THE EFFECTS OF PHONETIC TRAINING AND VISUAL FEEDBACK ON NOVEL CONTRAST PRODUCTION

Susan Lin[1], Margaret Cychosz[1], Alice Shen[1], & Emily Cibelli[2]

[1]University of California, Berkeley, [2]Ingenuity
susanlin@berkeley.edu

ABSTRACT

This study tested the effectiveness of different techniques to train the production of a novel consonant contrast. Native English speaker participants completed a syllable repetition task highlighting dental vs. retroflex consonants, followed by training via: 1) static, midsagittal diagrammatic visualization of the contrast, 2) online ultrasound sagittal imaging of the participant's tongue with the midsagittal visualization, or 3) online ultrasound imaging of the participant's tongue with midsagittal lingual ultrasound video of a native speaker's production.

A second repetition task followed training, and ultrasound and acoustic data collected during the two repetition tasks suggest that static training diagrams and live lingual ultrasound feedback are effective aids in the early stages of adult novel L2 contrast learning. However, there appear to be limitations to their combined effectiveness, as participants who were exposed to both training methods behaved similarly to control participants who received no training.

Keywords: L2 phonology, phonetic training, ultrasound feedback

1. INTRODUCTION

Phonological acquisition is a recurrent challenge in second language (L2) learning and instruction [5, 8, 11, 16, 25]; learners require both grammatical and phonological skills to become proficient in a new language. Explicit phonetic instruction, in which learners are taught the articulatory mechanics for producing new sounds, receives surprisingly little attention in foreign language pedagogy research [1, 6, 21]. In L2 speech research, there is a similar lack of focus on the mechanisms that drive phoneme acquisition, whether for novel or experienced contrasts. Inquiries tend to compare the predictive capacities of theories of non-native acquisition to one another [4, 12], or manipulate speech stimuli to train learners on a novel contrast/phoneme [17, 20].

Here we focus on the mechanics behind L2 speech acquisition, a vital component of a complete understanding of non-native acquisition. Does explicit phonetic training – via static diagrammatic visualizations of the articulators and/or ultrasound visual feedback – aid in the early formation of contrast categories? If so, which training method provides novice learners the strongest foundation?

In addressing these research questions, this work brings together two lines of research that have remained distinct: the applicability of learning paradigms such as implicit and explicit phonetic training to non-native speech acquisition [9, 14, 15, 26], and the effectiveness of live ultrasound feedback for speech mediation [3, 10, 22].

Prior to using ultrasound, the primary methods to relay articulatory information during L2 instruction were electropalatography (EPG) [13], a costly and time-consuming method, and electroglottography (EGG) [18], which has limited applicability. Our work adds to the growing number of studies examining the effectiveness of ultrasound feedback for L2 contrast learning [7, 27], especially the incorporation of online ultrasound feedback [1, 6, 19]. In these studies, learners receive live visual feedback of their lingual movement and positioning, information that would typically be unavailable to both learner and instructor. Learners exposed to such visualization often demonstrate increased articulatory and perceptual discrimination of a non-native contrast compared to learners who do not. Here, we expand on previous work by contrasting ultrasound feedback training with diagram-based phonetic training for the acquisition of a novel dental-retroflex stop contrast.

2. METHOD

2.1. Subjects and experimental groups

Forty-nine adult native speakers of American English (13M, 36F) who reported no speech or hearing disorders participated in this study. They were pre-screened to filter out participants with experience speaking any language with a dental vs. retroflex contrast.

Every participant took part in two identical repetition tasks, a pretest and a posttest, during which acoustic and articulatory data were collected. In both repetition tasks, the participant heard a single stimulus, then repeated it to the best of their ability. Stimuli were recordings of /CV/ and /VCV/ syllables produced by a native Marathi speaker, where C was

one of the four consonants [d̪, t̪, ɖ, ʈ], and V was one of the three vowels [i, a, u] (both vowels in any given /VCV/ stimulus were the same quality). Participants heard each stimulus 3 times per repetition task.

Between the pretest and posttest, participants engaged in a ten-minute training phase that depended on the condition to which they were randomly assigned during intake. Participants were never given any explicit verbal feedback from the researchers on the quality of their productions.

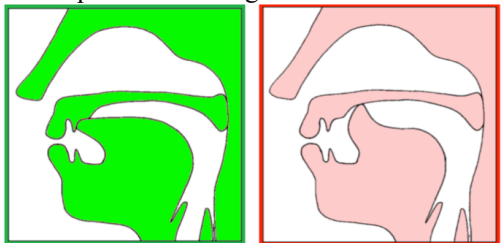
- Control (CTL): Researcher engaged in casual conversation with the participant.
- Phonetic training only (PT): Participant received a short lesson on the distinction between retroflex and dental articulation, with the support of mid-sagittal diagrams, shown in Figure 1, and were given time to practice.
- Phonetic training with visual feedback (PTVF): Participant received the same lesson as PT participants and had access to ultrasound imaging of their tongue while practicing.
- Visual feedback only (VF): Participants had access to mid-sagittal ultrasound videos of the model Marathi speaker producing retroflex and dental stops and live ultrasound imaging of their own tongue as visual aids while practicing.

2.2. Data Collection & Processing

Participants were seated in a sound-attenuated booth, before a computer screen. They were instructed to wait to hear the target stimulus, which was presented over speakers located inside the booth. After presentation of the stimulus item, participants were instructed to wait until the screen indicated that it was their turn to speak, and then to repeat the stimulus they just heard. In all four trial conditions, an experimenter joined the participant in the booth after pretest to administer the training phase.

Ultrasound data were gathered during the pretest and posttest phases, on an Ultrasonix SonixTablet with a frame rate of 113 fps. The ultrasound transducer was fixed to participants' heads with an ultrasound stabilization headset [24]. The acquisition computer and the ultrasound engine were located outside the booth, to reduce the extent that mechanical noise would interfere with the recordings. The ultrasound probe was passed through a small acoustical foam filled hole in the side of the booth. Audio from participants' speech was combined with hardware synchronization pulses generated by the ultrasound engine as a 2-channel wav file.

Figure 1. Mid-sagittal diagram of dental (left) and retroflex (right) articulations used as visual aids in phonetic training.



2.3. Ultrasound Analysis

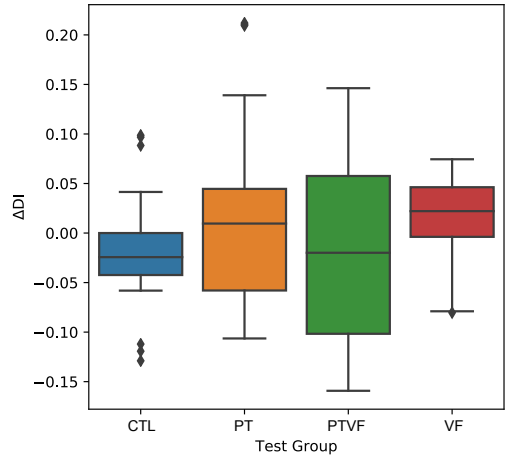
In this study, we were interested in learning what form of training would give learners the greatest "improvement" in novel contrast articulation. However, it's unclear precisely what constitutes an *improved* articulation of a novel contrast. For instance, a subject may learn a distinction between the stimulus items but be unable to work out precisely what that distinction is. In doing so, they may make unintended articulatory differentiation. In this study, we chose to focus on whether participants articulated any distinction at all (instead of whether those articulations were "correct") and further whether those articulations resulted in acoustic differences.

To this end, we calculated the change in discriminability between articulations that followed dental vs. retroflex prompts separately for both pre- and posttest trials. In this analysis, the acquisitions were separated by vowel quality due to coarticulatory influences from vowels overwhelming the influence of place of articulation on tongue shape. For each trial from each participant, the frame immediately preceding the release of the target stop was identified. Improvement from the pretest to posttest phase was evaluated as a positive change in discriminability between the set of retroflex attempts and the set of dental attempts from the pretest to posttest phase.

We first used principal component analysis (PCA) to reduce the dimensionality of the raw ultrasound data, using the Python scikit-kit-learn library [23]. A PC model was fit to each speaker's full set of stop constriction frames, and the loadings of the 20 PCs accounting for the most variance were calculated for each frame. Then, to determine discriminability, a linear discriminant model (LDA) was trained on each speaker's PCA-reduced frames, labelled for target place of articulation, separated by vowel quality and test phase.

These models were then used to predict place of articulation within the same set of data, and a discriminability index was calculated as the proportion of trials in which the predicted place of articulation matched the target place of articulation. Finally, the change in discriminability index (ΔDI) was calculated for each participant by subtracting their pretest DI scores from their posttest DI scores, representing the magnitude of "improvement" from pre- to posttest. Thus, a participant with a positive

Figure 2. Articulatory ΔDI between pretest and posttest phases, for the four test groups.



ΔDI produced dental and retroflex consonants more differently after training, and a participant with a negative ΔDI produced dental and retroflex consonants more similarly to each other after training.

2.4. Acoustic Analysis

In addition to analysing the ultrasound data, we also performed acoustic analyses on each subject's repetitions during the pre- and posttest trials. We suspected that our participants would make changes to their productions that would not be reflected in the ultrasound imaging. In particular, we thought it plausible that participants might perceive differences in the closure duration or VOT of the target consonants, whether or not those differences were real, and act on those distinctions rather than the targeted lingual contrast.

Towards this goal, we measured the duration of closure (for intervocalic consonants) and VOT, as well as the first three formants (F1/F2/F3) at 0/10/20ms following release (for all stops) and 0/10/20ms before closure (for intervocalic stops), and the relative amplitude of the stop release. Participants for whom either acoustic or ultrasound data could not be analysed were not included in the analysis (N=7). Table 1 lists the number of participants in each test condition whose data were included in the analysis.

3. RESULTS & DISCUSSION

3.1. Articulatory Results

Fig. 2 shows the overall ΔDI between pretest and posttest for the four test groups. We performed a linear mixed effects model, with ΔDI as the dependent factor, Test Group as the independent factor, and Vowel Quality as a random factor [2]. We found that PT participants, who received only explicit phonetic training, and VF participants, who received only visual feedback training, improved their articulatory discriminability after training, compared to CTL participants (Group PT: β=0.035, t=2.001, p=0.0476; Group VF: β=0.033, t=1.927, p=0.0564). Releveling the Group factor showed no significant difference in ΔDI between the PT and VF Groups (β=-0.002, t=-0.118, p=0.9065), suggesting that the two forms of guided articulatory instruction improved participants' ability to perceive and produce novel articulatory contrasts equally well.

Table 1: Participants included in analysis.

	No phonetic training	Phonetic training
No visual feedback	CTL: N=11	PT: N=11
Visual feedback	VF: N=12	PTVF: N=8

Surprisingly, PTVF participants, who received *both* explicit phonetic training and visual feedback during practice, showed no significant improvement over CTL participants (β=0.000, t=-0.017, p=0.9866). We note qualitatively, however, a great deal of between-subject variability in the ΔDI in the PT participants and especially in PTVF participants' data compared to the other three test groups. A Hartigans' dip test for unimodality is suggestive but not conclusive that PTVF participants' ΔDI values may be multimodal (D=0.094, p=0.0712), suggesting that the paired training may have been very effective for some PTVF participants, but not at all for others.

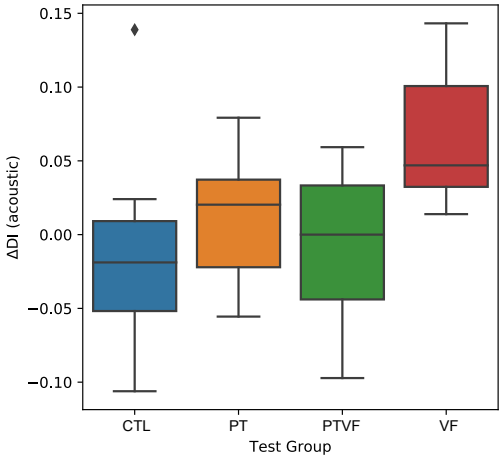
3.2. Acoustic Results

We performed a similar LDA on the 21 acoustic measures as we did on the (PCA-reduced) articulatory data: for each participant, the 21 acoustic measures taken from their pretest and posttest repetition tasks were used to train two separate LDA models, which in turn were used to classify productions as being dental or retroflex. ΔDI was calculated as the change between pretest to posttest discriminability.

A one-way ANOVA showed a significant effect of Group on acoustic ΔDI ($F(3,38)$=5.31, p=0.0037). As shown in Fig. 3, VF participants improved in their acoustic discrimination the most (Tukey post-hoc test: CTL vs. VF, p=0.0042; PT vs. VF, p=0.0841; PTVF vs. VF, p=0.0224). A separate t-test also showed that CTL participants' ΔDI values were not significantly less than zero.

To determine which acoustic measurements contributed most to the positive change in acoustic ΔDI, we gathered the posttest LDA coefficients for all 21 acoustic measures and ran a linear regression between the posttest coefficients and acoustic ΔDI. While this analysis revealed no significant relationship between the coefficients for individual acoustic measures and the propensity for positive ΔDI (r^2=-0.47, p=0.9848), we found noteworthy patterns in the effect sizes for *groups* of acoustic

Figure 3. Acoustic ΔDI from pretest to posttest phases, for the four test groups.



measurements: the combined measurements of release amplitude and pre-consonantal formant frequencies were more strongly correlated with improvement in ΔDI than post-consonantal formant frequencies, VOT, or closure duration. We hope to further scrutinize these relationships in future work.

3.3. Discussion

In sum, participants in the PT and PTVF Groups demonstrated similar articulatory ΔDI, but VF participants also demonstrated higher acoustic ΔDI. These results suggest that while both forms of articulatory training were beneficial to participants in improving their ability to *articulate* the distinction between retroflex and dental stops, visual feedback training alone was superior to phonetic instruction alone in terms of training a meaningful *acoustic* contrast. It is possible that this may be due to VF participants having access to the model Marathi talker's voice during their training, as it was embedded in the videos. This additional exposure to the model productions may have helped participants beyond the benefits of visual feedback alone.

Supposing though that this difference between the PT and VF Groups is due to the differences in training, we interpret these findings as suggesting that visual feedback training may be superior to diagram-based training in the acquisition of non-native lingual contrasts. This may be due to participants being better able to associate the visual feedback with their own auditory and somatosensory feedback during speech than the static midsagittal diagrams.

One curiosity is the behavior of PTVF participants. If PT participants in the PT and the VF Groups have high ΔDI values, it follows that PTVF participants, who received *both* forms of training, should have ΔDI values that are just as high, if not higher. Instead, these participants do not behave significantly differently from CTL participants. While we have no definitive answers, we speculate that the training PTVF participants received may have been "too much too fast;" they received effectively twice as much training as PT and VF participants without twice as much time for processing the new information. It is also plausible that PTVF participants found it difficult to connect the live ultrasound images to the static articulatory diagrams due to differences in the visual displays.

4. CONCLUSION

This work contrasted the effectiveness of different learning techniques to train native English speakers in the production of a novel consonant contrast. Improvement in production was operationalized as the change in discriminability between categories along both articulatory and acoustic dimensions. Results suggest that static and dynamic articulatory training via explicit phonetic instruction and ultrasound visual feedback can improve the short-term production of a novel contrast. Of these two techniques, visual feedback appears most effective. Still, our results suggest that the combined effectiveness of explicit phonetic instruction and ultrasound feedback may be limited – participants who received both types of training did not outperform the control group.

Importantly, our findings make no claims regarding *retention* of training effects; in the current study, each participant completed the task once, without follow-up. Nor can they determine if native speakers would judge participants' productions as more native-like following training. Further, while many challenges in L2 phonology are attributed to difficulty articulating new sounds and phonotactic combinations, language learners must also master perceptual differences between sounds in their L2. Given that our participants' ability to *produce* a contrast between sounds improved, we can perhaps infer that their ability to *perceive* the contrast also improved. However, perceptual outcomes of articulatory training should be studied explicitly.

Nevertheless, the results can still inform research and pedagogical practices. Articulatory instruction techniques appear to provide a solid foundation for novel contrast learning. This finding is all the more relevant given both the challenging nature of learning an L2 phonology as well as the fact that explicit articulatory training in the L2 classroom is uncommon [6, 21]. Future studies should continue to investigate the benefits of different training techniques. Important next steps are to study other contrasts and to measure learning outcomes in populations who are exposed to L2 in different environments, such as children or immigrants.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge Rachel Arsenault and Ben Papadopoulos for their assistance in recruiting and running participants, and Ronald Sprouse for his technical expertise. We are also indebted to members of the UC Berkeley PhonLab for their valuable feedback throughout this project.

6. REFERENCES

[1] Antolík, T. K., Pillot-Loiseau, C., & Kamiyama, T. 2019. The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training. *J. Second Lang. Pronunciation*, *5*(1), 72-97.

[2] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67(1), 1–48.

[3] Bernhardt, B., Gick, B., Bacsfalvi, P., Adler-Bock, M. 2005. Ultrasound in speech therapy with adolescents and adults. *Clin. Linguist. Phon.* 19(6–7), 605–617.

[4] Best, C. T. 1995. A Direct Realist View of Cross-Language Speech Perception. In Strange, W. (ed) *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York, 171–204.

[5] Best, C. T., McRoberts, G. W., Goodell, E. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *J. Acoust. Soc. Am.* 109(2), 775–794.

[6] Bliss, H., Abel, J., Gick, B. 2018. Computer-assisted visual articulation feedback in L2 pronunciation instruction. *J. Second Lang. Pronunciation*, 4(1), 129–153.

[7] Bliss, H., Bird, S., Cooper, P. A., Burton, S., & Gick, B. 2018. Seeing Speech: Ultrasound-based Multimedia Resources for Pronunciation Learning in Indigenous Languages, *Lang. Documentation and Conservation*, 12, 315-338.

[8] Briere, E. J. 1966. An Investigation of Phonological Interference. *Language* 42(4), 768-796.

[9] Cibelli, E. 2015. *Aspects of articulatory and perceptual learning in novel phoneme acquisition.* PhD Thesis, University of California, Berkeley.

[10] Cleland, J., Scobbie, J., Naki, S., Wrench, A. 2015. Helping children learn non-native articulations: the implications for ultrasound-based clinical intervention. In *Proc. of the 18th ICPhS* Glasgow, Scotland, 1-5.

[11] Flege, J. E. 1987. A critical period for learning to pronounce foreign languages? *Appl. Linguist.* 8(2), 162–177.

[12] Flege, J. E., Munro, M. J., MacKay, I. R. 1995. Factors affecting strength of perceived foreign accent in a second language. *J. Acoust. Soc. Am.* 97(5), 3125–3134.

[13] Gick, B., Bacsfalvi, P., Bernhardt, B. M., Oh, S., Stolar, S., Wilson, I. 2008. A motor differentiation model for liquid substitutions in children's speech: English /r/ variants in normal and disordered acquisition. In *Proc. Mtgs. Acoust.* 1(060003), 1–9.

[14] Goudbeek, M., Cutler, A., Smits, R. 2008. Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Commun.* 50(2), 109–125.

[15] Gulian, M., Escudero, P., Boersma, P. 2007. Supervision hampers distributional learning of vowel contrasts. *Proc. of the 16th ICPhS* Saarbrücken, Germany, 1893–1896.

[16] Ioup, G. 2008. Exploring the role of age in the acquisition of a second language phonology. *Phonol. Second Lang. Acquis.* 36, 41–62.

[17] Iverson P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C. 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87(1), B47–B57.

[18] Kitzing, P. 1990. Clinical applications of electroglottography. *J. Voice* 4(3), 238–249.

[19] Li, J. J., Ayala, S., Shiller, D., McAllister, T. m.s. in preparation. Do individual differences predict learning outcomes in biofeedback training?

[20] Lively S. E., Logan, J. S., Pisoni, D. B. 1993. Training Japanese listeners to identify English /r /and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.* 94(3), 1242–1255.

[21] Lord, G. 2005. (How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania* 88(3), 557–567.

[22] McAllister Byun, T., Swartz, M. T., Halpin, P. F., Szeredi, D., Maas, E. 2016. Direction of attentional focus in biofeedback treatment for /r/ misarticulation. *Int. J. Lang. Commun. Disord.* 51(4), 384–401.

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., *et al.* 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

[24] Scobbie, J., Wrench, A., van der Linden, M. 2008. Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proc. of the 8th Intl. Seminar on Speech Prod.*, 373–376.

[25] Tahta, S., Wood, M., Loewenthal, K. 1981. Foreign accents: Factors relating to transfer of accent from the first language to a second language. *Lang. Speech*, 24(3), 265–272.

[26] Vlahou, E. L., Protopapas, A., Seitz, A. R. 2012. Implicit training of nonnative speech stimuli. *J. Exp. Psychol. Gen*. 141(2), 363–381.

[27] Wilson, I., Gick, B., O'Brien, M., Shea, C., Archibald, J. 2006. Ultrasound technology and second language acquisition research. In *Proc. of the 8th Generative Approaches to Second Lang. Acq. Conference* Somerville, MA, 148–152.