NON-LINEAR DYNAMICS OF ADULT NON-NATIVE PHONEME ACQUISITION: PERCEPTION AND PRODUCTION

by

Gregory Anderson
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

Director

Department Chairperson

Program Director

Dean, College of Humanities and Social Sciences

Date: May 5, 2011

Spring Semester 2011
George Mason University
Fairfax, VA
Non-Linear Dynamics of Adult Non-Native Phoneme Acquisition
Perception & Production

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

By

Gregory G. Anderson
Master of Arts
George Mason University, 2007

Director: Carryl Baldwin, Professor
Department of Psychology

Spring Semester 2011
George Mason University
Fairfax, VA
Copyright 2011 Gregory G. Anderson

All Rights Reserved
DEDICATION

This work is dedicated to my nephew, Nicholas Anderson, who is a wellspring of inspiration and pride for me, and, being possessed of exceptional promise and intellect, will accomplish many great things in his life. I also dedicate this work to my parents, who are my best friends and a constant source of support and guidance. Thank you. Thank you. Thank you.
ACKNOWLEDGEMENTS

I would like to thank my committee member and research sponsor, Dr. Betty Tuller, for agreeing to advise me and for being a wonderful and supportive mentor through a somewhat lengthy dissertation process. Dr. Tuller is possessed of exceptional clarity of thought and masterful communication, even among the intellectual elite, and has been an excellent role model for me.

I would also like to thank my dissertation director, Dr. Carryl Baldwin, for helping me navigate the bureaucracy and minutiae of the dissertation process. Dr. Baldwin did what was necessary, without the luxury of time or precedent, to help me succeed at both comprehensive exams and the dissertation process. Her excellent advice and finely-honed editorial pen were most appreciated.

Finally, I would like to thank my committee member Dr. Michelle Ramos-Pellicia for her invaluable contributions to my understanding of the phonology (and sociology) of Spanish throughout Latin America, and my former advisor Dr. Chris Kello, who dramatically changed my world view by introducing me to complexity and the common thread underlying the vast array of natural organized systems throughout the universe.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1: Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Categorical Speech Perception</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Perceptual Assimilation</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Models of Speech Perception</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Individual Differences</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Dynamics of Speech Perception</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Categorical and Exemplar-Based Learning</td>
<td>11</td>
</tr>
<tr>
<td>1.7 Production</td>
<td>13</td>
</tr>
<tr>
<td>1.8 Generalization</td>
<td>16</td>
</tr>
<tr>
<td>1.9 Objectives</td>
<td>19</td>
</tr>
<tr>
<td>CHAPTER 2: Methods</td>
<td>22</td>
</tr>
<tr>
<td>2.1 Participants</td>
<td>22</td>
</tr>
<tr>
<td>2.2 Perceptual Mapping Stimuli</td>
<td>23</td>
</tr>
<tr>
<td>2.3 Perceptual Mapping Procedures</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Training Stimuli</td>
<td>28</td>
</tr>
<tr>
<td>2.5 Training</td>
<td>28</td>
</tr>
<tr>
<td>2.6 Production</td>
<td>29</td>
</tr>
<tr>
<td>CHAPTER 3: Data Analysis</td>
<td>31</td>
</tr>
<tr>
<td>3.1 Model Fitting</td>
<td>31</td>
</tr>
<tr>
<td>3.2 Dynamics of Learning</td>
<td>34</td>
</tr>
<tr>
<td>3.3 Production</td>
<td>37</td>
</tr>
<tr>
<td>CHAPTER 4: Results</td>
<td>40</td>
</tr>
<tr>
<td>4.1 Perceptual Profiles</td>
<td>40</td>
</tr>
<tr>
<td>4.2 Perceptual Training Results</td>
<td>44</td>
</tr>
<tr>
<td>4.3 Production</td>
<td>47</td>
</tr>
<tr>
<td>4.5 Learning Dynamics</td>
<td>60</td>
</tr>
<tr>
<td>CHAPTER 5: Discussion</td>
<td>77</td>
</tr>
<tr>
<td>5.1 Modes of Learning</td>
<td>77</td>
</tr>
<tr>
<td>5.2 Non-Assimilable Phonemes and Task Difficulty</td>
<td>80</td>
</tr>
<tr>
<td>5.3 Production</td>
<td>83</td>
</tr>
<tr>
<td>5.4 Conclusion</td>
<td>89</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>93</td>
</tr>
<tr>
<td>Manner of Production (Waveforms and Spectrograms)</td>
<td>94</td>
</tr>
<tr>
<td>Phonological Profile (MDS)</td>
<td>95</td>
</tr>
<tr>
<td>Refiner Profile (MDS)</td>
<td>96</td>
</tr>
<tr>
<td>Non-Learner Profile (MDS)</td>
<td>97</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>98</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1: JG Comparison for “too good” participants</td>
<td>42</td>
</tr>
<tr>
<td>Table 2: Cluster analysis of NS ratings of production</td>
<td>52</td>
</tr>
<tr>
<td>Table 3: Proportion of vibrant productions</td>
<td>56</td>
</tr>
<tr>
<td>Table 4: Acoustic measures of trill production</td>
<td>58</td>
</tr>
<tr>
<td>Table 5: Acoustic measures of tap production</td>
<td>59</td>
</tr>
<tr>
<td>Table 6: Distinguishing characteristics of three learning profiles</td>
<td>61</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure ............................................. Page
Figure 1: Two-dimension representation of perceptual space ........................................... 7
Figure 2: Representation of possible Hindi perceptual space ............................................. 8
Figure 3: Select waveforms from testing continuum ......................................................... 24
Figure 4: Averaged ID results for NS listeners ................................................................. 25
Figure 5: NS ID performance and theoretical model ......................................................... 32
Figure 6: NS JG performance and theoretical models ....................................................... 33
Figure 7: Diagrammatic explanation of calculation of RMSE ........................................... 34
Figure 8: Diagrammatic explanation of Perceptual Magnet effect ..................................... 35
Figure 9: Diagrammatic explanation of clustering ............................................................. 37
Figure 10: Waveform, spectrogram and annotation for trill production attempt ............... 38
Figure 11: ID results for NE and NS listeners ................................................................. 41
Figure 12: Averaged JG ratings for NE and NS listeners .................................................... 43
Figure 13: Standard deviation of ratings for one participant .............................................. 46
Figure 14: Standard deviation of ratings for tap and trill ends on JG-Trill tasks ............... 46
Figure 15 Correlation between initial production ratings and JG RMSD ....................... 48
Figure 16: Item analysis for production targets ............................................................... 49
Figure 17: Individual and group production progress ....................................................... 51
Figure 18: Distribution of manner of production for each target ..................................... 55
Figure 19: NS ratings by manner of production ............................................................... 57
Figure 20: MDS profile for Phonological Learner ............................................................ 63
Figure 21: MDS profile for 2nd Phonological Learner ..................................................... 64
Figure 22: MDS profile for Refiner ................................................................................ 65
Figure 23: MDS profile for Non-Learner ....................................................................... 66
Figure 24: Perceptual clustering for trills and taps ........................................................... 68
Figure 25: RMSE scores for ordinal rankings of stimuli .................................................... 68
Figure 26: JG and ID performance of the three training groups ..................................... 70
Figure 27: NS Ratings of production for three training groups ....................................... 71
Figure 28: Correlations between JG-Trill and Trill Production ........................................ 73
Figure 29: Correlations between perceptual clustering and JG performance ................. 74
Figure 30: Pre- and Post-training clustering and ordinals scores .................................... 75
ABSTRACT

NON-LINEAR DYNAMICS OF ADULT NON-NATIVE PHONEME ACQUISITION: PERCEPTION & PRODUCTION

Gregory G. Anderson, Ph.D.

George Mason University, 2011

Dissertation Directors: Dr. Carryl Baldwin, Dr. Betty Tuller

A number of formal models of non-native speech perception make predictions regarding ease of discrimination based on the phonetic similarity of non-native phonemes to phonemes from the learner’s native phonology (Best, McRoberts, & Goodell, 2001; J. Flege, 1995; Kuhl, et al., 2008). Recently, individual differences in non-native sound learning have been examined from a dynamical perspective, focusing on initial perceptual abilities of individuals, and how the structure of these perceptions influences learning (Tuller, Jantzen, & Jirsa, 2008).

Here, monolingual speakers of American English were trained in the perception of the Spanish tap and trill rhotics to examine if different types of native/non-native speech contrasts exhibit different structure of initial perception and learning dynamics. In addition, production of the target phonemes by the learners was recorded analyzed to
determine the effects of perceptual training on production and examine the relation between initial perceptual structure and production learning dynamics.

Results showed that initial perceptual structure for the rhotics was predictive of learning dynamics, though the types and distribution of initial patterns were different from those seen in prior work with different phoneme contrasts. Production improvement was seen in a number of participants, corresponding to changes in perceptual constraints over the course of training. However, articulatory (motor) constraints mitigated considerably the relation between perceptual ability and production ability.

These results are consistent with current models of non-native speech perception, and provide further evidence of non-linear perceptual learning processes. Production results suggest variable interaction between perceptual and articulatory constraints on speech production learning.
CHAPTER 1: Introduction

1.1 Categorical Speech Perception

Categorical perception is a phenomenon whereby listeners hear changes in the acoustic signal not as a linear continuum, but rather as sudden transitions from one perceptual category to another (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This is evidenced by more reliable discrimination of acoustic differences that cross a phonemic boundary than of equal acoustic differences within a phonemic category (Eimas, Siqueland, Jusczyk, & Vigorito, 1971). Categorical perception is crucial for the efficient processing of speech. Speech sounds are heard in a variety of contexts that alter the acoustic properties of the sounds, yet must be represented in the brain as discrete phonemes in order to recognize patterns of meaning from one context to the next. However, the development of categorical speech perception, by definition, means that listeners become less sensitive to allophonic variations within phonemic categories. Therefore, as people develop mastery of a particular sound system, their speech perception adapts to the demands of oral/aural communication by organizing into categories that increase communication effectiveness by prioritizing key phonemic features over context-dependent acoustic features.
In order to prioritize acoustic elements of the incoming speech signal, the listener must have some sense of what constitutes a phoneme. In other words, there must be some cognitive representation of the phoneme. When I imagine the sound of the letter “t”, I have a single sound in my head. This canonical representation seems independent of all the allophones I hear throughout the day, including “t”s that are aspirated, unaspirated, alveolar, dental and even glottal. The process by which this canonical representation is achieved has been the source of much debate, represented by two major approaches: the abstractionist approach and the exemplar approach. According to the abstractionist approach, the surface form of a speech sound undergoes a transformation that sheds the context-dependent acoustics that vary among speakers and situations, leaving only abstract phonological units that represent the phoneme across all contexts. One explanation of this transformation is that the representation is “underspecified,” containing only the minimally necessary acoustic elements for phoneme identification (Eulitz & Lahiri, 2004; Wheeldon & Waksler, 2004). This results in tolerance of, or insensitivity to, context-dependent variations that are not specified in the abstract representation of the phoneme. In the abstractionist view, the phonemic representation is by definition static, as unspecified acoustic features are “pruned” in the perception process, leaving the original representation intact.

In contrast, the exemplar approach is based on all details of each encounter with a phoneme being retained. The representation of the phoneme is therefore not static, but dynamically changing as an “average” of the acoustic elements of each instance of the
phoneme perceived by the listener. This representation is not pre-defined, but built from the ground up as the listener hears instances of the phoneme and associates each instance with meaning. Implied in this approach is that the acoustic elements that are present in a greater number of encounters with a phoneme, such as the stop silence present in every instance of /t/, become more salient, and less common elements, such as the high-frequency aspiration present in words like “top” but absent in words like “butter”, are less salient, with roll-off in saliency corresponding to their frequency and importance to successful communication.

1.2 Perceptual Assimilation

Whether through design or experience, the most relevant/prevalent characteristics of phonemes are thought to allow stable representations across myriad productions that contain any number of less relevant acoustic properties. For example, in American English (AE) productions, nasalization of vowels is common, due to coarticulation from surrounding nasal consonants. As a result, the nasalized /a/ in “mom” is perceived as the same phoneme as the oral /a/ in “pot.” Were the nasalization to take place outside the typical nasal phonemic context, for example, a nasalized /ɔ/ in “paw,” the AE listener would likely assimilate that production to the standard oral phoneme category, attributing the nasalization to non-linguistic factors (voice quality, upper respiratory abnormalities, etc.), if it were perceived at all. In this case, nasalization is not a key element that would change the identification of the phoneme when heard and would be expected to be
assimilated into a single phonemic category. While that particular example may be true for American English, it would not be for French, a language in which nasalization is a phonemic contrast (Trehub, 1976). In French, the presence of nasalization in some vowels is contrastive between two different phonemic categories. In that case, nasalization is a key element of the phoneme, the assimilation of which would obscure meaning.

1.3 Models of Speech Perception

Best’s Perceptual Assimilation Model (Best, 1994) and Flege’s Speech Learning Model (J. Flege, 1995) both explain differences in the learning of non-native speech contrasts in terms of perceptual assimilation. In the above example from French, both models would predict that learning to discriminate the nasalized vowel from the non-nasalized vowel in the same phonological context (such as “pa”) would be difficult for the AE listener, due to the ease with which the two vowels are assimilated into a single AE phoneme category. The relative differences in ease of learning are directly contrasted in the case of the Zulu click and the Hindi stop. It has been reported that Native English (NE) speakers can easily discriminate voice and place of articulation contrasts across Zulu oral clicks (Best, McRoberts, & Sithole, 1988), but have difficulty discriminating place of articulation contrasts between the unaspirated retroflex and dental stop consonants, which are different phonemes in languages such as Hindi (Tees & Werker, 1984).
According to Best (1994), the alveolar and retroflex stop consonants are more difficult for native English speakers to discriminate because both are allophonic variations of the English alveolar stop consonant /d/. This similarity results in the assimilation of the two non-native sounds into the single American English phonemic category. According to Best’s Perceptual Assimilation Model (PAM), (Best, et al., 2001), a non-native phoneme may be assimilated to the native phonemic structure as either a categorized or uncategorized exemplar. The above example is one of a single categorized exemplar, due to the fact that the retroflex stop is assimilated into the single English phonemic category of the alveolar stop. A non-native phoneme that is phonetically similar to more than one category, such as the Danish /ø/, which is similar to both the English /u/ and /ʊ/, would be an uncategorized exemplar. Zulu clicks, on the other hand, are considered non-assimilable - they are dissimilar to any American English phoneme categories, are not perceptually equivalent to the native English speaker, and are therefore not assimilated into a native category (Best, et al., 1988). Since these non-native sounds are not assimilated into existing phonemic categories, the perceiver is able to attend to the sub-segmental features of the contrast, leading to better discrimination.

Although PAM proposes multiple exemplar categories, in the case of a single-category exemplar, Best’s model’s predictions are very similar to Flege’s Speech Learning Model (SLM)(J. Flege, 1995). Flege’s original model concerned itself primarily with age limits on learning to produce L2 sounds, and postulated that the same mechanisms used for L1 learning are available throughout the lifespan for L2 learning.
Rather, it is the existence of the L1 that makes learning the L2 more difficult for adults, in that pre-existing phonetic categories and interfere with or even block formation of new categories in perceptually similar space. Without different exemplar categories, the SLM predicts the success of non-native phoneme acquisition based on the perceived phonetic distance between the native and non-native phonemes. This model would also predict that the Hindi stop consonants, being perceptually more similar to an English phoneme than the Zulu click, would be more difficult to learn.

A third relevant model is Kuhl’s Native Language Magnet (NLM) model (Kuhl, 1994). Predictions regarding perception of non-native phonemes are similar, in that non-native phonemes are easier to perceive the more dissimilar their acoustic properties are to native phonemes. The NLM model, however, defines a phoneme prototype that functions as the “center” of a phonemic category. This prototype functions as a perceptual magnet, with variable within-category discrimination increasing in difficulty the more similar the stimulus is to the prototype.

Though all three models (PAM, SLM, NLM) account for some type of assimilation of non-native phoneme perception based on its similarity to established native phonemes, they vary in the processes by which they assume similarity judgments are made, such as the general acoustic process assumed by NLM, the phonetic process assumed by SLP or the articulatory gestures assumed by PAM.
1.4 Individual Differences

Key to the vagaries of perceptual assimilation is the concept of perceptual space. If we were to map people’s perception of different speech sounds in a spatial representation, phonemes that sounded more similar to one another would be closer together. For example, to the AE listener, the aforementioned Hindi stops would be closer together than either sound would be to the palatal click, which is perceptually quite different, and would be represented spatially as farther away. A two-dimensional representation of that perceptual space can be seen in Figure 1.

![Figure 1](image)

**Figure 1.** Two-dimensional representation of perceptual space. The graduated grey area around the [d] represents the waning influence of the perceptual magnet of the existing phoneme category. The retroflex stop, falling inside this sphere of influence, is assimilated by the AE listener, whereas the palatal click, falling outside this range, is not.

While the perceptual space represented in Figure 1 may be representative of the typical AE listener, it is likely not reflective of the perceptual space of the typical Hindi listener. In the competent Hindi listener, the unvoiced retroflex stop and the dental stop are easily distinguished, and, since they are not assimilated into a single category, would
not be in the perceptual “sphere of influence” for either phoneme category. Such an arrangement of perceptual space is depicted in Figure 2.

Figure 2. Model of the two-dimensional representation of the perceptual space of a competent Hindi listener. Each of the three phonemes lies outside the phonemic boundaries of the other two, so there is no assimilation by native Hindi listeners. When Euclidean distances are presumed from this representation of perceptual space, it gives the impression that the palatal click is more similar to either of the other two phonemes than they are to one another. This is not necessarily true. Sounds can differ on any number of parameters, and an accurate representation of perceptual space would need to comprise as many dimensions as there are parameters. Therefore, this two-dimension plot, while illustrative, is inherently inaccurate.

The contrast between the presumed perceptual space of the native AE and native Hindi listeners points to an important assumption made by all three of the aforementioned speech perception models - similar perceptual space among listeners. The models rely on formal linguistic descriptions of native and nonnative phoneme sets for different languages as well as of different phonemes within a language. As such, they do not address differences among native speakers of the same language. However, significant individual differences in ability to discriminate non-native speech sounds have been exhibited by adults with similar language backgrounds, even after phonetic training (Jenkins & Yeni-Komshian, 1995; Polka, 1991; Pruitt, Strange, Polka, & Aquilar, 1990).
If the role of perceptual space in the ease with which non-native sound contrasts are perceived is as central as the formal models suggest, it would suggest that individual differences in speech discrimination among listeners with similar language backgrounds may be due to differences in each individual’s perceptual space. Drawing from the earlier example of the Hindi retroflex stop, some monolingual AE listeners may be more sensitive to the acoustic differences between the alveolar and retroflex stops than others, in which case the alveolar stop would act as a weaker attractor of the non-native sound, making perception less difficult. This difference could be due to some innate ability or different experience or exposure to non-native speech sounds, but would not be accounted for in models based solely on the phonology of the native language relative to that of the non-native language. Furthermore, if one’s ability to discriminate sounds changes as a result of training, it suggests that phonetic training is effecting a change in that individual’s perceptual space. If a listener is unable to discriminate two phonemes prior to training, the formal models would suggest that the non-native phoneme was assimilated by the native phoneme. However, some individuals can learn to discriminate phonemes with training. Post-training discrimination of the two phonemes would suggest a change in perceptual space, such that perception of the non-native phoneme had diverged to the point that it was no longer assimilated by the native phoneme. In other words, perceptual space can be malleable and dynamic, changing with individuals’ phonetic experience.
1.5 Dynamics of Speech Perception

The dynamics of perceptual space have recently been explored using two stop consonants that differ in place of articulation - the alveolar and dental stops, which are phonemically contrastive in the south Indian language of Malayalam (Tuller, et al., 2008). In that study, a continuum of acoustic stimuli ranging from the dental to the alveolar place of articulation of a Malayalam stop consonant was used to map the structure of perceptual space in native AE listeners. Perceptual space was mapped using three different low-dimension measures - a Judged Goodness (JG) task that measured the internal phonetic structure within the phonemic category, a Phoneme Identification (ID) task that measured the phonemic boundaries of the two stimuli, and a one-dimensional plot of difference ratings that measured perceptual distance among stimuli in the continuum.

Perceptual space was measured before, during and after phonetic training to assess changes over time. Results indicated strong individual differences among participants in both the patterns of learning and the initial perceptual space. One cluster of participants was unable to discriminate the two phonemes either before or after training. This suggests that, for those participants, the non-native phoneme was strongly assimilated into the native category, and training had no effect on perceptual space. The other two clusters of participants, however, did make changes in their perception as a result of training. Both of the learning groups were able to discriminate between the native and non-native phonemes prior to training. However one cluster of participants exhibited strong within-phoneme structuring, whereas the other did not. These individual
differences prior to training were predictive of two different learning dynamics. Over the course of training, the group with initial strong internal structuring of the non-native phoneme became less attuned to the fine acoustic distinctions, in essence, developing categorical perception. This pattern was dubbed Phonological Learning, as there was a progression over time that resulted in what looked to be the development of a phonological category for the non-native phoneme.

The cluster of participants who were able to discriminate between the two phonemes but had weak initial internal structuring of the non-native phoneme exhibited a different learning pattern. These participants became more attuned to the fine acoustic distinctions, without developing categorical perception. This pattern was dubbed Acoustic Learning.

1.6 Categorical and Exemplar-Based Learning

These patterns are consistent with the literature on the general issue of categorization, with the Acoustic Learning analogous to the rule-based category learning that is typical when presented with confusable stimuli. Conversely, the Phonological Learning is analogous to an exemplar-based model, which is typical when stimuli are distinctively clear (Rouder & Ratcliff, 2006). Rule-based and exemplar-based category learning are both useful in acquiring a new phonological category, in that acoustic thresholds act as rules that help distinguish among similar allophones in the beginning of training, whereas true phonological learning shows the “warped” correspondence
between acoustic properties and perceived similarity at the end of acquisition - a perceptual space typical of exemplar learning.

Tuller (2004) has proposed that this learning process is best framed within a nonlinear dynamical approach. Clearly, the initial state of the perceptual space has an impact on the learning of a new category, but as learning a new phonological category is a dynamic process that takes place over time, the effect of existing state is iterative, constantly changing in response to new learning, and in turn, constantly influencing the impact of new learning.

This nonlinear dynamical view does not differentially test competing hypotheses of the PAM, NLM and SLM models, but rather adds information about the dynamics of new phoneme acquisition, which are, as Tuller has shown, highly dependent upon the initial individual differences among learners. The common theme among the three models is the propensity for assimilation of different phonemes. In the non-linear dynamical view, this propensity for assimilation is determined by the strength of the native (L1) category. As L1 categories become more stable, they become more attractive regions of perceptual space, resulting in greater assimilation of acoustically similar non-native (L2) targets. The infant, of course, being one who has little experience with any particular sound set, has weak attractors, resulting in the ability to discriminate among a wider range of speech sounds than an adult, for whom the acoustic differences in new L2 sounds are obscured by the strong attraction of very stable native phonological categories.
The dynamics of learning to perceive the non-native phoneme center around the strength of the pre-existing attractors, as defined by an individual’s organization of perceptual space. A new phoneme that is not assimilated by existing attractors can fit easily within one’s perceptual space without any adjustments of the existing organization. However, learning to perceive an assimilated phoneme is achieved through the reorganization of existing perceptual space to allow the development of a new attractor. This bifurcation of a single category to two new categories is not the same as just adding the perception of a new phoneme to one’s perceptual space – the similarity of the new phoneme to the established phoneme requires a restructuring of initial perceptual space.

1.7 Production

Modest correlations between perception and production of L2 speech sounds have been shown repeatedly (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; Yamada, Strange, Magnuson, Pruitt, & Clarke, 1994).

As seen with perception, the ease with which production of non-native phonemes can be learned appears to be influenced by the same relations (similarity/dissimilarity) to native phonemic categories (Aoyama, et al., 2004). Individual differences in age and L2 exposure and experience have also been shown to be predictive of production learning success (Baker & Trofimovich, 2006; J. E. Flege, 1999). Many studies have looked at whether production precedes, succeeds or parallels perception in L2 learning (Best, 1995; J. Flege, 1995; Sheldon & Strange, 1982). Flege’s SLM model incorporates the idea that
perception accuracy is a limiting factor in production accuracy. However, it is not known whether initial intra- and inter-phonemic perceptual structuring is predictive of production learning in the same way it has been shown to be predictive of perceptual learning. Moreover, while it is a common research design to compare production pre- and post-training, little is known about the progression over the course of training. The strength of the link between perception and production in L1 has been debated extensively (Browman & Goldstein, 2008; Mitterer & Ernestus, 2008; Plaut & Kello, 1999; Rvachew, 1994), but it is widely accepted that there is some link. That, taken with the aforementioned perception-production links in L2 learning, would seem to suggest that production is influenced by perceptual training. The exact nature of that influence could range from the same phase transition present in perceptual learning to a linear path of improvement, to anything in between.

A couple of characteristics of speech production suggest that the learning path may mirror that of speech perception. The first is the goal-oriented, adaptable nature of production. There would be little value to categorical perception of variable acoustic parameters if there were no variability in speech production. Not only do different people produce phonemes differently, but individuals can make themselves understood across a wide range of variable constraints on production. People talk with missing teeth, numbed lips following a visit to the dentist and occasionally, though not ideally, with food in their mouth. People adapt their production patterns automatically to achieve the goal of communication via speech. This rapid reorganization of speech gestures to achieve a
communication goal in the presence of various physical constraints is considered one of the hallmarks of a self-organized, nonlinear dynamic system (Porter & Hogue, 1998). Evidence of 1/f scaling relations in intrinsic fluctuations of speech production (Kello, Anderson, Holden, & Van Orden, 2008) lends further support to the interpretation of speech production as a metastable complex system. As such, speech production would be influenced by attractors that adapt to constraints through a process of stabilization and destabilization such as was seen in Tuller’s perceptual learning data. This suggests that reliable production of a similar L2 phoneme would require a bifurcation of the single category of L1 sound productions to accommodate the new production category. Of course, production introduces an additional parameter on which to judge similarity - motoric patterns. In the case of the dental vs. alveolar stop, tongue position varies only slightly, so they are similar acoustically, perceptually and motorically. Conversely, the alveolar and bilabial stops /d/ and /b/, while they may sound similar to some listeners, are produced using different articulators. This motoric component means that even if production learning in general is influenced by the same assimilation effects as perception, we cannot assume that production and perception are affected in the same way for any given sound contrast.

There is evidence from L1 speech development that the production learning path is not necessarily tied directly to the perception path. A number of studies of articulation-delayed children have shown that many children are able to perceive phonemic contrasts that they are not able to produce (Broen, Strange, Doyle, & Heller, 1983; Hoffman,
Stager, & Daniloff, 1983; Mitterer & Ernestus, 2008; Rvachew, 1994), and at least one study shows evidence of correct production in the presence of inaccurate perception (Sheldon & Strange, 1982). These correlations between speech perception and production say little about the dynamics of production learning, whether the cause of changes in production are due exclusively to corresponding changes in perception or whether production-specific constraints, such as motor control, could possibly lead to a phase transition from one stable production to another that is independent of the phase transition seen in the destabilization and creation of perceptual phonemic categories.

1.8 Generalization

Tuller’s work on the dynamics of speech perception (Tuller, 2004, 2007; Tuller, Case, Ding, & Kelso, 1994; Tuller, et al., 2008), draws conclusions regarding a process generic to speech perception, rather than unique to the dental/alveolar distinction studied. According to Best’s PAM model, the dental and alveolar stops would be classified as a Single Category (SC) assimilation, as they are generally perceived by American English speakers as equally good exemplars of the phoneme /d/. However, there is a variety of possible non-native phoneme pairs that differ greatly in the relation of the individual phonemes to a listener’s native phonemes. Research has supported that these myriad differences can aid or hinder discrimination (Best, 1994, 1995). As a result, it’s possible that one or both of the learning dynamics seen with the Malayalam pair might be replaced by a different learning dynamic more suitable for learning other types of phoneme pairs.
In this study, we will be looking at the Spanish tap and trill rhotics. This pair has a great number of characteristics that add to the complexity of their relation to American English phonemes. First, the acoustic cues that differentiate the two Spanish rhotics are substantially different from those that differentiate the two Malayalam stops. The difference in the two Malayalam stops is place of articulation, which manifests acoustically in the spectrum primarily as different f2 and f3 contours. The Spanish rhotics differ by manner of articulation, manifesting both spectrally and temporally.

Second, the tap is acoustically an allophone of the AE dental stop /d/, usually found in the intervocalic position before an unstressed vowel (e.g., “butter”, “fitting”). However, phonologically, the tap is the equivalent of the AE alveolar approximant /ɹ/, so depending on a person’s exposure to Spanish, the tap could assimilate to the Single Category of either /d/ or /ɹ/, or could be classified as Uncategorized, falling somewhere in between the two AE categories.

In contrast, acoustically the trill is Non-Assimilable (i.e., not like any AE phoneme). Therefore, according to models of L2 phoneme perception, we might expect the trill rhotic to be easily distinguished from the NE phonemes, /ɹ/ and /d/.

Expectations regarding AE listeners’ discrimination between the two non-native rhotic phonemes are less certain. If we apply the PAM model to the tap and the trill, this pair should be classified as Categorized/Non-Assimilable, making discrimination between the two not difficult for native AE listeners. However, this distinction is complicated somewhat by the classification of both sounds as rhotics. If the tap is assimilated to the
AE /ɹ/ and the trill is known to be an “r” sound, it is possible that, despite the acoustic
differences between the two sounds, they could assimilate to the Single Category of the
native AE phoneme /ɹ/. The complex assimilation patterns among multiple native and
non-native phonemes will allow us to investigate the application of Best’s model to a
real-world phoneme set, with its inherent interactions.

In addition to applying prior methodology to other phonemic distinctive features,
the use of the Spanish rhotics has the added benefit of educational and cultural relevance
to American NE speakers/listeners. Spanish is the 2nd most common language spoken
in the United States, and is the primary language spoken at home by over 34 million
people aged 5 years or older (U. S. Census Bureau, 2007). There are roughly 4 million
Spanish-speaking children with Limited English Proficiency (LEP) in the United States
(DOE, 2008). Roughly 30% of all American students in grades 7-12 are enrolled in
Spanish classes (Draper & Hicks, 2002), and Spanish accounts for over half of language
enrollments at American 2- and 4-year colleges (Welles, 2004). Bilingualism is a highly-
valued skill in many jobs where communication is a significant component, and is crucial
to both teachers of Spanish and teachers of English. Yet, the Spanish tap and trill rhotics
are notoriously difficult for NE listeners to perceive and produce. Given the great need
for mutually intelligible English-Spanish bilinguals, a greater understanding of the
processes underlying acquisition of these phonemes could have a significant impact on
the education and training of English-Spanish bilingual professionals.
The Spanish rhotics are especially relevant to the production portion of this study. There is evidence that the Spanish rhotics are undergoing a sound change, especially in Latin America, where a variety of non-canonical productions of the trill are commonly used (T. Bradley, 2001; Bradley & Willis, 2008; T. G. Bradley, 2001; Inouye, Ladefoged, Andersen, Steriade, & Maddieson, 1995; Willis & Bradley, 2006). Dispersion theory posits that phonology evolves towards a minimization of effort while maintaining a maximum number of contrasts as well as maximum distinctiveness of those contrasts (Flemming, 1996). This is consistent with the type of changes taking place in Latin America, where productions in general feature a reduced number and completeness of occlusions in the vibrant trill. It is unknown if this dynamical process of sound change might be reflected in the learning patterns of individual novice speakers of Spanish.

1.9 Objectives

The objectives of this present study are two-fold – the first is extend the methods used in Tuller’s work to a different phonemic contrast. The working hypothesis is that the nonlinear dynamical patterns and effects of pre-training structuring of perceptual space are general principles that can be applied to all L2 speech perception learning. As a result, findings similar to those reported by Tuller (Tuller, et al., 2008) are expected. Failure to replicate earlier findings with a different set of phonemes would suggest that the earlier findings may be unique to the specific phonemes or phonemic contrast studied, and not a generic property of L2 speech perception.
The second objective is to measure changes in production of the nonnative phonemes. Acoustic measurements of production attempts will be analyzed to determine if changes in production are taking place without specific training on production. If production does change over time in response to perceptual training on L2 sounds, the pattern of changes will be analyzed to determine if non-linear reorganization is evident in production.

Given the variety of ways the vibrant trill is produced by Native Spanish (NS) speakers around the world, a number of different acoustic measures will be assessed. These measures were based on those used in the literature regarding production competency of both NS children and non-native learners of Spanish. At the most gross level, one can count the number of occlusions in the production of the trill. Reeder (1998) measured the mean number of trills in speakers with five levels of Spanish language competency, ranging from “beginning learners” to native speakers (“very advanced learners”). There was a linear progression of mean occlusions from 0.6 in the beginning group to 3.2 in the NS group, suggesting that mean number of occlusions is a valid measure of production competency. Face (2006) also counted occlusions, with a single occlusion being counted as a correct tap and a “series” of occlusions comprising a correct trill.

Precedent for more fine-grained spectral and temporal measures of rhotic production can be found in literature on the productions of native Spanish-speaking children. The rhotic phonemes (trill, tap and alveolar approximant) are among the latest
developing phonemes in both English and Spanish (Jimenez, 1987). A number of acoustic differences have been seen between NS-speaking children who have not yet mastered /r/ production and those who have – lower F1 mean frequency, higher F2 mean frequency, longer initial aperture period, fewer occlusions and fewer apertures (Carballo & Mendoza, 2000).

Of course the conditions resulting in mispronunciation by adult L2 learners are different from those for L1 children. Consonant duration, in particular, decreases with maturity and is seen as a direct correlate of motor control (Kent & Forner, 1980). Therefore, while duration of the trill was correlated with sound mastery in children, adults, who are presumed to have more mature motor control, may mispronounce trills or taps in a way that does not differ in terms of duration. Furthermore, the trill in particular utilizes an aerodynamic mechanism that is unique from English phonemes, so it is possible the NE speakers lack motor control for production of that particular phoneme, resulting in duration measures similar to those found in misarticulating L1 children.
CHAPTER 2: Methods

2.1 Participants

Participants in this study included both NS (Native Spanish) and NE (Native English) speakers. The NS group (n = 7) comprised four women and three men, ages 18-42. All NS participants learned Spanish as their first language and conversed with peers or family in Spanish daily. All NE participants were adult (ages 18-25) monolingual with no reported hearing loss or learning disorder. Candidates who were currently participating in foreign language instruction or who reported that they were able to carry on a simple conversation of two or more exchanges in a language other than English were excluded from participation.

There were 73 NE participants who underwent perceptual screening for inclusion in the training portion of the study. Of the 73 screened participants, 26 were not eligible for the training portion of the study because their perceptual structuring of the speech stimuli was such that they were unlikely to exhibit any changes in perception as a result of training. Of the eligible participants, 21 agreed to participate in the training. The training group (n = 21) comprised six men and 15 women, and the participants who were screened but did not participate in the training (n = 52) comprised 16 men and 36 women.
2.2 Perceptual Mapping Stimuli

Maps of the perceptual space surrounding the target L2 phonemes were generated for the 73 participants who met the above criteria for participation in the study. These maps were generated by participation in three tasks: identification (ID), judged goodness as a tap (JG 1) and judged goodness as a trill (JG 2). These tasks used a continuum of speech sounds comprised of progressive acoustic steps from a single exemplar of one target phoneme to a single exemplar of the other phoneme to which it might be assimilated in the novel listener. Prior research using Malayalam (Tuller, 2004) created a synthesized continuum between the alveolar and dental /d/ phonemes, by varying the acoustic parameter of the 2nd and 3rd formant onset frequencies.

A continuum between the trill and tap rhotics was created by modifying a recording of a single utterance of “perro” by a native speaker of Spanish. The trill in the original recording was a canonical vibrant trill with 3 occlusions comprising the rhotic phoneme. The continuum was created by removing 2-4 ms from the center of the medial occlusion. The audio was removed in both directions from the center, and the total time removed with each successive stimulus varied within the 2-4 ms range in order to make the cuts at zero crossings. A total of 60 ms was removed through this process to create the “tap” end of the continuum, resulting in a total number of 23 stimuli. Figure 3 shows the waveforms from various points along the continuum.

To verify that the resulting stimuli were perceived as a continuum from one target (trill) to the other (tap), the three perceptual mapping tasks were administered to NS
listeners. In the ID task, the stimuli from the continuum were presented in random order, and the participants were asked to indicate whether the sound they heard was Sound A or Sound B, which were randomly assigned to opposite ends of the continuum. If the right acoustic variable is manipulated in the right manner, we would expect the unaltered stimulus to be identified as a trill nearly all the time, and the manipulated single occlusion to be uniformly identified as a tap. One would expect the modified stimuli in between the two ends to be less uniformly identified as either a trill or a tap, with chance identification (50%) at some point in the continuum.

Figure 3. Waveforms for the testing continuum. A) Natural trill “perro” recording, 498 ms in duration. B) 11th stimulus in the continuum, 466 ms in duration. C) 17th stimulus, 450 ms, and D) 23rd stimulus, comprising the tap “pero” end of the continuum, 428 ms.
The original 23-stimulus continuum was validated using results from the ID task. Results showed that all stimuli greater than 466ms in duration were always identified as a trill. Therefore, stimuli at the “trill” end of the continuum were removed to center the transitional stimuli forming the unstable boundary between the two phonemes. This resulted in a 13-stimulus continuum with similar (inverse) identification values for the longest and shortest stimuli. Figure 4 shows the averaged results from four of the NS evaluators on the final 13-stimulus continuum.

Figure 4. Averaged ID results for the trill-tap continuum by six NS listeners. Error bars show greater consistency of identification at the ends of the continuum.

In this chart, the x axis represents the continuum of stimuli, and the y axis represents the percent of the time each stimulus was identified as a trill. As can be seen in the graph, the four shortest stimuli were identified as a tap 90-100% of the time, while the five longest stimuli were categorized as a trill 90-100% of the time. The middle four
stimuli were identified as one or the other phoneme from 37-83% of the time, representing the unstable phoneme boundary. This unstable region where even NS listeners are not 100% certain of which phoneme they are hearing forms a positive slope from one phoneme to the other.

2.3 Perceptual Mapping Procedures

All monolingual native English (NE) participants were tested individually and seated directly in front of a computer in a room treated with 3” acoustic foam on all walls, and bass traps in all corners to reduce reflections and ambient noise. All stimuli were presented binaurally through Sennheiser HD280 Professional headphones. Although the sound system was calibrated to a starting presentation level of 75 dB, participants were allowed to adjust the volume to a comfortable listening level.

All participants were played each of the two end stimuli from the 13-stimulus continuum four times, correctly labeled as Sound A and Sound B prior to each of the three tasks.

In the ID task, the participants were asked to label each stimulus as either a tap or a trill (two-alternative forced-choice: 2AFC). The stimuli consisted of randomized presentation of 10 of each of the 13 stimuli, resulting in 130 trials. This task indicates the degree to which the participants are able to parse the productions into two phoneme categories, and is graphed as a percent of presentations identified as one or the other phoneme.
Each JG task consisted of the participants rating the stimulus from each of the 130 trials as to its goodness as an exemplar of one of the two target sounds (tap or trill). For half the participants, the trill was labeled “Sound A” with the tap as “Sound B”, with the other half of the participants presented with the reverse naming scheme. The target sound was played for the participants three times before beginning each JG task, and again every 20 trials until task completion. Judgments were made using a 7-point Likert scale from 1 = Extremely Poor Example to 7 = Extremely Good Example. Half the participants performed the “JG as a tap” task (JG-tap) followed by the “JG as a trill” task (JG-trill), with the other half performing the tasks in the reverse order.

Participants who participated in the training performed the Difference Ratings task prior to each training session, as well as during mid- and post-training test sessions, totaling 10 over the course of the study. In the Difference Ratings task, subjects heard all possible pairs of stimuli from a 7-stimulus subset (stimuli 1, 3, 5, 7, 9, 11, and 13) of the test continuum. There are seven possible pairs of matched stimuli and 42 possible pairs of mismatched stimuli. The mismatched pairs were presented to the subjects 4 times and the matched pairs were presented 8 times, providing a total of 168 pairs. The pairs were rated on a scale from 1 to 7, with 1 being ‘no difference’ and 7 being ‘very different’. There was a 400 millisecond ISI between items in a pair and an 800 millisecond ITI between pairs. Participants who did not respond within a 3-second window were presented with a popup window with a reminder to respond more quickly.
2.4 Training Stimuli

Six native speakers of a variety of Spanish dialects, who consistently produce the trill /ɾ/ in conversation with multiple occlusions, were recorded producing a list of 60 common Spanish words containing the tap /ɾ/ and 60 containing the trill /ɾ/ in the intervocalic position. An online database was used to normalize the word lists by frequency of use (Davies, 2002). The tap and trill word lists were balanced according to frequency of use in Spanish, to control for any production bias resulting from differences in familiarity of words to the NS speakers. The 720 recorded words were evaluated by a different group of 10 NS listeners for their judged goodness of production. The evaluation task consisted of a 3-point rating scale indicating whether the recorded word was “easy,” “difficult,” or “impossible” to understand. Utterances with more than one rating of “impossible” were excluded from the training corpus. Due to some inconsistency in NS ratings, some utterances with a single rating of “impossible” were included, since the bulk of the other ratings were “easy.” In these cases, the “impossible” rating was treated as an outlier. The final training corpus consisted of 245 words containing the tap /ɾ/ and 245 containing the trill /ɾ/.

2.5 Training

Training consisted of eight 1-hour sessions conducted over a 2-3 week period. The training task was a 2AFC task with feedback. Each day 50 words from each corpus (trills
and taps) were randomly selected for training. As with the ID and JG tasks from the perceptual mapping, the trill and tap targets were labeled “Sound A” and “Sound B.” Each stimulus from the training set was presented for identification as Sound A and Sound B. Correct responses were reinforced with a cash register sound and in increment in “points earned.” A running total of points earned was visible throughout the training. Incorrect responses resulted in the participant hearing “Listen again” instead of the cash register sound, followed by a repeat of the missed word. No points were awarded for incorrect responses.

### 2.6 Production

Participants were recorded imitating Spanish words prior to the first day of training, between days four and five of training, and after the last day of training. The words included five each of the tap and trill stimuli at the ends of the continuum, followed by five words each from the tap and trill training corpora. The stimuli were presented at a fixed rate of 25 words per minute, with participants told to repeat each word they heard before the next one was played.

Stimuli were normalized individually and were played through Sennheiser HD280 headphones set initially for a 75dB loudness level. While listening to the first several stimuli, participants were allowed to adjust volume to a “comfortable” listening level. This also allowed participants to familiarize themselves with the rate of presentation.
Each recorded response was normalized then cropped to leave only the rhotic and the vowels occurring on either side. The cropping was done by hand, using visual and auditory cues to minimize the influence of coarticulation from other sounds before the onset vowel and after the offset vowel. Cuts were made at zero crossings to prevent audible clicks during file playback.
3.1 Model Fitting

The results of the ID and the two JG tasks were compared for each participant to a theoretical model of phonological perception. The theoretical model, by providing a point of comparison, allows for interpretation of changes in perception as having directionality. Both a strong and weak model were devised for each task. The difference between the two models is determined by ratings at both ends of the continuum. Even with strong phonological perception, the stimuli in the middle of the continuum are not easily discriminated, and should not differ significantly in their ratings from ratings by participants with weak phonological perception. Given that we are dealing with perceptual space, which may be quite different from acoustic space, boundaries between any two sounds on a continuum cannot be determined based on acoustics alone. Rather, performance by a variety of individuals determined to have stable 2-category perception of the target phonemes (i.e., native speakers) was used to determine where we would expect there to be stable perceptual judgments.

For the ID task, the bistable range was set at 95% identification. In other words, tokens that were identified as one or the other sound 95% of the time were considered to have stable perceptual ratings. The ratings by six of the NS listeners were averaged, and
the three shortest stimuli were identified as taps 95% of the time, while the four longest stimuli were identified as trills 95% of the time. Therefore, the theoretical model for ID was specified accordingly, as seen in Figure 5.

![Figure 5](image)

Figure 5. ID performance of the six NS listeners, and the resultant model used to define strong phonological perception. The data are graphed as a percentage identified as a trill. In a 2AFC design, anything not identified as one choice is necessarily identified as the other. Therefore, the three shortest stimuli, with percentage scores under 5% were identified as taps 95-100% of the time. The intermediary stimuli are not specified in the model, as they are expected to be “unstable,” and by definition, non-specific.

For the JG tasks, there was less consistency in the break points evidenced by the NS participants. As with ID, the trill end of the continuum was comprised of a greater number of stimuli than the tap end. However, even among the NS group, this ranged from 3-6 items for the trill, and either 2 or 3 for the tap. The most conservative of the ranges was chosen for the theoretical model, with the stimuli closer to the unstable boundary on either side reflecting the mixed boundary seen among the NS participants (Figure 6).
Figure 6. NS Group performance and corresponding theoretical models for the two Judged Goodness tasks. Goodness of Fit is measured by calculating the RMSD from the 7 points specified by the model, which for the JG-Trill task, would be 1, 1, 1.5, 6.5, 7, and 7. The ratings of 1.5 and 6.5 in the theoretical model capture the variability in performance of the NS group, while retaining seven values for model fitting.

In addition to providing directionality, the theoretical models were used to measure “distance”, which was useful for interpreting changes in perception as well as for categorizing participants’ perception of the non-native phonemes as strong or weak. For these purposes, the goodness-of-fit was assessed for participant performance relative to the theoretical models. Root Mean Square Deviation (RMSD) is commonly used to measure goodness-of-fit in situations where there is a need to quantify deviation from
exact values (Schunn & Wallach, 2005). RMSD was calculated for the difference in the ratings for the seven specified points in the model, providing a single measure of similarity to the theoretical model. An error value of 0 would be attained if all seven points were exactly the same as those in the theoretical model. While the actual values are derived from the original raw score and are therefore meaningful units with the same scale (percent; 7-point) as the data they are used to describe, they are used here only in a relative context, to contrast greater or lesser correspondence with the strong phonological model (Figure 7).

![Figure 7. An example of each of the JG tasks, and the distance from the specified values in the theoretical model. These difference scores are used to calculate the RMSD. The phonological model is superimposed, and the RMSD for the JG-Tap task on the left is 1.74. The RMSD for the JG-Trill task (right) is 1.12, representing less discrepancy from the theoretical model.](image)

### 3.2 Dynamics of Learning

The similarity matrices generated by the Difference Ratings task provide a measure of perceived difference between two stimuli or groups of stimuli. While the
acoustic continuum is linear and evenly-spaced, perceived differences among the stimuli may be quite different. Multi-dimensional scaling is a method of converting difference ratings so that they can be expressed in terms of Euclidean distance. This allows interval comparisons of perceptual ratings for measuring change. Dimension scores can be plotted on a 1-dimensional axis. In such a configuration, the closer together two points are, the more similar the stimuli were perceived to be. Figure 8 is an example of how the output from a MDS analysis might be graphed.

![Figure 8](image)

Figure 8. Example of how MDS results could be plotted. The top row represents the acoustic stimuli. Acoustically, the stimuli in this continuum are evenly-spaced in a specific order, represented by the numbered balls. The bottom row represents the results of the MDS analysis of a set of difference ratings. This particular example is what one might expect when the two ends of the continuum exhibit the Perceptual Magnet effect - stimuli at the ends of the continuum are perceived as more similar to one another than stimuli in the middle. In this case, stimulus 3 was also perceived as being more similar to stimulus 1 than was stimulus 2.

For this analysis, similarity matrices were converted to dissimilarity objects and subjected to a one-dimensional MDS analysis using the Kruskal algorithm with stress formula 1, with the speech analysis software package Praat (Boersma & Weenink, 2010).

Difference ratings were collected each day of training, resulting in 10 one-dimensional lines of data. Three elements of the MDS solutions were measured - ordering of stimuli as perceived by the participant, clustering of stimuli, and day-to-day
variability in distance among stimuli. MDS solutions are inherently non-directional, so the same perceived relation among stimuli could on one day show up in the order 1,2,3,4,5,6,7, and the next 7,6,5,4,3,2,1. This would, of course, result in high day-to-day variability of ordering that would be an artifact of the measurement tool, rather than a reflection of changes in the participant’s perception. Therefore, direction for each day’s data was corrected to minimize variability, such that any remaining day-to-day variability was reflective of actual perceptual changes, rather than the idiosyncrasies of the method of analysis.

Day-to-day variability of the MDS scores was calculated by subtracting each day’s dimension score for each stimulus from the dimension score for the subsequent day. The squared difference scores were then summed and the standard deviation for each set of change scores calculated.

Ordering was measured by calculating the RMSD for the difference between each stimulus’s acoustic ordinal and their perceptual ordinal. Clustering was expressed as the percent of the total range occupied by the three stimuli at each end of the continuum. Therefore tight, distinct clustering of the three stimuli at each end of the acoustic continuum that are far apart would comprise a small percentage of overall perceptual space. The same clustering closer together would comprise a larger percentage of overall perceptual space. Figure 9 shows how clustering scores would be determined for results from a single session.
Figure 9. A sample day’s MDS data. Acoustic ordering is represented by the numbers, and perceptual ordering by the physical location of the circles on the axis. With only one stimulus differing between acoustic and perceptual ordering, RMSD is low (0.53). The range for Sound A is roughly a third of the overall range and shows no warping of perceptual space. The range for Sound B is evidence of tighter clustering, indicating perceptual warping.

3.3 Production

The pre-, mid-, and post-training recordings of each participant’s productions were rated by NS listeners and subjected to acoustic analyses. The recordings were trimmed so that only the rhotic and adjacent vowels remained. They were then played to the NS listeners who were asked to identify the rhotic as either a trill, a tap, or a non-Spanish sound, such as a alveolar approximant r or a dental /d/. Recordings were randomized and presented to the NS listeners through headphones. They were asked to identify each stimulus as a Spanish tap, a Spanish trill, or some other, non-Spanish sound. Since the terms “tap” and “trill” are not familiar to all NS speakers, they were also described as “single r” and “double r” and example words were written on the buttons on the computer screen –
“mira” and “carro.” Ratings data are expressed as percent correct, reflecting the portion of total ratings that correctly identified the intended target.

Acoustic analyses were conducted using the Praat software. Auditory and visual displays of each VCV recording were used to label the acoustic components in a TextGrid object. The acoustic elements were comprised of the vowels on either side of the rhotic (V1, and V2), occlusions (O), apertures (A), constrictions (CON), and sibilance (SIB). Figure 10 shows an example of the labels for one recorded utterance.

Figure 10. An attempted trill production (“arri”, excised from “arrisco”) by an NE speaker. The manner of production was a vibrant trill, with the final “occlusion” constricting but not occluding the airflow. It was thus labeled “CON” for constriction. The intensity contrast between occlusion (“O”) and aperture (“A”) can be seen in both the lighter area in the spectrogram (middle) and the smaller cycles in the waveform representation (top).
The labels were used to identify the manner of production (fricative, continuant, etc.) and to identify segments for further acoustic analyses. Specifically, if there were any apertures in a trill production attempt, the duration of the 1st aperture and the mean frequency of the first three formants across all apertures were measured.
CHAPTER 4: Results

4.1 Perceptual Profiles

Perceptual profiles were created for 73 student volunteers, using the ID and two JG tasks. To participate in the training portion of the study, participants needed to demonstrate the ability to discriminate between the tap and trill ends of the continuum, yet have room for improvement in their perception of within-phoneme category structure. As described earlier, the theoretical model for strong identification was comprised of seven points - the three shortest stimuli (0% identified as trill) and the four longest stimuli (100% identified as trill). An RMSD value of zero represents no deviation from the strong theoretical model. RMSD values greater than 0.45 were considered to indicate no ability to parse the tap and trill ends of the continuum. Eight NE participants had RMSD values greater than 0.45. Three of these participants showed similarly poor discrimination on the JG tasks, and were excluded from participation in the training portion of the study. Prior research has shown that participants who lacked initial structuring of the two sounds exhibited no improvement as a result of training. Figure 11 shows the ID results for the strong and weak NE participants, as well as the NS evaluators.
Figure 11. Performance on the ID task for the weak NE performers, strong NE performers and NS participants

Of the 70 participants who showed good performance on at least one of the three tasks, 18 were judged to have little room for improvement in their performance on the JG tasks. An independent samples t-test was performed to determine whether there was a significant difference in the JG performance of the 18 participants judged to be “too good” and the rest of the screened participants. The performance of the two groups was significantly different on both the JG-Tap \(t = 6.68, p < 0.01\) and JG-Trill \(t = 6.67, p < 0.01\) tasks. An independent samples t-test was also performed to compare the “too
good” NE participants with the NS participants. The “too good” NE participants and the NS participants were not significantly different in their distribution (t = 0.64, p > 0.1).

The mean RMSD values for the three groups are listed in Table 1. Graphs of group performance on the two JG tasks for the “too good” participants, the remaining participants and the NS evaluators can be seen in Figure 12.

Table 1. Means and standard deviations of RMSD on the two JG tasks for three groups of participants: NS listeners, “too good” NE listeners and eligible NE listeners.

<table>
<thead>
<tr>
<th>Task</th>
<th>NS Group</th>
<th>“Too Good” NE Group</th>
<th>Eligible NE Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>JG-Tap (RMSD)</td>
<td>M = 0.68, SD = 0.44</td>
<td>M = 0.79, SD = 0.29</td>
<td>M = 1.69, SD = 0.82</td>
</tr>
<tr>
<td>JG-Trill (RMSD)</td>
<td>M = 0.96, SD = 0.56</td>
<td>M = 0.86, SD = 0.29</td>
<td>M = 1.59, SD = 0.60</td>
</tr>
</tbody>
</table>

From the pool of 52 eligible participants, 21 completed the training portion of the study. An independent samples t-test was conducted to compare the training group to the eligible candidates who did not participate in the training. The results revealed no significant difference in distribution of RMSD on either the JG-Tap (t = 1.39, p > 0.1) or the JG-Trill (t = 1.33, p > 0.1) task.
Figure 12. Averaged JG ratings for the training-eligible NE participants, the “too good” NE participants and the NS evaluators. The shapes of the ratings by the “too good” group are very similar to those of the NS group. In contrast, the eligible participants’ performance is characterized by a shallower slope, indicating less defined internal structure for the two sounds.
4.2 Perceptual Training Results

The RMSD values on the perceptual mapping tasks used to select training participants were also used to measure changes over time. Correspondence with the strong phonological model was measured before training, on day 6 of training, and after training. Progress was defined as a reduction in RMSD between the participants’ performance and the theoretical phonological model.

Another metric was used to determine if there was significant discrimination between the stimuli at each end of the continuum. To determine significant discrimination, the mean response for each of the three stimuli at one end of the continuum was compared with the distribution of responses for the three stimuli at the opposite end of the continuum. If that mean score fell outside the 95% confidence range for the opposite end population, it was considered to be significantly discriminated from the other stimuli.

Seven training participants showed significant discrimination for all 6 end stimuli on both the JG tasks. An additional five participants showed significant discrimination for at least three of the six stimuli for each JG task. Four participants showed significant discrimination on one or no stimuli for each JG task.

Figure 13 shows the JG-Trill results for one participant for whom the stimuli at the tap end were significantly different from the stimuli at the trill end, but for whom the reverse was not true. The distance between the means for the end stimuli seems substantial, but the z-scores for all three trill stimuli were not significantly different from
the distribution of tap ratings. This is due to the variance in responses at the tap end of the continuum. This participant was consistent in judging the trill stimuli as good examples of a trill, but was quite inconsistent in her ratings for the stimuli at the tap end of the continuum, which were sometimes rated as very poor examples of the tap and sometimes rated as very good examples. This is consistent with the hypothesis that the native phonology is acting as an attractor for the the more similar tap than the less similar trill, resulting in less stable perception of the tap.

This effect was seen in five participants on the JG-Trill task, always with the greater variability on the tap end of the continuum. Of the six values for each participant (126 total), 50 of the taps were significantly discriminated from the trills, while only 35 of the trills were significantly discriminated from the taps.

The effect was seen in three participants for the JG-Tap task as well, with two of the three showing more significant taps than trills, and one showing more significant trills than taps. However, for the JG-Tap task, this was an isolated phenomenon, and the group totals for this task show the same number of taps and trills (38) as being significantly discriminated from the other end of the continuum.

In general, there was significantly more variability in ratings at the tap end of the continuum than the trill end, regardless of which JG task was being conducted (Figure 14). This shows that despite considerable individual differences, there was a significant group effect of the destabilizing influence of the NE on the assimilable tap rhotic phoneme.
### Table

<table>
<thead>
<tr>
<th>“tap” Stimuli</th>
<th>“trill” Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>428ms</td>
<td>461ms</td>
</tr>
<tr>
<td>432ms</td>
<td>465ms</td>
</tr>
<tr>
<td>435ms</td>
<td>466ms</td>
</tr>
</tbody>
</table>

**Mean:**
- Mean “tap” Stimuli: 2.10, 4.20, 3.70
- Mean “trill” Stimuli: 7.00, 6.60, 6.90

**SD:**
- SD “tap” Stimuli: 2.34
- SD “trill” Stimuli: 0.73

Figure 13. Mean ratings for this participant show greater variability at the stimuli at the “tap” end of the continuum than at the “trill” end of the continuum. This suggests stronger definition of the trill phoneme than of the tap phoneme.

Figure 14. Average standard deviations for the end stimuli in both JG tasks for group of 21 training participants. When averaged across tasks, the mean standard deviation of ratings for the three tap-end stimuli (428ms, 432ms, 435ms) was 1.26, versus 1.08 for the trill-end stimuli (461ms, 465ms, 466ms). While the main effect of stimulus set was significant (F = 9.965, df = 1, p < 0.05), the difference was more pronounced for the JG-Trill task.
4.3 Production

Along with the ID and JG ratings, tap and trill production attempts were recorded before, during and after training. This was to see if there were any changes in production that accompanied changes in perception. First, there is the question of whether there is any correlation between initial production and perception competence. As mentioned earlier, perception competence is measured using RMSD values, so the lower the value, the closer the fit to a theoretical ideal of phonological perception. The most basic measure of production competence was NS categorization of NE production attempts. These scores are reflected as a percentage of NS listeners who rated each production as its intended target - the tap or the trill. Therefore, if production and perception competence were closely related, we would expect to see a negative correlation between the two measures. Figure 15 shows the correlations between these two measures taken prior to the onset of training. None of the correlations is significant in either direction.
Figure 15. Success of initial production attempts for taps and trills was not significantly correlated with performance on either of the two JG tasks.

**Item Analysis**

Half of the productions for each phoneme were of the minimal pair “pero/perro,” while the other half were of miscellaneous Spanish words containing the intervocalic tap or trill. All minimal pairs were relatively closely matched for frequency in the Spanish language with the exception of “pero/perro” and “mira/mirra.” However, potential familiarity effects on NS raters were controlled by the fact that NE productions were trimmed to just the VCV portion of the word before NS listeners heard and rated the production attempts. Figure 16 shows the results of an item analysis on individual words.
Of the taps, performance on two of the words ("arisco" and "encerada") was significantly worse than on the other tap stimuli at all three time points ($p < 0.05$). Conversely, performance on two of the trills was significantly better than the other trill stimuli at all three time points ($p < 0.05$). The “pero/perro” minimal pair was in the high performance group of stimuli for both phoneme sets. This is to be expected, given the four extra trials for those stimuli per recording session.

The minimal pair “encerada/encerrada” was very rarely perceived correctly by NS listeners. This is possibly due to the fact that these were the only 4-syllable words in the corpus, making them harder to pronounce regardless of the rhotic used. For all other word pairs, those that had a high correct production rate for the taps had a low correct production rate for the trills, and vice versa. This suggests that phonological context may
have been functioning as an attractor to one or the other phoneme for each of the words. In other words, NE listeners may have categorized each minimal pair as either a tap or a trill (but not both, with the exception of “pero/perro”), making it more difficult to produce the word using the alternative phoneme.

Production Progress

Progress on production differed for the tap and the trill (Figure 17). For production of the tap, small but significant improvement in production was seen between pre-training and post-training (F = 6.765, P < 0.05). All gains that were seen by the end of training had been achieved by day six. There was no significant difference in performance between mid- and post-training recordings (p > 0.1). In contrast, there was no significant change in trill production performance among any of the three time comparisons (p > 0.1). Although final mean performance on the taps and trills was not substantially discrepant (tap-39%, trill-37%), there was substantially more variability in trill production performance at all three times. This wide range of scores by different participants invites analysis of other measurements that may be related to performance on production of the trill.

As can be seen in Figure 17, the range of scores for trill production pre-training was greater than the range for tap production. Both sounds seem to be bimodally distributed. A K-Means cluster analysis was conducted, and the results show a greater distance between cluster centers for the trill than for the tap. Results are in Table 2.
Figure 17. Individual and group progress on NS ratings of production. Progress on the tap (left) was significant between pre-training and mid-training. Progress on the trill was highly variable and not significant.

Final performance on trill production was analyzed for correlation with initial performance, final performance and improvement on the two JG tasks. All correlations
were between -0.23 and 0.54, and were not significant. Correlations with MDS results were not significant for clustering or acoustic ordering of taps and trills.

Table 2. K-means cluster analysis shows the cluster centers for each mode of production ratings for the two targets.

<table>
<thead>
<tr>
<th></th>
<th>Tap Production</th>
<th>Trill Production</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low (n)</td>
<td>High (n)</td>
</tr>
<tr>
<td>Pre</td>
<td>20% (11)</td>
<td>48% (8)</td>
</tr>
<tr>
<td>Post</td>
<td>21% (8)</td>
<td>51% (11)</td>
</tr>
</tbody>
</table>

Acoustic Analyses

Six acoustic analysis were conducted on the trill production attempts, and two on the tap attempts. Because there was a wide range of ability in producing the target Spanish sounds, not all tokens were produced with the same manner of production. This is important for two reasons. The first is that the manner of production in large part dictates the acoustic analyses that can be conducted. For example, the canonical vibrant production of the trill includes periods of occlusion that can be measured for duration. However, many trill utterances were produced as approximants or fricatives, which have no measurable period of occlusion. In addition to determining which acoustic measurements are appropriate, identifying the manner of production may show production changes over time that do not manifest in higher ratings by NS listeners.
Therefore, the distribution of manner over time, as well as the ratings by manner were analyzed.

Based on auditory, waveform and spectrogram information, each utterance was broadly classified as a fricative, stop, approximant or vibrant. Appendix A shows examples of the four different manners of production.

For the purposes of this analysis, the term “vibrant” refers to productions containing two or more occlusions or constrictions. This is in contrast to the phoneme “trill”, which would be perceptual and contextually-determined, regardless of its manner of production. Similarly, the term “stop” in this analysis is inclusive of “taps,” which have been described in the literature as being rapid, incomplete stops. Fricatives were those productions that had sibilance with no accompanying occlusion.

For both tap and trill targets, the largest number of pre-training production attempts were produced as approximants. This is consistent with the type of transfer error expected in the early stages of L2 phonological learning (Major, 1986), assuming that the NE speakers were approaching the targets from the initial space of the American English /l/. While this would be expected for the trill, which is not part of any American English phoneme category, it is interesting that this was also true for the tap. The tap is an American English allophone of the intervocalic alveolar stop (e.g., “butter”, “bidding”), but is not an allophone of the approximant /l/. It could be reasonably assumed that the NE speakers would hear the target word as containing the alveolar stop, resulting in the bulk of productions being stops, rather than approximants. However, the
fact that the majority of tap production attempts were approximants suggests that context played a substantial role in initial classification of the tap rhotic. Both the task context and the phonological context within each target word were likely to have attracted the participants to the approximant rather than the stop. The taps and trills were intermingled for all perceptual and production tasks. Since the trill is widely thought to be a variant of /ɾ/, that could have influenced participants to assume the tap was a variant of /ɾ/ as well. Phonologically, the alveolar tap in American English is always followed by an unstressed syllable containing one of the following lax vowels: [ɪ, ə, ɜ]. In contrast, the target Spanish words followed the tap with [o, i, a], all of which constitute new phonological contexts for the alveolar tap. The novel phonological context for the alveolar tap in the Spanish words would likely influence participants away from the acoustically equivalent American English phoneme.

Figure 18 shows the proportions of productions of each type over time. Initially, both targets were produced as approximants more often than any other manner. Over the course of training, the number of approximants declined, with a corresponding increase in the number of vibrant trills for the trill targets and stops for both targets.
Figure 18. Distribution of manner of production across time for each of the two phonological targets. Most notable is the large number of approximants at pre-training for both targets. By post-training, approximants have decreased, and significant increases are seen in the canonical manner of production for the trill (vibrant) and tap (stop).

A Wilcoxon Signed Ranks Test was conducted to determine the significance of any changes in the proportions of manner of production over time. As seen in Table 3, there were significant changes in manner of production between pre- and post-training samples. With significant increases in production of trill targets as vibrants and tap targets as stops. Significant declines were seen in production of both targets as approximants.
Table 3. Trill attempts that were produced as vibrant trills rose significantly from 18% to 24% over the course of training. Those that were produced as approximants fell significantly from 42% to 29% over the course of training. The increase in stop productions was not significant, and there was no change in fricatives.

<table>
<thead>
<tr>
<th>Manner of Production</th>
<th>Percent Pre</th>
<th>Percent Post</th>
<th>Direction</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trill Target</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>42%</td>
<td>29%</td>
<td>Decrease</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Vibrant</td>
<td>18%</td>
<td>24%</td>
<td>Increase</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Stop</td>
<td>37%</td>
<td>46%</td>
<td>Increase</td>
<td>p = 0.08</td>
</tr>
<tr>
<td>Tap Target</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>63%</td>
<td>47%</td>
<td>Decrease</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td>Stop</td>
<td>36%</td>
<td>53%</td>
<td>Increase</td>
<td>p &lt; 0.05</td>
</tr>
</tbody>
</table>

A one-way ANOVA was conducted to determine if the manner of production of trill targets was predictive of correct identification by NS listeners. Results were significant (F = 40.62, df = 3, p < 0.05). In the pre-training recordings, vibrants were identified by the NS raters as being trills substantially more often than any of the other manners of production. Approximants were least often identified as trills. For tap targets, all but one of the utterances were categorized as stops or approximants. An independent samples t-test was conducted to determine if stops were significantly more likely to be identified as taps than were approximants. Results were significant, with taps being substantially more likely than approximants to be identified as taps (T = 4.66, df = 178, p < 0.05). The correct identification levels for both targets can be seen in Figure 19.

Although both production targets show increases over time in the proportion of attempts produced with the canonical manner of production, only the tap targets showed a significant increase in correct identification by NS evaluators. This could be related to the size of the change, since there was only a 6% increase in trills, compared to a 17%
increase in stops. Even though there was no significant increase in NS ratings of NE production attempts for the trill, this shift in manner of production suggests that there were changes over time in production that brought participants closer to correct trill production, even if these changes were not sufficient to manifest in the gross perceptual categorization of attempts as trills.

Figure 19. Pre-training relations between manner of production and identification as a trill (left), and tap (right). Trill targets that were produced as approximants were least likely to be correctly identified as a trill, while, unsurprisingly, vibrants were the most likely to be correctly identified as a trill. Tap targets produced as a fricative were the least likely to be correctly identified as taps. No tap targets were produced as a vibrant.

Many of the acoustic measures of trill production success reported in the literature (Carballo & Mendoza, 2000; Diaz-Campos, 2008; Johnson, 2008; Reeder, 1998; Weech, 2009) presume periods of occlusion and aperture, and therefore, can only be conducted
on productions that contain such acoustic elements. For these measurements, only those targets produced as vibrant trills are included (n: pre=33, mid=43, post=43).

In contrast to the shifts in manner of production, those productions that were vibrant trills did not show any acoustic shifts towards more competent production. A repeated measures ANOVA was conducted to determine if there were changes in the duration of the first aperture, the number of apertures and the mean frequencies of the first 2 formants. As seen in Table 4, there were no significant changes in any of these measures over the course of training. In other words, while there was an increase in the number of utterances produced using the most intelligible manner of production (vibrant), there was no change in the quality of those vibrant trills to make that subset of productions more competent. This is partly due to the already high intelligibility rate for the trills (79%).

Table 4. Acoustic measures shown to differentiate among levels of competency among NS and L2 learners show no significant changes from pre- to post-training levels.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre Mean SD</th>
<th>Mid Mean SD</th>
<th>Post Mean SD</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 Duration</td>
<td>25.52 0.013</td>
<td>25.33 0.012</td>
<td>27.64 0.013</td>
<td>0.423</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td># Apertures</td>
<td>Mean 1.73 1.33</td>
<td>Mean 1.22 0.57</td>
<td>Mean 1.16 0.48</td>
<td>4.948</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Mean F1</td>
<td>Mean 550.42 252.54</td>
<td>Mean 554.36 219.96</td>
<td>Mean 603.31 184.62</td>
<td>0.740</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Mean F2</td>
<td>Mean 1.833 412.72</td>
<td>Mean 1.879 381.77</td>
<td>Mean 1.785 368.86</td>
<td>0.625</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>
Duration and intensity of the occlusion during stop production were measured to determine if either was related to correct NS ratings as a tap. Given that a tap is generally faster, with less occlusion than a stop, one might expect there to be a negative correlation between duration and NS ratings and a positive correlation between intensity and NS ratings. As seen in Table 5, neither correlation was significant. Table 5 also shows that there was no significant change in duration or mean intensity of stops over the three time periods.

The lack of the correlation between production ratings and acoustic characteristics of the stops produced is likely due to the specific acoustic measurements taken and the phonetic context in which the phonemes were produced. For example, NE speakers tended to produce the word “pero” using the medial vowel /ɛ/ or the diphthong /ɛI/, whereas NS speakers tend to use the flat /e/. It is likely that the nonnative preceding vowel influenced NS listeners’ perception of the subsequent consonant as a nonnative phoneme different from the Spanish tap rhotic.

Table 5. Duration and intensity were not significantly correlated with NS ratings for either approximants or stops (upper table). There were also no significant changes in average duration or intensity of stops over time (lower table).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Approximants</th>
<th>Stops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>r = -0.12</td>
<td>r = -0.10</td>
</tr>
<tr>
<td>Intensity</td>
<td>r = -0.03</td>
<td>r = 0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre</th>
<th>Mid</th>
<th>Post</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.22</td>
<td>16.58</td>
<td>23.70</td>
<td>14.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22.16</td>
<td>10.50</td>
<td></td>
<td></td>
<td>1.027</td>
</tr>
<tr>
<td>Intensity</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>74.12</td>
<td>4.12</td>
<td>74.00</td>
<td>3.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>73.79</td>
<td>4.03</td>
<td></td>
<td></td>
<td>0.146</td>
</tr>
</tbody>
</table>
4.5 Learning Dynamics

MDS analyses were conducted daily and provide a global view of perceptual space for each day of training, as well as the day-to-day changes in perceptual space that occur between the pre- and post-training perceptual profiles. This is the measure that provides insight into the dynamics of the learning process for nonnative phoneme acquisition. The MDS data serve to illuminate patterns that differentiate participants based on their learning dynamics.

Previous research (Tuller, et al., 2008) saw a clear categorization of learning dynamics that differentiated three groups of learners. These groups were differentiated in part by the dynamics seen in the MDS analyses and by changes in the JG and ID tasks before and after training. A similar differentiation of learning is apparent in the MDS data from this study. There were three broad categories of learning, based on the variability of their difference ratings throughout the training. One group closely resembles the Phonological learning group. This is the only group that made significant improvements in their performance on the JG tasks. The other two groups can be dubbed “non-learners” for their lack of perceptual changes by the end of training. They differ, however, in their initial performance on the perceptual tasks, as well as their dynamical patterns over the course of training. The group with strong initial perceptual skills is referred to as the “Refiners,” while those with weak initial perceptual skills are the “Non-Learners.” When participants are broken out into these three groups (n = 7 each),
we see significant differences in a number of measures. Table 6 outlines the primary characteristics of each group.

Table 6. Three learning profiles and their distinguishing characteristics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Phonological Learners</th>
<th>Refiners</th>
<th>Non-Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial MDS Clustering</td>
<td>Weak</td>
<td>Strong</td>
<td>Weak</td>
</tr>
<tr>
<td>JG Improvements</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ID Improvements</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Final JG RMSD</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Final Ordinals</td>
<td>Strong</td>
<td>Strong</td>
<td>Weak</td>
</tr>
<tr>
<td>Final MDS Clustering</td>
<td>Strong</td>
<td>Strong</td>
<td>Weak</td>
</tr>
<tr>
<td>MDS Variability</td>
<td>Single Burst, then stable</td>
<td>Consistently Low</td>
<td>Consistently High</td>
</tr>
<tr>
<td>Production</td>
<td>Variable</td>
<td>Variable</td>
<td>Variable</td>
</tr>
</tbody>
</table>

**Phonological Learners**

Seven of the 21 participants showed a profile consistent with what Tuller has termed the “Phonological Learner.” The learning dynamics for this group follow the same pattern seen in previous work with other phonemes: a period of increased variability in MDS dimensions followed by consistently low variability for the remainder of the training.

Day-to-day variability in difference ratings indicates the stability of perceptual categories over time. An increase in variability has been interpreted as a sign of a phase transition, where there is a sudden shift in perception. In this case, the phase transition represents the destabilization of a single perceptual category as part of the bifurcation into two distinct perceptual categories (Tuller, 2008). There is evidence of a similar
process here. Figure 20 shows the dimension scores and standard deviation of day-to-day difference scores.

Figure 20. MDS dimension scores (top) and variability scores (bottom) throughout training for one participant in the phonological group. All dimension scores were normalized to a scale ranging from -0.75 to 0.75.

On day 1, the stimuli are perceived to be equidistant from one another and ordered with little respect to the acoustic ordering. The variability chart shows that between Day 1 and Day 2 there was not much change in this participant’s difference ratings. But as the learning process continued, the participant exhibited highly variable ratings in days 3-6. By Day 7, perceptual space had stabilized, and was maintained throughout the remainder of the training. This is consistent with variability patterns seen
with Phonological Learners in prior research. All seven participants in the phonological group exhibited a similar pattern of phase transition - a single burst of instability followed by stability, though the time period of increased variability differed among participants.

![Graph showing MDS results from another participant in the phonological group, with dimension scores on the top graph, and variability ratings on the bottom. This participant showed greater perceptual separation between stimuli at either end of the continuum.]

Figure 21. MDS results from another participant in the phonological group, with dimension scores on the top graph, and variability ratings on the bottom. This participant showed greater perceptual separation between stimuli at either end of the continuum.

Similarly, another Phonological Learner (Figure 21) starts out with little grouping and a lack of correspondence with acoustic ordering. Days 1-3 exhibit high instability as the single undifferentiated phoneme category bifurcates. From day 4 on, there is relative stability of the two perceptual categories. During the final two days of training
perceptual ordering becomes closely aligned with acoustic ordering, and it appears that perception of stimulus #3 is becoming attracted to the other group of stimuli.

**Refiner**

The defining characteristic of the Refiner profile is consistently low variability throughout training. Figure 22 shows a participant who appears to perceive the acoustic differences well on Day 1. Perceptual order was very closely aligned with acoustic order and the differences were relatively evenly spaced. By Day 4, we begin to see a warping of perceptual space, and by Day 6 there is evidence of very strong phonological perception of the trill. In spite of these changes over time, there is almost no change in the standard deviation. Although there is tightening of one phoneme category and some movement in the ordering of stimuli, neither of these changes is contingent upon the bifurcation of an existing phoneme category. As a result, there is no burst of variability as seen in the Phonological Learner.
Non-Learners

In spite of attempts to screen out participants unlikely to make any changes over the course of training, a third of the training participants exhibited no improvements in any of the perceptual tasks, and fit the definition of the “non-learner.” These participants exhibited high variability throughout the training, suggesting that they never achieved stable representations of the two phonemes.
Figure 23. MDS Variability for Non-Learners. Variability is high throughout the training, the end stimuli are never clustered together, and there is no improvement in ordinals.

Figure 23 shows an example of the MDS results for one of the non-learners. Although there is some clustering in the beginning, Stimulus 6 is in the tap cluster, and stimulus 1 is in the middle between the two clusters. There are many changes from day to day, and the final dimension scores show no improvement in ordinal rankings and no clustering of the end stimuli.

**Group Comparisons**

In addition to patterns of variability, there are seven measures that serve to differentiate the three groups. In a number of respects, the Phonological Learners resemble the non-learners prior to the onset of training and resemble the Refiners at the end of training. At
the start of training, all three groups exhibited weak clustering of the tap end of the continuum. While the phonological and non-learners also exhibited weak clustering at the trill end of the continuum, the Refiners had significantly tighter clustering - tighter than the acoustic model would predict, suggesting more categorical perception. By the end of training, both the Phonological and Refiners had tight clustering of both phonemes, in contrast to the Non-Learners, who exhibited no change in clustering (Figure 24).

Pre-training, all three groups had similar RMSD values relative to the acoustic ordering of the stimuli. Though no groups made significant changes over time in their ordinals, by the end of training, the non-learners had significantly larger RMSD values relative to the acoustic order relative to both the phonological and Refiner groups (Figure 25).

Figure 24. Perceptual clustering of the stimuli at each end of the continuum for the Non-Learners (NL), Phonological Learners (Phon) and Refiners (Ref).
There were no significant differences among the groups in the pre-training performance on the JG or ID tasks. Paired t-tests were conducted on the pre- and post-training RMSD values for each group to see if significant change occurred over the course of training (Table 7).

Table 7. Paired t-test results for all three training profiles.

<table>
<thead>
<tr>
<th>Group</th>
<th>Task</th>
<th>Mean Change</th>
<th>t</th>
<th>df</th>
<th>sig</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonological Learners</strong></td>
<td>ID</td>
<td>0.21</td>
<td>2.70</td>
<td>6</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>JG-Tap</td>
<td>1.34</td>
<td>2.6</td>
<td>6</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>JG-Trill</td>
<td>0.83</td>
<td>3.32</td>
<td>6</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td><strong>Competent Non-Learners</strong></td>
<td>ID</td>
<td>0.23</td>
<td>3.54</td>
<td>6</td>
<td>p &lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>JG-Tap</td>
<td>0.78</td>
<td>2.06</td>
<td>6</td>
<td>p = 0.07</td>
</tr>
<tr>
<td></td>
<td>JG-Trill</td>
<td>0.56</td>
<td>1.94</td>
<td>6</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td><strong>Non-Learners</strong></td>
<td>ID</td>
<td>0.10</td>
<td>1.39</td>
<td>6</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td></td>
<td>JG-Tap</td>
<td>0.05</td>
<td>0.17</td>
<td>6</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td></td>
<td>JG-Trill</td>
<td>0.01</td>
<td>0.08</td>
<td>6</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>
The Phonological Learners made significant reductions in RMSD on all three tasks. The Refiners showed significant RMSD reductions on the ID task but not the two JG tasks, and the Non-Learners made no significant changes on any of the three tasks (Figure 26). A repeated measures ANOVA was conducted to compare the scores of the three groups for the three perceptual profile tasks. Pre-training differences among groups were non-significant. Post-training differences between the Non-Learners and the other two groups were significant at the p < 0.05 (df = 2) level for all three tasks (ID: F=4.85, JG-Tap: F=14.77, JG-Trill: F=6.71).
Figure 26. Identification and Judged Goodness ratings for the tap and the trill pre- and post-training.
Production was highly variable within each group, and there were no significant differences between the groups on production success. Figure 27 shows NS ratings for production attempts broken down by the three groups.

![Figure 27. Changes in production ratings over time. There were no significant differences across groups or times.](image)

**Prediction**

Tuller’s prior work with Malayalam showed two distinct groups of participants, based on the significant discrimination from the median score (4, on a 7-point Likert scale). Here, the initial perceptual mapping results on the tap and trill are not as clearly categorized, as there seems to be a continuum of performance, regardless of the measure used. With the introduction of a number of new metrics to this task, it is important to
look at the value of each metric in predicting changes in perceptual space as a result of training, as well as dynamical patterns as shown by the Difference Ratings analysis.

Pre-training, RMSD on the ID task was not correlated with either of the two JG tasks, nor was it predictive of any of the post-training tasks. The RMSD on the pre-training JG-Tap and JG-Trill tasks were highly correlated with one another (0.74), and were moderately correlated with post-training JG-Trill RMSD, but not with post-training JG-Tap RMSD. Both were highly correlated with their respective change scores from pre- to post-training. Taken together, these correlations suggest that those participants with the most room for improvement made the most improvement, but not enough to be even with those who started with the least room for improvement.

Pre-training RMSD scores were also used to categorize participants into strong and weak groups. This was done by contrasting RMSD relative to the theoretical model with RMSD relative to the flat model, which is a rating of 4 (the mean) for all stimuli. Participants with a RMSD relative to the theoretical model that was lower than their RMSD relative to the flat model were in the “strong” group, and those with a lower flat RMSD were in the “weak” group. In spite of weak correlation between the pre-training JG-Trill RMSD scores and production ratings (r = 0.32), the relation between the categorical grouping of JG-Trill performance and post-training production ratings was significant (Figure 28).
Figure 28. Participants in the weak JG-Trill RMSD group performed better on the Trill Production task both pre-training and post-training. However, only the post-training was significant ($t = 2.158$, $p < 0.05$), and correlations were weak for both times.

MDS Analyses - MDS analyses of Difference Ratings results gave several different metrics. The first is the amount of clustering for each end of the continuum as a percentage of the total range. Tap clustering is measured by the total range of the first three stimuli, and trill clustering is measured by the total range of the last three stimuli. With seven stimuli, the evenly-spaced acoustic distances from stim1-stim3, stim3-stim5, and stim5-stim7 would each be a third of the overall range. Therefore, percentage of the
total range occupied by each end of the continuum gives a sense of how much warping of perceptual space there is for each phoneme.

Tap clustering was not significantly correlated with either of the JG tasks, the production tasks, or the MDS variability patterns. Trill clustering, however, was significantly correlated with post-training performance (RMSD) on both the JG-Tap and JG-Trill tasks (Figure 29). Trill clustering was also the only pre-training measure that was associated with a particular learning pattern as measured by day-to-day MDS variability. This suggests that more stable non-native phonological perception prior to the onset of training is predictive of successful and stable perception of both phonemes after training.

Figure 29. Participants with tighter clustering (i.e., lower percent of overall range) of the trill before training exhibited lower RMSD relative to the theoretical model on both JG tasks after training.
Clustering and ordinal figures for the three training groups are in Figure 30. Both phonological and Non-Learners exhibited a significantly wider perceptual range for the last three stimuli than did the Refiners.

An independent samples t-test showed no significant difference in pre-training trill clustering between the phonological and Non-Learners. This is consistent with the idea that tight clustering is evidence of phonological perception, and those who already
exhibit that type of clustering prior to the onset of training have no need to go through the phonemic category destabilization process, indicated by a period of high variability in the Difference Ratings results. The pre-training loose clustering of the trill for both the phonological and Non-Learners suggests that neither group initially perceived the non-native phoneme in a phonological manner. The Phonological Learners, however, developed a stable category as a result of training, as evidenced by the post-training perceptual profiles. The Non-Learners, on the other hand, never developed a new category, as evidenced by the weak clustering relative to the other two groups on the post-training perceptual tasks. The contrast in MDS patterns over the course of training is related then to having learned phonological perception of the target phonemes.

Participants who exhibited day-to-day stability after a phase transition experience a change in their perception, while those who exhibited continuous instability throughout training, made no such perceptual change.
CHAPTER 5: Discussion

The analyses conducted as part of this study were designed to test predictions made by multiple theories of speech perception and production. The tap and trill rhotics were used to examine a number of issues related on non-native phoneme learning in adults, including a) the effect of native phoneme categories on the learning of a non-native phonemic distinction, b) if different modes of learning were used by different training participants, c) if those modes of learning could be predicted from pre-training perceptual profiles, d) if production is affected by perceptual learning and e) what metrics were most sensitive to that learning and the processes involved therein.

5.1 Modes of Learning

Prior research on NE listeners’ discrimination between the dental and alveolar stops in Malayalam showed two distinct patterns of initial perceptual space among learners of the phonemic distinction. What distinguished the two patterns was variability in Difference Ratings as well as performance on the JG tasks.

Using the same methodology on a different phoneme pair with different relations to AE attractors, a third of the present participants exhibited all the characteristics of what was termed the “Phonological Learner.” As with the Malayalam pair, Phonological
Learners as a group evidenced initial discrimination between the tap and trill on both JG tasks and the ID task. Also evident was the distinctive pattern of variability in the Difference Ratings, where at some point in the training there is a period of increased variability followed by low, stable variability for the remainder of the training. This has been interpreted as evidence of a phase transition in perception, and is a sign seen in other non-linear dynamical systems. These learners all made significant changes in their JG and ID performance to more closely correspond with the phonological model by the end of training. In addition to the above, Phonological Learners also showed significant changes in the new metric of clustering, i.e., the percentage of perceptual space occupied by the three stimuli at each end of the continuum. Prior to training, the Phonological Learners exhibited no warping of perceptual space for the trill, and very weak clustering of the tap.

These measures taken together support Tuller’s interpretation of the dynamics of phonological learning involving the bifurcation of existing perceptual space to accommodate a new phonemic distinction. Prior to training, these participants had moderately good acoustic discrimination of the continuum as evidenced by their performance on the JG and ID tasks. However, after the phase transition, marked by the period of high variability in the Difference Ratings, they showed significant changes on every perceptual training measure with the exception of ordinals. These changes, combined with the post-training tight clustering at both ends of the continuum suggest strong categorical perception of both phonemes as well as good within-category
structuring. In essence, these participants developed phonological perception of a native/nonnative phoneme contrast over 10 days of exemplar training.

In contrast, the Acoustic Learner described by Tuller was not fully evident with the present sound pair. The Acoustic Learners and the Refiners both exhibit consistently low variability in the Difference Ratings throughout training. There are, however, two key differences between the two groups. The acoustic learners began training with JG scores that indicated little discrimination between the two ends of the continuum, with a post-training shift to better discrimination, but not as good as that of the Phonological Learners. In contrast, there was no difference in the pre-training JG tap/trill performance between the Phonological Learners and the Refiners. In addition, the Refiners’ JG performance did not show the same shift towards the phonological model seen in the acoustic learners. The high pre-training discrimination seen in the JG tasks was also evidenced by the tight clustering seen in the Difference Ratings analysis for the trill. In many respects the Refiners resemble the “too good” participants that were excluded from the training portion of the study. They had the strongest pre-training performance, no evidence of phonological destabilization and very strong, but only slightly improved, post-training performance on all perceptual measures. In contrast, the acoustic learners, though they had similarly stable dynamics, did make substantial changes in their perception over time, improving within-category perception.
5.2 Non-Assimilable Phonemes and Task Difficulty

The fact that both learning groups in the present study showed strong initial perceptual structuring of the tap-trill phonemes is consistent with predictions by the PAM model of the ease with which Non-Assimilable (NA) phonemes like the trill for NE listeners are discriminated relative to Categorized (C) phonemes like the dental stop in Malayalam. Due to the lack of phonological similarity, the native phoneme exerts no perceptual influence on the nonnative phoneme, leaving the listener with “unwarped” perceptual space with which to make judgements. Prior to the onset of training, NE listeners actually had more consistent ratings of the non-native sound than the native sound. This is seen in the standard deviation scores for each end of the continuum on the JG tasks showing significantly more consistent responses across learning groups for the trill stimuli than for the tap. The PAM model predicts that this is due to the fact that the native phoneme “suffers” from categorical perception, and that warping of perceptual space makes it difficult to rate sound variants purely on their acoustic properties.

Given the ease with which this NA-C phoneme pair was discriminated relative to the Single Category (SC) Malayalam pair, it was a bit surprising that a full third of the participants failed to develop stable perception after 10 days of training. This group of Non-Learners was still different from Tuller’s non-learners - the non-learners for Malayalam were distinguishable from the other two groups by their initial lack of structuring of perceptual space. Although the rhotic Non-Learners had less tight clustering of the trill than the Refiner group, their clustering was no different from that of
the Phonological Learners, and there were no significant differences among the groups on any of the three perceptual profile tasks (ID, JG-Tap, JG-Trill). RMSD comparisons with a theoretical model of phonological perception and significant differences between stimuli at each end of the continuum both failed to distinguish among the training groups. The only pre-training measure that was sensitive to the three training patterns was clustering of the NA phoneme in the Difference Ratings. This suggests that the Judged Goodness and Identification tasks may not be sensitive to pre-training differences among learners of an easier phoneme pair such as the tap and trill rhotics. Although the Difference Ratings task did illuminate a difference for one of the training groups (Refiners), it may take an even more fine-grained discriminatory task to differentiate potential learners from potential non-learners.

While it may be the case that all non-assimilable phonemes are equally well perceived by non-native listeners, there are factors other than assimilability that may contribute to the NE participants’ pre-training perception of the trill. The first factor is that NE speakers are generally familiar with the Spanish trill. Even those with no exposure to functional communication in Spanish have probably heard in the media the cry “¡Arriba!” produced with an exaggerated trill. That familiarity may be enhanced by the unique mechanics of trill production, which prompt many people with no understanding of Spanish to attempt to produce a trill. This may help explain why internal structuring of this nonnative sound was as good as the native tap for these NE listeners.
Another factor to consider is the phonological role of the tap. In Spanish, it is spelled with the letter “r” and is considered an ‘r’ sound. The acoustically very similar tap in American English is spelled with the letters “t” or “d” (usually doubled) and is considered a ‘t’ or ‘d’ sound. Perceptual mapping tasks in this study did not provide linguistic context that would color the sounds as “r”s or “d”s, so it is possible that participants experienced perceptual influence from either of those two AE phonemes. Furthermore, the phonological context of “pero” is one that is not found in AE. The tap in AE is always followed by an unstressed syllable in which the vowel is mostly reduced to either /ə/ or /ɪ/. Therefore, for some participants the tap might not have even been perceived as a NE phoneme. Due to the incongruity of the acoustics of the sound (native) and the phonological context in which it was presented (nonnative), the tap may have had no assimilatory effect, making the tap-trill distinction, using Best’s coding scheme, NA/U (Uncategorized) or even NA/NA.

It is possible that the pre-training production attempts could provide clues as to the initial perception of the tap by the NE listeners. Interestingly enough, only one participant produced all initial taps using the same manner of production (the approximant). There were no group differences in manner of production. All three groups initially produced “pero” using a stop an average of 35%-37% of the time, and an approximant 61%-65% of the time. Looking at participants who initially produced tap words using an approximant 70%-100% of the time, three were in the Phonological Learner group, three were in the Non-Learner group and three were in the Refiner group.
Grouping participants by predominant initial manner of production also yielded no significant differences in performance on any of the perceptual tasks at any of the three time points. This could mean one of two things - either production does not provide insight into which is the assimilative native phoneme, or it does, but which phoneme the tap assimilates to does not affect any of the measures used in this study.

5.3 Production

Production was analyzed to measure the degree to which it paralleled changes in perception over time. As a whole, participants made improvements on all perceptual tasks. There was also a significant improvement in ratings of tap production attempts, and significant changes in the manner of production for both taps and trills. So, in general, production changes do seem to accompany perception changes. However, when looked at individually, or when broken into groups based on learning style, there is only one correlation between perception and production improvements. Interestingly, both measures of tap production, NS ratings and the number of tap production attempts produced as stops, at the end of training were correlated with performance on the JG-Trill task. Since the positive correlation is with the RMSD value of the model fit, this means that those participants who ended up with weaker structuring on the JG-Trill task had better production of the tap than those with stronger structuring.

What does this say about production? We see that tap production improved more than trill production. On the surface, this could be assumed to be due to greater
familiarity with the “native” tap than with the nonnative trill. Or it could be due to the fact that trill production is motorically a more complex task (Solé, 2002). However, for these explanations to be true, the differences should be most apparent prior to training, shrinking over the course of training, which is not the pattern evidenced in these data. Both phonemes were initially produced accurately (as rated by NS listeners) 32% of the time. By the end of training, they differed by a non-significant 2%. However, at all times, trill production ratings were more variable than tap ratings, and that variability is a contributing factor to the lack statistical improvement over time.

While articulatory complexity may not directly explain the differences between tap and trill production success, it can explain the high variability at all three times for trill production. The highest and lowest percent correct production scores were both for trills. In tap production, the range was smaller and the groupings more uniform. Since all participants had very similar levels of exposure to the Spanish rhotics, we are left with possible explanations of perceptual attraction and articulatory complexity. Perceptual attraction and articulatory complexity together can explain both the high degree of individual differences in trill production and the lower scores for the best performers on tap production.

While theories of speech perception deal primarily with the acoustics of the target phoneme, no one would argue that the phonological context does not play a role as well. Since the tap in AE is usually followed by a reduced vowel, the unreduced /o/ in “pero” could be an attractor for the alveolar approximant /ɹ/ or the alveolar stop /d, t/, both of
which occur far more frequently in that phonological context than the tap. That attraction would tend to reduce overall success of tap productions. Conversely the trill, being a Non-Assimilable nonnative phoneme, is by definition, not constrained by prior experience with phonological context. That is, novel speakers are not going to take cues from the surrounding vowels as to how to produce the trill. Phonological context is similar for all participants, and any effect would be a downward bias, depressing scores overall. Motor coordination for speech production, on the other hand, is more susceptible to individual differences, without the “unifying” bias in one direction of an attractive phonological context. This explanation rests on the idea that the forces of perceptual assimilation apply to production in a manner similar to that applied to perception. It also rests on the idea that different phonemes have different motoric requirements that can mitigate the effects of perceptual assimilation.

The item analysis of production data in this study suggests that individual words may vary in the way the sounds in them are assimilated. For one stimulus minimal pair, “arrisco/arisco,” production of the trill was always significantly more successful than production of the tap. For three of the other minimal pairs with similar phonological contexts, it was production of the tap that was always significantly more successful. This suggests that production stimuli became paired with one end of the continuum, in some sense, preventing successful simultaneous pairing with the contrasting phoneme.

One key variable is that there was no explicit production instruction during the perceptual training. Of course, the mere act of saying the utterances for the purpose of
recording is a form of training, as it provides immediate feedback about the success of the articulatory gestures in duplicating the desired acoustics. The effect of this practice can be seen in the difference in the successful production of “pero” and “perro” relative to the miscellaneous taps and trills. There were five trials of each of those targets in each recording session, and only one trial of each of the other five targets (per phoneme; 10 total). Both “pero” and “perro” were in the significantly better group of targets, though not significantly different from the other targets in that group.

In spite of differences among participants, stimuli and target phonemes, there exists a general relatedness of changes in perception and production. The individual data, however, clearly show that these changes are not in lockstep. Individual analyses indicate that while some participants were improving just their perception, others were improving just their production, and yet others were doing both simultaneously.

This is not unexpected, but also not very well explained by existing models of speech production. In general, models of speech production acquisition can be divided into articulation models and perception models. The perception models discussed here (SLM, PAM, NLM) address production in terms similar to those of perception - namely, the more perceptually distinct phoneme is easier to discriminate, and therefore, easier to produce than the phoneme more likely to be assimilated into an existing L1 category. Flege’s SLM model of speech perception addresses production in one of the stated hypotheses: “The production of a sound eventually corresponds to the properties represented in its phonetic category representation.” Although this does provide for
delay of production acquisition relative to perceptual discrimination, there is still the
prediction that the trill would be easier to produce than the tap, which was not the case in
this study.

Articulation models predict a different pattern. Eckman’s Structural Conformity
Hypothesis (SCH) invokes typological markedness to explain acquisition differences
among different phonemes (Eckman, 1991). In essence, because the tap is produced in
AE as an allophone of the alveolar stops, it is “unmarked”, relative to the trill, which does
not occur as a production in AE. Similarly, Brown’s theory of segmental perception
looks at the articulatory features of a phoneme (Brown, 2000). The AE and Spanish tap,
in spite of phonological context differences, consist of the same articulatory features, and
accordingly, should be easy for MNE speakers to produce. The trill, on the other hand, is
produced in the vibrant manner, which is not part of the AE articulatory feature set. This
lack of familiarity and experience would make the trill more difficult for the adult MNE
speaker to produce.

The Direct-Realist approach to speech perception (Fowler, 1996), suggests that
the acoustics of speech provide a direct representation of the gestures of the speaker.
However, there are multiple gestures that can be used to produce the same acoustics, such
as formant patterns (Ohala, 1996; Riordan, 1977). It is possible that some acoustic
properties provide a more direct link to gestures than others. Take, for example, the AE
alveolar approximant /ɹ/. It is the most common sound for which NE children need extra
instruction to master. If there are speech primitives, like occlusions, then full occlusions,
as are part of stops, may be more primitive in that they are motorically easier to master, and easier to hear acoustically. Therefore the acoustics of the /ɹ/ are less directly mapped to the speech gestures than the earlier developing sounds.

In contrast, Kello (Plaut & Kello, 1999) posits that the link between perception and production is learned through a feedback mechanism. The /ɹ/ problem is similar within this framework. If a learner tries to produce an acoustic target and fails, that feedback directs them to adjust their gestures to see if they do better the next time. If the motor constraints of that sound are difficult, it will take more trials before the correct gesture is found.

Regardless of the approach, motor constraints or poor acoustic-motoric mapping can account for the lack of synchrony in development of perception and production. Recently there have been calls for a more complex model that accounts for more than differences between phonemes or individuals. Colantoni (Colantoni & Steele, 2008), used explicit instruction to contrast NE learners’ acquisition of the French rhotic and the Spanish tap. Differential hypothesis testing of different models of L2 phonology failed to capture the influence of positional, contextual and differing phonetic parameter constraints in predictions of production success.

The current data are not inconsistent with a model that combines both perceptual and articulatory constraints. Neither the tap nor the trill was successfully produced more than 40% of the time, so clearly there are constraints on both that interfered with “native-like” production. Articulatory complexity and inexperience offer possible explanations as
to why successful trill production was below 40%, in spite of being an easier perceptual
distinction. Conversely, perceptual assimilation provides an explanation as to why
successful tap production was below 40%, in spite of being comprised of articulatory
gestures that are well honed in native (AE) speakers.

5.4 Conclusion

This study has tested the predictions of a number of perceptual and articulatory
models against a specific Categorized/NonAssimilable phoneme pair. A comparison of
initial perceptual profiles between this study and prior work using a pair that could be
labeled Single Category or Category Goodness supports Best’s hypothesis that a Non-
Assimilable phoneme provides a more easily discriminable contrast with a native
phoneme than either of the other two types (SC or CG). Though the codification of these
contrasts is unique to Best’s PAM, these findings are not inconsistent with other speech
perception models, such as NLM and SLM, that make predictions based on the phonetic
relations between native and nonnative phonemes.

Analysis of the dynamics of learning also support Tuller’s model of phonological
learning, where learning a new phonemic distinction requires a phase transition during
which existing phoneme categories are destabilized and bifurcate to accommodate the
new one. Acoustic learning, where within-category perception improves, but which does
not entail a phase transition was not evident with this phoneme contrast. The acoustic
learning process may be unique to SC phoneme contrasts, where some learners are unable
to perceive enough of a distinction to begin the phonological learning process. A phoneme pair involving a NA phoneme may provide enough clarity that all participants capable of learning within-category structure are also capable of creating a new phoneme category.

Analysis of production attempts for the two phonemes showed evidence of both perceptual and articulatory constraints. Initial trill production success seemed to contradict perceptual models suggesting that the NA phoneme would be easier to produce. However, the best performers on trill production were more successful than the best performers on tap production and the variability among participants was greater. Therefore, NA phonemes may be easier to produce in the absence of other articulatory constraints that differentially affect individuals. Initial tap production attempts, on the other hand, seemed to contradict articulatory models suggesting that the Spanish tap would be no more difficult to produce than the AE tap. Taken together, these differences indicate that articulatory and perceptual constraints were both in play, and that production success is only as good as the most constraining factor, which for the trill was articulatory and for the tap was perceptual.

Over the course of 10 days, perceptual training consisted of 5,000 trials, in contrast to just the 40 production trials. With this heavy emphasis on perceptual training over articulatory training, one would expect the phoneme more constrained by perception to improve more than the phoneme more constrained by articulation complexity, which is what occurred.
In addition to the methods replicated from prior studies, some new metrics were introduced here. Comparing RMSD of participant JG and ID responses to a theoretical model of phonological perception provided a single value that could be used to compare participants and measure directionality of change. The RMSD value at post-training differentiated Non-Learners from both Phonological Learners and Refiners, and changes in RMSD over time differentiated Phonological Learners from the other two groups. However, pre-training RMSD was not able to distinguish any of the three groups from each another. This suggests that the components of perceptual space useful in predicting learning patterns for SC (and CG) distinctions are less useful for other types of contrasts. One useful new metric was tightness of clustering of the NA phoneme in the Difference Ratings, which differentially distinguished Refiners from Phonological Learners and Non-Learners.

There are limitations to this study as well. Dispersion Theory posits that the phonological space of a language is derived from the frequently conflicting need for maximal perceptual contrast and minimal articulatory effort. This suggests a strong interplay of the entire phonological landscape. As a result, the discrimination of nonnative phonemes should be influenced by more than the perceptual space of the one native phoneme presumed to be the most similar. In addition, the development of a new phonological category, as seen in a third of the participants in this study, may involve adjustments in, if not outright bifurcations of, more than a single native phoneme category. As a result, mapping a greater range of perceptual space involving a number of
native phonemes may shed light on processes that are not readily apparent using a 1-dimension continuum between a nonnative and selected native phoneme. An example is the potential influence of both the alveolar stop and approximant on perception and production of the Spanish tap.

Another area of need is the development of a theory of speech production and perception that accommodates the interplay of articulation, phonetics, and phonology in the learning of new phonemes. The results of this study are best explained by the differential effect of different constraints on each of the phonemes. This includes the articulatory effort of the trill and the phonological context (surrounding vowels and prosody) of the tap, as well as the acoustic properties of each sound. The beginnings of several theories have been postulated (Colantoni & Steele, 2008; Diehl, Lotto, & Holt, 2004; Plaut & Kello, 1999), and they hold great promise for uniting these various constraints without the need for generative universals.

Finally, there is a need for continued development of innovative metrics that can be used to define the complexities of perceptual space, encompassing its myriad interactions and dimensions. The use of MDS to illuminate one-dimensional relations among stimuli over time, borrowed from others for this research, provides an exciting way to visualize and quantify the dynamics of perceptual learning. The application of similar methodologies to higher dimensions of speech perception as well as the dynamics of production learning can only lead to the continued maturation of the study of this most basic and essential function of human cognition.
APPENDIX A

Manner of Production (Waveforms and Spectrograms)
APPENDIX B

Phonological Profile (MDS)
APPENDIX C

Dimension Scores

Variability

Ordinals

Clustering

Refiner Profile (MDS)
APPENDIX D

Non-Learner Profile (MDS)
REFERENCES
REFERENCES


CURRICULUM VITAE

Gregory G. Anderson received his Bachelor of Arts in Speech & Hearing Sciences from the University of California, Santa Barbara in 1985, and his Master of Science in Speech-Language Pathology from the University of Arizona in 1989. After 16 years as a licensed speech-language pathologist, he returned to school and received his Ph.D. in Psychology from George Mason University in 2011.