tasks at least 1 day prior to training and at least 1 day following training while controls had a 2- to 3-week break between the first and second set of tasks. All participants completed the pretest tasks and the posttest tasks within a 2- to 3-week window. The tasks were presented in the same order to all participants.

The perception pretest and posttest were identical forced-choice perceptual identification tasks presented via Paradigm (Tagliaferri, 2011). The task included 30 minimal pairs read by a native Spanish speaker from Peru (F1) for a total of 60 tokens. Participants first heard an auditory stimulus that contained [ɾ], [r], or [ð] intervocalically, and then they saw two words on the computer screen, the orthographic representation of the word they heard and the other word in the minimal pair. For example, participants might hear *cara* "face" [kaɾa] and then see *cara* and *cada*. Their task was then to mouse-click the word they heard. The stimuli were presented in random order.

In addition to the perception posttest, participants completed a generalization to new speakers task and a generalization to new words task. Both generalization tasks were identical in presentation to the perception pretest and posttest. In the generalization to new speakers task, participants identified the 60 minimal pairs previously used in the pre/posttest as read by M1, a new speaker from Peru, and M5, a new speaker from Spain. In the generalization to new words task, participants identified 90 new minimal pairs read in equal parts by F1, M1, and M5. The posttest and generalization stimuli were presented as one experiment, but the stimuli were blocked by speaker, with the posttest and generalization stimuli randomized together. The speakers were always presented in the following order: F1, M1, and M5. The combined posttest and generalization identification task included three speakers reading 60 minimal pairs each for a total of 360 stimuli.

During production pretest and posttest tasks, participants read the 30 minimal pairs used in the perception pretest and posttest in a randomized list including 50 additional words as fillers. During the posttest, participants read a randomized list that included the 30 minimal pairs from pretest, an additional 30 minimal pairs taken equally from the lists read by F1, M1, and M5 for the generalization to new words task, and 100 additional words as fillers.

TABLE I. Summary of which Spanish-speaking participants and how many unique minimal pairs were presented in each task.

| Tasks | Participants | Minimal Pairs |
|---|---|---|
| Pretest (Pre) | F1 | 30[a] |
| Training | F2, F3, F4, M2, M3, M4 | 90 |
| Posttest (Post) | F1 | 30[a] |
| Generalization Tasks | | |
| New Speakers (Gen-S) | M1, M5 | 30[a] |
| New Words (Gen-W) | F1, M1, M5 | 90 |

[a]Pre, Post, and Gen-S stimuli were identical.

*a. Perception training.* One group of 10 participants (perception trainees) underwent perception training following the procedure described in Logan *et al.* (1991) and refined in Bradlow *et al.* (1997). The participants were trained using 90 minimal pairs recorded by six different speakers. During the training sessions, which lasted between 20 and 30 min, the participants completed forced choice identification tasks similar to the perception pretest and posttest. After hearing a stimulus that contained either [ɾ], [r], or [ð] over a pair of headphones, participants saw two orthographic choices on a computer screen. Participants then chose the item they heard by mouse-clicking their response. After choosing an item, participants either saw the message, "Right! That was *token*. Let's hear *token* again," or "Oops! That was *token*. Let's hear *token* again," at which point the auditory stimulus was replayed. Participants attended six training sessions during a period of 2 to 3 weeks, practicing one pair of sounds (i.e., [ɾ] vs [ð], [ɾ] vs [r], or [ð] vs [r]) read by two different speakers each day. Two sessions were spent on each contrast, and the contrasts and speakers were never repeated in consecutive sessions.

*b. Production training.* Ten participants (production trainees) underwent production training following a procedure based on that described in Hirata (2004). As was the case with perception training, production trainees were presented with 90 minimal pairs read by six different speakers. Trainees practiced one contrast per session for a total of six sessions completed within 2–3 weeks. Two sessions were spent on each contrast, and the contrasts and speakers were never repeated in consecutive sessions.

During training, participants were presented with the waveform, spectrogram, and orthographic representation of a native speaker's production of a word via Praat (Boersma and Weenink, 2011). Each participant was prompted to inspect the native speaker's production, and then to click

"continue" when ready to record a version of the word. The program would record the participant for 1.5 s, and then the participant's waveform and spectrogram would appear. The participant would next be prompted to compare the two versions of the stimulus, and then to press "continue" when ready to see a new word. Participants were instructed to attempt different pronunciations in order to match their waveforms and spectrograms to those of the native speakers and to continue using a pronunciation once the waveforms and spectrograms matched. Production trainees were never allowed to hear the native speakers' stimuli.

The first training session lasted 60–75 min, half of that time devoted to a tutorial during which the first author taught participants how to identify and distinguish [ɾ], [r], and [ð] using waveforms and spectrograms. Participants were taught that a tap [ɾ] consists of one short closure while a trill [r] consists of a series of two to ten closures. The [ð] was visually distinguished from the [ɾ] and [r] by the presence of frication or the approximation of frication instead of one complete closure. After completion of the first session, the other five production training sessions lasted 35–45 min each.

*c. Combination training.* The third group included 11 participants (combination trainees), all of whom completed both perceptual and production training. This group completed three perceptual training sessions and three production training sessions within 2–3 weeks, rotating each modality from session to session. As was the case with perception and production trainees, these participants practiced each paired contrast twice, once through perception training and once through production training. The same contrasts and speakers were never trained on consecutive days.

*d. Controls.* A fourth group of 11 Spanish learners (controls) completed the pretests and posttests but did not undergo training. These participants completed the posttests 2 to 3 weeks after the pretests.

### 2. Native-speaker identification

In order to evaluate the trainees' improvement in production from pretest to posttest, identification data were collected from eight native Spanish speakers in Chile. The purpose of the identification task was to find if native Spanish speakers could correctly identify the phoneme intended by the Spanish learner and if intelligibility and pronunciation improved as a result of training.

Native Spanish speakers were presented with the pretest and posttest productions of one minimal pair from each of the three paired contrasts read by the 42 learners of Spanish, resulting in 504 tokens (2 tests × 2 words × 3 contrasts × 42 speakers = 504 tokens). The minimal pair used to represent each contrast was randomly selected. During the identification task, native Spanish speakers heard words produced by Spanish learners, and then chose the words they thought they heard from three choices presented orthographically on the computer screen. For example, if a Spanish learner intended *moda* "fashion," the native speaker would choose from *moda*, *mora*, or *morra*.

### 3. Waveform and spectrogram inspection

In addition to a portion of stimuli being presented to native Spanish speakers for identification, all of the pretest, posttest, and generalization productions were analyzed via waveform and spectrogram. Each Spanish learner produced 30 minimal pairs at pretest, the same 30 minimal pairs at posttest, and 30 new minimal pairs as a generalization test, resulting in 180 stimuli. Each stimulus was then analyzed and scored based on visual inspection of the waveform and spectrogram using Praat (Boersma and Weenink, 2011). Stimuli received a "0" if the target Spanish phoneme was replaced by an English one, a "0.5" if the production approached the intended target, and a "1" if the intended target was pronounced correctly. A more detailed explanation of how each contrast was scored is presented below. This gradient scoring scale was designed to capture the improvement of Spanish learners who produce the trained contrasts in a more target-like manner without reaching native-like pronunciation. These scores were then used to conduct the statistical analyses that follow.

If Spanish learners intended to produce a /ɾ/, they received a "1" if the waveform contained one brief and complete closure, which was defined as a less than 50 ms absence of $F_1$, $F_2$, and $F_3$ formants, a lack of a release burst following the closure, and a decrease in intensity in the corresponding waveform. A combination of a Spanish [ɾ] and an American English [ɹ] resulted in a "0.5." The addition of the American English [ɹ] was identified based on a steep decline in $F_3$ preceding the closure or a steep incline in $F_3$ following the closure accompanied by near steady-state $F_1$ and $F_2$ and by a periodic shift in the corresponding waveform. Substituting a Spanish [r] for a /ɾ/ was also scored as "0.5." The Spanish [r] was defined as a rapid succession of two or more [ɾ] closures. The use of an American English [ɹ] was scored "0." The American English [ɹ] was identified based on an absence of closure, a dip in $F_3$, and near steady-state $F_1$ and $F_2$. This scoring system reflects that producing a combination of an American English [ɹ] and a tap [ɾ] (i.e., [ɾɹ]) is better than producing an American English [ɹ] without a closure and that substituting a [r] for a /ɾ/ is a "native-like" error.

Likewise, when an intended trill /r/ consisted of two or more complete occlusions, it received a score of "1." Replacing the /r/ with a [ɾ], an error occasionally reported in the speech of native Spanish speakers, resulted in a "0.5." However, producing the Spanish phoneme /r/ as the English [ɹ] was scored "0." If an intended intervocalic /d/ was produced as a voiced dental fricative [ð] or a voiced dental approximant [ð̞], the production was scored as "1." Voiced dental fricatives were defined as constrictions lacking formant structure and accompanied by the noisy waveforms typical of a fricative. Voiced dental approximants were defined as minimal constrictions that contained steady state $F_1$, $F_2$, and $F_3$ formants accompanied by periodic waveforms. If the Spanish learner produced /d/ as a voiced alveolar or dental

TABLE II. Summary of statistical results from ANCOVAs for the effect of Group on mean adjusted RAU scores for posttest (Post), generalization to new speakers (Gen-S), generalization to new words (Gen-W) and native speaker identifications (NSIDs).

| Test | Contrast | Perc M[a] (SE) | Prod M[a] (SE) | Combo M[a] (SE) | Controls M[a] (SE) | $F$ | $p$ | $\eta_P^2$ | *post hoc* (Bonferroni) |
|---|---|---|---|---|---|---|---|---|---|
| Perception Results | | | | | | | | | |
| Post | Overall | 88 (1.3) | 86 (1.3) | 85 (1.2) | 81 (1.2) | 5.543 | 0.003 | 0.33 | Perc > Controls[c]; Prod > Controls[b] |
| | /ɾ/-/r/ | 97 (4.0) | 103 (4.0) | 94 (3.9) | 84 (3.9) | 3.953 | 0.016 | 0.26 | Prod > Controls[b] |
| | /d/-/ɾ/ | 71 (2.5) | 63 (2.5) | 62 (2.8) | 62 (2.6) | 3.198 | 0.036 | 0.22 | Perc > Controls[a] |
| Gen-S | Overall | 87 (1.9) | 89 (1.8) | 85 (1.8) | 82 (1.8) | 2.975 | 0.045 | 0.21 | Prod > Controls[b] |
| | /ɾ/-/r/ | 90 (3.9) | 103 (3.9) | 88 (3.8) | 80 (3.8) | 6.416 | 0.001 | 0.36 | Prod > Controls[c] |
| Gen-W | /ɾ/-/r/ | 91 (2.8) | 92 (2.8) | 85 (2.7) | 78 (2.7) | 4.970 | 0.006 | 0.31 | Perc > Controls[b]; Prod > Controls[b] |
| Production Results | | | | | | | | | |
| NSIDs | Overall | 77 (4.9) | 84 (4.9) | 84 (4.7) | 65 (4.7) | 3.503 | 0.025 | 0.22 | Combo > Controls[b]; Prod > Controls[a] |
| | /r/ | 75 (9.7) | 79 (9.8) | 75 (9.3) | 39 (9.3) | 3.924 | 0.016 | 0.24 | Prod > Controls[b]; Perc, Combo > Controls[a] |
| Post | Overall | 53 (2.9) | 52 (3.1) | 63 (2.9) | 47 (2.7) | 5.613 | 0.003 | 0.33 | Combo > Controls[c] |
| | /r/ | 57 (3.5) | 59 (3.7) | 80 (3.7) | 51 (3.8) | 3.685 | 0.021 | 0.25 | Combo > Controls[b] |
| Gen | Overall | 56 (2.5) | 56 (2.7) | 64 (2.5) | 52 (2.3) | 4.592 | 0.008 | 0.29 | Combo > Controls[c] |
| | /r/ | 57 (6.6) | 66 (6.9) | 80 (6.9) | 56 (6.0) | 2.792 | 0.055 | 0.20 | n.s. |

[a] $p < 0.075$.
[b] $p < 0.05$.
[c] $p < 0.01$.

stop [d], identified as a complete closure followed by a burst, it was scored "0.5." This reflects that pronouncing *moda* "style" as [moda] would sound very unnatural but that [d] cannot be confused with any other Spanish phonemes. On the other hand, producing the intervocalic /d/ as a tap [ɾ] resulted in a "0" because it involves replacing /d/ with another phoneme in Spanish and producing a different word.

In order to evaluate the consistency of the scoring criteria, a 5% subset of the stimuli (378 tokens) randomly selected to equally represent participants, contrasts, and tests was scored by a second coder. The correlation between the two sets of measurements was high (Pearson's $r = 0.890$, $p < 0.0001$). This high degree of inter-coder reliability suggests the scoring criteria were applied consistently. In both the original scoring of the entire dataset and the rescoring of the subset, coders were naive as to participants' group assignments and whether the stimuli were collected before or after training.

### D. Data analysis

All pretest, posttest, and generalization accuracy scores were converted to rationalized arcsine units (RAUs) using the method detailed in Studebaker (1985). This conversion allows accuracy scores to be compared on a linear and additive scale that ranges from −23 RAUs to 123 RAUs. Posttest and generalization RAUs were then submitted to Analyses of Covariance (ANCOVAs) with Group (perception trainees, production trainees, combination trainees, and controls) as a between-subjects fixed factor and corresponding pretest RAUs as a covariate. The pretest covariates were significant in all cases at the $p < 0.02$ level. Bonferroni *post hoc* tests were conducted to further analyze the relationships between groups when a main effect of Group was found. Only significant and marginally significant effects are reported. Based on partial eta-squared ($\eta_P^2$) values, significant and marginally significant effect sizes ranged from medium (0.19–0.22) to large (0.26 and above). Figures illustrate the original pretest, posttest, and generalization RAU scores rather than the covariate-adjusted means while Table II summarizes the adjusted mean RAU scores for reported effects and Table III presents raw mean pretest and posttest data.

## III. RESULTS

### A. Speech perception

Figure 1 illustrates the mean pretest and posttest RAUs for each contrast organized by Group. Four separate ANCOVAs were conducted to determine whether Group differences between mean adjusted posttest RAUs on overall accuracy, /d/ vs /ɾ/ accuracy, /ɾ/ vs /r/ accuracy, and /d/ vs /r/ accuracy reached significance. Corresponding pretest RAUs were used as covariates. Groups differed significantly in overall accuracy [$F(3,37) = 5.543$, $p = 0.003$], /d/ vs /ɾ/ accuracy [$F(3,37) = 3.198$, $p = 0.036$], and /ɾ/ vs /r/ accuracy [$F(3,37) = 3.953$, $p = 0.016$]. Bonferroni *post hoc* comparisons revealed perception trainees ($p = 0.003$) and production trainees ($p = 0.038$) performed significantly better overall than controls. With respect to perceiving specific contrasts, perception trainees outperformed controls at a near-significant level ($p = 0.074$) on the /d/ vs /ɾ/ contrast, and production trainees outperformed controls at a significant level ($p = 0.012$) on the /ɾ/ vs /r/ contrast.

To measure whether gains from training generalized to new speakers (Gen-S) and new words (Gen-W), a series of ANCOVAs were performed on overall accuracy, /d/ vs /ɾ/ accuracy, /ɾ/ vs /r/ accuracy, and /d/ vs /r/ accuracy in the Gen-S and Gen-W conditions. Corresponding pretest RAUs were used as covariates. Figure 2 illustrates the relationships between pretest and generalization tests for the different contrasts organized by Group. Adjusted mean accuracy scores across the three contrasts reached significance for the Gen-S condition [$F(3,37) = 2.975$, $p = 0.045$]. Bonferroni *post hoc* results revealed that production trainees outperformed

TABLE III. Summary of raw data from pretest (Pre), posttest (Post), generalization to new speakers (Gen-S), and generalization to new words (Gen-W).

| Test | Contrast | Perc M (SE) | Prod M (SE) | Combo M (SE) | Controls M (SE) |
|---|---|---|---|---|---|
| Perception Results | | | | | |
| Pre | Overall | 81 (3.0) | 83 (1.4) | 84 (1.3) | 82 (1.9) |
| | /ɾ/-/r/ | 80 (5.3) | 82 (3.6) | 80 (3.3) | 83 (4.3) |
| | /d/-/ɾ/ | 67 (3.7) | 72 (2.2) | 74 (2.0) | 65 (2.9) |
| | /d/-/r/ | 95 (2.0) | 96 (1.0) | 97 (1.0) | 97 (1.0) |
| Post | Overall | 85 (1.8) | 85 (1.2) | 85 (1.0) | 80 (1.7) |
| | /ɾ/-/r/ | 90 (2.4) | 93 (2.0) | 90 (2.1) | 82 (4.8) |
| | /d/-/ɾ/ | 70 (3.9) | 65 (2.1) | 66 (3.2) | 61 (2.0) |
| | /d/-/r/ | 96 (1.6) | 97 (1.1) | 99 (0.7) | 96 (1.4) |
| Gen-S | Overall | 84 (1.8) | 87 (1.6) | 84 (1.0) | 81 (2.0) |
| | /ɾ/-/r/ | 86 (2.5) | 94 (1.5) | 84 (2.4) | 79 (4.7) |
| | /d/-/ɾ/ | 72 (2.9) | 72 (2.7) | 71 (2.8) | 67 (2.3) |
| | /d/-/r/ | 82 (1.9) | 83 (1.4) | 82 (2.1) | 80 (1.6) |
| Gen-W | Overall | 84 (1.5) | 86 (0.9) | 86 (1.1) | 82 (1.6) |
| | /ɾ/-/r/ | 86 (2.7) | 87 (1.5) | 82 (1.9) | 78 (4.2) |
| | /d/-/ɾ/ | 69 (3.1) | 74 (1.7) | 76 (1.5) | 72 (2.7) |
| | /d/-/r/ | 98 (0.8) | 97 (0.7) | 96 (2.8) | 96 (1.3) |
| Native Speaker Identification Results | | | | | |
| Pre | Overall | 66 (4.3) | 65 (5.4) | 73 (3.9) | 72 (2.8) |
| | /ɾ/ | 73 (8.7) | 81 (6.1) | 81 (6.8) | 74 (6.5) |
| | /r/ | 43 (11.2) | 39 (7.9) | 56 (10.7) | 55 (6.8) |
| | /d/ | 82 (5.8) | 75 (9.4) | 81 (5.4) | 89 (4.8) |
| Post | Overall | 73 (4.7) | 84 (4.9) | 83 (4.4) | 67 (4.3) |
| | /ɾ/ | 79 (6.1) | 80 (6.1) | 88 (8.4) | 76 (7.4) |
| | /r/ | 67 (8.5) | 68 (9.6) | 76 (10.1) | 45 (10.2) |
| | /d/ | 72 (9.0) | 85 (6.4) | 85 (7.5) | 82 (7.3) |
| Production Results | | | | | |
| Pre | Overall | 41 (5.5) | 51 (3.3) | 51 (4.1) | 41 (7.6) |
| | /ɾ/ | 42 (8.8) | 45 (5.3) | 49 (6.6) | 32 (8.5) |
| | /r/ | 39 (6.7) | 54 (4.6) | 60 (6.2) | 41 (9.6) |
| | /d/ | 41 (4.5) | 52 (6.8) | 44 (5.7) | 49 (7.2) |
| Post | Overall | 48 (7.1) | 60 (5.3) | 66 (2.9) | 42 (7.4) |
| | /ɾ/ | 49 (9.8) | 56 (8.4) | 61 (4.6) | 32 (7.8) |
| | /r/ | 47 (8.7) | 67 (8.1) | 80 (5.5) | 44 (9.8) |
| | /d/ | 48 (5.6) | 55 (8.9) | 58 (6.9) | 49 (7.9) |
| Gen | Overall | 49 (7.9) | 63 (4.5) | 69 (3.4) | 47 (7.9) |
| | /ɾ/ | 54 (11.2) | 63 (8.5) | 66 (4.1) | 39 (8.8) |
| | /r/ | 48 (7.7) | 72 (7.8) | 83 (6.0) | 48 (9.8) |
| | /d/ | 47 (8.3) | 54 (10.9) | 57 (8.6) | 51 (8.6) |

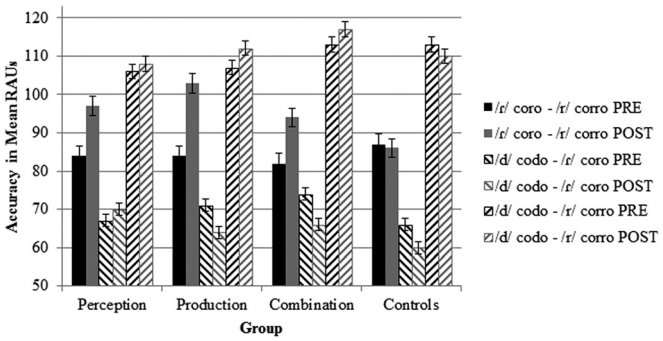


FIG. 1. Mean identification accuracy in RAUs from pretest to posttest for each contrast organized by Group. Error bars indicate the standard error of the mean.

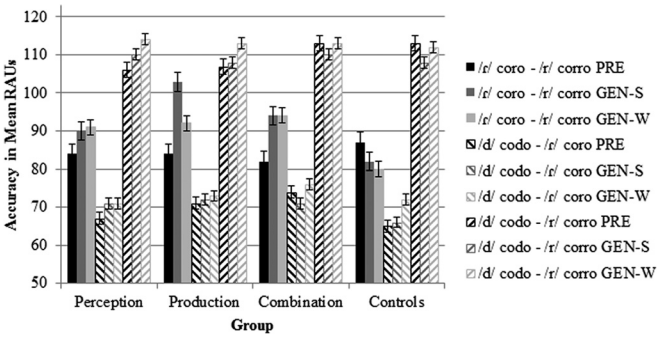


FIG. 2. Mean identification responses in RAUs from pretest to generalization tests (new speakers–GEN-S and new words–GEN-W) for each contrast organized by Group. Error bars indicate the standard error of the mean.

controls overall in the Gen-S condition ($p = 0.041$). Main effects of Group were also obtained for the /ɾ/ vs /r/ contrast in the Gen-S condition [$F(3,37) = 6.416$, $p = 0.001$] and the Gen-W condition [$F(3,37) = 4.970$, $p = 0.006$]. Bonferroni *post hoc* results indicated that production trainees significantly outperformed controls in the Gen-S ($p = 0.001$) and Gen-W ($p = 0.016$) conditions. Perception trainees also significantly outperformed controls in the perception of the /ɾ/ vs /r/ in the Gen-W condition ($p = 0.011$).

## B. Speech production

### 1. Native speaker identification

Figure 3 shows the mean pretest and posttest scores for the three trained contrasts (/ɾ/, /r/, and /d/) organized by Group. Four separate ANCOVAs were conducted on overall native speaker identification RAUs, /ɾ/ identification RAUs, /r/ identification RAUs, and /d/ identification RAUs with corresponding pretest RAUs as covariates and Group as a between-subjects fixed factor. The main effect of Group reached significance overall [$F(3,37) = 3.503$, $p = 0.025$] and for identification of the /r/ [$F(3,37) = 3.924$, $p = 0.016$]. Bonferroni *post hoc* comparisons found that all training groups performed (near-)significantly better than controls. Combination trainees performed significantly better overall ($p = 0.048$) and marginally better on /r/ identification ($p = 0.058$) than controls. Production trainees performed marginally better overall ($p = 0.055$) and significantly better on /r/ identification ($p = 0.035$) than controls. Perception

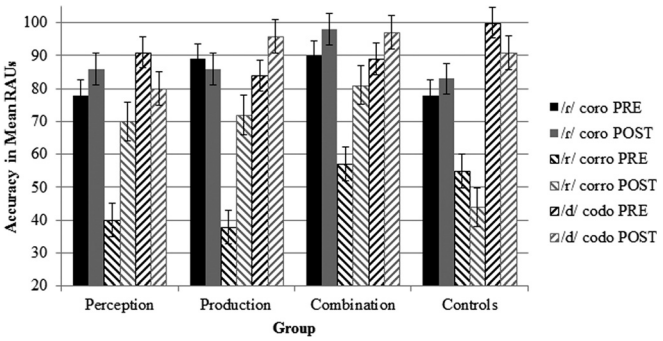


FIG. 3. Pretest and posttest RAUs of native speaker identification scores for each contrast organized by Group. Error bars indicate the standard error of the mean.

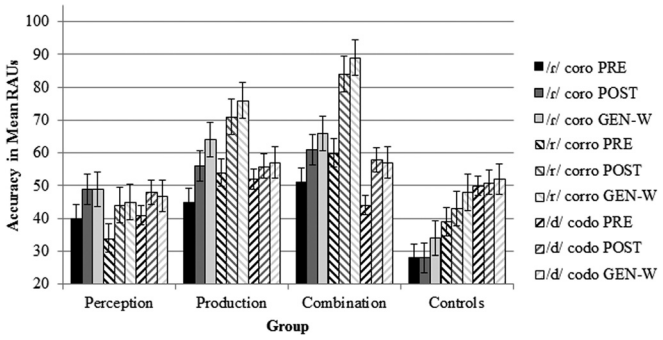


FIG. 4. Waveform and spectrogram inspection scores in RAUs from pretest, posttest and generalization test for each contrast organized by Group. Error bars indicate the standard error of the mean.

trainees also performed marginally better than controls on /ɾ/ identification ($p = 0.064$).

### 2. Waveform and spectrogram inspection

Figure 4 illustrates the mean pretest, posttest, and generalization test RAUs for each contrast organized by Group. Four separate ANCOVAs were conducted to determine whether Group differences between mean adjusted posttest RAUs on overall accuracy, /ɾ/ accuracy, /r/ accuracy, and /d/ accuracy reached significance. Corresponding pretest RAUs were used as covariates. Groups differed significantly in overall accuracy [$F(3,37) = 5.613$, $p = 0.003$] and /r/ accuracy [$F(3,37) = 3.685$, $p = 0.021$]. Bonferroni *post hoc* comparisons revealed combination trainees performed significantly better overall ($p = 0.002$) and in /r/ accuracy ($p = 0.018$) than controls.

To measure whether gains from training transferred to new stimuli, four additional ANCOVAs were performed on overall accuracy, /ɾ/ accuracy, /r/ accuracy, and /d/ accuracy in the generalization condition. Corresponding pretest RAUs were used as covariates. Adjusted mean accuracy scores across the three contrasts reached significance overall [$F(3,37) = 4.592$, $p = 0.008$] and neared significance for /r/ accuracy [$F(3,37) = 2.792$, $p = 0.055$]. Bonferroni *post hoc* results revealed combination trainees outperformed controls overall ($p = 0.008$), but no differences between groups were found for /r/ accuracy.

## C. Results summary

Due to the large number of comparisons, the significant results have been summarized in Table II, which includes adjusted mean posttest or generalization test RAUs ($M^a$), standard error (SE), $F$ statistic values, $p$ values, partial eta-squared ($\eta_p^2$) values as a measure of effect size, and results of Bonferroni *post hoc* comparisons. Raw mean (M) pretest, posttest, and generalization scores with standard error (SE) have also been included in Table III to aid in the interpretation of the transformed data used in figures and statistical comparisons.

## IV. DISCUSSION

Based on posttest and generalization data, all three training paradigms resulted in significantly higher adjusted mean posttest and/or generalization RAU scores than those of the controls. This offers strong support for the efficacy of training nonnative contrasts in general. Specifically, perception and production trainees performed better in the overall perception of the three contrasts, with perception trainees performing better on the /d/-/ɾ/ contrast and production trainees performing better on the /ɾ/-/r/ contrast. In addition, production trainees generalized these improved abilities to new speakers. Likewise, both perception and production trainees generalized the improved perception of the /ɾ/-/r/ contrast to new words.

All three training groups also performed better on measures of post-production accuracy than controls. Native speaker identifications indicated that the intelligibility of combination and production trainees improved overall and that perception, production, and combination trainees all improved in the intelligibility of the /r/. According to inspection of waveforms and spectrograms, combination trainees outperformed controls overall and in /r/ accuracy with the overall improvement transferring to new words. It is notable that all three training groups performed better than controls on posttest production of the /r/, a sound that causes American English learners continued difficulty even at advanced stages of Spanish acquisition (Face, 2006; Johnson, 2008; Rose, 2010).

One aim of this study was to evaluate the effectiveness of perception and production training, and, more specifically, to determine which training type transferred most effectively to the other modality. Inspection of Table II evinces that both perception and production training transferred to opposite modalities. Moreover, the improvement of the group exhibiting transfer (i.e., production trainees' performance on the perception task and perception trainees' performance on the production task) was comparable to gains made by the group trained specifically in that modality (i.e., perception trainees' performance on the perception task and production trainees' performance on the production task). Perception trainees and production trainees making equivalent gains in both modalities suggests both training types transfer equally well to the other modality.

It is also interesting to look at the specific contrasts where training was most effective. While perception trainees exhibited improvement in the perception of the /d/-/ɾ/ contrast, production trainees improved in the /ɾ/-/r/ contrast. This suggests how well perception or production transfers depends on the relationship between the sounds being trained. The /d/-/ɾ/ contrast exhibited the lowest average pretest perception scores, so one could have predicted that it would show the most improved posttest scores when comparing training groups to controls; however, that was not the case. Instead, more improved posttest scores were recorded for the /ɾ/-/r/ contrast across all training modalities. This suggests the /d/-/ɾ/ contrast differs from the /ɾ/-/r/ contrast.

To distinguish the /ɾ/-/r/ contrast, American English speakers only need to acquire the /r/ as a new phonemic category, because the /ɾ/ already exists as part of the allophonic inventory of English. However, when distinguishing the /d/-/ɾ/ contrast, American English speakers must acquire the /ð/ as an allophone of /d/ and reassign the /ɾ/, an allophone

of /d/ in American English, to a separate phonemic category. It appears perception training is the more effective training paradigm for teasing apart two allophonic variants of the same phoneme while production is more effective for learning a new contrast. Alternately, production training may have only improved the perception of the /ɾ/-/r/ contrast because the difference between the /ɾ/ and /r/ is such a visually salient distinction when looking at waveforms and spectrograms whereas the corresponding distinction between the /d/ and /ɾ/ is more subtle. In short, the contrast being trained determines which training method is most effective.

It is also possible that differences in improvement arose due to differences in the type of feedback trainees received. Perception trainees were provided explicit feedback in the form of correction from the computer while production trainees received self-reflective feedback when they compared their productions to those of native speakers. It would be helpful in future research to devise a method of giving production trainees explicit corrective feedback during production training, allowing more direct comparisons between perception and production training.

Another aim of this study was to investigate the effectiveness of perception and production training in combination. With respect to perception accuracy, combination trainees made no gains. This lack of improvement may be due to the number of perception training sessions available to combination trainees. In order to hold the total number of training sessions constant across groups, all training groups participated in six training sessions, meaning combination trainees only participated in three perception sessions and three production sessions. This suggests that more than three perception training sessions are necessary to obtain gains in perception and, since combination trainees' gains in production did not transfer to perception, that more than three production training sessions are necessary to transfer gains to another modality.

However, compared to the other training paradigms, combination trainees showed the largest number of gains in production accuracy. In spite of participating in only three production training sessions, combination trainees exhibited more improvement in production accuracy than production trainees who participated in six such sessions. Table II illustrates that perception and production trainees' improvement was largely limited to the perception domain while combination trainees' improvement was in the production domain. This pattern provides evidence that training in both perception and production most effectively improves production, further suggesting that perception and production training in combination are necessary in order to evince production gains.

## V. CONCLUSIONS

This study investigated whether native speakers of American English could be trained to perceive and produce the three-way /ɾ, r, d/ contrast in Spanish. The perception, production, and generalization results strongly indicate that all three training types (perception, production, and combination) improve trainees' ability to perceive and/or produce contrasts in the L2. This study also sought to tease apart the effects of perceptual and production training with respect to which modality transfers more effectively to the other and to evaluate which training paradigm (i.e., perception, production, or combination) proved most effective. Perception and production training proved most effective for training perception while combination training, which notably included only half of the exposure to each modality, proved most effective for training production. Rather than determining whether perception or production training transferred more effectively to the other modality, it was determined that both training types transferred equally well and that the type of contrast being trained determined which training type was most effective. Perception training more effectively trained the perception of the /d/-/ɾ/ contrast, production training more effectively trained the perception of the /ɾ/-/r/ contrast, and the two training types resulted in similar gains in the production of the /r/. The findings of this study, the first to systematically control and compare the modality of training, suggest that, while all three training types resulted in trainees performing significantly better than controls, the effectiveness of training type ultimately depended on the type of contrast being trained and the modality in which trainees were tested.

## ACKNOWLEDGMENTS

This research was part of the first author's dissertation, supervised by Allard Jongman and Joan Sereno and supported by the National Science Foundation (#0843653). We are grateful to Marcela Quintana-Lara for assisting with data collection and to Stephen Politzer-Ahles for assisting with statistical analyses.

Boersma, P., and Weenink, D. (**2011**). "Praat: doing phonetics by computer [Computer program]," Version 5.3.03, from http://www.praat.org/ (Last viewed 12/6/2011).

Boomershine, A., Hall, K. C., Hume, E., and Johnson, K. (**2008**). "The impact of allophony versus contrast on speech perception," in *Contrasts in Phonology: Theory, Perception, Acquisition*, edited by P. Avery, B. Dresher, and K. Rice (de Gruyter, Berlin), pp. 143–172.

Bradlow, A., Akahane-Yamada, R., Pisoni, D., and Tohkura, Y. (**1999**). "Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production," Percept. Psychophys. **61**, 977–985.

Bradlow, A., Pisoni, D., Akahane-Yamada, R., and Tohkura, Y. (**1997**). "Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production," J. Acoust. Soc. Am. **101**, 2299–2310.

Byrd, D. (**1994**). "Relations of sex and dialect to reduction," Speech Commun. **15**, 39–54.

Connine, C. (**2004**). "It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition," Psychonomic Bull. Rev. **11**, 1084–1089.

Davies, M. (**2002**). "Corpus del Español [Online Corpus]," http://www.corpusdelespanol.org/ (Last viewed 12/6/2011).

Face, T. L. (**2006**). "Intervocalic rhotic pronunciation by adult learners of Spanish as a second language," in *Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, edited by C. Klee and T. Face (Cascadilla Proceedings Project, Somerville, MA), pp. 47–58.

Herd, W., Jongman, A., and Sereno, J. (**2010**). "An acoustic and perceptual analysis of /t/ and /d/ flaps in American English," J. Phonetics **38**, 504–516.

Hirata, Y. (**2004**). "Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts," Comp. Assisted Lang. Learning **17**, 357–376.
Johnson, K. E. (**2008**). *Second Language Acquisition of the Spanish Multiple Vibrant Consonant*, Ph.D. thesis, University of Arizona, pp. 63–113.
Logan, J., Lively, S., and Pisoni, D. (**1991**). "Training Japanese listeners to identify English /r/ and /l/: A first report," J. Acoust. Soc. Am. **89**, 874–886.
Patterson, D., and Connine, C. M. (**2001**). "Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production," Phonetica **58**, 254–275.
Rose, M. (**2010**). "Intervocalic tap and trill production in the acquisition of Spanish as a second language," Stud. Hisp. Lusophone Ling. **3**, 379–419.
Studebaker, G. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Tagliaferri, B. (**2011**). "Paradigm: Perception research systems [Computer Program]," http://www.paradigmexperiments.com/ (Last viewed 12/6/2011).
Waltmunson, J. (**2005**). *The Relative Difficulty of L2 Spanish /d,t/, Trill, and Tap by L1 English Speakers: Auditory and Acoustic Methods of Defining Pronunciation Accuracy*, Ph.D. thesis, University of Washington, pp. 92–241.
Wang, Y., Jongman, A., and Sereno, J. (**2003**). "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training," J. Acoust. Soc. Am. **113**, 1033–1043.
Zampini, M. (**1993**). *Spanish Voiced Stop Phonemes and Spirantization: A Study in Second Language Acquisition*, Ph.D. thesis, Georgetown University, pp. 165–249.
Zampini, M. (**1994**). "The role of native language transfer and task formality in the acquisition of Spanish spirantization," Hispania **77**, 470–481.
Zue, V. W. and Laferriere, M. (**1979**). "Acoustic study of medial /t,d/ in American English," J. Acoust. Soc. Am. **66**, 1039–1050.