Assessing L2 vowel production gains after high-variability phonetic training: acoustic measurements vs. perceptual judgements

Joan C. Mora¹

¹Universitat de Barcelona, Spain

Email: mora@ub.edu

Abstract

The current study examined vowel production data from a high-variability phonetic training project (HVPT) focusing on the English vowel contrast /æ/-/a/ to compare the sensitivity of various acoustic measures of vowel overlap and distinctiveness (Pillai scores and Euclidean and Mahalanobis distances) in capturing L2 learners’ training gains in (a) qualitatively distinguishing between /æ/ and /a/ in production, and (b) reducing qualitative differences with respect to native speakers’ productions of /æ/ and /a/. A subset of these data was then perceptually judged by a group of native listeners for pronunciation accuracy.

The results showed that all acoustic distance measures consistently captured phonetic training gains in measures of contrast distinctiveness and nativelikeness, correlating strongly with one another, and predicting perceptual distance measures. However, measures also differed substantially in effect sizes for training gains (pre- vs- post-test distance scores), training group differences (lexical vs. non-lexical training) and vowel (/æ/ vs. /a/). Acoustic distance measures were found to predict perceptual distance scores based on native English listeners’ ratings of pronunciation accuracy. These findings suggest that when assessing gains in the pronunciation of L2 vowel contrasts it is important to measure both acoustic distances between vowels and acoustic distances between learners’ and native speakers’ productions.

Keywords: HVPT, Pillai scores, Euclidean distance, Mahalanobis distance, L2 vowel production.

1. Introduction

Pronunciation proficiency may be assessed in terms of global and specific constructs of L2 speech [1]. Global constructs are perceptual dimensions of L2 speech that can be assessed subjectively through listeners’ ratings, such as comprehensibility, intelligibility, fluency and accentedness [2, 3]. Such perceptual judgments have been shown to depend on pronunciation and phonological features, but may also be affected by linguistic, task and listener factors, such as the occurrence of grammatical errors, whether speaking tasks elicit spontaneous or controlled speech samples, or the listeners' familiarity with the accent they are asked to judge [4, 5]. Specific constructs like segmental accuracy, lexical stress, intonation, or breakdown fluency, on the other hand, can be operationalized through subjective or objective measures. Subjective ratings of specific features of speech, like judging how native-like a given consonant or vowel is in an identification task [6, 7], may also be influenced by contextual, listener, and task factors, but mainly depend on the acoustic dimensions that characterize the speech samples. Objective measures typically involve acoustic analysis of phonetic and phonological properties of speech like pitch, or measurements of phonologically contrastive acoustic dimensions, such as voice onset time (VOT) or vowel formant frequency and duration [8]. Interestingly, in a meta-analysis of 77 studies [1] found that the effectiveness of pronunciation instruction depended to a certain extent on the type of assessment method employed. In general, studies that assessed pronunciation instruction through acoustic measures, those that focused on specific speech dimensions, and those that used controlled tasks, found pronunciation instruction to be more effective than those that had used perceptual judgments, had had a global focus or had elicited spontaneous speech. This suggests that both speech elicitation tasks and assessment methods need to be carefully considered when researching L2 pronunciation development.

Assessing L2 pronunciation development, however, does not only imply choosing among several methodological choices of target dimensions, stimuli, elicitation tasks and measures (see [9] for an overview). In instructed SLA, where target populations may be relatively homogeneous, measurement choices also entail considering how to best capture pronunciation gains over time in target learner groups of a specific L2 proficiency, learning history, L2 experience and pronunciation proficiency level. In fact, learning contexts and learner
populations may determine to a large extent how much pronunciation can improve over time. For example, long-term immersion in the country where the L2 is spoken would not affect L2 pronunciation development in the same way as classroom instruction in a foreign language learning context. Similarly, the amount of pronunciation improvement one may expect from pedagogical interventions in classrooms likely differs from pronunciation gains learners might obtain from individual phonetic training on specific pronunciation targets in a lab context. In addition, one may expect large inter-subject variability in L2 pronunciation development irrespective of learning context and training condition. From a research perspective, assessing pronunciation gains after pronunciation-focused pedagogic interventions or after phonetic training sessions remains a methodological challenge, especially as regards the sensitivity of the measures to capture pronunciation gains.

In pronunciation-focused communicative interventions both perceptual judgments [10, 11] and acoustic measures [6, 12] have been used. In phonetic training, gains in production have also been assessed perceptually, often through identification and goodness ratings scores [13], acoustically, or both at the same time [14]. All measures, however, present strengths and weaknesses. For example, when using perceptual measures of identification accuracy and goodness ratings, it is difficult to tell what exactly listeners are paying attention to, and it may be difficult for listeners to focus on a single segment while hearing a whole word or sentence. In addition, if gains are small in magnitude, they may be difficult to detect perceptually. However, while it may be easier to detect changes over time when measuring speech sounds acoustically than perceptually, acoustic measures may be said to be less ecologically valid than native listeners' perceptual judgements and more time-consuming. Performing acoustic analyses would also entail taking decisions about which measures, which baseline data and which normalization procedures to use.

Acoustic measures of vowel pronunciation improvement typically involve measuring formant frequencies and then computing pronunciation accuracy scores based on the qualitative distance between contrastive vowels, i.e., how much more distinct they have become through training [15], a measure of vowel distinctiveness. They may also involve computing the distance between learners' vowel productions and those of native speakers, i.e., how much learners' vowels approximate the quality of native speakers' vowels [16], a measure of vowel nativelikeness. Such distances have been assessed in previous research through Pillai scores [17, 18, 19], a measure of vowel overlap or merger between two vowel distributions, Euclidean distances [20, 21], a measure of acoustic distance (e.g., in Hz, Bark, or Mel) between contrastive or nonnative-native vowel categories on a two-dimensional vowel space, or Mahalanobis distances, the number of standard deviations between a point (a vowel token) and the centre of a vowel distribution [15, 16]. We are unaware of any studies comparing these three measures of acoustic distance in assessing pronunciation accuracy gains after phonetic training. In addition, the few existing studies that have used acoustic measures of vowel quality distance to predict listeners' judgments of pronunciation accuracy have obtained mixed results [22, 23].

In the current study we focus on L2 pronunciation assessment through acoustic measures of vowel quality distinctiveness and nativelikeness after 4 sessions of HVPT on the English vowel contrast /æ/-/ʌ/ and compare three acoustic measures (Pillai scores, Euclidean distances and Mahalanobis distances) in their sensitivity to capture phonetic training gains and to predict native listeners' judgements of pronunciation accuracy.

2. The HVPT study

HVPT is a training paradigm that trains L2 learners in the perception and production of difficult L2 sound contrasts by exposing them to the target sounds as produced by multiple talkers in a variety of phonetic environments. It typically uses identification and discrimination tasks based on minimal-pair syllables or words to train specific L2 phonological contrasts. Research has shown that HVPT is effective at improving the perception and production of difficult L2 phonological contrasts irrespective of learners' L1 and L2 proficiency level (see [24, 25] for recent overviews of methods and findings). Training gains are often shown to generalize to new lexical items and speakers [14, 26, 27, 28], and across modalities, so that perception gains transfer to production and vice-versa [29, 30, 31, 32, 33].

In the current study we implemented a short 4-session comprehensive HVPT paradigm that included AX discrimination, identification and immediate repetition tasks (in this order) in every training session to train the perception and production of the English vowel contrast /æ/-/ʌ/ (see [34] for details). L1-Spanish/Catalan advanced EFL learners (N=105) were randomly assigned to either a lexical training condition (based on real English EFL learners, N=54) or a non-lexical training condition (based on English CVC nonwords). A group of untrained participants served as controls. Training gains in production were tested through a delayed word repetition task (DWR)
administered before (pre-test, T1) and after (post-test, T2) training.

2.1. DWR elicitation task

The DWR task contained 76 CVC test items (38 /æ/ + 38 /a/), including trained and untrained words and nonwords always produced by novel untrained voices. Participants were asked to repeat each word as accurately as they could after hearing a beep sound that was presented with 1.5-second delay. They had 2 seconds to repeat each item. Each trial was presented twice (once by a male and once by a female voice) and repeated twice. Learners’ productions were recorded through Shure SM-58 voice microphones onto Marantz PMD-661 mkii solid-state digital recorders (44.1kHz) in a quiet lab. The second repetition of each trial from each testing time (76 trials x 2 testing times x 105 participants = 15,960 vowel tokens) was selected for acoustic analysis.

2.2. Acoustic measurements

We obtained acoustic measures of vowel frequency (f0, F1, F2) in Praat from a 20ms window centered at the steady-state period of the second formant of the learners’ CVC productions from the DWR test items. Normalization was accomplished by converting frequency measures to a Bark scale and then computing a Bark-distance metric, such that the difference of subtracting Bark-converted f0 from F1 (B1-B0) was used as an estimate of vowel height and the difference of subtracting Bark-converted F1 from F2 (B2-B1) was used as an estimate of degree of vowel frontness [20, 35].

Training gains on vowel production were gauged by three different acoustic distance measures between learners (L) and native speakers (NS) as well as between the /æ/ and /a/ vowel productions: Pillai scores, Euclidean distances and Mahalanobis distances.

Pillai scores are obtained through multivariate analysis of variance (MANOVA) and measure the degree of overlap between two vowel clusters or distributions on a 0-1 scale, where 1 represents no overlap and 0 complete overlap [17, 23], producing a single measure of overlap between two vowel distributions per speaker. We therefore expected training to result in higher Pillai scores between the contrasting vowels (/æ/-/a/ distance) and in lower Pillai scores between learners and native speakers (L-NS distance).

Euclidean distances are used to measure the distance between two vowels, the distance between two average points representing two vowel clusters (two distribution means), or the distance between one vowel and the mean of a distribution representing a vowel cluster [14, 20, 36, 37]. This measure allows to compute distances by item (in addition to by speaker), but when distances between distribution means are computed, they don’t take into account the shape of the distribution. For this measure we expected training to result in smaller L-NS Euclidean distances and in larger /æ/-/a/ Euclidean distances.

Mahalanobis distances compute the distance in standard deviations between a point and the mean (or centroid) of a distribution. Unlike the Euclidean distance, the Mahalanobis distance is a unitless, scale-invariant measure that takes into account the shape of the distribution, reflecting token variability [15, 16, 38]. Distances can be computed by item. We expected training to result in larger /æ/-/a/ Mahalanobis distances and in smaller L-NS Mahalanobis distances.

2.3. Perceptual assessment measures

Ten native English listeners performed a pronunciation accuracy rating task administered in Praat on a sub-sample of 22 learners (out of the 105 learners of English in this study). A selection of 12 minimal pairs from the DWR task that each learner performed (back-buck, bad-bud, bag-bug, cap-cap, cat-cut, fan-fun, hat-hut, lack-luck, mad-mud, match-much, pan-pun, sack-suck) were normalized for peak and mean amplitude and included in the rating task. Listeners were asked to first identify the word they had heard (e.g., fan) by clicking on one of 6 labelled options appearing in orthographic form on the computer screen (e.g., back, buck, fan, fun, match, much) and then they were asked to rate it on a 9-point scale for how well they thought it matched their own pronunciation (1=very bad match; 9=very good match). Native listeners’ responses were coded to obtain a 1-to-18 perceptual distance score between learners’ productions and listeners’ phonological representations of these productions, such that 1 would represent a perfect match response (i.e., 9) for a correctly identified word and 18 a perfect match response for an incorrectly identified word. For example, when auditorily presenting fan, correctly identifying fan and selecting 9 on the rating scale would yield a perceptual L-NS distance score of 1, whereas wrongly identifying fun and selecting 9 on the rating scale would yield a perceptual L-NS distance score of 18.

Perceptual distance scores (1-18), therefore, represent a native listeners’ perceptual measure of L-NS distances. We explored the association between these perceptual distance scores and the L-NS and /æ/-/a/ acoustic distance scores we obtained through the Pillai, Euclidean and Mahalanobis measures.

2.4. Research questions

Using the DWR production data and the acoustic distance measures described above, the present study
aimed at answering the following research questions (RQ):

RQ1: Do Pillai, Euclidean and Mahalanobis distance measures between learners and native speakers (nativelikeness) and between the contrastive vowels (distinctiveness) capture training and group effects differentially?

RQ2: Are L-NS acoustic distances (nativelikeness) and /æ/-/ɐ/ acoustic distances (distinctiveness) related to one another?

RQ3: Do Pillai, Euclidean and Mahalanobis distance measures differentially predict native English listeners' judgments of pronunciation accuracy?

2.5. Results

The amount of overlap between the vowel clusters of /æ/ and /ɐ/ from the DWR task, as expected, was larger in learners than in native speakers. The degree of overlap between learners’ distributions of /æ/ and /ɐ/ was generally large, indicating learners' difficulty in qualitatively distinguishing /æ/ from /ɐ/ in production. Changes in the amount of overlap between distributions resulting from training (differences between the distributions before (T1) and after (T2) training) also varied greatly among learners. Figure 1 provides a representative example of an individual learner with much larger overlap between vowel clusters before training than after training (top panel), another learner whose vowel productions hardly change between testing times (middle panel), and the averaged vowel productions of the native speakers that provided the stimuli for the DWR task (bottom panel). As expected, there was also much more dispersion in the vowel clusters from learners than in those from native speakers.

2.5.1. Acoustic distance measures (nativelikeness)

The acoustic distances between learners’ (L) vowel productions and those of native speakers (NS) were found to shorten between testing times as a result of training, whereas those of the control group did not, suggesting that HVPT was overall effective at improving learners' ability to distinguish the target vowels in production. We present the pre-test (T1) and post-test (T2) L-NS distances (Pillai, Euclidean, Mahalanobis) separately by Vowel and by Training Group, as we were expecting learners to be more target-like in their production of /æ/ (perceptually closer and acoustically more similar to Spanish and Catalan /a/; [39]) than in their production of /ɐ/, and we were also expecting non-lexical training (nonwords) to be more effective than lexical training (words), as training learners through nonwords would promote a focus on phonetic-level information rather than meaning [40] while avoiding the activation of words that could be misrepresented phonologically with an L1 accent in the learners' L2 lexicon [34].

Figure 1: Productions of /æ/ and /ɐ/ from the DWR task by individual learners (top and middle panel) and native speakers (bottom panel) according to Time (T1, T2).

Figure 2 below plots the Pillai L-NS distance scores (degree of overlap between the distributions of /æ/ and /ɐ/). These results indicate that the lexical (word) training group was overall more target-like than the non-lexical (nonword) training group and that, as expected, /æ/ productions were more target-like than /ɐ/ productions. Training effects can only be observed for the learner group trained on nonwords.
In order to test for training and group effects, the Pillai overlap scores were submitted, separately by vowel, to a linear mixed-effects model (in SPSS 25) with Test (T1, T2), Group (Nonwords, Words) and the Test x Group interaction as fixed effects, and Subject as a random intercept. For /æ/, these analyses revealed significant main effects of Test and Group, as well as a significant Test x Group interaction (see Table 1 below). The interaction arose because the learners in the nonword training group were significantly less target-like in their productions of /æ/ at T1 (t(15955)=2.44, p=.015) but not at T2 (t(15955)=1.52, p=.127) than those in the word training group and, crucially, because it was only for the nonword training group (/æ/: t(15955)=18.15, p<.001), but not for the word training group (/æ/: t(15955)=-.769 p=.442), that the training significantly shortened L-NS distances. That is, the nonword training group significantly increased the overlap between L and NS distributions, which was larger (lower Pillai scores) at post-test than at pre-test. For /ʊ/ there was a main effect of Test, but neither the main effect of Group, nor the Test x Group interaction reached significance (see Table 1). These results suggest that /ʊ/, which was less target-like than /æ/ at pre-test, improved irrespective of training condition, whereas /æ/ only improved for learners trained with nonwords, possibly because they were less target-like at pre-test than the word training group and this allowed them to have more room for improvement.

The pattern of results for the L-NS Euclidean distances is similar to that of the Pillai scores in that the nonword training group is less target-like than the word training group, and in that /æ/ is more target-like than /ʊ/. In general differences between testing times are more clearly observable (see Figure 3) with Euclidean distances than with Pillai scores.

We tested for training and group effects by submitting the Euclidean distances to a linear mixed-effects model with Test (T1, T2), Group (Nonwords, Words), Vowel (/æ/, /ʊ/) and their interactions as fixed effects, with random intercepts for subject and item. The main effects of Test, Group and Vowel all reached significance (see Table 1). The Group x Test and the Group x Vowel interactions reached significance because although both training groups significantly reduced Euclidean distances after training (nonword: t(15949)=6.64, p<.001; word: t(15949)=2.60, p=.009), the nonword training group was significantly less target-like at pre-test (t(15949)=2.80, p=.005) than the word training group, and because overall, the nonword training group was significantly less target-like than the word training group on /ʊ/ (t(15949)=2.72, p=.007), but not on /æ/ (t(15949)=1.95 p=.051). Interestingly, the only significant interaction involving Test (Group x Test) could not be attributed to differential training effects on the nonword and word training groups, suggesting that unlike Pillai scores, Euclidean distances captured training gains more generally (for both training groups and for both vowels).

Finally, the Mahalanobis distance metric reflects the same effects of training group and vowel on training gains as the Pillai scores and the Euclidean distances do (see Figure 4), but this measure appears to enhance the effect of vowel on the L-NS acoustic distance.
Training and group effects were tested through a linear mixed-effects model with Test (T1, T2), Group (Nonwords, Words), Vowel (/æ/, /ʌ/) and their interactions as fixed effects (including subject and item as random intercepts). This model yielded significant main effects of Test and Vowel and a significant Group x Vowel interaction that arose because it was only for /ʌ/ ($t(15952)=2.49$, $p=.013$), but not for /æ/ ($t(15952)=1.24$, $p=.216$), that the nonword training group performed less target-like than the word training group. None of the other main effects or interactions reached significance (see Table 1 below). Therefore, the analysis based on Mahalanobis distances captured general main effects of Time and Vowel irrespective of training group differences.

To sum up, all three measures (Pillai, Euclidean, Mahalanobis) capture L-NS T1-T2 and training group differences similarly, but they do so in a slightly different way. Group effects were small and not all measures captured group differences. For example, the training group effect did not reach significance on the Mahalanobis measure, nor did any of the interactions involving testing time, unlike the Pillai scores and Euclidean distances, where Time effects were found to depend on either Group or Vowel. As Pillai scores cannot be computed on individual vowel tokens and Euclidean distances do not take into account token variability, we follow previous research [15, 16] in favoring measures of L-NS vowel production accuracy based on Mahalanobis distances. However, in answering RQ1, we could not observe very substantial differences between distance measures, except that the Mahalanobis distance appears to be more sensitive in capturing vowel changes over time in L-NS distance independently of vowel and training group.

2.5.2. Acoustic distance measures (distinctiveness)

We next explored potential measure-dependent differences in acoustic distances between the target vowels /æ/ and /ʌ/. As shown in Figure 5, the three /æ/-/ʌ/ acoustic distance measures differ substantially in how sensitive they are to differences between testing times (T1 vs. T2) and training groups (Nonwords vs. Words). Euclidean distances are the least sensitive, whereas Pillai and Mahalanobis appear to capture training and group effects similarly.

### Table 1: Tests of fixed effects for L-NS acoustic distances ($df= 1, 15956$; asterisks indicate significance at the $α=0.05$ level).

<table>
<thead>
<tr>
<th></th>
<th>Pillai</th>
<th>Euclidean</th>
<th>Mahalanobis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test /æ</td>
<td>155.77 *</td>
<td>40.06 *</td>
<td>25.06 *</td>
</tr>
<tr>
<td></td>
<td>/ʌ</td>
<td>17.77 *</td>
<td>5.11 *</td>
</tr>
<tr>
<td>Group /æ</td>
<td>3.94 .047*</td>
<td>2.91 .088</td>
<td>3.60 .058</td>
</tr>
<tr>
<td></td>
<td>/ʌ</td>
<td>2.35 .125</td>
<td>8.07 .005*</td>
</tr>
<tr>
<td>Group * Test</td>
<td>183.69 *</td>
<td>172.60 *</td>
<td>712.90 *</td>
</tr>
<tr>
<td>/æ</td>
<td>2.42 .120</td>
<td>2.14 .230</td>
<td>11.33 .001*</td>
</tr>
<tr>
<td>/ʌ</td>
<td>2.42 .120</td>
<td>5.47 .019*</td>
<td>0.03 .860</td>
</tr>
</tbody>
</table>

*Figure 4: Mahalanobis distance scores as a function of training Group (Nonwords, Words), Test (T1, T2) and Vowel (/æ/, /ʌ/). Error bars: 95% CI.*

*Figure 5: Distance scores between /æ/ and /ʌ/ according to training Group (Nonwords, Words) and Test (T1=light, T2=dark). Error bars: 95% CI.*
These analyses (Table 2) revealed main effects of Test for all three measures, suggesting that phonetic training was effective in making learners produce the /æ/-/ᴧ/ contrast more distinctly at post-test. The Group x Test interaction in the Pillai scores arose because the nonword training group produced substantially smaller /æ/-/ᴧ/ distances than the word training group at T1 (.308 vs. .392; t(15955)=-1.82, p=.069) and T2 (.392 vs. .437; t(15955)=-1.42, p=.154), while differences between testing times were significant (p<.001) for both groups. Test did not interact with Group for the Euclidean and Mahalanobis /æ/-/ᴧ/ distances, nor did the main effect of Group reach significance for any of the measures, suggesting that the significant main effect of training was independent of whether learners had been trained with non-lexical or lexical materials.

Table 2: Tests of fixed effects for /æ/-/ᴧ/ acoustic distances (df = 1, 15956; *= significance: α=0.05 level).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Distance</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td>2.64</td>
<td>.104</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>2661.06</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Group * Test</td>
<td></td>
<td>16.70</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Euclidean</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td>0.95</td>
<td>.331</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>11.37</td>
<td>.001*</td>
</tr>
<tr>
<td>Group * Test</td>
<td></td>
<td>0.33</td>
<td>.564</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td></td>
<td>1.44</td>
<td>.230</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>45.12</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Group * Test</td>
<td></td>
<td>2.44</td>
<td>.118</td>
</tr>
</tbody>
</table>

2.5.3. Comparing L-NS and /æ/-/ᴧ/ distances

The summary of the outcome of the statistical analyses in sections 2.4.1 and 2.4.2 above presented in Table 3 below show that all L-NS and /æ/-/ᴧ/ distance measures captured differences between testing times (Test) and L-NS distances between vowels (Vowel) similarly. Group effects (Group) were small and only present for the L-NS measures.

Table 3: Main effects for L-NS and /æ/-/ᴧ/ distances.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Distance</th>
<th>Test</th>
<th>Group</th>
<th>Vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillai</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-NS</td>
<td>/æ/-ᴧ/</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euclidean</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-NS</td>
<td>/æ/-ᴧ/</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mahalanobis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-NS</td>
<td>/æ/-ᴧ/</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To explore the consistency of the three measures within learners, we computed Pearson-r correlation coefficients among them for L-NS and /æ/-/ᴧ/ distances (based on data aggregated by subject). As shown in Table 4, all measures were significantly associated with one another, suggesting that they captured inter-learner differences in acoustic distance scores similarly. However, whereas correlations among measures were very strong for L-NS distance scores, which shared 32-80% of the variance, they were very weak for /æ/-/ᴧ/ distance scores, sharing only about 7.5% of the variance. In addition, Euclidean and Mahalanobis distance scores were more strongly associated with one another than they were with Pillai scores.

Table 4: Pearson-r correlation coefficients between acoustic distance measures (**p<.01)

<table>
<thead>
<tr>
<th>Distance</th>
<th>Euclidean</th>
<th>Mahalanobis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillai</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-NS</td>
<td>.570**</td>
<td>.589**</td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td>.864**</td>
<td>.809**</td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td>.275**</td>
<td>.722**</td>
</tr>
<tr>
<td>Euclidean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-NS</td>
<td></td>
<td>.895**</td>
</tr>
<tr>
<td>/æ/-ᴧ/</td>
<td></td>
<td>.270**</td>
</tr>
</tbody>
</table>

To summarize, in answering RQ1, Pillai, Euclidean and Mahalanobis distance measures captured training effects (differences between testing times) similarly both for L-NS and /æ/-/ᴧ/ acoustic distances. Group effects were small and only surfaced for Pillai and Euclidean L-NS distances. Correlational analyses suggest that the three measures are more consistent in capturing inter-learner differences in L-NS acoustic distance than in /æ/-/ᴧ/ acoustic distance.

Our second research question (RQ2) asked to what extent L-NS and /æ/-/ᴧ/ acoustic distances were related to one another for the Pillai, Euclidean and Mahalanobis measures. It is often assumed [15] that for learners who find it difficult to produce an L2 vowel contrast accurately, pronunciation training gains in distinguishing two vowels (distinctiveness) would correspond to gains in production accuracy, with the two vowels becoming more target-like (nativelikeness). However, as a result of phonetic training focusing on a specific contrast (/æ/-/ᴧ/), learners might try to make an acoustic distinction based on the wrong acoustic cues, or they might be exaggerating the acoustic distance between /æ/ and /ᴧ/, producing vowels that may be acoustically more distinct but at the same time less target-like. To assess this possibility, we computed Pearson-r correlation coefficients between L-NS and /æ/-/ᴧ/ acoustic distances separately for each measure. Two significant weak-to-moderate correlations arose. L-NS and /æ/-/ᴧ/ Pillai scores correlated significantly and negatively for /ɑ/ (r=-.339, p<.001) but not for /ʌ/ (r=.046, p=.641), suggesting that the better learners could distinguish /ɑ/ from /ʌ/ (i.e., the larger the Pillai score was, representing less overlap, for the /ɑ/-/ʌ/ distance), the more target-like /ɑ/ was (i.e., the smaller the Pillai score was, representing more overlap, for the /ɑ/ L-NS distance). A similar association in the same direction (larger /ɑ/-/ʌ/ distances corresponding to smaller L-NS distances)
was found between the Mahalanobis L-NS and /æ/-/ᴧ/ distances, but it did not reach significance ($r=\cdot185$, $p=.059$). These findings are consistent with the hypothesis that training effectiveness in helping learners produce more distinct realizations of /æ/ and /ᴧ/ led to more target-like productions. However, in the case of the Euclidean measure, larger /æ/-/ᴧ/ distances were associated with larger (i.e., less target-like) L-NS distances ($r=.489$, $p<.001$), thus suggesting that the nature of the association between /æ/-/ᴧ/ and L-NS distances may depend on the acoustic measure used.

2.5.4. Relating acoustic and perceptual distance measures.

Our third research question (RQ3) asked whether Pillai, Euclidean and Mahalanobis distance measures would differentially predict native English listeners’ judgments of pronunciation accuracy, as measured though perceptual distance scores. Perceptual distance scores (1-18) based on 12 /æ/ and 12 /ᴧ/ minimal-pair words from 22 learners largely reflected, for T1-T2 differences, what we had found in the acoustic measures based on 38 /æ/ and 38 /ᴧ/ minimal-pair items from 105 learners. These perceptual data show smaller perceptual distance scores at post-test than at pre-test, indicating improvement as a result of phonetic training (Figure 5). Training group and vowel effects were also observable, but of a much smaller magnitude than those observed thorough the acoustic distance scores.

![Figure 6: Perceptual distance scores according to Test (T1, T2) Group (Nonwords, Words) and Vowel (/æ/-/ᴧ/). Error bars: 95% CI.](image)

To assess whether Pillai, Euclidean and Mahalanobis distance scores were related to perceptual distance scores we computed Pearson-\(r\) correlation coefficients between the three acoustic measures and the perceptual measure for both L-NS and /æ/-/ᴧ/ distances. These results showed that perceptual distances were unrelated to L-NS acoustic distances, but they were significantly negatively associated with all /æ/-/ᴧ/ acoustic distances. This suggests that the more distinctly the learners produced /æ/ from /ᴧ/ (larger acoustic distance), the smaller the perceptual distance native listeners perceived between learners’ productions and their own, that is, the more accurate learners’ productions were judged to be by native listeners.

**Table 5: Pearson-\(r\) correlation coefficients between acoustic and perceptual distances (*\(p<.05\), **\(p<.01\)).**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Distance</th>
<th>Perceptual Distance</th>
<th>(r)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pillai</td>
<td>L-NS /æ/</td>
<td>.054</td>
<td>.612</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/ᴧ/</td>
<td>-.344</td>
<td>.001**</td>
<td></td>
</tr>
<tr>
<td>Euclidean</td>
<td>L-NS /æ/</td>
<td>.119</td>
<td>.260</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/ᴧ/</td>
<td>-.301</td>
<td>.004**</td>
<td></td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>L-NS /æ/</td>
<td>.073</td>
<td>.491</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/ᴧ/</td>
<td>-.225</td>
<td>.031*</td>
<td></td>
</tr>
</tbody>
</table>

3. Discussion

Assessing improvement in vowel production accuracy after a short training intervention using acoustic distance measures is challenging. This is partly due to large inter-learner variability within and between testing times and training conditions. In the current study, acoustic (Pillai, Euclidean, Mahalanobis) and perceptual distance measures have been shown to capture training effects. However, such effects are not easily interpretable in terms of pronunciation improvement because they partly depend on which measure is being used, and whether improvement is interpreted in terms of reduction of L-NS distance (nativelikeness) or extension of /æ/-/ᴧ/ distances (distinctiveness). For example, for the Pillai and Mahalanobis measures (but not for the Euclidean distance measure) the production of more distinct realizations of /æ/ and /ᴧ/ also corresponded to more target-like productions of /æ/ and /ᴧ/.

The three measures included (Pillai, Euclidean, Mahalanobis) captured phonetic training effects and differences between vowels in terms of L-NS distance similarly (RQ1), whereas small training group effects were not captured by all measures. Interestingly, the training effects on the Mahalanobis distance measure were general and independent of training group condition or vowel, whereas such training effects depended on interactions with training group and vowel in the case of Pillai scores and Euclidean distances. As regards differences in acoustic distances between the target vowels /æ/ and /ᴧ/, a similar picture emerged. Although training effects appeared to be larger on Pillai scores and Mahalanobis distances than on Euclidean distances, they reached significance on all three measures, whereas training group differences did not. Consistent with these findings is the outcome of the correlational analyses
that related acoustic distance measures to one another. Here acoustic distance measures were found to be more consistent for L-NS distances than for /æ/-/ə/ distances and in general Pillai scores and Mahalanobis distance scores correlated more strongly with one another than with Euclidean distances.

As regards the relationship between acoustic distance measures of nativelikeness (L-NS) and distinctiveness (/æ/-/ə/) (RQ2), our findings appear to be inconclusive in that they depend on the acoustic distance measure used. For Pillai scores and Mahalanobis distances, in accordance with common assumptions in phonetic learning [15], an inverse relationship between /æ/-/ə/ and L-NS distances was found, such that larger /æ/-/ə/ distances would correspond to smaller L-NS distances, indicating pronunciation improvement both in vowel distinctiveness and nativelikeness. However, these inverse relationships were very weak (3-11% of shared variance) and for Euclidean distances the relationship that emerged was stronger and positive, larger /æ/-/ə/ distances corresponding to larger L-NS distances (24% of shared variance), suggesting that /æ/-/ə/ were produced more distinctly but less-target like. Further research would be necessary to specifically test the hypothesis that in the assessment of phonetic training gains, increased distinctiveness between contrastive vowel categories such as /æ/-/ə/ corresponds to improved accuracy (i.e., increased nativelikeness).

Finally, we assessed whether Pillai, Euclidean and Mahalanobis distances would predict perceptual distance scores based on native speakers’ judgments of pronunciation accuracy (RQ3). Our findings, based on a sub-sample of test items (n=24) and learners (n=22), showed a positive association between the size of the acoustic distance learners could produce between /æ/ and /ə/ (irrespective of the acoustic measure used) and how accurately their productions were perceived to be by native English listeners. Acoustic distance measures were found to predict a small (but significant) amount of variance in perceptual acoustic distances (5-12% depending on the acoustic distance measure).

4. Conclusions

Although all acoustic measures investigated were generally consistent in capturing phonetic training gains in measures of pronunciation nativelikeness and distinctiveness for the target L2 vowel contrast /æ/-/ə/, they correlated strongly with one another, and they predicted perceptual distance measures based on native listeners’ ratings, they varied in their ability to capture training gains and in detecting differences between training groups and vowels. The variability between measures observed, based on the current phonetic training data set (DWR stimuli), is consistent with the use of the Mahalanobis distance measure of vowel pronunciation accuracy adopted in recent phonetic training studies [15, 16]. However, it would be important to extend this research to vowel productions elicited through production tasks other than DWR, such as sentence repetition tasks or tasks eliciting more extemporaneous types of speech, such as oral narratives or picture-based descriptions.

The perceptual distance measure we obtained through a rating task may be interpreted as a measure of distance between learners’ productions and native speakers’ phono-lexical representations for the words the learners produced. Other types of perceptual judgment tasks, such as one based on the discrimination and dissimilarity rating of learners’ vowel productions may be more sensitive in perceptually assessing gains in pronunciation accuracy resulting from phonetic training. Finally, given the differences observed between L-NS and /æ/-/ə/ distance measures, it seems important to measure both when assessing pronunciation improvement in the production of vowel contrasts. Future research should explore the reliability of measures that include both types of distances (distinctiveness and nativelikeness) in the assessment of pronunciation training gains.

5. Acknowledgements

I would like to thank Mireia Ortega, Ingrid Mora-Plaza, Cristina Aliaga-García for their work in data collection; Natalia Kartushina and Gerda Melnik-Leroy for sharing their Mahalanobis scripts; Joan Borràs for R-scripting and acoustic measures, and James Waltz for collecting and analysing the perceptual judgment data. This research was funded by grant PID2019-107814GB-I00 from the Spanish Ministry of Science, Innovation and Universities.

6. References


