Phonetic Training Strategies for Non-Native Speech Perception

Katherine Phelps Ridgeway

University of Colorado at Boulder, k1000ridgeway@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/ling_gradetds

Part of the Cognitive Psychology Commons, and the Linguistics Commons

Recommended Citation

https://scholar.colorado.edu/ling_gradetds/59

This Dissertation is brought to you for free and open access by Linguistics at CU Scholar. It has been accepted for inclusion in Linguistics Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Phonetic training strategies for non-native speech perception

by

Katherine Phelps Ridgeway

B.A., California State University, Chico, 2005

M.A., University of Colorado, Boulder, 2013

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

2016
This thesis entitled:
Phonetic training strategies for non-native speech perception
written by Katherine Phelps Ridgeway
has been approved for the Department of Linguistics

Prof. Rebecca Scarborough

Prof. Lewis O. Harvey Jr.

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

IRB protocol #15-0048
Native language experience affects the perception of non-native sounds and can interfere with learning sounds in a new language. Targeted training with acoustic dimensions that underlie phonemic contrasts can help improve non-native speech perception. The assumption of most phonetic training studies to date is that training strategies should mirror the perceptual strategies of native speakers. The ways in which a learner’s first language may influence their ability to utilize specific acoustic cues is rarely taken into account when deciding on a phonetic training strategy. This dissertation research directly addresses this oversight. Two perceptual training experiments were run in order to test the efficacy of acoustic training dimensions that differed in regards to prior perceptual experience of the learners. Specific mechanisms of category learning are demonstrated through a variety of statistical approaches, including a novel application of multi-dimensional scaling. The results of these experiments show that it is important to consider both prior perceptual experience and the specific features of target contrasts when developing a phonetic training strategy. These results have implications for models of speech perception based on attention to acoustic cues, and can inform theories of how native language experience affects non-native speech perception.
Dedication

To my husband Karl and our parents, Cindy & Kelly Phelps and Gretchen & Peter Ridgeway, whose love and support have made this dissertation possible.
Acknowledgements

This research was partially funded by grants from the Institute of Cognitive Science, University of Colorado, Boulder.

I would like to thank my advisers, Dr. Rebecca Scarborough and Dr. Lewis O. Harvey Jr., for their guidance and support throughout my doctoral research. Many thanks also to my research assistants, Story Kiser, Jane Chai, Matt Becker, and Jayne Williamson-Lee. I am grateful to Gary McClelland for good discussions about multi-dimensional scaling. Finally, I would like to acknowledge Kit LaTouche, Allie McCarthy, Stras Acimovic, Ben Warren, Nick Williams, and Lee Altamirano, for encouraging me to write even when (and especially when) I didn’t want to.
# Contents

## Chapter

1. **Introduction and Background**  
   1.1 Non-native speech perception  
   1.2 An Attention-to-Dimensions model of speech perception  
      1.2.1 Selective attention  
      1.2.2 Expansion and compression of relevant and irrelevant dimensions  
      1.2.3 Acquired similarity and distinctiveness  
   1.3 A2D training studies  
      1.3.1 What is the “right” dimension?  
   1.4 Accounting for L1 experience: models of non-native speech perception  
      1.4.1 The Perceptual Assimilation Model  
      1.4.2 Native Language Magnet Model  
      1.4.3 The Speech Learning Model  
      1.4.4 Similarities between PAM, NLM, and SLM  
      1.4.5 Comparing models of non-native speech perception and category learning  
   1.5 Toward an integrated model of non-native speech perception and more effective training strategies  
   1.6 Dissertation outline
Experiment 1: Xorx and Bjaran

2.1 Study design and procedure
   2.1.1 Xorx training
   2.1.2 Xorx difference rating
   2.1.3 Bjaran condition assignment
   2.1.4 Bjaran training
   2.1.5 Bjaran difference rating

2.2 Stimuli
   2.2.1 Xorx stimuli
   2.2.2 Bjaran stimuli
   2.2.3 “Same dimension” condition
   2.2.4 Familiar Dimension condition

2.3 New Dimension condition
   2.3.1 Stimuli creation

2.4 Subjects

2.5 Predictions for Experiment 1

2.6 Results
   2.6.1 Xorx category training
   2.6.2 Xorx difference ratings
   2.6.3 Category learning in Xorx: Acquired Similarity and Acquired Distinctiveness
   2.6.4 Bjaran training
   2.6.5 Bjaran difference ratings
   2.6.6 Category learning in Bjaran

2.7 Discussion
   2.7.1 Xorx results summary
   2.7.2 Bjaran results summary
2.7.3 Which training dimensions are most effective? ......................... 66
2.7.4 Linguistic relevance of stimuli in Experiment 1 ......................... 68
2.7.5 Aliens vs. Humans (or, The relationship between Bjaran training con-
ditions and English) ......................................................... 69
2.7.6 Implications of Experiment 1 for phonetic training in real languages . 71

3 Experiment 2: Training adults to hear a Korean voicing contrast ............. 72
3.1 Korean stop consonants .................................................................. 73
  3.1.1 Acoustic dimensions underlying the Korean stop contrast .............. 75
  3.1.2 Cue weighting for Korean stop contrast ...................................... 78
3.2 Relationship of Korean stop cues to English .................................. 79
3.3 Study Design and Procedure ........................................................... 81
  3.3.1 Subjects ................................................................................. 82
  3.3.2 Training ................................................................................. 82
  3.3.3 Testing .................................................................................... 85
3.4 Stimuli ........................................................................................... 87
  3.4.1 Test Stimuli ............................................................................. 87
  3.4.2 Training Stimuli ...................................................................... 92
  3.4.3 Category structure for training stimuli ....................................... 93
  3.4.4 Creation of training stimuli ....................................................... 94
3.5 Results .......................................................................................... 100
  3.5.1 Korean category training ........................................................... 101
  3.5.2 Korean identification task ........................................................ 105
  3.5.3 Korean difference ratings ......................................................... 112
  3.5.4 Category learning in Korean: Acquired Similarity and Acquired Dis-
tinctiveness ................................................................................. 120
3.6 Discussion ..................................................................................... 133
3.6.1 Overview of results ........................................ 133
3.6.2 Which training dimensions are most effective? ............ 138
3.6.3 Phonetic training for Korean stops ........................ 140
3.6.4 Voice quality as a training dimension ..................... 143
3.6.5 L1-L2 relationships and phonetic training .............. 145

4 General Discussion and Conclusions ............................. 147
4.1 Research questions, revisited .................................. 148
4.2 Will perceptual training result in better perception of non-native sounds? ........................................ 148
4.3 Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others? Given a learner’s L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language? ........................................ 149
4.3.1 The acoustic dimensions learners are already familiar with from their L1 ........................................ 150
4.3.2 The acoustic dimensions that native speakers use ......... 153
4.3.3 Are the predictions of Experiment 1, based on artificial languages, borne out in cases of training with real languages? (RQ 3) 155
4.4 Theoretical implications ......................................... 157
4.4.1 A “dimensional approach” to the development of new sound categories ........................................ 157
4.5 Methodological contributions for phonetic training studies .................................................................. 161
4.6 Future work and applications .................................... 161
4.6.1 Subjects with different native languages .................... 161
4.6.2 Time course of training ....................................... 162
4.6.3 Motivation and learning context ............................. 162
4.6.4 Type of training ............................................... 163
4.6.5 Production .................................................. 164
4.6.6 Applications ............................................... 164
Tables

Table

1.1 Comparison of four speech perception models. Dashes (—) indicate cases where a characteristic or theoretical position is not an explicit part of a model. 13

2.1 Experiment 1 Design 20
2.2 Bjaran training conditions and L1-L2 relationship 21
2.3 Regression model for within-category shift of Xorx stimuli 44
2.4 Regression model for between-category shift of Xorx stimuli 46
2.5 Regression model for Bjaran category training, p values for z-statistics in the table are not corrected, see text for F-statistics with corrected p values 48
2.6 Bjaran Acquired Similarity regression model; p values for t-statistics in the table are not corrected. See text for F-statistics and corrected p values 61
2.7 Effectiveness of training dimensions 67
3.1 Minimal triads for Korean stops in initial position (Cho et al., 2002; Francis and Nusbaum, 2002; Kim et al., 2002) 75
3.2 Contrastive dimensions for Korean stops 76
3.3 Overview of design for Experiment 2 81
3.4 Training conditions and dimensions for experiment 2 82
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Fixed effects from Korean category training regression model, predicting accuracy with block number and contrast codes for training conditions. The coefficients in a logistic regression model represent the change in log odds as values increase along a continuous predictor, or from one level of a contrast coded predictor to another level. P values for z-statistics are not corrected. See text for F-statistics and corrected p values.</td>
</tr>
<tr>
<td>3.6</td>
<td>Fixed effects from regression model for Korean identification task, predicting accuracy categorical variables for training condition. The regression coefficients in this model represent the change in log odds when moving from one level of a contrast coded predictor to another level.</td>
</tr>
<tr>
<td>3.7</td>
<td>Coefficients from individual regression models where acoustic variables are regressed on D1 and D2 coordinates from the group space scaling solution.</td>
</tr>
<tr>
<td>3.8</td>
<td>Intercepts for post-hoc models of within-category shift for each training condition and stop type.</td>
</tr>
<tr>
<td>3.9</td>
<td>Intercepts from post-hoc models of between-category shift for each training condition and centroid pair.</td>
</tr>
<tr>
<td>B.1</td>
<td>Fixed effects from mixed-effects regression model for VOT, with pair 2 as the reference level.</td>
</tr>
<tr>
<td>B.2</td>
<td>Fixed effects from mixed-effects regression model for VQ (F0-VQ array), with pair 6 as the reference level.</td>
</tr>
<tr>
<td>B.3</td>
<td>Fixed effects from mixed-effects regression model for VQ (VQ-VOT array), with pair 6 as the reference level.</td>
</tr>
</tbody>
</table>
Figures

2.1 Xorx stimuli ................................................................. 26

2.2 Stimuli arrays for the three Bjaran conditions .................. 27

2.3 Xorx training accuracy by block (mean accuracy across subjects). Error bars represent $SEM_{betw}$. ................................................ 36

2.4 Xorx training accuracy, first 36 Trials (mean accuracy across subjects). Blue line represents a linear fit; shaded area represents between-subjects standard error. ................................................................. 37

2.5 MDS solution for Xorx difference ratings ............................. 39

2.6 Schematic for within-category distance, where $d =$ the Euclidean distance between the category centroid and a single stimulus. Overall within-category distance was found by taking the mean of all individual distances. For the Xorx stimuli, Euclidean distance was calculated in three dimensions. ...... 42

2.7 Xorx within-category distance, pre- vs. post-training. Asterisks indicate significant difference between pre- and post-training within-category distance. “AS” indicates an Acquired Similarity effect (within-category distances are smaller, post-training). Error bars represent $SEM_{betw}$. ......................... 43
2.8 Xorx between-category distance, pre- vs post-training. Asterisks indicate significant difference between pre- and post-training between-category distance. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent the between-subjects standard error of the mean (hereafter, $SEM_{betw}$).

2.9 Bjaran training accuracy by block (mean accuracy across subjects) for three training conditions. Error bars represent $SEM_{betw}$.

2.10 Training accuracy for the Same Dimension condition, comparing shifted and non-shifted stimuli. Error bars represent $SEM_{betw}$.

2.11 MDS solution for Bjaran difference ratings, “Same Dimension” condition.

2.12 MDS solution for Bjaran difference ratings, “Familiar Dimension” condition.

2.13 MDS solution for Bjaran difference ratings, “Familiar Dimension” condition.

2.14 Bjaran within-category distance, pre- vs post-training. Asterisks indicate significant difference between pre- and post-training in post-hoc tests; periods represent marginal differences. “AS” indicates an Acquired Similarity effect (within-category distances are smaller post-training). Error bars represent $SEM_{betw}$.

2.15 Bjaran between-category distance, pre- vs post-training. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent $SEM_{betw}$.

3.1 Korean letter-color combination for responses in training tasks. Information in parentheses are the IPA and consonant type for that Korean letter, which were not seen by subjects.

3.2 Distribution of test stimuli along acoustic dimensions of VOT, F0, and spectral tilt.
3.3 Category structure for training stimuli in the VOT training condition . . . . 93
3.4 Category structure for training stimuli in the F0 training condition . . . . . 94
3.5 Category structure for training stimuli in the VQ training condition . . . . . 95
3.6 Korean category training: Mean accuracy by training block, split by training condition. Error bars represent $SEM_{betw}$ . . . . . . . . . . . . . . . . . . . . . 102
3.7 Korean identification training: Mean accuracy, split by training condition. Dashed line represents “chance” (.33). Error bars represent $SEM_{betw}$ . . . . . . 106
3.8 Korean identification training: Mean accuracy, split by training condition and consonant type. Dashed line represents “chance” (.33). . . . . . . . . . . . . . . . . . 107
3.9 Confusion plot for VOT condition, showing the proportion of subject responses as a function of stimulus consonant type. . . . . . . . . . . . . . . . . . . . . 108
3.10 “Confusion” plot for F0 condition, showing the proportion of subject responses as a function of stimulus consonant type. . . . . . . . . . . . . . . . . . . . . 110
3.11 “Confusion” plot for Voice Quality condition, showing the proportion of subject responses as a function of stimulus consonant type. . . . . . . . . . . . . . . . . . . . . 111
3.12 Biplot of group space from INDSCAL of Korean difference ratings. . . . . . 113
3.13 INDSCAL results for VOT (Same) condition. Points represent average stimulus coordinate, pre-training. Endpoints of vectors represent average stimulus coordinate, post-training. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 117
3.14 INDSCAL results for F0 (Familiar) condition. Points represent average stimulus coordinate, pre-training. Endpoints of vectors represent average stimulus coordinate, post-training. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
3.15 INDSCAL results for Voice Quality (New) condition. Points represent average stimulus coordinate, pre-training. Endpoints of vectors represent average stimulus coordinate, post-training. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
3.16 Schematic for within-category distance, where \( d = \) the Euclidean distance between the category centroid and a single stimulus. Overall within-category distance was found by taking the mean of all individual distances.

3.17 Within-category distance, pre- vs. post-training, split by training condition.

Error bars represent \( SEM^{betw} \).

3.18 **VOT training condition**: within-category distance, pre- vs. post-training, split by consonant type. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AS” indicates an Acquired Similarity effect (within-category distances are smaller post-training). Error bars represent \( SEM^{betw} \).

3.19 **F0 training condition**: within-category distance, pre- vs. post-training, split by consonant type. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. Error bars represent \( SEM^{betw} \).

3.20 **Voice Quality training condition**: within-category distance, pre- vs. post-training, split by consonant type. Periods indicate marginal difference between pre- and post-training in post-hoc tests. Error bars represent \( SEM^{betw} \).

3.21 Schematic for between-category distance, where \( d_x = \) the Euclidean distance between two category centroids. Overall between-category distance was found by calculating the area of the triangle formed by the three pair-wise distances.

Error bars represent \( SEM^{betw} \).

3.22 Area between category centroids, pre- vs. post-training, split by training condition.

3.23 **VOT training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent \( SEM^{betw} \).
3.24 **F0 training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent $SEM_{betw}$.  

3.25 **Voice Quality training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Error bars represent $SEM_{betw}$.  

B.1 Stimuli used in the pilot experiment (red boxes). Figure shows F0-VOT array; the same middle column and row tested for the other two stimuli arrays (F0-VQ, VQ-VOT).

B.2 D-prime for each dimension in F0-VOT array, averaged across subjects  

B.3 Maximum D-prime by pair (counts), F0-VOT array  

B.4 Mean d-prime for F0 and VOT dimensions, F0-VOT array. Error bars represent $SEM_{betw}$.  

B.5 D-prime for each dimension in F0-VQ array, averaged across subjects  

B.6 Maximum D-prime by pair (counts), F0-VQ array  

B.7 Mean d-prime for F0 and VQ dimensions, F0-VQ array. Error bars represent $SEM_{betw}$.  

B.8 D-prime for each dimension in VQ-VOT array, averaged across subjects  

B.9 Maximum D-prime by pair (counts), VQ-VOT array. Note that pair 7 is missing in both dimensions because in neither case did any subject have pair 7 as their maximum dprime.  

B.10 Mean d-prime for VQ and VOT dimensions, VQ-VOT array. Error bars represent $SEM_{betw}$.  


Chapter 1

Introduction and Background

Distinguishing non-native sounds can be challenging. Adults are often unable to hear non-native phonetic distinctions that are easily heard by native speakers. The ability to perceive non-native contrasts is a foundational step in learning a new language, and is necessary for fluid comprehension, successful production, and the reduction of foreign-accented speech. Even when second language learners become better at hearing sound contrasts in a target language, the achievement of near-native competence in phonetic discrimination and production lags behind mastery of higher order linguistic structures such as morphology and syntax (Flege and Davidian, 1984; Birdsong, 2003; Best and Tyler, 2007).

A central goal of this dissertation research is to train adults to hear sounds in a new language through a process of targeted phonetic training. Phonetic training allows learners to focus on the phonetic details of non-native sounds, which can result in better perception. Most phonetic training methods to date teach learners to use the same contrastive cues as native speakers, without much regard to how a learner’s first language (L1) experience will affect their ability to use these cues effectively. This is despite a large body of evidence which indicates that difficulty in non-native speech perception is largely due to experience with L1 sound categories. The detrimental effect of L1 experience in non-native speech perception is particularly strong for adults, whose deeply ingrained knowledge of L1 sound categories can cause interference when perceiving new speech sounds. Nonetheless, most phonetic training programs designed for adult learners generally lack an explicit focus on L1-L2 relationships.
as a guide to training strategies.

In this dissertation research, the particular relationship between the learner’s L1 and the target language will be taken into account when deciding on a phonetic training strategy. It is hoped that this approach will lead to advances in phonetic training for language learners. A second, but equally important, goal of this dissertation research is to integrate models of non-native speech perception from linguistics and cognitive psychology, which will result in a fuller understanding of non-native speech perception as a form of general category learning.

1.1 Non-native speech perception

The idea that non-native speech perception is difficult for adults is not a new one, and is supported by over 50 years of empirical research in linguistics and cognitive psychology. This phenomenon has been explored with a variety of L1/L2 combinations. Miyawaki et al. (1975), in a well-known study of L2 speech perception, showed that native Japanese speakers cannot reliably hear the distinction between the English alveolar approximate /ô/ and the lateral approximate /l/ (see also, e.g., Best and Strange, 1992; Iverson et al., 2003; Mann, 1986). Several studies have shown that the ability to hear the English contrast between tense and lax vowels (e.g. /i/ vs. /ɪ/) is a function of language background; the tense/lax contrast is difficult for Spanish, French, and German speakers, but to varying degrees, depending on both the learner’s English experience as well as the particular relationship between the learner’s L1 and English (Flege et al., 1997; Kondaurova and Francis, 2008). English speakers have trouble perceiving Korean voicing contrasts (Francis and Nusbaum, 2002), Polish sibilant distinctions (McGuire, 2007), tone contours in Mandarin (Wang et al., 1999) and other non-native sounds (see, e.g., Levy and Strange, 2008; McAllister et al., 2002). There are many other studies that have investigated this effect (see Strange, 1995 for a comprehensive review).

Though adults (particularly novice learners with no experience in the target language) often have difficulties perceiving non-native sounds, infants are quite good at this task.
Lasky et al. (1975) showed that infants raised in a Spanish speaking environment could reliably distinguish between English stop consonants on the basis of voice-onset time, but that Spanish monolingual adults could not. Werker and Tees (1984) found that young infants (age 6-8 months) raised in English-speaking environments could hear phonemic contrasts in both Hindi and Nthlakampx (a Salish language spoken in the Pacific Northwest), but older English infants (around 10 months) and English-speaking adults could not. These and other studies (e.g. Kuhl et al., 1992; Trehub, 1976; Best et al., 1988; Aslin et al., 1981; Best et al., 1995; Rivera-Gaxiola et al., 2005; Tsushima et al., 1994) demonstrate how, over the course of development, children lose the ability to easily distinguish phonemes in a new language. This phenomenon has been characterized as a “reorganization” of perceptual space (Werker and Tees, 1999), constrained by the phonological system of one’s native language. This reorganization does have its advantages: we are “tuned in” to the phonetic details of our native language, which allows for efficient L perception. There is a growing body of evidence to suggest that native language experience facilitates the discrimination of native sounds (Tsao et al., 2006; Kuhl et al., 2006; Sundara et al., 2006). Nonetheless, tuning into native sounds also effectively “tunes out” the phonetic details of non-native sounds, which results in poor non-native speech perception. Native language experience makes learning non-native sounds more difficult, especially in adulthood, because adults have more accumulated experience with native language sounds compared to young children.

1.2 An Attention-to-Dimensions model of speech perception

How does this accumulated native language experience work against adults when they begin to learn the sounds of a new language? One potential explanation is that, when perceiving sounds in a new language, adults selectively attend to contrastive acoustic cues that underlie similar sound contrasts in their L1. This strategy may be problematic for several reasons.

First, the learner’s “native” cues aren’t always useful. For instance, Japanese speakers
fail to attend to the high frequency formant transitions that English speakers use when differentiating between /u/ and /l/. Instead, Japanese speakers seem to be more sensitive to transitions along F2 (mid-range formant frequencies), which may be less relevant for the English distinction (Iverson et al., 2003).

A second problem arises when a common acoustic cue underlies sound contrasts in two languages, but the boundary between sound categories is different in each language. A well-known example is voice-onset time (the time between the burst of a stop consonant and the onset of vocalic voicing) as a cue to stop consonants. The VOT boundary separating voiced and voiceless stops is language-dependent (Lisker and Abramson, 1964). In English, stops with a VOT of 20-40 ms are perceived as “voiceless” (i.e. /p/, /t/, or /k/) while stops with a VOT lower than 20 ms are perceived as “voiced” (i.e. /b/, /d/, or /g/). In other languages, e.g. Spanish, the VOT boundary between voiced and voiceless stops is different, such that significantly shorter VOTs are perceived as voiceless by native Spanish speakers (Flege and Eefting, 1988). An English speaker listening to Spanish, but using the “English VOT boundary”, would be wrong when discriminating between Spanish voiced and voiceless stops (in some cases, at least).

Adults experience interference when perceiving non-native sounds due to perceptual “habits” regarding specific acoustic cues. The idea that selective attention to L1 acoustic cues can result in poor non-native speech perception has been formalized in the Attention-to-Dimensions (A2D) model. The A2D model is a general model of category learning developed in cognitive psychology, in the context of visual learning (see, e.g., Goldstone, 1994, 1998; Livingston et al., 1998; Nosofsky, 1986). It has been subsequently extended to account for effects of L1 experience in non-native speech perception.

One of the fundamental ideas of the A2D approach is that perceptual category learning results in a warping of psychological space. Psychological space is conceptualized as a multi-dimensional similarity space with axes defined by “dimensions”, which are related to the physical features of perceptual stimuli (Goldstone, 1998). For instance, in a similarity space
for simple shapes, such as rectangles, the “dimensions” might be height and width. In speech perception, the “dimensions” are acoustic cues such as VOT, duration, vowel quality, or degree of fricative noise. Note that not all of these cues are necessarily unitary physical dimensions: for instance, a “vowel quality” cue is potentially made up of several dimensions, e.g. frequency continua of individual vowel formants. “Perceptual warping” occurs as a result of category learning via attention to the dimensions of the psychological space. In a “warped” psychological space, perceptual stimuli that are of the same learned category are closer together, and stimuli from different categories are further apart.

1.2.1 Selective attention

Perceptual warping is a consequence of attending to some dimensions while drawing attention away from others. The weighting of attention to these dimensions is determined, in part, by learning. In the process of learning perceptual categories, attention is shifted to those dimensions that are important for distinguishing between categories (the category-relevant dimensions), while attention to category-irrelevant dimensions is decreased. Attention, in this case, is therefore selective: not all possible dimensions are attended to equally. That selective attention is critical in perception is not a novel concept. In the late nineteenth century, William James noted “My experience is what I agree to attend to. Only those items which I notice shape my mind— without selective interest, experience is utter chaos” (James [1890]).

A simplified example of selective attention in speech perception can be demonstrated with vowel discrimination in English. English speakers primarily use vocalic formant cues to determine the quality of a vowel (e.g. /i/ vs. /ɪ/). Variation along other acoustic dimensions, such as sound level, also exists, but is not informative in terms of distinguishing vowel categories. English listeners selectively attend to formant frequencies over other acoustic dimensions, precisely because they are informative for contrasting vowels in word pairs such as “sheep” and “ship”.
1.2.2 Expansion and compression of relevant and irrelevant dimensions

As attention is paid to category-relevant dimensions, they are “expanded”: psychological distance between stimuli in different categories increases along the category-relevant dimension. Irrelevant dimensions, those that are not useful for category distinctions, may be “compressed”: psychological distance between stimuli decreases as attention is drawn away from these dimensions. In the English vowel example, as people acquire the phonetic categories of /i/ and /ɪ/, the psychological distance between instances of the two vowel categories increases along the formant dimension(s). At the same time, psychological distance along the sound level dimension (or any other irrelevant dimension) become smaller.

1.2.3 Acquired similarity and distinctiveness

A consequence of expansion and compression of dimensions is that within-category items become similar to each other. In the A2D model, this “tightening up” of categories is known as “acquired similarity”. In the process of learning the /i/ and /ɪ/ contrast, instances of /i/ will become more similar to other instance of /i/, and instances of /ɪ/ will become more similar to other instances of /ɪ/.

At the same time, between-category items become less similar. This separating of categories in psychological space is known as “acquired distinctiveness”. In the vowel example, all instances of /i/ will become more different from all instances of /ɪ/.

1.3 A2D training studies

The A2D approach is a general model of category learning which has been extended to account for adult non-native speech perception. Much of the speech perception research in the A2D framework involves training adults to hear sounds in a non-native language. The goal of these studies is to improve non-native speech perception by training subjects to attend to the “right” acoustic dimensions. Various training methods are employed, but
most involve a type of perceptual training where subjects are not explicitly told which acoustic dimensions they should attend to in order to distinguish a non-native contrast. Rather, subjects undergo a type of perceptual training based on accuracy feedback. This method helps subjects implicitly learn which dimensions are important for categorization. A2D phonetic training studies often make use of training stimuli that have been manipulated to provide a range of values along various acoustic dimensions. High variability training stimuli, combined with accuracy feedback during the category training, effectively highlights specific acoustic dimensions.

Phonetic training in the A2D framework has been undertaken in a variety of L1/L2 contexts. For instance, Kondaurova and Francis (2010) trained Spanish speakers to attend to spectral quality cues while ignoring duration cues in order to hear the contrast between English tense and lax vowels (/i/ vs. /ɪ/). English speakers were trained to hear a contrast between Polish sibilants by directing attention toward spectral noise, and away from vowel transition cues (McGuire, 2008). Japanese speakers were trained to distinguish between English /ô/ and /l/ by attending to F3 transition cues rather than F2 transition cues (Iverson et al., 2003). English speakers were trained to attend to a combination of acoustic cues for distinguishing between Korean aspirated, strong, and weak stops (Francis and Nusbaum, 2002), a contrast that is also the focus of Experiment 2 in this dissertation.

In general, the results of these studies suggest that targeted phonetic training with acoustic dimensions can lead to modest improvements in non-native speech perception. In all of these studies, subjects are trained to hear a non-native sound contrast by directing attention to the “right” acoustic dimension underlying the target contrast. However, one common assumption underlies these studies: the “right” dimension (or set of dimensions) are those used by native speakers as primary cues for a target contrast. Thus, phonetic training with “native cues” has been the strategy of most attention-based phonetic training thus far.
1.3.1 What is the “right” dimension?

The assumption of most A2D training studies is that the best way to train adults to perceive a non-native contrast is to use the perceptual strategy of native speakers. However, this “what’s good for the goose...” approach overlooks a crucial point: adults learning to hear a non-native contrast have their own L1, which makes them different from a native speaker in terms of their accumulated phonetic experience. So, for an adult learner with a robust phonological system already in place, are the “native speaker” dimensions always the right ones for effective phonetic training?

There are actually several possible phonetic training strategies. First, it may be the case that the most effective training dimension is, as most have assumed, the “native” dimension, i.e. the primary contrastive cue used by native speakers of the target language. However, a more important consideration may be to choose a training dimension that doesn’t interfere with a learner’s L1 experience. A training strategy that requires a listener to make new boundaries along an already-attended dimension (as in the case of English speakers try to learn a Spanish stop contrast based on VOT) might cause interference. In other words, perhaps there is a cost associated with shifting a native language boundary. The most effective training dimension may be one that is familiar to learners because it occurs in their native language, but is not used as a primary cue to a similar native phonetic contrast. Finally, completely novel dimensions—those that don’t occur as a contrastive cue in the L1—might make for very effective training dimensions.

The motivation for these hypotheses is based on the fact that language learners indeed have strategies for hearing non-native sounds that naturally differ from native speakers. For instance, Spanish speakers seem to attend to duration as a contrastive cue for the tense/lax distinction in English instead of vowel quality cues ([Kondaurova and Francis, 2010]). Rather than training Spanish speakers to hear English sounds in the same manner as an English speaker, phonetic training with other cues (like duration) might result in better perception.
This approach leverages the perceptual habits of learners in order to help them become better at perceiving non-native sounds, rather than working against natural tendencies.

1.4 Accounting for L1 experience: models of non-native speech perception

The A2D model as applied to non-native speech perception has many strengths, such as a focus on domain-general cognitive processes that underlie speech perception and other types of category learning, as well as an exploration of how perceptual dimensions relate to physical ones. However, one potential weakness is its failure to explicitly address how the relationship between the phonetic categories in a learner’s L1 and those of the target language will affect learning in specific ways. Three other models of non-native speech perception reviewed here, the Perceptual Assimilation Model (Best 1994), the Native Language Magnet Model (Kuhl 1992), and the Speech Learning Model (Flege 1995), do address the systematic phonetic relationships between L1 and non-native speech categories in the context of learning. These models all make explicit predictions about how a new sound or phonemic contrast will be learned, based on L1-L2 interactions.

1.4.1 The Perceptual Assimilation Model

The “Perceptual Assimilation Model” (PAM) was developed by Catherine Best. Best et al. (1988) found that some phonemic contrasts, e.g. Zulu clicks, were more easily distinguished by non-native speakers compared to other non-native contrasts. This finding led eventually to the formulation of PAM (Best 1994), which makes predictions about how non-native contrasts will be perceived, based on their ability to be assimilated into the L1 phonological system. According to PAM, experience with L1 sound categories will aid discrimination of a non-native contrast when the non-native sounds are separated by native phonological boundaries (Best et al. 2001). This is called a “Two Category” contrast in the PAM model, because each pair of the non-native contrast assimilates to two robust L1 categories. Discrimination of a non-native contrast may be hindered if two non-native sounds are
heard as equally acceptable exemplars of the same L1 category, or when a pair of sounds are heard as better and worse exemplars of the same L1 category (called “Same Category” and “Category Goodness” contrast types, respectively). Finally, some non-native contrasts are “Nonassimilable” (Best et al. 2001), meaning that they bear no resemblance to any native sound categories. These types of contrasts will be easily distinguished by non-native speakers (as in the case of Zulu clicks). PAM is primarily concerned with “naive” non-native speech perception (i.e. novice learners).

1.4.2 Native Language Magnet Model

The Native Language Magnet Model (NLM) was developed by Patricia Kuhl and Paul Iverson. The NLM proposes that phonetic prototypes, or representative instances of phonetic categories, act as perceptual “magnets” which shrink psychological space around them. Because of this, phonetic variation is difficult to discriminate around the category prototypes (Kuhl et al. 1992, Kuhl 1994). This is similar to the idea of “acquired similarity” in the A2D model: within-category similarity increases around native language “magnets”. Another similarity with the A2D model is that the NLM tries to characterize psychological space with reference to the physical (acoustic) dimensions of speech stimuli.

The main goal of the NLM is to characterize the reorganization of infants’ perceptual systems due to native language exposure in the early stages of life, so research within the NLM framework mainly focuses on speech first language acquisition. However, the NLM has been extended to account for non-native speech perception (Iverson et al. 2003). For instance, non-native sounds may be heard as better or worse exemplars of native language prototypes. The formation of a new, non-native category prototype depends on how similar the non-native sound is to native prototypes. According to the model, there is no hope of forming a new prototype if the non-native sound is too similar to a native prototype, because the non-native sound will be “pulled” to the native language “magnet”.
1.4.3 The Speech Learning Model

The Speech Learning model (SLM) [Flege, 1995], attributed mainly to James Flege, was developed as a model of second language speech production, rather than speech perception. Therefore, it is concerned primarily with the production abilities of fairly experienced learners of a second language, but it does address non-native speech perception to some degree. The SLM maintains that the creation of new L2 phonetic categories is crucial for accurate perception of non-native sounds. According to the SLM, the greater the perceived dissimilarity between an L1 and an L2 sound, the more likely it is that new L2 phonetic categories will be developed. The development of L2 categories can be “blocked by the mechanism of equivalence classification” (p. 239), where a single phonetic category is used to process perceptually similar L1 and L2 sounds. The L2 categories developed by a learner may differ from a native speaker’s phonetic categories, in that they may be based on different features or features weights, compared to the phonetic categories of native speakers. According to the SLM, production eventually corresponds to the phonetic properties of L2 representations, so effective perception must precede native-like production.

1.4.4 Similarities between PAM, NLM, and SLM

There are notable similarities between these three models of non-native speech perception. First, all acknowledge that the perception of non-native speech is related to L1 experience; there is a shared notion of the “native language filter” through which new sounds are perceived.

All three models also agree on the idea that learning to hear or produce a new speech sound is not constrained by age or a biologically defined “critical period” (though some of the NLM’s claims about the strength of the native language magnet are reminiscent of critical period hypotheses, see critique in [Best et al., 2001]). Learning can (and does) happen well into adulthood. Perception is thus (somewhat) malleable, though the three models disagree
on the extent to which new perceptual categories can be formed.

In general, in these three models, there tends to be a focus on phonetic categories or contrasts, rather than underlying acoustic dimensions or features. L1 effects on non-native speech perception are explained in terms of the interaction between phonetic categories without considering the structure of those categories. This is particularly true for PAM, whereas NLM and SLM both recognize the role of underlying features or dimensions, to varying degrees.

All have predictions about the extent to which specific non-native sounds will be assimilated into the native phonological system. PAM makes detailed predictions about how non-native contrasts will assimilate to native contrasts. The NLM makes predictions about how new sounds will be perceived in relation to native language “magnets”: either as better or worse exemplars, or potentially as new, non-native prototypes. SLM assumes that some L2 sounds will be fully assimilated into L1 categories if they are similar enough to those categories. In general, according to all three theories, some sounds may be easier or harder to learn, depending on their relationship to sounds in the native language. The more similar a target sound is to a native sound, the harder it will be to hear. Similarity results in assimilation, which is thought to be a major source of L1 interference.

1.4.5 Comparing models of non-native speech perception and category learning

Table 1.1 compares major characteristics and theoretical positions of the three models of non-native speech perception reviewed thus far (A2D, PAM, NLM, and SLM), along with the A2D model of category learning.

There are several points of convergence when comparing the four models, and there are important divergences as well. NLM is perhaps the most similar to A2D, to the point of being grouped with the A2D approach under the umbrella of “attentional models” (see, e.g., Francis and Nusbaum [2002]). In both models, perception is conceptualized with regard to
Table 1.1: Comparison of four speech perception models. Dashes (—) indicate cases where a characteristic or theoretical position is not an explicit part of a model.

<table>
<thead>
<tr>
<th></th>
<th>A2D</th>
<th>NLM</th>
<th>PAM</th>
<th>SLM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Developing Field</strong></td>
<td>Psychology</td>
<td>Speech science</td>
<td>Linguistics</td>
<td>Linguistics</td>
</tr>
<tr>
<td><strong>L1 or L2 focus</strong></td>
<td>—</td>
<td>L1</td>
<td>L2</td>
<td>L2</td>
</tr>
<tr>
<td><strong>Age focus</strong></td>
<td>adults</td>
<td>infants, children, adults</td>
<td>adults</td>
<td>older children, adults</td>
</tr>
<tr>
<td><strong>Type of learner</strong></td>
<td>novice</td>
<td>—</td>
<td>novice</td>
<td>experienced</td>
</tr>
<tr>
<td><strong>Perception or production?</strong></td>
<td>perception</td>
<td>perception</td>
<td>perception</td>
<td>production</td>
</tr>
<tr>
<td><strong>Cognitive mechanisms</strong></td>
<td>selective attention, category learning processes</td>
<td>perceptual warping, prototype effects</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Speech specific?</strong></td>
<td>No; the A2D model applies to other cognitive domains</td>
<td>General-domain cognitive mechanisms, but native-language magnets are phonetic</td>
<td>Yes; units of speech perception are articulatory</td>
<td>—</td>
</tr>
<tr>
<td><strong>Training studies?</strong></td>
<td>Yes, a main focus of this approach</td>
<td>Yes, mostly extensions of the core NLM model</td>
<td>—</td>
<td>Yes, mostly production training</td>
</tr>
<tr>
<td><strong>Predictions of learning based on L1 experience</strong></td>
<td>—</td>
<td>L2 sounds may be perceived as (good or bad) exemplars of L1 prototypes or may form new prototypes</td>
<td>Depends on how new sounds are perceived relative to L1 contrasts</td>
<td>Depends on similarity to L1 categories</td>
</tr>
</tbody>
</table>
a psychological “space” that is defined by acoustic dimensions. Native language experience “warps” this perceptual space. Both NLM and A2D focus on cognitive mechanisms that may underlie speech perception, for instance, on the role of attention in category learning. To some extent, both the NLM and A2D models assume that the cognitive mechanisms operating in speech perception are similar to those operating in other domains (for instance, visual categorization). The NLM is perhaps more speech specific, maintaining that perceptual changes due to language experience are “phonetic rather than purely auditory” (Iverson et al., 2003, p. B49).

One major difference between NLM and A2D is that when the latter is applied to speech perception, it tends to focus on L2 acquisition. Conversely, the main focus of the NLM is to characterize speech perception in L1 acquisition, and the exploration of L1 effects in non-native speech perception is an extension of NLM rather than a core focus. As such, NLM studies generally involve infants and young children, whereas A2D phonetic training studies involve adult learners.

Like the other three models reviewed here, the A2D model does assume that the native language “filter” — which is a result of differential attention to L1 acoustic dimensions—will affect non-native speech perception. However, the A2D model does not explicitly consider the hypothesis that some sounds will be easier or harder to learn, and may be learned in different ways, depending on how they are related to the L1. In contrast, PAM, NLM, and SLM do make explicit predictions about patterns of assimilation and how these patterns may influence perception and the learning of new sounds.

All of the models reviewed here have contributed to our understanding of non-native speech perception. All four models also have their limitations. PAM and SLM don’t consider cognitive mechanisms to great extent. These two models are fairly speech specific, and as a result don’t easily fit with domain-general theories related to perception and category learning. In terms of phonetic training, SLM focuses mainly on production, and phonetic training is not a focus of studies in the PAM framework. NLM is generally less concerned
about characterizing the perceptual experience of adult learners and the development of new phonological categories. While the A2D model, as extended to speech perception, addresses some of these limitations, it doesn’t seriously consider the relationship between specific L1 categories and target contrasts. It therefore lacks the ability to make predictions about how a learner’s L1 experience will interact with training and learning.

1.5 Toward an integrated model of non-native speech perception and more effective training strategies

The focus of this dissertation research is to empirically determine which training dimensions are most effective for helping adults to hear non-native sounds. By considering how a specific acoustic dimension operates in a learner’s L1 and the target language, it is possible to make predictions regarding phonetic category learning. These predictions can help determine which acoustic dimensions will be most effective for phonetic training, given a learner’s previous language experience.

It is clear that L1 experience exerts a strong influence on non-native speech perception. Therefore, when training people to hear new sounds, it’s important to be thoughtful about the influence of the L1—not just that L1 experience will influence non-native speech perception, but that it will do so in systematic ways. The relationship between the underlying dimensions in both L1 and the target language can inform phonetic training strategies, resulting in more effective learning.

This consideration will also serve to integrate models of non-native speech perception, by combining the explicit predictions of PAM, SLM, and NLM regarding patterns of assimilation and native language interference with the aspects of the A2D model that focus on category learning via attention to acoustic dimensions. It is hoped that this approach will lead to more effective phonetic training strategies and a fuller understanding of non-native speech perception.
1.5.1 Research Questions

Two of the research questions addressed in this dissertation are:

Research Question 1 Will phonetic training result in better perception of non-native sounds?

Research Question 2

2A Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others?

2B Given this L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language?

Predictions for each of these questions are addressed in subsequent chapters, in the context of specific experiments.

1.6 Dissertation outline

Chapter 2 In Chapter 2 (begins page 18), Experiment 1 is presented. Experiment 1 utilized pseudo-speech stimuli to explore the effects of different training dimensions on category learning. Descriptions of the design and procedure of Experiment 1 begin on page 19, followed by the results (page 33). A summary of the results from Experiment 1 begins on page 64, followed by a discussion regarding effective training dimensions and the implications of Experiment 1 with respect to language learning.

Chapter 3 In Chapter 3, (begins page 72), Experiment 2 is presented. Experiment 2 explored the effects of using different acoustic dimensions to train English speakers on a Korean consonant contrast. Training stimuli consisted of Korean syllables that were manipulated to highlight specific acoustic dimensions. Descriptions of the design and procedure of Experiment 2 begin on page 81. The results section begins on
The discussion section begins with a summary of results on page 133, followed by a discussion of effective training strategies for Korean stops and a broader consideration of the relationship between L1 and L2 in phonetic training.

**Chapter 4** In chapter 4 (begins page 147), the main research questions are revisited with respect to the results from Experiments 1 and 2. Following this, the theoretical implications of this research, as it relates to models of non-native speech perception, are discussed (page 157). This is followed by an overview of future work and potential applications of this research (page 161). Chapter 4 ends with remarks about the broader impacts of this research in terms of non-native speech perception, phonetic training, and understanding the relationship between category learning and perception.
Chapter 2

Experiment 1: Xorx and Bjaran

Experiment 1 investigated the effect of phonetic training on perceptual learning and categorization of auditory stimuli. The main focus of Experiment 1 was to determine if some types of training dimensions are more effective than others (see discussion in Chapter 1, section 1.3.1). The stimuli in Experiment 1 were synthetic, spectrally complex “words” from two “alien” languages.

The stimuli used in Experiment 2 are synthetic, pseudo-speech stimuli. They are not entirely synthetic (e.g. created with the use of a Klatt synthesizer), but instead were created using a method of source-filter synthesis, where the filter was based on syllables produced by a human. The source, however, was entirely synthetic.

Previous work indicates that perceptual and training effects achieved with these types of stimuli are comparable to those achieved with real speech tokens. For instance, Mirman et al. (2004) found that training with synthetic speech-like sounds resulted in perception and categorization effects that were similar to the perception and categorization of consonants and vowels. In a study where subjects were trained to categorize non-speech acoustic stimuli, Holt and Lotto (2006) found that subjects weighted cues in a manner similar to the weighting of phonetic cues in real speech. Research with non-speech and pseudo-speech stimuli allows for a high level of control over stimuli characteristics, and they are generally simpler and more easily manipulated. For these reasons, research with non-speech stimuli can play an important role in the development of attentional models of speech perception models.
Experiment 1 addresses Research Questions 1 and 2 (in the context of laboratory created “languages”):

**Research Question 1** Will phonetic training result in better perception of non-native sounds?

**Research Question 2**

2A Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others?

2B Given this L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language?

### 2.1 Study design and procedure

To approximate the experience of learning a second language, the study had two main sections. In the first section, subjects learned to distinguish between two “words” from an alien language called Xorx, which served as the quasi-“L1” in this experiment, because it was the subjects’ “first alien language”. In the second part, subjects learned to distinguish between two words from a different language called Bjaran, so Bjaran was their quasi-“L2”, their “second alien language”. This two-part approach was done to carefully control the relationship between interacting acoustic dimensions (rather than to try and create alien stimuli that interacted in a particular way with English or any other human language).

In both the Xorx and Bjaran sections, subjects completed training and testing. Training consisted of category training with feedback, and testing consisted of a difference rating task, which was done before and after training to determine if training had any effect on difference ratings. See table 2.1 for an overview of the experiment design.

In the training task for both Xorx and Bjaran, subjects learned to categorize stimuli by using one particular dimension—this was the “relevant dimension” for that language. In
Table 2.1: Experiment 1 Design

<table>
<thead>
<tr>
<th>Section</th>
<th>Alien Language</th>
<th>Phase</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Xorx</td>
<td>pre-training</td>
<td>difference rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>training</td>
<td>category identification with feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td>post-training</td>
<td>difference rating</td>
</tr>
<tr>
<td>2</td>
<td>Bjaran</td>
<td>pre-training</td>
<td>difference rating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>training</td>
<td>category identification with feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td>post-training</td>
<td>difference rating</td>
</tr>
</tbody>
</table>

Xorx, F0 is the “relevant” dimension—it is a contrastive cue in that it indicates the difference between two Xorx “words”. All subjects were trained and tested with the same Xorx stimuli. This established the acoustic dimensions in Xorx as the alien-L1 dimensions. Therefore, all subjects shared Xorx as their “native alien language”.

After training in Xorx, subjects were trained in Bjaran. The Bjaran section had three between-groups training conditions which differed in terms of the relationship between the category relevant dimension in Bjaran (the alien-L2) and the category relevant dimension in Xorx (F0). This manipulation was done to determine which kind of training dimension results in the most successful learning of novel sound categories (as in Bjaran), taking into account prior knowledge from previously learned categories (as in Xorx).

These training conditions can be defined either in terms of the actual training dimension, which is an acoustic property of the stimuli, or in terms of how the category relevant dimension in Bjaran relates to the category relevant dimension in Xorx. Table 2.2 shows the training conditions in Bjaran and their properties, both acoustic and in terms of an L1-L2 relationship with Xorx.

Though the Bjaran training conditions were designed relative to Xorx, the relationship between the training conditions and English (the native language for most of the subjects in the Experiment) was also considered. This is discussed in section 2.7.5, page 69.

All experiment tasks were presented using Psychopy (Peirce 2007). Stimuli were played through Bose speakers in the sound booth of the CU phonetics lab. Subjects used a keyboard
Table 2.2: Bjaran training conditions and L1-L2 relationship

<table>
<thead>
<tr>
<th>Bjaran condition</th>
<th>Training Dim.</th>
<th>Use in L1 (Xorx)</th>
<th>Use in L2 (Bjaran)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Dimension</td>
<td>F0</td>
<td>category-relevant</td>
<td>category relevant, different boundary</td>
</tr>
<tr>
<td>Familiar Dimension</td>
<td>Duration</td>
<td>category-irrelevant, adds complexity</td>
<td>category relevant</td>
</tr>
<tr>
<td>New Dimension</td>
<td>Pitch-rise</td>
<td>not a dimension in Xorx</td>
<td>category relevant</td>
</tr>
</tbody>
</table>

Prior to starting the experiment, subjects answered several questions about their native language, language experience, and musical training (see section 2.4 page 31).

2.1.1 Xorx training

Xorx training consisted of a two-alternative, forced choice category identification task with feedback (see “training procedure” below for more details on the task itself). All subjects in the experiment did the same Xorx training task (i.e. there were no training conditions).

The stimuli for the Xorx training were 36 syllables of the form /ba/, from a 6 X 6 stimulus array that varied along two acoustic dimensions: F0 and duration (see figure 2.1 page 26). These stimuli were created using source-filter synthesis (see section 2.3.1 page 30).

Training was divided into blocks. In each block, subjects heard all 36 Xorx stimuli, presented randomly. There were 5 blocks of training, for a total of $5 \times 36 = 180$ trials. There were no breaks between training blocks. Subjects completed the training blocks immediately after the pre-training difference task (see below).

2.1.1.1 Xorx training procedure

The instructions for the Xorx training task were as follows:

You will now learn the Xorx words for “pilot” and for “vegetable.” You will hear examples of the two Xorx words. Press ‘v’ if you think it is “vegetable”. Press ‘p’ if you think it is “pilot”. Press ‘space’ to start the training.
On each trial during the training task, subjects heard a single Xorx stimulus play. After the stimulus played, a screen with text instructions was shown, asking them to choose if the sound was the Xorx word for *pilot* or *vegetable*. After making their choice, subjects received feedback: the sound played again, with a message that informed subjects if they were correct (e.g. “Correct! The word was ‘pilot’ ” or “Wrong! The correct answer was “vegetable’ ”). The next trials started automatically after feedback had ended on the previous trial with a 500 ms ISI.

Before the trials began, 10 randomly selected Xorx stimuli were played, with their category label (*pilot* or *vegetable*) displayed on the screen. This was to allow subjects to familiarize themselves with the sounds and their category labels.

### 2.1.2 Xorx difference rating

In the Xorx difference rating task, subjects were presented with pairs of stimuli from the 6X6 Xorx stimulus array, which they rated in terms of their difference on a scale from 1 (very similar) to 9 (very different).

The entire pairwise comparison matrix for 36 stimuli is quite large: 630 comparisons, excluding reciprocal comparisons and “identity” comparisons where a stimulus is compared with itself. So, in the Xorx difference rating task, each subject judged a random subset of 200 pairs from the complete pairwise matrix (the techniques used to analyze the difference judgments were sufficiently robust to handle missing data).

The Xorx difference rating task was repeated twice: once before Xorx training and once after. The random subset of 200 pairs presented to the subject in the post-training task differed from that in the pre-training task.

### 2.1.2.1 Xorx difference rating procedure

The instructions for the difference rating task were as follows:

> On each screen, you will hear examples of two sounds in Xorx, and decide how similar
they are, on a scale from 1 (very similar) to 9 (very different). After hearing both examples play, make your choice by pressing the number keys.

On each trial in the difference rating task, subjects heard a pair of stimuli play. Then they were presented with a screen that asked them to rate the similarity of the stimuli. Subjects used the numbers on a keyboard to enter in their numeric response. Following the response, there was an ISI of 500 ms before the start of the next trial.

In the pre-training difference task, subjects heard a block of practice trails. All 36 Xorx stimuli were played in random order, to familiarize subjects with the sound of Xorx. In the post-training task, subjects did not have a block of practice trials.

2.1.3 Bjaran condition assignment

After a short break, subjects moved on to the part of the study where they learned Bjaran. Subjects were randomly assigned to one of three Bjaran conditions, which differed in terms of the dimensions that subjects were trained with, and, therefore, the stimuli they heard during the training AND difference rating tasks. These training conditions were F0 ("same dimension"), duration ("familiar dimension"), and pitch rise ("new dimension"). See table 2.2, page 21 for a listing of the Bjaran training conditions. See section 2.2, page 25 for details on these dimensions and how the stimuli were created. In the Bjaran section, as in Xorx, subjects first completed a pre-training difference rating task, then category training, ending with a post-training difference rating task.

2.1.4 Bjaran training

Bjaran word category training proceeded in the same way as the Xorx training: subjects completed a two-alternative forced choice task where they categorized a single stimulus as belonging to one of two Bjaran word categories: rock or mechanic. Subjects received feedback in the form of "correct" or "incorrect" messages after entering their response.

As in Xorx, the Bjaran training was organized into blocks where each block contained
all 36 Bjaran stimuli from the condition-dependent 6 X 6 stimulus array, presented randomly. There were 5 blocks for a total of 180 Bjaran training trials.

2.1.4.1 Bjaran training procedure

The instructions for the Bjaran training task were:

*You will now learn the Bjaran words for “rock” and for “mechanic.” . . . You will hear examples of the two Bjaran words. Press ‘r’ if you think it is “rock”. Press ‘m’ if you think it is “mechanic”. Press ‘space’ to start the training.*

The procedure for the Bjaran training task was the same as in Xorx. On each trial, subjects heard a single stimulus play and decided if the sound was the Bjaran word for *mechanic* or *rock*. Subjects used the “m” and “r” keys to enter in their response. After making their choice, subjects received feedback: the sound played again, with a message that informed subjects if they were correct (e.g. “Correct! The word was ‘mechanic’ ” or “Wrong! The correct answer was ‘rock’ ”).

As in Xorx, before the training trials began, 10 example Bjaran stimuli were played, with their category label (*mechanic* or *rock*) displayed on the screen. This was to allow subjects to familiarize themselves with the sounds and their category labels.

2.1.5 Bjaran difference rating

The Bjaran difference rating task proceeded in the same way as in the Xorx section of the experiment. Subjects were presented with pairs of Bjaran stimuli, which they rated on a scale from 1 (very similar) to 9 (very different). As in Xorx, subjects were presented with a random subset of 200 stimuli from the full pairwise comparison matrix for all 36 Bjaran stimuli.

The Bjaran difference rating task was repeated twice: once before Bjaran training and once after. The subset of stimuli in the post-training task differed from that in the pre-training task.
2.1.5.1 Bjaran difference rating procedure

The instructions for the Bjaran difference rating task were:

*On each screen, you will hear examples of two sounds in Bjaran, and decide how similar they are, on a scale from 1 (very similar) to 9 (very different). After hearing both examples play, make your choice by pressing the number keys.*

The procedure for the Bjaran difference rating task was the same as in Xorx: on each trial, subjects heard a pair of stimuli play. Then they rating the difference of the pair on a scale from 1 to 9, using the numbers on a keyboard to enter in their response.

In the pre-training difference task, subjects heard a block of practice trails. All 36 Bjaran stimuli were played in random order, to familiarize subjects with the sound of Bjaran. In the post-training task, subjects did not have a block of practice trials.

2.2 Stimuli

Stimuli in the Xorx-Bjaran experiment were syllables created using the source-filter method of speech synthesis (see section 2.3.1 for details on stimuli creation). There was one set of Xorx stimuli (36 syllables), and three different sets of Bjaran stimuli (36 syllables each), one for each of the Bjaran training conditions.

Stimuli in each of these arrays varied along two acoustic dimensions. In Xorx, these dimensions were fundamental frequency (F0) and duration. In the Bjaran training arrays, the dimensions were F0, duration, pitch-rise, and relative amplitude. The relationship of these acoustic dimensions to phonetic features in human languages is discussed in section 2.7.4 on page 68.

2.2.1 Xorx stimuli

The Xorx stimulus array consisted of 36 /ba/ syllables that varied along two acoustic dimensions: overall F0 and duration (see figure 2.1). All Xorx stimuli were synthesized using
the same /ba/ filter (see section 2.3.1), which means that every Xorx stimulus had the same consonant-to-vowel formant transitions and the same formant values in the vowel portion.

Figure 2.1: Xorx stimuli.

The F0 dimension for Xorx ranged from 100-110 Hz, with stimuli evenly spaced along the dimension in 6 steps of 2 Hz each. The duration dimension for Xorx ranged from 350-600 ms, with stimuli evenly spaced in 6 steps of 50 ms each.

The category boundary that separated the Xorx pilot and vegetable stimuli was halfway along the F0 dimension, between 104 and 106 Hz. F0 is a category relevant dimension for Xorx, because it can be used to discriminate between Xorx word categories. Duration varies systematically and adds complexity to the stimuli, but it does not function as a cue to category membership. Duration is an irrelevant dimension—it does not help to distinguish pilot stimuli from vegetable stimuli.

2.2.2 Bjaran stimuli

There were three stimulus arrays for Bjaran (see figure 2.2), each consisting of 36 /bi/ syllables that varied along two acoustic dimensions: a category-relevant dimension that varied between training conditions, and a category-irrelevant dimension (relative amplitude, see
section 2.3.1) which was the same for all three conditions. All Bjaran stimuli were synthesized using the same /bi/ filter, so every Bjaran stimulus had the same formant transitions and formant values in the vowel portion.

(a) Same Dimension condition

(b) Familiar Dimension condition

(c) New Dimension condition

Figure 2.2: Stimuli arrays for the three Bjaran conditions
2.2.3  “Same dimension” condition

The two acoustic dimensions for the “Same dimension” condition were F0 and relative amplitude (see figure 2.2a). All stimuli in this condition had a constant duration of 600 ms. The F0 dimension, as in Xorx, ranged from 100-110 Hz, with stimuli evenly spaced in 6 steps every 2 Hz.

The relative amplitude dimension ranged from 4-24 dB, with stimuli evenly spaced along the dimension in steps of 4 relative-amplitude-dB each. “Relative amplitude” refers to the difference in overall amplitude between a lower and upper frequency band (with the split at 1000 Hz), where smaller relative amplitude values indicate a smaller relative difference in amplitude between the two frequency bands, and larger values indicate a larger difference. The perceptual effect of this acoustic manipulation is a “buzzy” quality for stimuli with low relative amplitude values, because in these stimuli, both frequency bands are around 70 dB. Stimuli with high relative amplitude values, on the other hand, have a more “muffled” or “muted” quality, because in these stimuli, low frequencies are louder than high frequencies.

In the “same dimension” condition, the category boundary distinguishing between the categories of rock and mechanic is along the F0 dimension, between 106-108 Hz, making F0 the category-relevant dimension in this Bjaran condition. F0 is also the category-relevant dimension in Xorx. In Bjaran, however, the category boundary has been shifted along the F0 dimension. (This is analogous to stop voicing in English vs. Spanish, where both Spanish and English speakers utilize VOT as the primary cue to contrast voiced and voiceless stops, but make different cuts along the VOT dimension.) So, although F0 is relevant in both Xorx and the Bjaran-“Same” condition, there are different category boundaries in the two languages.

As in the case of the Xorx stimuli, the second Bjaran dimension (relative amplitude) adds complexity to the stimuli, but is not helpful in distinguishing between stimuli in the rock category and those in the mechanic category. Relative amplitude is a category-irrelevant dimension.
2.2.4 **Familiar Dimension condition**

The acoustic dimensions for the stimuli in the Familiar Dimension condition were duration (in ms) and relative amplitude (see figure 2.2b). All stimuli in this condition had a constant F0 of 100 Hz. The duration dimension, as in Xorx, ranged from 350-600 ms, with stimuli spaced evenly along the dimension in 6 steps of 50 ms each. Relative amplitude was the same as in the Same Dimension condition.

The category-relevant dimension in this condition is duration. Recall that in Xorx, duration was the irrelevant dimension. It was perceptually available to subjects, but not helpful in distinguishing between words. So, in this condition, duration is familiar, but has not been used as a primary cue to category membership.

As in the other Bjaran conditions, relative amplitude in the Familiar Dimension condition does not function as a cue to category membership, so it is an irrelevant dimension.

2.3 **New Dimension condition**

The two acoustic dimensions for the stimuli in the “New Dimension” condition were pitch-rise and relative amplitude (see figure 2.2c). All stimuli in this condition had a constant duration of 600 ms.

The pitch-rise dimension ranged from 2-12 Hz. Pitch-rise refers to a rise in F0 from 100 Hz to a target F0, which ranged from 102-112 Hz. A pitch rise of 2 means that the stimulus started with an F0 of 100 Hz, and rose to a target pitch of 102 Hz over the duration of the stimulus. A pitch rise of 12 means that the stimulus started with an F0 of 100 Hz, and rose to a target pitch of 112 Hz. Relative amplitude was the same as in the other Bjaran conditions.

Pitch-rise, the category-relevant dimension in this condition, was not one of the dimensions of the Xorx stimuli. While it is the case that subjects used F0—a component of pitch-rise—as a cue to category membership in Xorx, they have not been exposed to
the pitch-rise dimension as such. In this condition, pitch-rise is a completely new category-relevant dimension: it has not been previously used to distinguish stimulus categories, nor have subjects been exposed to it as a systematically varying irrelevant dimension.

As in the other Bjaran conditions, relative amplitude is not helpful in distinguishing between rock and mechanic stimuli, so it is an irrelevant dimension.

2.3.1 Stimuli creation

All stimuli were made in Praat (Boersma & Weenink, 2014), a software package used for acoustic analysis and sound manipulation. The stimuli were created using source-filter synthesis according to the following general procedure:

(1) **Filter extraction.** Several tokens of syllables (/ba/ for Xorx, /bi/ for Bjaran) were recorded (produced by the author), at a level of 80dB. Recording took place in a double-walled sound booth using the built-in microphone of a Mac Book Pro. From the recorded set, a single syllable was chosen based on the following criteria: steady formant structure during vowel portion, duration over 600 ms (to allow for syllable durations up to 600 ms). The acoustic filter then was extracted from the /b/ or /bi/ token using Praat’s linear prediction analysis (LPC).

(2) **Glottal source creation.** The glottal source consisted of synthesized glottal pulses. These were created using Praat’s pitch-target source synthesis. F0 and duration (measured in ms) of the source signals were defined by the parameters of each individual stimulus, according to the stimuli arrays used in the experiment. These manipulations resulted in a set of spectrally rich source signals (i.e. not just a single tone). Unfiltered, the source signals sound like a “buzz” with a specific F0 and duration.

(3) **Filtering.** Each source signal (with its specific F0 and duration values) was passed through the extracted /b/ or /bi/ filter. This resulted in acoustic stimuli with varying
F0, duration, and other acoustic parameters (see below for descriptions of the pitch-rise and relative amplitude dimensions).

For stimuli where only F0 and duration varied (Xorx stimuli), the F0 and duration dimensions were created during the process outlined above. The Pitch-rise dimension (used in the Bjaran-New Dimension condition) was manipulated by specifying pitch parameters during the creation of the source signal, prior to applying the /bi/ filter. The rise in pitch began at 100 ms and rose steadily, hitting the pitch target at 500 ms.

Relative amplitude, the category-irrelevant dimension in all three Bjaran conditions, was implemented by filtering each source signal—already with values for duration and F0 specified, but not yet run through the /bi/ filter—into a high and low frequency band, with the split at 1000 Hz. Next, overall intensity of the high frequency band was lowered, and overall intensity of the low frequency band was increased, in even steps centered around 70 dB. For illustration, a relative amplitude value of 4 means that the high frequency band had an overall intensity of 68 dB, while the low frequency band had an overall intensity of 72 dB (72 – 68 = 4). A relative amplitude of 24 means that high frequencies=58 dB and low frequencies=82 dB (82 – 58 = 24) Finally, the two frequency bands were recombined to create a single source, and passed through the extracted /bi/ filter, yielding a stimulus with specific values for duration, F0 (or pitch-rise, depending on condition), and relative amplitude.

2.4 Subjects

Subjects were 60 normal hearing (by self-report) adults between the ages of 18 and 70. Forty-six subjects were recruited via the CU Boulder undergraduate psychology subject pool; these subjects received course credit for participating in the study. Ten subjects, all linguistics department graduate students, were paid $20/each for participating. The rest were volunteers from the linguistics department or the Boulder community who expressed
an interest in the research.

Prior to the experiment, subjects filled out a survey with questions about their native language and musical ability. Most subjects were native English speakers, but there were 5 native Mandarin speakers, 2 native Spanish speakers, 2 native Cantonese speakers, 2 native Arabic speakers, and one native speaker each of the following languages: Korean, Swedish, Japanese, Burmese. There was also one true bilingual English/Burmese speaker (the subject had learned both languages from birth). These non-native English speakers all had significant experience with English, most having learned English early in childhood. Of the native English speakers, 4 had significant experience with another language.

Subjects were also asked if they had received formal music training. Some research suggests that people with musical training are better able to hear sounds in a foreign language (e.g., Delogu et al. 2010; Cooper and Wang 2012). Of the 60 subjects, 8 had had recent formal music training.

2.5 Predictions for Experiment 1

It was predicted that training with the Xorx category relevant dimension—F0—would affect subsequent learning in Bjaran, which had three different training conditions where the category-relevant dimensions varied, systematically, in relation to Xorx. Differences between Bjaran training conditions, both in terms of training accuracy and measures of category learning derived from the results of the difference tasks, were expected due to the interplay between the “alien-L1” and the “alien-L2”.

The “Same Dimension” condition was designed to determine if there is a cost associated with utilizing the same category relevant dimension (F0) in the alien-L1 and the alien-L2, but with different category boundaries in each language. This is analogous to the VOT example in English and Spanish: if English speakers use the English VOT boundary for voiced and voiceless stop consonants when listening to Spanish, they will be wrong in determining the phonemic category of some sounds. Likewise, Xorx “speakers” will be wrong in determining
the phonemic category of Bjaran sounds if they continue to use the Xorx F0 boundary, and it may be difficult to train these subjects to use the Bjaran F0 boundary instead.

In the “Familiar Dimension” condition, the category-relevant dimension (duration) is familiar to the learner: they’ve been exposed to duration as a category-irrelevant dimension of their alien-L1 sounds, but in the alien-L2, duration is now a category-relevant dimension. This kind of training dimension may not be the best type for learners either: it may be difficult for subjects to switch from ignoring duration in Xorx to using it to make category distinctions in Bjaran. On the other hand, familiarity with the duration dimension might be advantageous: the previous experience with duration in Xorx may allow the subjects to “tune-in” to duration cues more easily, as soon as they learn that duration is now category-relevant.

In the “New Dimension” condition, the category-relevant dimension (pitch-rise) is completely new: it was not a category-relevant or irrelevant dimension of the alien-L1 sounds. Subjects in this condition do not have to learn to shift a category boundary along a previously attended dimension, nor do subjects have to learn to attend to a dimension they have previously been trained to ignore. Because the sounds are more dissimilar from Xorx in terms of their acoustic dimensions, there may be less alien-L1 interference. This may result in more effective learning for subjects in this condition, compared to the other two conditions.

2.6 Results

Results from Experiment 1 are presented in this order: Xorx category training (section 2.6.1 page 34), Xorx difference rating (section 2.6.2 page 37), Bjaran category training (section 2.6.4 page 46), Bjaran difference tasks (section 2.6.5 page 51). There is an overview of all of the results from Experiment 1 at the beginning of section 2.7 page 64. Unless otherwise noted, all statistical analyses were done using R (version 3.2.3) (R Core Team 2015).

In cases where an analysis included categorical variables (e.g. in the analysis of training
accuracy for each of the Bjaran training conditions), orthogonal contrast coding was used to specify mean comparisons between levels of the categorical variable(s). Orthogonal contrast coding was used because it allows for specific planned comparisons between the levels of categorical variables (for instance, comparing mean accuracy in specific training blocks, or comparing mean accuracy for each of the three Bjaran training conditions). Contrast coding is more flexible than, e.g. treatment (i.e. “dummy”) coding. Unlike treatment coding, where one reference level of a categorical variable is compared to all other levels, contrast codes can be specified for desired comparisons. For example, for a categorical variable with three levels (as in the case of the three Bjaran training conditions), it may be the case that a mean comparison of interest is between level 1 and the other two levels. To allow for this comparison, two orthogonal contrast coded predictors can be included in a model that will code for (1) a mean comparison between level 1 and the other two levels, and (2) a mean comparison between level 2 and level 3. This type of contrast coding can be included in a model with, e.g., continuous variables, to explore interactions between a continuous variable and contrast coded categorical variables. If significant interactions between a continuous and a contrast coded variable exist, this is evidence that the effect of the continuous variable is different for the different levels of a categorical variable, providing motivation for, e.g., post-hoc tests to examine the effect of the continuous variable for each of the levels of a categorical variable.

2.6.1 Xorx category training

For the Xorx category training task, subjects heard the Xorx stimuli (36 synthetic “ba” syllables, varying in F0 and duration), and identified the word category (“pilot” or “vegetable”).

Data from the category training tasks were analyzed using logistic mixed-effects regression models, which allowed for a binary dependent variable (because accuracy on a single trial was coded as 0 or 1), and random effects for subject and/or item, which can account for
subject- and item-level variance in the case of a repeated-measures design. Unless otherwise noted, all logistic mixed-effects models were run using the R packages lme4 \cite{Bates2015} and lmerTest \cite{Kuznetsova2016}, and all model comparisons to test for the significance of fixed effects were done using the "anova()" function in the R stats package. Contrast coded variables were used for specific mean comparisons (coding schemes depended on the models, see descriptions below).

Generally, subjects improved in accuracy over the course of the Xorx training task (from Block 1 to Block 5). Figure 2.3 shows the subject mean accuracy in the Xorx training task as a function of block.

Accuracy improved over the course of training, i.e. as block number increased, mean accuracy also increased. To evaluate if accuracy improved over the course of training, a logistic, mixed-effects model was fit to the Xorx training data. The dependent variable was accuracy on a single trial (coded as 1 or 0). Subjects and items were treated as random variables (allowing for within-subject and within-item variance). Block number was considered as a continuous, fixed effect, and was centered so that the intercept in the model represented mean accuracy in block 1. In this model, the slope for the block number variable was positive, indicating that as block number increased, accuracy also increased.

In order to assess the significance of the block number variable, a form of model-comparison was used. In this method, a log-likelihood ratio test compares the full model (including the block number variable) with a reduced model (in this case, an intercept-only model without the block number variable, but with the same random-effects components). This method allows for significance testing of a single fixed effect variable of interest \cite{Raudenbush2002}, in this case, the block number variable.

This test indicated that including block number in the model as a fixed effect resulted in a significant improvement in log-likelihood ($\chi^2(1) = 4.64, p = .03$). In other words, the effect of block number was significant. Furthermore, in a logistic mixed-effects model comparing the mean accuracy in block 1 with the mean accuracy in block 5 by means of a contrast coded
block variable (“Block 1 vs Block 5”), and subject and item set as random effects, accuracy in Block 5 ($M = .82$) was significantly higher than in Block 1 ($M = .79$), as indicated by a significant contrast-coded block variable ($\chi^2(1) = 7.32, p = .007$).

![Figure 2.3: Xorx training accuracy by block (mean accuracy across subjects). Error bars represent $SEM^{betw}$.
](image)

One thing to note is that modeling accuracy with block number obscures what is happening in a single block. In fact, analysis of the first block of training indicated that the Xorx categories were learned fairly quickly over the course of the first block by most subjects. Accuracy rates then stayed fairly high (note the high mean accuracy in 2.3). Figure 2.4 shows mean accuracy as a function of trial number for the first 36 trials, which constituted Block 1. In order to assess changes in accuracy over the course of the first 36 trials, a logistic mixed-effects model was fit to a subset of the training data (those from the first block of training). The dependent variable was accuracy on a single trial (coded as 1 or 0). Subjects, items, and each subjects’ trial number effect were treated as random variables. Trial number was considered as a continuous, fixed effect, and was centered so that the intercept in the model represented mean accuracy on trial 1. Trial number had a positive slope, and was found to be
a significant predictor of accuracy (by log-likelihood ratio test, $\chi^2(1) = 26.71, p < 0.0001$).

In other words, subjects improved in accuracy over the course of the first 36 trials.

Figure 2.4: Xorx training accuracy, first 36 Trials (mean accuracy across subjects). Blue line represents a linear fit; shaded area represents between-subjects standard error.

Taken together, these results suggest that subjects were successful in learning the Xorx word categories over the course of the training task. Though accuracy improved quickly within the first trial, it continued to improve throughout the rest of the Xorx training, with significantly higher accuracy in Block 5 compared to Block 1.

2.6.2 Xorx difference ratings

For the difference rating task, subjects heard all pairwise combinations of the Xorx stimuli (36 synthetic “ba” syllables, varying in F0 and duration) and provided a difference rating for each pair on a scale from 1-9.

Results from the Xorx pre- and post-training difference rating tasks were analyzed using a combination of Multi-dimensional scaling (MDS) and multiple regression models.

MDS is a technique that allows for subjects’ judgments about pairs of stimuli (in this case, difference judgments) to be represented as distances in a multi-dimensional space.
This multi-dimensional space can be thought of a representation of psychological similarity space, where stimuli judged as “similar” appear closer together in space, and stimuli judged as “different” appear further apart. The dimensions in MDS analyses emerge as a result of a scaling algorithm, and represent psychological dimensions that subjects may utilize to make their difference judgments. Though these psychological dimensions are NOT the physical dimensions of the stimuli per se, it is possible interpret the MDS dimensions in light of actual stimuli dimensions by visual inspection of the MDS plots and by using multiple regression models where MDS dimensions are used to predict the physical attributes of stimuli.

The Xorx difference judgments were analyzed with an MDS technique called Individual Scaling Differences (INDSCAL) (Carroll and Chang, 1970). This technique allows for the fact that subjects may have different weights for each of the psychological dimensions in the scaling solution. INDSCAL models the differences between individual subjects by computing an individual’s weights for each of the dimensions in the scaling solution. These individual weights act as multipliers on group-space stimuli coordinates, which are based on the average pair-wise difference judgments across all subjects. Thus, an individual’s weights for each dimension, multiplied by the group-space coordinates for a particular stimulus, yields the individual’s coordinates for that stimulus. The INDSCAL analysis was implemented in R, using the smacofIndDiff() function from the SMACOF package (De Leeuw and Mair, 2011), assuming an interval distance model.

Figure 2.5 shows the scaling results for the pre- and post-training Xorx difference ratings. In this plot, points represent average stimulus coordinates across all subjects, pre-training, and are colored to represent the word category of the stimulus. From each point extends a vector, the endpoint of which is the average stimulus coordinate, post-training. Therefore, each vector represents both the magnitude and direction of pre- to post-training “shift” for a single stimulus. These vectors give a global picture of how difference ratings may change as a result of training.
Figure 2.5 shows D1 (dimension 1) plotted against D3 (dimension 3). Note that these are simply labels for the dimensions in the scaling solution, and do not reflect any particular hierarchy of importance (they could just as easily be labelled “X” and “Y”).

![Graph showing MDS solution for Xorx difference ratings](image)

Figure 2.5: MDS solution for Xorx difference ratings

It was found that 3 dimensions, rather than two, best represented the similarity space for the Xorx data, as evidenced by a lower stress value for the 3 dimension solution, with fairly interpretable dimensions. “Stress” is a goodness-of-fit measure for MDS solutions, which captures how well a particular configuration reproduces the original distance matrix ([Lewicki and Hill, 2006](#)). As the number of dimensions increases, stress generally goes down, but the interpretability of the dimensions becomes more difficult, and there is a risk of over-
fitting. Deciding how many dimensions to include in a scaling solution is thus a process of balancing low stress with interpretability.

As discussed previously, the dimensions in the scaling results are not the physical dimensions of the stimuli. Rather, they represent emergent psychological dimensions along which subjects make their difference judgements. It is, however, possible to interpret these dimensions in light stimulus characteristics. In this case, D1 was found to be related to F0 (see below) and D3 to duration (see below). D2 represented something akin to “distance from the category boundary” (this dimension was less interpretable in terms of the physical dimensions of stimuli, so further analysis for D2 is not presented here).

Visual inspection of the scaling solution supports the claim that D1 is related to both F0 and word category: all of the pilot stimuli are on the left hand side of the black dotted line, while all of the vegetable stimuli are on the right hand side.

One way to interpret dimensions in an MDS solution is by modeling the extent to which the dimensions in the scaling plot correlate with the physical dimensions of the stimuli. Individual linear regression models can be run for each stimulus characteristic of interest (e.g. F0 or duration), where a single acoustic property is regressed on the stimuli coordinates from the group scaling space.

To evaluate the extent to which the scaling dimensions in the group space correlated with the F0 values of the stimuli, a linear regression model was fit to the results of the Xorx INDSCAL group space, using the R Stats package [R Core Team 2015]. The dependent variable was the F0 values (z-scored) of the Xorx stimuli, which were regressed on the scaling coordinates of the stimuli for each of the three dimensions from the group-space. F0 was found to load heavily on D1, as indicated by a significant effect for D1 in the regression model $F(1,33) = 180.08, p < .00001$. The effect for D3 was not significant ($p = .83$), and D2 was marginally significant ($p = .05$).

Moreover, D1, and not D2 or D3, is related to word category. To determine if the scaling dimensions in the Xorx group space were predictive of word category, a logistic
regression model was fit to the results from the Xorx group space using the R Stats package. The dependent variable was word category (pilot or vegetable), which was regressed on the scaling coordinates of the stimuli for D2 and D3 (D1 was left out, because it perfectly predicts word category, i.e. pilot always occurs with certain coordinates along D1, and never with other coordinates of D1). In this model, neither D2 nor D3 were significant ($p = .28, .93$, respectively). So, the best interpretation of D1 is the “category relevant” dimension—it is related to F0 and distinguishes pilot stimuli from vegetable stimuli.

Duration loaded most heavily on D3. To evaluate the extent to which the scaling dimensions in the group space correlated with the duration values of the stimuli, a linear regression model was fit to the results of the Xorx INDSCAL group space. The dependent variable was the duration values (z-scored) of the Xorx stimuli, which were regressed on the scaling coordinates of the stimuli for each of the three dimensions from the group space. In this model, the effect for D3 was significant ($F(1, 33) = 2049.37, p < .00001$). Neither the effect for D1 nor D2 was significant ($p = .65$ and $p = .5$, respectively), indicating that these dimensions are not related to duration.

2.6.3 Category learning in Xorx: Acquired Similarity and Acquired Distinctiveness

Analysis of the scaling results can provide some insight into if, and how, subjects are learning the Xorx word categories. One way to evaluate if category learning is taking place is by looking for evidence of within-category acquired similarity and between-category acquired distinctiveness, two category learning processes described in detail in chapter 1 (page 6).
Within-category distance was obtained by finding each subject’s category centroids, which were the average spatial coordinates for all pilot stimuli and the average spatial coordinates for all vegetable stimuli. Then the Euclidean distance between the category centroid and each stimulus in that category was calculated, following the Euclidean distance formula for three dimensions:

$$\sqrt{(D_{1\text{mean}} - D_{1\text{stimulus}})^2 + (D_{2\text{mean}} - D_{2\text{stimulus}})^2 + (D_{3\text{mean}} - D_{3\text{stimulus}})^2}$$

The overall within-category distances for each subject were found by taking the mean of all within-category distances for each stop category. See figure 2.6 for a schematic showing within-category distance.

![Schematic for within-category distance](image)

Figure 2.6: Schematic for within-category distance, where \(d\) = the Euclidean distance between the category centroid and a single stimulus. Overall within-category distance was found by taking the mean of all individual distances. For the Xorx stimuli, Euclidean distance was calculated in three dimensions.

Each subject had four within-category distances: overall within-category distance for pilot stimuli, pre- and post-training, and overall within-category distance for vegetable stimuli, pre- and post-training. Figure 2.7 shows the within-category distances (averaged across subjects) for the Xorx stimuli, pre- and post-training.

A measure of “within-category shift” was constructed by subtracting post-training distances from pre-training distances. This “within-category shift” variable thus represents
the degree and direction of change between post-training and pre-training within-category distance.

To evaluate the effect of training on judgments of similarity and in order to test for a significant change in within-category distance after training, a linear model was fit to the results of the pre- and post-training Xorx INDSCAL scaling configurations using the R Stats package. The dependent variable was the “within-category shift” variable (the difference between the pre- and post-training within-category distance, which was the Euclidean distance between each subject’s category centroid and the positions of each stimulus in each subject’s individual configuration, see above for full description of how this variable was computed). The independent variable was a contrast coded predictor for stimulus type (“Pilot vs vegetable”). This variable compared the mean within-category shift for the pilot stimuli with the mean within-category shift for the vegetable stimuli, and was included to determine if any change in within-category distance (from pre- to post-training) was due to the stimulus type.

Figure 2.7: Xorx within-category distance, pre- vs. post-training. Asterisks indicate significant difference between pre- and post-training within-category distance. “AS” indicates an Acquired Similarity effect (within-category distances are smaller, post-training). Error bars represent $SEM_{betw}$. 
Table 2.3 shows the output for the intercept and stimulus type variable from the linear model. In order to obtain an F-statistic to test for the significance of the intercept in this model, the full model was compared with a reduced model where the intercept was left out. The full and reduced models were compared using the anova() function in the R Stats package. This model comparison test showed that including the intercept resulted in a significant reduction in error ($F(1, 118) = 39.05, p < .00001$); in other words, indicated that the mean within-category shift, controlling for stimulus category, was non-zero.

Furthermore, the intercept was positive (.02), which shows that the within-category shift was in the predicted direction for an AS effect, i.e. post-training within-category distances were smaller after training. Finally, the stimulus category variable was not significant ($p = .86$), indicating that the AS effect was not statistically different for pilot stimuli compared to vegetable stimuli.

Table 2.3: Regression model for within-category shift of Xorx stimuli

|                     | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------------|----------|------------|---------|----------|
| (Intercept)         | 0.027367 | 0.004379   | 6.249   | 6.78e-09 *** |
| “Pilot vs vegetable”| -0.001505| 0.008758   | -0.172  | 0.864    |

### 2.6.3.2 Acquired distinctiveness

Acquired distinctiveness is the process by which between-category distances become larger as a result of category learning. The comparison of pre- and post-training between-category distance can provide evidence of Acquired Distinctiveness (AD): if post-training between-category distances are larger, this suggests that categories are “moving apart” in psychological space, and becoming less similar to each other.

Between-category distances were obtained for each subject by calculating the mean Euclidean distance between each stimulus category centroid, using the Euclidean distance formula for three dimensions. Figure 2.8 shows the between-category distance for the Xorx stimuli, pre- and post-training.
Figure 2.8: Xorx between-category distance, pre- vs post-training. Asterisks indicate significant difference between pre- and post-training between-category distance. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent the between-subjects standard error of the mean (hereafter, $SEM_{betw}$).

A measure of “between-category shift” was computed for each subject by subtracting pre-training between-category distance from post-training between-category distance. This “between-category shift” variable thus represents the degree and direction of change between post-training and pre-training between-category distances.

To evaluate the effect of training on judgments of similarity and in order to test for a significant change in between-category distance after training, a linear model was fit to the results of the pre- and post-training Xorx INDSCAL scaling configurations. The dependent variable was the ”between-category shift” variable (the difference between the pre- and post-training between-category distance, which was the Euclidean distance between each subject’s category centroids, see above for full description of how this variable was computed). This model was a simple, intercept-only model, which was used to test if the between-category shift was significantly different from zero (in other words, if there was a significant shift in between-category distance, from pre- to post-training).

Table 2.4 shows the output for the intercept in the simple linear model of between-
category distance. In order to obtain an F-statistic to test for the significance of the intercept in this model, the full model was compared with a reduced model where the intercept was left out. The full and reduced models were compared using the anova() function in the R Stats package. This model comparison test showed that including the intercept in the full model resulted in a significant reduction in error ($F(1,59) = 20.2, p < .0001$), in other words, the mean between-category shift from pre- to post-training was significantly different from zero.

Moreover, the intercept was positive (.07), which shows that the between-category shift was in the predicted direction for an AD effect, i.e. post-training between-category distances were larger after training.

Table 2.4: Regression model for between-category shift of Xorx stimuli

|                | Estimate | Std. Error | t value | Pr(>|t|)   |
|----------------|----------|------------|---------|------------|
| (Intercept)    | 0.07293  | 0.01623    | 4.494   | 3.32e-05 ***|

2.6.4 Bjaran training

For the Bjaran category training task, subjects in each condition heard a set of Bjaran stimuli (36 synthetic “bi” syllables, varying along one of three training dimensions and relative amplitude), and identified the stop category of the initial consonant.

Figure 2.9 shows the mean accuracy in the Bjaran task as a function of block number for the three Bjaran training conditions.

To evaluate if accuracy improved over the course of training, a logistic, mixed-effects model was fit to the Bjaran training data (see table 2.5). The dependent variable was accuracy on a single trial (coded as 1 or 0). Block number was considered as a continuous, fixed effect, and was centered so that the intercept in the model represented mean accuracy in block 1. Subjects, items, and each subject’s block number effect were treated as random variables. The model also contained two categorical fixed effects variables that were coded for condition.
These two variables allowed for comparisons in mean accuracy among the training conditions: one variable compared the mean accuracy for the Same Dimension condition to the mean accuracy of the other two conditions (“Same vs others”), and the other variable compared the mean accuracy of the Familiar Dimension condition with that of the New Dimension condition (“Familiar vs New”). Interactions between block number and each of the contrast coded predictors were included as well, to assess the extent to which the block number effect depended on training condition.

To account for multiple testing with the Bjaran training data for the family of mixed models where accuracy was predicted by block number as a continuous fixed effect (the omnibus test, described in the previous paragraph, and three post-hoc tests, one for each training condition, described below), alpha levels were adjusted using the Holm-Bonferroni correction, which is more powerful than the classic Bonferroni correction (Holm 1979). The correction is noted (e.g. “p<.001”, H-B corrected) when applicable.

Accuracy in the Bjaran training task improved over the course of training, but only for certain training conditions. Controlling for training condition, subjects improved over the course of the training task, there was a significant effect of block number, \(\chi^2(1) = \)
34.08, p < .0001, H-B corrected). However, the block number effect depended on the training condition. The interaction between block number and the coded variable “Same vs others” was significant ($\chi^2(1) = 17.49, p < .0001$, H-B corrected), indicating that the block effect was different for the Same condition, such that the Same condition did not improve in accuracy over the course of training, while the other conditions did. The block number effect was also different for the Familiar Dimension compared to the New Dimension condition: it was smaller for the Familiar Dimension, as indicated by a significant interaction between block and the “Familiar vs new” variable ($\chi^2(1) = 14.96, p = .0003$, H-B corrected).

Table 2.5: Regression model for Bjaran category training, p values for z-statistics in the table are not corrected, see text for F-statistics with corrected p values

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 0.85486 | 0.13765 | 6.210 | 5.29e-10 *** |
| Block number | 0.19097 | 0.02830 | 6.749 | 1.49e-11 *** |
| “Same vs others” | -0.76591 | 0.47931 | -1.598 | 0.110 |
| “Familiar vs new” | -0.46555 | 0.36483 | -1.276 | 0.202 |
| Bl. no.*“Same vs others” | 0.31161 | 0.07520 | 4.144 | 3.42e-05 *** |
| Bl. no.*“Familiar vs new” | 0.45090 | 0.09929 | 4.541 | 5.59e-06 *** |

Post-hoc tests for individual training conditions were run, consisting of logistic mixed-effects models, where the dependant variable was accuracy in the Bjaran training task, and block number (centered) was treated as a continuous, fixed effect. There were random effects for item, subject, and each subject’s block number effect.

For the Same Dimension condition, there was no significant block number effect ($p = .46$, H-B corrected), indicating that subjects in the Same Dimension condition did not improve in accuracy over the course of training. In the model for the Familiar Dimension condition, there was a significant effect of block number ($\chi^2(1) = 18.57, p < .0001$, H-B corrected), indicating that subjects in this condition did improve. Likewise, there was a significant effect of block number for the New Dimension condition ($\chi^2(1) = 12.06, p = .001$, H-B corrected) indicating that these subjects also improved over the course of training.
There were differences in accuracy between the training conditions at the start of the training task, but by the end of the task, the training conditions did not differ in accuracy. To assess differences in accuracy among the three training conditions in Block 1, a logistic, mixed-effects model was fit to the Bjaran training data. The dependent variable was accuracy on a single trial (coded as 1 or 0). There were two independent training condition variables that were considered as categorical fixed effects. The model included random effects for subject and item. The two condition variables allowed for comparisons in mean accuracy among the training conditions ("Same vs others" and "Familiar vs new"). In Block 1, subjects in the Same Dimension condition were more accurate than the other two conditions (by log-likelihood ratio test, $\chi^2(1) = 9.37, p = .002$). Additionally, subjects in the New Dimension condition were more accurate than those in the Familiar Dimension condition ($\chi^2(1) = 6.4, p = .01$). By Block 5, there was no difference in accuracy between the three training conditions: in a logistic, mixed-effects model predicting accuracy in Block 5 with the contrast coded condition variables described above as fixed effects (and random effects for subject and item), there was no difference in accuracy of the Same Dimension condition compared to the other two conditions ("Same Dimension vs others", $p = .3$). Furthermore, mean accuracy in the Familiar Dimension condition did not differ from the New Dimension condition ("Familiar vs new", $p = .25$).

It is worth noting that subjects in the Same Dimension condition were trained with a Bjaran category structure that was nearly the same as in Xorx, with F0 as a category-relevant dimension, but where the F0 boundary between word categories was shifted, relative to Xorx. As a result, there were only 6 stimuli whose category needed to be relearned: those with an F0 of 106 Hz. In the Xorx category structure, these stimuli belonged in the same category as stimuli with F0 of 108 & 110 Hz, but in Bjaran, they belonged with stimuli that had an F0 of 100-104.

If the category training in Bjaran is effective, subjects should “shift” this F0 boundary, and should be just as accurate with the shifted stimuli as the other stimuli. However, the
mean accuracy on these 6 Bjaran stimuli ($M = .57$) was lower than all other stimuli ($M = .76$). See figure 2.10, which shows the mean accuracy for the 6 “shifted” stimuli in the Bjaran-Same Dimension condition, compared to the accuracy for the other 30 Bjaran stimuli.

![Figure 2.10: Training accuracy for the Same Dimension condition, comparing shifted and non-shifted stimuli. Error bars represent $SEM_{betw}$.

Not only was the mean accuracy lower on the shifted stimuli, but accuracy did not improve for these stimuli over the course of training. Furthermore, even when leaving out the non-shifted stimuli, subjects still did not improve over the course of training. To assess the accuracy of the six “shifted” stimuli vs. the “non-shifted” stimuli over the course of training, a logistic mixed-effects model was fit to the Bjaran training data for the Same Dimension condition. The dependent variable was accuracy on a single trial (coded as 1 or 0). Block number (centered) was treated as a continuous, fixed effect. The model also included a contrast-coded variable for stimuli type, treated as a categorical fixed effect, and an interaction term (“Block number * stimuli type”). The categorical variable (“Shifted vs. not shifted”) compared the mean accuracy for the shifted stimuli (with an F0 of 106) to the mean accuracy of the other stimuli, controlling for block number. The interaction tested to see if the shifted variables had a different block effect compared to the non-shifted
stimuli. Controlling for block number, the accuracy of the shifted stimuli was significantly lower than that of the non-shifted stimuli (significant “Shifted vs. not shifted” variable, $F(1, 7) = 41.1, p < .00001$). Overall, subjects did not improve in accuracy (controlling for stimuli type, there was no effect of block number, $p = .75$). Furthermore, the interaction between block number and stimuli type was not significant ($p = .39$). This indicates that non-significant block effect did not differ for the shifted stimuli.

Taken together, these results suggest that, though the Same Dimension group had higher accuracy in Block 1 compared to the other groups, they did not improve over the course of the training task. Subjects in the other two training conditions did improve, in a sense “catching up” to the Same Dimension condition. The Same Dimension had an advantage at the outset of training because they were able to use the same relevant dimension to categorize the stimuli. However, for the stimuli which were essentially in the same category as in Xorx, the Same Dimension subjects started off with fairly high accuracy, but further training did not help them improve their accuracy for these stimuli. Moreover, for the six shifted stimuli in the Same Dimension condition, subjects started off with fairly poor accuracy (55%) and didn’t improve at all. This suggests that training did not help the subjects in the Same Dimension condition to shift their F0 boundary: they were utilizing the contrastive cue of F0 in a “Xorx” way instead of a “Bjaran” way.

### 2.6.5 Bjaran difference ratings

For the Bjaran difference rating task, subjects heard all pairwise combinations of the Bjaran stimuli (36 synthetic “bi” syllables, varying along one of three training dimensions and relative amplitude) and provided a difference rating for each pair on a scale from 1-9.

The Bjaran pre- and post-training difference ratings were analyzed using MDS (INDSCAL), following the same procedure as the scaling of the Xorx similarity judgments. As in the analysis for the Xorx data, the INDSCAL analysis was implemented in R, using the smacofIndDiff() function from the SMACOF package [De Leeuw and Mair 2011], assuming
an interval distance model.

Figures 2.11, 2.12 and 2.13 show the scaling results for each of the Bjaran training conditions. As in the Xorx scaling plot, points represent average stimulus coordinates across all subjects, pre-training, and are colored to represent the word category of the stimulus. The endpoint of the vectors extending from the points is the average stimulus coordinate, post-training.

The best-fitting scaling solutions for all three training conditions were in two dimensions. As in the analysis of the Xorx scaling results, emergent psychological dimensions can be interpreted both by visual inspection of the scaling plots, and by modeling the extent to which the scaling dimensions are correlated with physical properties of the stimuli.

### 2.6.5.1 Scaling dimensions in the Same Dimension condition

Subjects in the Bjaran Same Dimension condition were trained with F0, which was also the category relevant training dimension in Xorx. In Bjaran, however, the F0 category boundary was shifted, relative to the Xorx F0 boundary.

As in Xorx, one of the scaling dimensions (in this case, D2) for this training condition is related to both F0 and word category. Visual inspection of 2.11 shows that D2 is related to both word category and F0: as D2 increases, so does the F0 of stimuli (all “Mechanic” stimuli, with a higher F0, are to the right of the plot).

To evaluate the extent to which the scaling dimensions in the Same Dimension group space correlated with the F0 values of the stimuli, a linear regression model was fit to the results of the Bjaran-Same Dimension INDSCAL group space, using the R Stats package (R Core Team, 2015). The dependent variable was the F0 values (z-scored) of the Bjaran-Same Dimension stimuli, which were regressed on the scaling coordinates of the stimuli for each of the two dimensions from the Bjaran-Same Dimension group space. F0 was found to load heavily on D2, as indicated by a significant effect for D2 in the regression model ($F(1,33) = 1713.38, p < .00001$). The effect for D1 was not significant ($p = .12$).
Moreover, D2, and not D1, is related to word category: all *rock* stimuli are on one side of the plot, and all *mechanic* stimuli are on the other side. This was confirmed by a logistic regression model fit to the results from the Bjaran-Same Dimension group space using the R Stats package. The dependent variable was word category (*mechanic* or *rock*), which was regressed on the scaling coordiantes of the stimuli for D1 (D2 was left out, because it perfectly predicts word category, i.e. *rock* always occurs with some coordinates along D2, and never with other D2 coordinates). In this model, coordinates along D1 did not predict word category \( (p = .72) \). The best interpretation of D2, therefore, is the “category-relevant
dimension”—it is related to F0 and distinguishes stimuli of one word category from the other.

Relative amplitude is related to D1. This was confirmed with a linear regression model that was fit to the results of the Bjaran-Same Dimension INDSCAL group space. The dependant variable was the relative amplitude values (z-scored) of the Bjaran-Same Dimension stimuli, which were regressed on the scaling coordinates of the stimuli for each of the two dimension in the Bjaran-Same Dimension group space. This model also included a contrast-coded categorical variable for stimulus category (“Mechanic vs rock”) and an interaction term, to control for effects of stimuli category. Relative amplitude was found to load heavily on D1, as indicated by a significant effect for D1 in the regression model \( F(1,30) = 1308.10, p < .00001 \). The effect for D2 was not significant \( (p = .28) \). Controlling for any effects of stimuli category, relative amplitude loads on D1 and not D2.

It should be noted that the vectors in the scaling plot for the “Same Dimension” condition are quite short, which indicates that there was not much change in the difference ratings after training.

### 2.6.5.2 Scaling dimensions in the “Familiar” Condition

Subjects in the Bjaran Familiar Dimension condition were trained with duration, which was present in the Xorx stimuli, but in Xorx it was not a contrastive cue.

The scaling results for this condition (see section 2.12) are harder to interpret. Neither dimension seems to be related to the category-relevant dimension for this condition, which is duration.

The rock and mechanic stimuli are intermixed in the plot. To evaluate the extent to which the scaling dimensions in the Familiar Dimension group space correlated with the duration values of the stimuli, a linear regression model was fit to the results of the Bjaran-Familiar Dimension INDSCAL group space, using the R Stats package [R Core Team, 2015]. The dependent variable was the duration values (z-scored) of the Bjaran-Familiar Dimension stimuli, which were regressed on the scaling coordinates of the stimuli for each of the two
Figure 2.12: MDS solution for Bjaran difference ratings, “Familiar Dimension” condition

dimensions from the Bjaran-Familiar Dimension group space. Duration was not found to load on either dimension; as indicated by non-significant effects for D1 \((p = .86)\) nor D2 \((p = .82)\).

Furthermore, neither dimension was related to word category. This was confirmed with a logistic regression model fit to the results from the Bjaran-Familiar Dimension group space using the R Stats package. The dependent variable was word category (mechanic or rock), which was regressed on the scaling coordinates of the stimuli for D1 and D2. In this model, coordinates along D1 did not predict word category \((p = .93)\), and neither did coordinates
Relative amplitude was found to load on D1, but not D2.

To evaluate the extent to which the scaling dimensions in the Familiar Dimension group space correlated with the relative amplitude values of the stimuli, a linear regression model was fit to the results of the Bjaran-Familiar Dimension INDSCAL group space. The dependent variable was the relative amplitude values (z-scored) of the Bjaran-Familiar Dimension stimuli, which were regressed on the scaling coordinates of the stimuli for each of the two dimensions from the Bjaran-Familiar Dimension group space. In this model, D1 was significant \( F(1, 30) = 214.33, p < .0001 \), but not D2 \( (p = .2) \). Relative amplitude was the category-irrelevant dimension for this (and all) Bjaran training conditions. However, relative amplitude does not seem to be entirely “irrelevant” for subjects in the Familiar Dimension condition. It seems that subjects in this condition may have been using relative amplitude as a basis for rating the differences between stimuli, rather than duration. However, note that the stimuli at the bottom of the plot seem to be moving toward the stimuli at the top of the plot, post-training. This suggests that if subjects in this condition were using relative amplitude as a basis for making difference ratings prior to training, training may have functioned to draw their attention away from this dimension to some extent (because, after training, this dimension was somewhat “compressed”, i.e. stimuli moved toward each other along this dimension).

### 2.6.5.3 Scaling dimensions in the New Dimension condition

Subjects in the Bjaran New Dimension condition were trained with pitch-rise, which was not present in the Xorx stimuli, either as a relevant or irrelevant dimension.

In the New Dimension condition, D2 is related to both pitch-rise and word category. In figure 2.13 the stimuli with the lower pitch-rise are represented with lower coordinates along D2, and as D2 increases, so does pitch-rise.

The relationship between D2 and pitch-rise was confirmed with a linear regression
model, which was fit to the results of the Bjaran-New Dimension INDSCAL group space. The dependant variable was the pitch-rise values (z-scored) of the Bjaran-Same Dimension stimuli, which were regressed on the scaling coordinates of the stimuli from the Bjaran-Same Dimension group space, for each of the two dimensions. Pitch-rise was found to load on D2: there was a significant D2 effect \( (F(1, 30) = 11.38, p = .002) \). Pitch-rise did not load on D1, however: the effect for D1 was not significant \( (p = .94) \).

D2 in this condition is also related to word category—most of the rock stimuli are grouped together on the left side of the plot, while most of the mechanic stimuli are on the
right side. This separation is not perfect: some of the *rock* stimuli are closer in distance to the *mechanic* stimuli. However, the relationship between D2 and word category was confirmed with a logistic regression model fit to the results from the Bjaran-New Dimension group space. The dependent variable was word category (*mechanic* or *rock*), which was regressed on the scaling coordinates of the stimuli for D1 and D2 (D2 can be included in this model, because D2 doesn’t perfectly predict word category). There was a significant effect for D2 ($\chi^2(1) = 30.12, p < .0001$), but not for D1 ($p = .95$). Therefore, the best interpretation of D2, in the “New Dimension” training condition, is the “category relevant” dimension—it is related to pitch-rise, and distinguishes (most) *rock* stimuli from *mechanic* stimuli.

D1, for the New Dimension condition, was related to relative amplitude. This was confirmed with a regression model which was fit to the results of the Bjaran-New Dimension INDSCAL group space. The dependant variable was the relative amplitude values (z-scored) of the Bjaran-Same Dimension stimuli, which were regressed on the scaling coordinates of the stimuli from the Bjaran-Same Dimension group space, for each of the two dimensions. Relative amplitude was found to load on D1, and not D2: there was a significant effect for D1 ($F(1,33) = 3071.19, p < .00001$), but not for D2 ($p = .11$).

### 2.6.6 Category learning in Bjaran

As in Xorx, the Bjaran scaling data was evaluated for evidence of Acquired Similarity and Acquired Distinctiveness.

#### 2.6.6.1 Acquired similarity

If within-category acquired similarity is taking place, then the average within-category distances should become smaller after training. This was found to be the case in the Bjaran scaling data, but only for the Familiar and New conditions.

Within-category distance was obtained by finding each subject’s category centroids (same procedure as in Xorx). Then, the Euclidean distance between the category centroid
and each stimulus in that category was calculated, following the Euclidean distance formula for two dimensions (because there were two, rather than three dimensions in the Bjaran scaling solutions):

$$\sqrt{(D_{1\text{mean}} - D_{1\text{stimulus}})^2 + (D_{2\text{mean}} - D_{2\text{stimulus}})^2}$$

Finally, an overall within-category distance measure was found for each subject by taking the mean of all within-category distances for each word category.

Each subject had four within-category distances: within-category distance for rock stimuli, pre- and post-training, and within-category distance for mechanic stimuli, pre- and post-training. Figure 2.14 shows the within-category distances (averaged across subjects) for the Bjaran stimuli, pre- and post-training, split by training condition.

A measure of “within-category shift” was constructed by subtracting post-training distances from pre-training distances.

To account for multiple testing with the Bjaran difference rating data for the family
of models that investigated within-category shift (the omnibus test, described in the next paragraph, and three post-hoc tests, one for each training condition, described below), alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “p<.001”, H-B corrected) when applicable.

To evaluate the effect of training on judgments of similarity and in order to test for a significant change in within-category distance after training, a linear model was fit to the results of the pre- and post-training Bjaran INDSCAL scaling configurations using the R Stats package. The dependent variable was the “within-category shift” variable (the difference between the pre- and post-training within-category distance, which was the Euclidean distance between each subject’s category centroid and the positions of each stimulus in each subjects individual configuration, see above for full description of how this variable was computed). The independent variables were two categorical condition variables. These were contrast-coded to compare the mean within-category shift among the training conditions. One variable compared the mean shift for the Same condition with the mean shift for the other two training conditions (“Same vs others”), and the other variable compared the mean shift for the Familiar condition with that of the New condition (“Familiar vs new”).

As in the analysis of the Xorx data, in order to obtain an F-statistic to test for the significance of the intercept in this model, the full model was compared with a reduced model where the variable of interest was left out. The full and reduced models were compared using the anova() function in the R Stats package. This model comparison test showed that including the intercept in the full model resulted in a significant reduction in error \( F(1,116) = 22.6, p < .0001, \) H-B corrected). In other words, controlling for training condition, the mean within-category shift from pre- to post-training was significantly different from zero (see \( \text{2.6} \)).

However, in the omnibus test, the training effect was not different for the Same Dimension condition compared to the other two conditions: the condition variable “Same vs others” was not significant \( (p = .1, \) H-B corrected). There was also no difference in the training effect
in the omnibus test for the Familiar Dimension condition compared to the New Dimension condition (“Familiar vs new” was N.S., \( p = .77 \), H-B corrected).

Table 2.6: Bjaran Acquired Similarity regression model; p values for t-statistics in the table are not corrected. See text for F-statistics and corrected p values

|                          | Estimate | Std. Error | t value | Pr(>|t|) |
|--------------------------|----------|------------|---------|----------|
| (Intercept)              | 0.030778 | 0.006480   | 4.749   | 5.934e-06 *** |
| “Same vs others”         | 0.045711 | 0.022256   | 2.054   | 0.04226 *   |
| “Familiar vs new”        | -0.004938| 0.016956   | -0.291  | 0.7714    |

Post-hoc linear models were fit to the results of the pre- and post-training Bjaran INDSCAL configurations for each of the three training conditions. The dependent variable in each of these post-hoc models was the “within-category shift” variable. These models were simple, intercept-only models, which were used to test if the within-category shift was significantly different from zero for each condition (in other words, if there was a significant shift in within-category distance, from pre- to post-training).

For the Same condition, there was no significant shift in within-category distances (intercept was N.S., \( p = .62 \), H-B corrected). For the Familiar condition, there was a marginally significant shift \( (F(1, 39) = 6.23, p = .06) \), H-B corrected), and the intercept was positive, indicating that the (marginal) shift was in the correct direction for an AS effect. For the New condition, there was a significant shift \( (F(1, 39) = 13.26, p = .004, \) H-B corrected), and the intercept in the simple model was positive, the correct direction for an AS effect.

### 2.6.6.2 Acquired distinctiveness

If acquired distinctiveness (AD) is taking place, then the average between-category distances should become larger after training. There was evidence of an AD effect in the Bjaran scaling data, but only for the Familiar and New conditions.

The distance between category centroids was calculated for each subject, pre- and post-training, using the Euclidean formula for distance in two dimensions. Figure 2.15 shows
Figure 2.15: Bjaron between-category distance, pre- vs post-training. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent $\text{SEM}_{\text{betw}}$.

the average between-category distance for pre- and post-training, split by condition.

Each subject’s pre-training between-category distance was subtracted from their post-training between-category distance to obtain a measure of “between-category shift” for each subject.

To account for multiple testing with the Bjaron difference rating data for the family of models that investigated between-category shift (the omnibus test, described in the next paragraph, and three post-hoc tests, one for each training condition, described below), alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “$p<.001$”, H-B corrected) when applicable.

To evaluate the effect of training on judgments of similarity and in order to test for a significant change in between-category distance after training, a linear model was fit to the results of the pre- and post-training Bjaron INDSCAL scaling configurations using the R Stats package. The dependent variable was the “between-category shift” variable (described above). The independent variables were two categorical condition variables. These
were contrast-coded to compare the mean between-category shift among the training conditions. One variable compared the mean shift for the Same condition with the mean shift for the other two training conditions ("Same vs others"), and the other variable compared the mean shift for the Familiar condition with that of the New condition ("Familiar vs new"). In this model, the intercept was significant \( F(1, 56) = 437, p < .0001, \text{H-B corrected} \). This indicates that, controlling for condition, there was a significant difference between pre- and post-training between-category distances. However, this effect was different for each training condition, as indicated by significant contrast-coded predictors for training condition. The effect was reversed for the Same Dimension compared to the others ("Same vs others, \( F(1, 56) = 1154.74, p < .0001, \text{H-B corrected} \)), and the magnitude of the effect was bigger for the New condition compared to the Familiar condition ("Familiar vs new", \( F(1, 56) = 168.84, p < .0001, \text{H-B corrected} \)).

Post-hoc models were run for each training condition, consisting of linear models fit to the results of the pre- and post-training Bjaran INDSCAL configurations. The dependent variable in each of the post-hoc models was the "between-category shift" variable. These models were simple, intercept-only models, which were used to test if the between-category shift was significantly different from zero for each condition (in other words, if there was a significant shift in between-category distance, from pre- to post-training).

For the Same condition, there was a significant shift in the between-category distances (significant intercept, \( F(1, 19) = 607.93, p < .0001, \text{H-B corrected} \)). However, the intercept in this model was negative, indicating that the shift was not in the expected direction for an AD effect: distances actually became smaller after training, which suggests that for subjects in this condition, the categories were becoming more similar to each other. There was a significant shift for both the Familiar Dimension (significant intercept: \( F(1, 19) = 48.69, p < .0001, \text{H-B corrected} \)) and the New Dimension conditions (significant intercept: \( F(1, 19) = 526.52, p < .0001, \text{H-B corrected} \)). For both of these conditions, the intercept in the simple model was positive, meaning that the shift was in the expected direction for
an AD effect. For both the Familiar and New conditions, between-category distances were larger after training, which suggests that for these subjects, the categories were becoming more distinct from each other.

2.7 Discussion

2.7.1 Xorx results summary

In general, subjects learned to distinguish between the two Xorx word categories fairly quickly (within the first 36 trials). However, they continued to improve over the course of training, and were more accurate in the last block of trials compared to the first block.

The dimensions in the MDS solutions for Xorx corresponded to F0/word-category and duration. Subjects judged similarities based on F0, even in the pre-training task (within-category stimuli are closer together along D1, even pre-training). This suggests that F0 may be more perceptually “available” to the subjects as a dimension along which perceptual judgments are made, even without any training on the F0 dimension. When comparing the MDS solutions from the pre- and post-training difference tasks, results show that training did affect difference judgments: there was evidence of an Acquired Similarity effect, where within-category distances were smaller after training. There was also evidence of an Acquired distinctiveness effect, where the distance between category centroids was larger after training.

2.7.2 Bjaran results summary

After the first block of training in Bjaran, subjects in the Same Dimension condition (trained with F0) had higher accuracy (75%) compared to the other two conditions. This is not terribly surprising, given that the structure of Bjaran for these subjects was similar to that of Xorx—there were only six stimuli (those with F0 of 106 Hz) that needed to be “shifted” in Bjaran in order to be correctly categorized. However, the subjects in this condition did not improve over the course of the training task. Subjects started out with poor
accuracy on the 6 “shifted” stimuli, and did not improve over the course of training. It is possible that subjects were categorizing Bjaran sounds on the basis of the Xorx F0 boundary, not the Bjaran F0 boundary. Furthermore, for the other non-shifted stimuli, accuracy was (marginally) worse in the last block compared to the first block. Subjects in the other two training conditions started off with lower accuracy compared to the Same Dimension subjects. They improved over the course of the training task, but to different degrees: the Familiar Dimension group improved more. Subjects in the Familiar condition had the lowest accuracy of all the training conditions to begin with (60%), but by the end of training they had caught up with the other two conditions in terms of accuracy.

The dimensions in the MDS solutions for each Bjaran training group corresponded (in most cases) to the physical dimensions of the stimuli. For both the Same Dimension and New Dimension conditions, D2 corresponded to the category-relevant dimension for that training condition (F0 and pitch-rise, respectively). None of the dimensions in the scaling results corresponded to the category-relevant dimension for the Familiar Dimension group (duration). For all three training groups, D1 corresponded to relative amplitude, which was the irrelevant dimension in Bjaran.

There was evidence of category learning when comparing the scaling results of pre- and post-training difference judgements. Two category-learning effects were examined: Acquired similarity and Acquired distinctiveness. There was evidence of both effects in the New Dimension and Familiar Dimension conditions (though for this condition, the acquired similarity effect was marginally significant). There was no evidence for either effect in the Same Dimension condition. In fact, for the Same Dimension condition, the category centroids were closer together after training, instead of further apart (as expected for the effect of Acquired distinctiveness).

While subjects in the Familiar Dimension did show some evidence of category learning, these results should be interpreted with caution, because it is unclear just which psychological dimensions this group was using as a basis for difference judgments. It appears as though the
Familiar Dimension subjects used relative amplitude (the supposed “irrelevant” dimension for this condition) as a basis for similarity, because stimuli with similar relative amplitude values were grouped together in the scaling results. There is no direct evidence to support the idea that this group used duration as a basis for difference ratings after training; neither of the scaling dimensions corresponded with duration or word-category. This is despite the fact that these subjects did improve in accuracy over the course of the training task—presumably, attending to differences in duration allowed them to be successful in categorizing the stimuli. So, there seems to be a difference between the two tasks (training vs difference ratings) in terms of subjects’ ability to attend to duration.

There are (at least) two possibilities for why these subjects did not attend to duration as a basis for making difference judgments. The first has to do with interference from Xorx: perhaps for these subjects, it was difficult to switch from ignoring duration in Xorx to attending to it in Bjaran, particularly in a task where subjects were not (implicitly) “told” to pay attention to duration (as in the case of the training task). Or, there may have been some interference from the so-called “irrelevant” dimension: the fact that relative amplitude seemed to be important to these subjects when judging similarity shows that subjects were having a hard time ignoring this dimension. This may be due to the relative salience of the two dimensions: it may be harder to distinguish small changes in duration compared to the changes in relative amplitude. In the Familiar Dimension condition, the “irrelevant” dimension may have been more salient than the “diagnostic dimension”. In fact, both of these sources of interference (from Xorx and from the irrelevant dimension) may have contributed to subjects’ inability to attend to duration when making difference judgements.

2.7.3 Which training dimensions are most effective?

Taking all results into account, subjects in the New Dimension condition were the most successful. Subjects in this group learned to attend to the relevant dimension (pitch-rise) and ignore the irrelevant dimension of relative amplitude during both tasks (training
and difference rating). They improved over the course of the training task, and analysis of the scaling results showed effects of Acquired Similarity and Acquired Distinctiveness (see table 2.7).

Table 2.7: Effectiveness of training dimensions

<table>
<thead>
<tr>
<th>Bjaran condition</th>
<th>Attention to relevant dimension</th>
<th>Training accuracy</th>
<th>Acquired Similarity</th>
<th>Acquired Distinctiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Dimension</td>
<td>yes</td>
<td>no improvement</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Familiar Dimension</td>
<td>?</td>
<td>improved</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>New Dimension</td>
<td>yes</td>
<td>improved</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Subjects in the Familiar Dimension condition were somewhat successful—though they had the lowest accuracy at the beginning of the training task, they improved the most compared to the other conditions. These subjects did show effects of Acquired Similarity and Acquired distinctiveness, though there was no evidence to support the idea that they were attending to the category-relevant dimension of duration, at least in the difference rating task. It may have been hard for these subjects to attend to duration in Bjaran after having been trained to ignore it in Xorx, or because duration was not particularly salient, compared to relative amplitude.

Finally, there seemed to be a real cost associated with using the same training dimension (with different category boundaries) in both Xorx and Bjaran. Subjects in the Same Dimension condition did not improve over the course of training, nor did they show much evidence of category learning.

Altogether, these results align with the hypotheses made on page 32, which predicted that the Same Dimension condition would fare best, because for this condition there was little interference from the (alien)L1.
2.7.4 Linguistic relevance of stimuli in Experiment 1

Though Xorx and Bjaran are not real languages, the acoustic dimensions of the Xorx and Bjaran stimuli were chosen to reflect dimensions that are phonetically relevant in human languages. In some form, F0, duration, pitch-rise, and relative amplitude are used as primary or secondary cues to indicate segmental contrast in a variety of linguistic contexts.

F0 is used as a primary cue in “register-tone” languages, where there is a distinction between level tones with higher and lower pitch. Many African languages have a two-way register tone contrast (e.g. Gur, Chadic, and Cushitic languages), and some have three-, four-, or even five-way tone contrasts (Odden, 1995). F0 also functions as a secondary cue for distinguishing sound contrasts in many languages including English (Francis and Nusbaum, 2002), Korean (Francis and Nusbaum, 2002), Japanese (Kinoshita et al., 2002), and Arabic (Al-Tamimi and Khattab, 2011). Pitch-rise, as implemented in Experiment 1, is related to tone contour, a contrastive feature in over 100 world languages, e.g. Mandarin, Thai, Mixtec, and Hausa (Gordon, 2002).

Duration is a contrastive feature in languages with vowel length distinctions, e.g. Swedish, Estonian, Navajo, and Japanese, and in languages which distinguish singleton and geminate consonants, e.g. Arabic and Italian (Al-Tamimi and Khattab, 2011). Duration is utilized as a secondary cue in many languages, for instance in English as both a cue to vowel and stop voicing contrasts (Kondaurova and Francis, 2010).

Both pitch rise and F0 are related to intonation, which is a suprasegmental feature of spoken languages. The intonation pattern of an utterance can indicate different grammatical or pragmatic meanings. For instance, in many languages, including English, rising intonation can signal a question, while falling intonation can signal a declarative statement (compare the intonation patterns for “She baked a cake.” vs “She baked a cake?”). F0 and duration both contribute to the suprasegmental feature of stress (lexical prominence), which can play a contrastive roll in some languages, for instance English (e.g. “record” as a noun, “record”
as a verb).

Relative amplitude, as implemented in Experiment 1, does not have a direct correlate in human languages. However, the relative amplitude of single harmonics during voicing (also known as spectral tilt) distinguishes between “creaky”, “modal”, and “breathy” voice quality in languages with contrastive voice quality types, such as Newar, Gujarati, and many Northwest Amer-Indian languages, e.g. Kwakw’ala (Gordon and Ladefoged, 2001). Spectral tilt also underlies nasality, which is contrastive in number of languages, e.g. French, Portuguese, Lua, and Yoruba (Maddieson and Disner, 1984), and other phonemic contrasts including the Korean stop contrast which is the focus of Experiment 2 (chapter 3).

2.7.5 Aliens vs. Humans (or, The relationship between Bjaran training conditions and English)

It is important to acknowledge that “Xorx” is, of course, not a true L1 for any of the subjects in the study. There is a possibility that any “interference” when learning Bjaran comes from the influence of the subject’s real L1, rather than Xorx.

Most of the subjects in the study (44/60) had English as a first language. Regarding the relationship between English and the Bjaran training dimensions, both F0 and duration could be considered as “Familiar Dimensions”. There are no segmental phonemic contrasts in English that have F0 or duration as the primary contrastive cue. English does not have contrastive vowel length (as in Swedish or Finnish), nor do segments with different fundamental frequencies form contrastive pairs (as in languages with tone distinctions). Nonetheless, there is evidence to suggest that English speakers utilize these dimensions as secondary contrastive cues. For instance, F0 is used as a secondary cue in determining the voicing of initial stops, particularly when VOT information (which is the primary cue) is degraded or unavailable. Duration functions as a secondary cue for discriminating between English tense and lax vowels (e.g. /bit/ and /bɛt/), and also aids listeners in discriminating between voiced and voiceless stops that are allophonically both voiceless in certain contexts, like at the end of
words (e.g. \[be:d\] and \[bet\]). Furthermore, both duration and F0 are components of stress, a suprasegmental feature in English that can, in a few cases, function as a contrastive cue (for instance in word pairs like “record”, as verb, and “record”, as a noun).

In sum, regarding the relationship between Bjaran conditions and English, both F0 and duration are “Familiar Dimension” cues, because though they can function as secondary cues to English contrasts, in Bjaran they are the primary contrastive cue. Thus, if there was only interference from English, and none from Xorx, we might expect to see no difference between the “Same” and “Familiar” training conditions. However, there were differences between the conditions in terms of subjects’ ability to learn Bjaran effectively: the “Same Dimension” subjects did not learn to correctly categorize the six “shifted” stimuli during training, nor did they change their difference ratings based on F0 to reflect the new category structure of Bjaran. The “Familiar Dimension” subjects were able to learn to utilize duration as a diagnostic cue, at least during the training task, because they improved over the course of the training task by learning to attend to duration (which is the only way they could improve in the training task). It stands to reason, therefore, that at least some of the interference experienced by subjects in the “Same Dimension” condition is related to interference from Xorx, and not (entirely) from English.

In any case, the fact that the “New Dimension” group was the most effective, by all measures, is interesting, no matter if we are considering relationships between Bjaran and Xorx or Bjaran and English. Pitch-rise is “New Dimension” in both cases: it was not a part of the Xorx stimuli at all, and it is not used contrastive cue in English (it is utilized suprasegmentally, e.g. in question formation, but not contrastively as it is in tone languages). Even if the relationship between Bjaran and English is the most important relationship to consider here, it is still the case that the “New Dimension” subjects out-performed other training groups.
2.7.6 Implications of Experiment 1 for phonetic training in real languages

The results of Experiment 1 show that it is important to take prior experience into account when choosing a training dimension for learning a new sound contrast. The “alien” languages in Experiment 1 are certainly less complex than human languages, but the acoustic dimensions of the stimuli do have a basis in human languages. Although Xorx and Bjaran are not real, the learning and categorization processes examined in Experiment 1 are real, and can be interpreted in the context of models such as PAM and SLM, which predict that sounds that are less similar to L1 sounds will be easier to process. Likewise, the Bjaran condition that was the least similar to Xorx was the condition that seemed to have the most success.

In the next experiment, English speakers are trained to hear a set of Korean sounds. Much of the methodology from Experiment 1, including the training tasks and scaling analysis of difference rating, is utilized in Experiment 2 (though there are some differences in terms of overall design). Most importantly, in Experiment 2 there are different training conditions which take into account the relationship between English and Korean, and these are comparable to the training conditions in Experiment 1 ("Same Dimension", "Familiar Dimension", and "New Dimension"). In this way, the results from Experiment 1 can be used to guide predictions about the effectiveness of particular training dimensions in the context of real languages.
Chapter 3

Experiment 2: Training adults to hear a Korean voicing contrast

The goal of Experiment 2 is to train native English speakers to hear a set of Korean consonants by using a specific acoustic dimension as a contrastive cue. Experiment 2 directly addresses Research Questions 1 & 2:

Research Question 1 Will phonetic training result in better perception of non-native sounds?

Research Question 2

2A Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others?

2B Given this L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language?

Experiment 1 explored how training with acoustic dimensions that differed in their relationship to a “quasi-L1” may result in better (or worse) learning. The experiment, in addition to providing an opportunity to pilot experimental methodology, yielded several predictions about the kinds of training dimensions that might be best for learners, given their L1. Experiment 2 compares the effect of phonetic training with dimensions that have different relationships to English (the subjects’ L1), similar to how subjects in Experiment 1
were trained with Bjaran dimensions that differed in particular ways from Xorx (the subjects’ “quasi-L1”). So, one further research question addressed in Experiment 2 is:

**Research Question 3** Are the predictions of Experiment 1, based on artificial languages, borne out in cases of training with real languages?

The results from Experiment 1 indicate that training with some dimensions may be more effective than with others, namely, dimensions that don’t interfere with native dimensions, or that are completely novel. Likewise, a general prediction for Experiment 2 is that subjects trained with some dimensions will exhibit a stronger training effect than subjects trained with other dimensions, precisely because of how those training dimensions operate in the phonological system of English.

This chapter is organized as follows: first, there is an overview of the Korean sounds that are the object of training in the experiment (page 73), followed by a discussion of how these sounds are related to English in terms of contrastive acoustic dimensions (page 79). Details pertaining to experimental design, procedure (page 81) and stimuli (page 87) are next. Then, the results of the experiment are presented (page 100). A general discussion of the experiment begins on page 133.

### 3.1 Korean stop consonants

In Experiment 2, subjects were trained to hear a set of Korean stops, which are consonants that are articulated with a complete constriction in the vocal tract. The contrast between Korean stops made at the same place of articulation (e.g. at the velum, alveolar ridge, or lips) is often described as a voicing contrast. Many of the world’s languages have a two-way voicing distinction between stops. For example, in English there are voiceless stops, e.g. /k/, and voiced stops, e.g. /g/, both of which share the same place and manner of articulation (in this case, they are both velar stops). In Korean, however, there is a three-way distinction between stops at the same place of articulation. Though the Korean contrast is
described as a distinction in voicing, the distinction between Korean stops is more complicated than a three-way difference in “voicing”, as it is commonly characterized. In other languages with a three-way voicing distinction, for instance Thai, the voicing categories are often “voiced”, referring mainly to a negative voice-onset time (where voicing begins before the stop release), “voiceless”, referring to positive voice-onset time, and “aspirated”, referring to longer positive voice-onset time (Lisker and Abramson, 1964). However, in Korean, the three types of stops are technically all “voiceless”, that is, they all have positive values of voice-onset time. Moreover, they are made with the same pulmonic egressive airstream mechanism (Cho et al., 2002). So, the voicing distinction for stop consonants in Korean is unlike those found in other languages, even those with a three-way voicing contrast.

The three types of Korean stops are aspirated, strong, and weak. Strong stops in Korean are also referred to as “fortis”, “tense”, and “pressed”. Likewise, weak stops have alternate labels: “lenis” and “spread”. In this and following chapters, the labels used to refer to the Korean stops will be aspirated, strong and weak. Aspirated stops, when written using the phonetic alphabet, will be denoted with a superscript \(^h\) diacritic (/k\(^h\)/), strong, with an asterisk diacritic (e.g. /k*/), and weak, with no diacritic (/k/).

The three types of Korean stops occur at three places of articulation: bilabial (/p, p*, p\(^h\)/), dental (/t, t*, t\(^h\)/), and velar (/k, k*, k\(^h\)/). Full three-way contrasts of the Korean stops exist mainly in word-initial position, and there are a number of minimal triads that illustrate the contrast between initial aspirated, strong, and weak stops (see table 3.1). Minimal triads in other positions (e.g. medial or final) are less common, particularly due to the allophonic voicing of weak stops between vowels (Cho et al., 2002).

The Korean stop contrast is a good candidate for phonetic training, in part because the three-way contrast is so difficult for English speakers to hear, even those who have some experience learning Korean (e.g. college students in Korean language courses). English does have a distinction between so-called “voiced” and “voiceless” stops in initial position. Phonetically, this English contrast is actually a distinction between aspirated and unaspirated
Table 3.1: Minimal triads for Korean stops in initial position \cite{Cho et al., 2002, Francis and Nusbaum, 2002, Kim et al., 2002}

<table>
<thead>
<tr>
<th>Stop type</th>
<th>Korean IPA</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>pʰ</td>
<td>pʰul</td>
<td>grass</td>
</tr>
<tr>
<td>p*</td>
<td>p*ul</td>
<td>horn</td>
</tr>
<tr>
<td>p</td>
<td>pul</td>
<td>fire</td>
</tr>
<tr>
<td>tʰ</td>
<td>tʰata</td>
<td>to ride</td>
</tr>
<tr>
<td>t*</td>
<td>t*ata</td>
<td>to pick</td>
</tr>
<tr>
<td>t</td>
<td>tata</td>
<td>to be sweet</td>
</tr>
<tr>
<td>kʰ</td>
<td>kʰida</td>
<td>to turn on</td>
</tr>
<tr>
<td>k*</td>
<td>k*ida</td>
<td>insert</td>
</tr>
<tr>
<td>k</td>
<td>kida</td>
<td>crawl</td>
</tr>
</tbody>
</table>

stops, as both stop types have positive voice-onset time. There are no phonemic categories in English that correspond neatly to the three types of Korean stops. One goal of Experiment 2 is to help English speakers develop the ability to hear the three-way voicing contrast in Korean, which is potentially based on a number of underlying acoustic dimensions, despite the fact that they already have a heavily ingrained two-way voicing contrast in English, based mainly on voice-onset time.

The Korean contrast is also ideal for training in the context of this dissertation research because of the acoustic dimensions that underlie Korean stops and how those acoustic dimensions are related to the phonological structure of English. This provides an opportunity to test the predictions of Experiment 1, regarding the efficacy of different types of dimensions (i.e. “same”, “familiar”, and “new” dimensions, as characterized in Experiment 1).

### 3.1.1 Acoustic dimensions underlying the Korean stop contrast

Three contrastive cues that differentiate the Korean stops are voice-onset time (VOT), fundamental frequency (F0), and voice quality (VQ). These three cues can be conceptualized as (potentially) orthogonal dimensions along which the three stops can be separated (see table 3.2).
Table 3.2: Contrastive dimensions for Korean stops

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Stop pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT</td>
<td>strong &lt; weak &lt; aspirated</td>
</tr>
<tr>
<td>F0</td>
<td>weak &lt; strong &lt; aspirated</td>
</tr>
<tr>
<td>VQ (spectral tilt)</td>
<td>strong &lt; weak &lt; aspirated</td>
</tr>
</tbody>
</table>

Voice-onset time (hereafter, VOT) refers to the time between the “burst” indicating the release of the stop closure and the onset of voicing. In both Korean and English, VOT values for stops are positive, meaning that voicing begins after the release of the stop closure.\(^1\) In Korean, strong stops have the shortest VOT, weak stops have a medium VOT, and aspirated stops have the longest VOT (Cho et al., 2002).

Fundamental frequency (F0) is the acoustic correlate to the perceptual quality of pitch, and is a contrastive cue to stop identity that is carried on the subsequent vowel. In Korean, F0 at the onset of the vowel following a stop is highest for aspirated stops, medium for strong stops, and lowest for weak stops (Cho et al., 2002).

Voice quality, in the broad sense of the word, has to do with the spectral characteristics (i.e. the harmonic structure) of speech sounds. Differences in voice quality are sometimes described as differences in “phonation”, an articulatory term that refers to the state of the glottis, or vocal folds, during voicing. “Voice quality”, as used here, applies to the perceptual experience related to spectral characteristics other than F0. Terms to describe voice quality in linguistic phonetics include “modal”, “breathy” and “creaky”. Neither “voice quality” or “phonation” are terms that directly describe an acoustic property of a speech sound, but there is an acoustic property called “spectral tilt” which is thought to be an acoustic correlate of voice quality. Spectral tilt is a direct measure of spectral/harmonic structure, and refers to the amount and direction of amplitude attenuation of harmonics relative to the amplitude of the first harmonic. Several measures of spectral tilt are common in studies of voice quality.

\(^1\) In some languages, stops can have negative VOT values, where voicing begins before the stop burst. Languages with stops that have this “voicing lead” include Spanish, Hungarian, Dutch, and Tamil (Lisker and Abramson, 1964). Neither Korean nor English have stops with negative VOT values.
such as the difference in amplitude between the first and second harmonic (H1-H2), and the difference in amplitude between the first harmonic and the harmonic closest to the first formant (H1-F1). Voice quality, like F0, is an acoustic feature of Korean stops that occurs at the onset of the vowel following the stop.

In terms of voice quality, strong stops are said to be “creaky”. Regarding spectral tilt, “creaky” voice is associated with negative tilt values. During creaky phonation, the vocal folds are tense, thus closed for a longer amount of time, which results in a dampening of lower frequency harmonics, and negative values for spectral tilt measures such as H1-H2 and H1-F1 (Cho et al., 2002).

There is some disagreement regarding voice quality of the other two Korean stops. According to several studies, aspirated stops are said to be “breathy”, and have a positive spectral tilt (Lee and Jongman, 2012; Ahn, 1999; Park and BA, 2002; Kang and Guion, 2008). During breathy voicing, the glottis is open for longer amounts of time because the vocal folds are more relaxed, resulting in a spectrum dominated by the first harmonic (F0) (Cho et al., 2002), and positive spectral tilt. The characterization of aspirated stops as “breathy” makes sense: aspiration, often described as a “puff of air” following the stop release, is essentially a period of breathiness, which may overlap with vowel onset. In these same studies that characterize aspirated stops as “breathy”, weak stops are describes as having “modal” voice quality, i.e. neither breathy nor creaky, which correlates to spectral tilt values that are somewhere around 0 (for measures of H1-H2) or slightly positive (for measures of H1-F1). Based on these studies, the pattern of spectral tilt for the Korean stops is strong < weak < aspirated.

There are two studies (Cho et al., 2002; Kenstowicz and Park, 2006) that posit a different pattern for the spectral tilt of Korean stops, where weak stops have the most positive spectral tilt, and aspirated stops are more modal (strong < aspirated < weak). It has been proposed that these inconsistencies may be due to dialectical differences. However, this is unlikely given that conflicting spectral tilt patterns have been found in studies of
the same Korean dialect: both (Cho et al., 2002) and (Ahn, 1999) investigated Korean stop production in speakers of the standard Seoul dialect, but found different results regarding voice quality/spectral tilt. The most likely explanation for the discrepancy in patterns of spectral tilt may be methodological, for instance, a difference in stimuli or stimulus context (e.g. said in isolation or a carrier sentence), or a difference in the ages of speakers across studies (Cho et al., 2002) (e.g. “creak” increases as people age).

In the present study (see discussion on test stimuli, beginning on page 87), aspirated stops were found to be the most “breathy”, with the most positive spectral tilt, and weak stops were more modal, resulting in the spectral tilt pattern strong < weak < aspirated. This aligns with the majority of studies that have investigated voice quality/spectral tilt as an acoustic dimension underlying the Korean stop contrast (four out of the six cited here).

3.1.2 Cue weighting for Korean stop contrast

Though these three acoustic dimensions have been found to separate the three acoustic stops in production studies, the extent to which native Korean speakers use each of these three dimensions in the perception of the stop contrast is up for debate. Studies investigating the stop perception by native Korean speakers have suggested that Koreans use a two-dimensional solution to optimally separate the three stop categories, and that these dimensions likely correspond to VOT and F0 (Francis et al., 2000; Lee et al., 2013; Lee and Jongman, 2012; Kim et al., 2002; Kim, 2004). Francis and Nusbaum (2002) argue that the two contrastive cues used by Koreans may actually be “integral”, i.e. made up of several correlated acoustic dimensions. Lee et al. (2013) showed that the weighting of VOT and F0 as acoustic cues depended on dialect, where speakers of the Seoul dialect use F0 as a primary cue and VOT as a secondary cue, and speakers of the Kyungsang dialect (spoken in the south eastern part of the country) primarily use VOT, and use F0 as a secondary cue (Lee et al., 2013). Regarding voice quality, Lee and Jongman (2012) argue that Korean speakers of the both the Seoul and Kyungsang dialect do not use harmonic structure as a
primary cue to distinguish the Korean stops, while Kim et al. (2002) argue that spectral tilt can serve as a “reasonably strong” cue for stop discrimination by native Korean speakers, particularly for separating the strong stop from the other two stop types.

Regardless of what cues Korean speakers use to distinguish between Korean stops, it is still unclear which dimensions English speakers use when perceiving Korean stops. Which acoustic cues do English speakers rely on when confronted with the Korean stop contrast? Do English speakers use more than one acoustic cue, and if so, is one cue more primary than the other cues? The models of second language speech acquisition discussed in chapter 1 predict that English speakers will not perceive Korean stops in the same way that Koreans do, and are likely to map Korean stops to English stop categories. In a study that investigated the perception of Korean stops by both Korean and English speakers, Francis et al. (2002) argued that English speakers, when first confronted with the problem of perceiving the Korean stops, use only one cue (likely VOT), but there is some evidence that after training with naturalistic Korean stimuli, English speakers can learn to use more than one cue (though it is still unclear how much weight they assign to this newly learned dimension). Finally, it remains to be seen if targeted training on a single acoustic cue will influence the way English speakers perceive the Korean stops, perhaps making them more effective in learning to distinguish the stop categories.

3.2 Relationship of Korean stop cues to English

The acoustic dimensions underlying the Korean stop contrast discussed here (VOT, F0, and voice quality) each have a different relationship to the phonological structure of English, and the extent to which English speakers rely on these acoustic dimensions to make native contrasts differs considerably.

In English, VOT is used as a primary cue to distinguish voiced and voiceless stops. VOT also seems to be, if not the only primary cue, one of the main cues to stop identity in Korean. However, English speakers and Korean speakers necessarily must have different
boundaries separating the stop categories along the continuum of VOT values, because in English there are two stop categories along the VOT dimension (so one VOT boundary), and in Korean there are three stop categories along the VOT dimension (so two boundaries). The L1-L2 relationship of VOT in English and Korean, in the terminology of the previous chapter, is a “same” dimension: it is the same acoustic dimension, used as a primary cue to a very similar contrast (i.e. one involving stops at the same place of articulation), but with different boundaries in L1 and L2.

Regarding the status of F0 in English phonology, F0 is a component of stress, which can function as a contrastive feature. Also, there are intrinsic differences in the F0 of different vowel qualities (which may be cross-linguistically universal, see Whalen et al. 1993). There are no phonemic contrasts in English that have F0 as the primary contrastive cue, i.e. there are no phonemic contrasts based solely on the fundamental frequencies of individual segments as in languages with phonemic tone distinctions (e.g. Mandarin Chinese). However, F0 is used as a secondary cue in several English contexts. It is used to distinguish between initial voices and voiceless stops in English, but mainly when VOT information is ambiguous or degraded (Francis et al. 2008; Abramson and Lisker 1985), though see Whalen et al. (1993), who argue that even the redundant feature of F0 is always taken into account in the perception of English stops. It may also be used as a secondary cue to vowel identity. The L1-L2 relationship of F0 in English and Korean is a “familiar” dimension: it is not used as a primary contrastive cue in the L1, but it may be used as a secondary cue for a similar contrast, or for other types of contrasts.

Finally, though voice quality can indicate social information to English speakers, there are no segmental contrasts in English that have voice quality as a primary (or secondary) cue. The L1-L2 relationship of voice quality in English and Korean is a “new” dimension, because it is not used in the L1 as a contrastive dimension.
3.3 Study Design and Procedure

The training study consisted of two main parts: training and testing, which each used different sets of stimuli. The testing portion of the study was further subdivided into a difference rating task (which was done pre- and post-training to compare effects) and an identification task (which was done post training). Table 3.3 provides an overview of the experiment design.

<table>
<thead>
<tr>
<th>Part</th>
<th>Task</th>
<th>Stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>testing</td>
<td>difference ratings (pre-training)</td>
<td>Test stimuli (Kor. syllables, unmanipulated)</td>
</tr>
<tr>
<td>training</td>
<td>category identification with feedback</td>
<td>Training stimuli (Kor. syllables with manipulated dimensions)</td>
</tr>
<tr>
<td>testing</td>
<td>difference ratings (post-training)</td>
<td>Test stimuli</td>
</tr>
<tr>
<td>testing</td>
<td>category identification with no feedback</td>
<td>Test stimuli</td>
</tr>
</tbody>
</table>

To assess the effect of training with the three Korean acoustic dimensions of VOT, F0, and VQ, there were three between-groups training conditions. In each condition, subjects were trained to use a single dimension as a cue to category membership. These conditions can be defined either in terms of the actual training dimension, or their relationship to the subjects’ first language. Defining the condition in terms of its L1-L2 relationship (as in the above discussion in section 3.2) allows for a comparison of the results from Experiment 1 and Experiment 2 (in spite of the fact that there are important differences between the two experiments, addressed in Chapter 4). See table 3.4 for an outline of the training conditions in Experiment 2, and the L1-L2 relationship for each training dimension.

All experimental tasks were presented using Psychopy (Peirce, 2007). Stimuli were played through Audio-technica ATH-M40 Studiophones at a comfortable listening volume (64 dB C). The experiment took place in the sound booth of the CU phonetics lab. Subjects

---

2 This sound pressure value was obtained by measuring the headphone output with a sound level meter (general radio company type 1565-A), for six adults with normal hearing (self reported).
Table 3.4: Training conditions and dimensions for experiment 2

<table>
<thead>
<tr>
<th>Training condition</th>
<th>Training dim.</th>
<th>L1-L2 relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same dimension</td>
<td>VOT</td>
<td>used as a primary contrastive cue for stops in both languages, with different boundaries in L1 and L2</td>
</tr>
<tr>
<td>Familiar dimension</td>
<td>F0</td>
<td>used as a secondary contrastive cue for stops in L1 (and elsewhere)</td>
</tr>
<tr>
<td>New dimension</td>
<td>VQ</td>
<td>not used as a contrastive cue in the L1</td>
</tr>
</tbody>
</table>

used either an i/o labs button box or a keyboard to enter in their responses, depending on the task (see “procedure” sections below).

3.3.1 Subjects

There were 60 subjects in the Korean training study. These subjects were undergraduates at CU Boulder, drawn from the linguistics and psychology subject pools. All subjects were monolingual native English speakers. Most of the subjects had studied other languages in school, but none had significant experience with another language from birth or early childhood. None of the subjects had any experience with learning Korean.

At the end of the experiment, subjects completed an exit survey that asked them about their native languages, any experience with linguistics or musical instruction, and their experience during the experiment and qualitative judgements regarding the experimental stimuli (see Appendix C for a list of the exit survey questions). A qualitative discussion of the exit surveys is in section 3.6.

3.3.2 Training

Training consisted of a three-alternative, forced choice category identification task with feedback (see section 3.3.2.1 below for more details on the task itself).

The training portion of the study had three conditions which differed in terms of the dimensions that subjects were trained with (and, therefore, the stimuli they heard during
training). These training conditions were VOT (in terms of relationship to L1: “same dimension”), F0 (“familiar dimension”), and voice quality (“new dimension”).

The stimuli for the training portion were Korean syllables of the form /ka/ from 9 x 9 stimulus arrays that varied along 2 of the three training dimensions, with the third held constant. These stimuli were created by manipulating tokens of Korean syllables (see section 3.4.4 on stimuli creation). The three training arrays were F0-VOT, F0-VQ, and VQ-VOT. Subjects in each condition were trained on the two stimulus arrays that contained “their” training dimension: e.g. subjects in the VOT training condition heard stimuli from the F0-VOT array and the Voice Quality-VOT array, and subjects in the F0 training condition heard stimuli from the F0-VOT array and the F0-Voice Quality.

Training was divided into blocks. In each block, subjects heard all 81 stimuli from a single 9 x 9 stimulus array, presented randomly. The order of stimulus arrays presented in each block was counterbalanced within condition, such that half of the subjects heard one order of stimulus arrays (e.g. ABAB...) while the other half heard the opposite order (e.g. BABA...). There were 10 blocks of training, for a total of 10x81 = 810 individual trials. Subjects completed the first 6 blocks of training immediately after the pre-training difference rating task (see below) and then took a short 5-10 minute break (enforced for 5 minutes with a timer) before finishing the last 4 blocks of training and the post-training test tasks.

3.3.2.1 Training procedure

The instructions for the training task were as follows:

In this training task, you’ll learn to identify the three Korean consonants. On each trial, you will hear a Korean syllable play. Then, you will identify the consonant sound at the beginning of the syllable by choosing the Korean letter for that sound.

Before beginning the training task, subjects were told explicitly that the training stimuli were different stimuli from the set used in the difference rating task (a task which the completed prior to training).
On each trial during training, subjects heard a single Korean syllable play. Then they were presented with an image of three Korean letters and told to choose which one corresponded to the consonant sound at the beginning of the syllable they had just heard. Because Korean letters can be hard to tell apart for English speakers, the letters were presented inside colored circles that matched the colors of three buttons on an i/o labs button box. Thus, subjects had two potential mappings: sound-to-letter AND sound-to-color. All letter-color combinations were consistent across all subjects and trials, and the same letter-color combinations were used in the post-training identification task (see figure 3.1). For each stimulus, subjects indicated which Korean stop category they heard by pressing the button that corresponded to their choice of letter (and color). Subjects were instructed to press the response buttons using the index finger of their dominant hand.

Figure 3.1: Korean letter-color combination for responses in training tasks. Information in parentheses are the IPA and consonant type for that Korean letter, which were not seen by subjects

After subjects entered their responses, they received feedback. If subjects were correct, they received a message that read “CORRECT! The sound was…” along with a picture of the correct Korean letter-color image. If subjects were wrong, they received a message that read “WRONG. The sound was…” along with a picture of the correct Korean letter-color image. So, in both cases (right or wrong), the correct answer was reinforced with the appearance of the correct letter-color image.

It is important to note here that the “correct” category of the consonant sound was determined experimentally, in that the category structure of training stimuli varied according to the training condition. See section 3.4.3 on page 93.
After feedback and an ISI of 500 ms, the next trial began automatically. Subjects were required to take a timed, five minute break after the 6th block of training.

Before the trials began, 20 example stimuli were played (randomly chosen from the training array that subjects heard first) with their correct letter-color image displayed on the screen. This was to allow subjects to familiarize themselves with the sounds and their category labels.

### 3.3.3 Testing

The testing portion of the study was subdivided into two tasks: difference ratings (done pre- and post-training), and post-training identification. For all three testing tasks, all subjects heard the same set of 18 stimuli. Because each condition heard a different set of stimuli during training, using the same set of stimuli in testing allowed for a more direct assessment of the effects of training across the three training conditions. The 18 test stimuli were different from those used during training: while they were Korean syllables of the form /ka/, these test syllables were natural, unmanipulated Korean syllables produced by a native Korean speaker (see section on stimuli creation, page 94).

#### 3.3.3.1 Difference rating task

In the difference rating task, subjects were presented with pairs of stimuli, which they rated on a scale from 1 (very similar) to 9 (very different).

All pair-wise combinations of stimuli from the set of 18 test stimuli were presented during the task, for a total of 153 stimuli. This set included only the top diagonal half of a pairwise comparison matrix, excluding reciprocal comparisons and “identity” comparisons where a stimulus is compared with itself. Subjects completed the difference rating task twice: once directly before training, and once directly after training.
3.3.3.2 Procedure for difference rating task

The instructions for the difference rating task were as follows:

*On each trial, you will hear a pair of syllables. Your task is to decide how different the two syllables are, on a scale from 1 (very similar) to 9 (very different).*

On each trial in the difference rating task, subjects heard a pair of stimuli play. Then they were presented with a screen that asked them to rate how similar the stimuli were on a scale from 1 (very similar) to 9 (very different). Subjects used the numbers on a keyboard to enter in their numeric response, and were instructed to make their response using the index finger of their dominant hand. The next trial began automatically after an ISI of 500 ms.

In both the pre-training and post-training difference rating task, subjects heard a block of practice trails. All 18 test stimuli were played in random order, to familiarize subjects with the sounds. Before beginning the post-training task, subjects were told that they would hear the same set of stimuli as they had heard in the pre-training difference rating task.

3.3.3.3 Identification task

In the identification task, subjects were presented with a single stimulus, and chose the stop category of the consonant at the beginning of the syllable, the same procedure as in the category training task. However, in the identification task, there was no feedback, and subjects heard a different set of stimuli (the 18 unmanipulated test stimuli) from those they were trained with. Subjects identified each of the test stimuli once, for a total of 18 trials in the identification task.

3.3.3.4 Procedure for identification task

The instructions for the identification task were as follows:

*This task is similar to previous training tasks where you learned to identify the three Korean consonants. On each trial, you will hear a Korean syllable play. Then, you will*
identify the consonant sound at the beginning of the syllable by choosing the Korean letter for that sound.

Before beginning the identification task, subjects were told explicitly that the stimuli were different from the set used in the training task (which had a similar procedure).

The procedure for the identification task was the same as in the training task: on each trial during training, subjects heard a single Korean syllable play. Then they were presented with an image of three Korean letters (presented in colored circles as in the training task, with the same letter-color mapping) and told to choose which letter/circle corresponded to the consonant sound at the beginning of the syllable they had just heard. Unlike during training, subjects did not receive feedback after they entered in their response. Instead, they simply moved on to the next trial, which started automatically after an ISI of 500 ms.

Before beginning the identification trials, subjects heard a block of practice trails. All 18 test stimuli were played in random order. Subjects completed the identification task once, following the last block of training.

3.4 Stimuli

There were two sets of stimuli in Experiment 2: the test stimuli (used in the identification and difference rating tasks), and the training stimuli, used in the category training task. The test stimuli are described first (page 87), followed by a description of the training stimuli (page 93) and how they were created (page 94).

3.4.1 Test Stimuli

Test stimuli consisted of 18 syllables produced by a native Korean speaker, and were unmanipulated except for intensity scaling. The syllables were consonant-vowel (CV) syllables of the form velar stop + /a/, where the velar stop was one of three Korean stop voicing categories (strong, weak, or aspirated). The syllables were spoken by a native Korean speaker, and recorded in a double-walled sound booth in the CU phonetics lab. After
recording, all syllables were scaled to an overall intensity of 70 dB.

The test syllables were recorded in 6 different contexts, e.g. alone, in a carrier phrase “ige _____ hago kata.” (the word is _____), and in citation or conversational form (see Appendix A for a list of the test syllables and their recording contexts). “Citation form” is slow and clearly enunciated, and the subject was instructed to speak “as if speaking to a person learning Korean”. “Conversational” form is faster, less carefully enunciated, and the subject was instructed to speak “as if speaking to a friend”.

The subject who recorded all stimuli material was a 25 year old male native Korean speaker, from Daejon, South Korea, a large city in Chungcheong Province. Though the rural areas of Chungcheong province are home to a non-standard dialect known as the Chungcheong dialect, people who live in larger cities such as Daejon speak the standard variety of Korean spoken in Seoul and elsewhere in central South Korea (Nordhoff, 2013). The subject began learning English in school at age 9.

3.4.1.1 VOT, F0, and voice quality of test stimuli

Figure 3.2 shows the distribution of the 18 test stimuli along three acoustic dimensions: VOT, F0, and VQ (quantified as spectral tilt, H1-F1).

VOT was measured from the stop burst to the onset of the vowel, which was characterized by complex periodic oscillations in a waveform and accompanied by the appearance of a steady first formant and regular glottal pulses in a spectrogram. As expected, aspirated stops (/kʰ/) had the highest VOT, ranging from 102–146 ms. Strong stops (/k*) had the lowest VOT, ranging from 22–29 ms, and weak stops (/k/) were in the middle, ranging from 55–96 ms.

In order to compare mean VOT values of specific consonant types, a linear regression model was fitted to the VOT measurements of the 18 test stimuli. The dependent variable was VOT, and the independent variables consisted of sets of categorical variables that coded for specific mean comparisons among the test stimuli, based on the type of consonant.
Unlike in previous analyses with categorical variables that used orthogonal contrast coding, in the analysis of the test stimuli, forward difference coding was used. Forward difference coding allows for a comparison of means for adjacent levels of a categorical variable. As in the case of other types of contrast coding (e.g. dummy coding or orthogonal contrast coding), the coding scheme used in the model depends on which mean comparisons are desired. Forward difference coding was used in this case (instead of, e.g. orthogonal contrast coding, which is the coding scheme used in most of the other analyses in this dissertation) because the desired comparisons were between adjacent levels of a categorical variable; i.e. in the case of Korean consonant type, for which there are three levels (strong, weak, and aspirated), the comparisons of interest for the VOT analysis were level 1 vs. level 2 (strong vs. weak), and level 2 vs. level 3 (weak vs. aspirated), as these are the consonant types that are adjacent along the VOT continuum. So, using forward-difference coding in the linear model for the analysis of VOT for the test stimuli allowed for two coded variables: “strong vs. weak” and “weak vs. aspirated”. Essentially, forward difference coding can be thought of as a type of contrast coding that was used in this case because it allowed for the correct mean comparisons in a more efficient manner than would be afforded by, e.g., orthogonal contrast coding.

The regression model confirmed that the mean VOT for strong stops is significantly lower than for weak stops (significant contrast coded variable, “strong vs. weak”, $F(1, 15) = 39.36, p < .0001$), and that the mean VOT for weak stops is significantly lower than for aspirated stops (significant contrast coded variable, “weak vs. aspirated”, $F(1, 15) = 23.21, p = .0002$). This VOT pattern, where $\text{strong} < \text{weak} < \text{aspirated}$, aligns with previous findings regarding VOT for Korean stops.

$F_0$ was measured at the onset of the vowel (measured at a single point). As expected, aspirated stops had the highest $F_0$, ranging from 163–179 Hz. Weak stops had the lowest $F_0$, ranging from 103–114 Hz, and strong stops were in the middle, ranging from 128–154 Hz.
Figure 3.2: Distribution of test stimuli along acoustic dimensions of VOT, F0, and spectral tilt

In order to compare mean F0 values of specific consonant types, a linear regression model was fitted to the F0 measurements of the 18 test stimuli. The dependent variable was F0, and the independent variables consisted of sets of categorical variables that coded for specific mean comparisons among the test stimuli, based on the type of consonant. As in the case of the VOT analysis, forward difference coding was used for the specific mean comparisons desired in this case: “weak vs. strong” and “strong vs. aspirated” (as these are the consonant types that are adjacent along the F0 continuum). This model confirmed that the mean F0 for weak stops is significantly lower than for strong stops (significant contrast coded variable, “weak vs. strong”, $F(1,15) = 71.52, p < .00001$), and that the mean F0 of strong stops is significantly lower than for aspirated stops (significant contrast coded variable, “strong vs. aspirated”, $F(1,15) = 57.21, p < .0001$). This F0 pattern, where weak < strong < aspirated, aligns with previous findings regarding vowel onset F0 for the Korean stops.

Spectral tilt provides a quantifiable acoustic measure of voice quality. For spectral tilt measures of the test stimuli, the amplitudes for the first harmonic (H1) and the amplitude of the harmonic nearest the first formant (F1) were measured at the onset of the vowel,
using FFT spectra with a 25 ms window. These amplitudes were used to calculate spectral tilt (H1–F1). Aspirated stops had the most positive tilt (ranging from 2.5–13.7 dB), which indicates a steep drop-off in spectral energy from the first harmonic to subsequent harmonics. This steeply positive tilt is associated with “breathy” voice quality. Strong stops had the most negative tilt (ranging from -12.9–-5), which indicates an increase in spectral energy from the first harmonic to subsequent harmonics. This negative tilt is associated with “creaky” voice quality. Weak stops were in the middle of the spectral tilt continuum (ranging from -4.1–8.5 for H1–F1). Middle values of spectral tilt are associated with “modal” voice quality—neither “breathy” or “creaky”.

In order to compare mean spectral tilt values of specific consonant types, a linear regression model was fitted to the spectral tilt measurements (H1–F1) of the 18 test stimuli. The dependent variable was the H1–F1 value for each stimulus, and the independent variables consisted of sets of categorical variables that coded for specific mean comparisons among the test stimuli, based on the type of consonant. As in the case of the previous test stimuli analyses, forward difference coding was used for the specific mean comparisons desired in this case: “strong vs. weak” and “weak vs. aspirated”, as these are the consonant types that are adjacent along the spectral tilt continuum. This model confirmed that the mean spectral tilt for strong stops is significantly lower than for weak stops (significant contrast coded variable, “strong vs. weak”, $F(1, 15) = 23.96, p = .0002$), and that the mean spectral tilt for weak stops is significantly lower than for aspirated stops (significant contrast coded variable, “weak vs. aspirated”, $F(1, 15) = 7.07, p = .02$). This spectral tilt pattern, where strong < weak < aspirated aligns with the majority of previous findings regarding spectral tilt of vowels following the Korean stops (see page 77). However, it is important to note that despite a significant mean difference in the comparison of spectral tilt for weak and aspirated stops, there is some overlap in the spectral tilt values for these two stop types. It remains to be seen if VQ alone can be a strong enough cue to separate aspirated and weak stops in this set of test stimuli.
3.4.2 Training Stimuli

Training stimuli consisted of three separate 9 x 9 stimulus arrays made up of CV syllables consisting of a velar stop + /a/. The training stimuli were made by manipulating specific acoustic properties of three representative syllables from the set of 18 test stimuli described in the previous section. Each stimulus array varied along two of the three training dimensions, with the third held constant.

There were 9 “steps” along each training dimension, resulting in a stimulus array of 81 individual stimuli. Steps along the F0 dimension ranged from 97 to 183 Hz. These values refer to the fundamental frequency at the onset of the vowel in the syllable. The range of F0 values in the training stimuli is an expanded range based on the full range of F0 values in the test stimuli, and the mean F0 of each stop type in the test stimuli. Steps along the VOT dimension ranged from 7 to 135 milliseconds, where the values refer to the voice onset time from stop burst to the onset of the vowel in the syllable. The range of VOT values in the training stimuli, like F0, is an expanded range based on the full range of VOT values in the test stimuli as well as the mean VOT of each stop type in the test stimuli. Details regarding the calculation of the expanded F0 and VOT ranges for the training stimuli can be found on page 97. The steps along the voice quality dimension consisted of syllables with blended vowels taken from three test stimuli (one from each of the stop categories), ranging from 100% “strong vowel” to 100% “aspirated vowel”. Details regarding the creation of the voice quality dimension can be found on page 95.

Hereafter, each training array is referred to by its two varying dimensions: in the F0-VOT array, F0 and VOT vary, but voice quality is held constant at 100% “weak” vowel, which is the middle step of the voice quality dimension. In the F0-VQ array, F0 and voice quality vary, but VOT is held constant at 71 ms, the middle step of the VOT dimension. Finally, in the VQ-VOT array, voice quality and VOT vary, but F0 is held constant at 140 hertz, the middle step of the F0 dimension.
3.4.3 Category structure for training stimuli

Subjects in each training condition heard stimuli from the arrays where “their” training dimension varied. For example, subjects in the F0 condition were trained with stimuli from the F0-VOT and the F0-VQ arrays. The feedback provided during training was based on training categories, the structure of which were determined by the training condition itself. In other words, even though both the F0 and the VOT training condition were trained with stimuli from the F0-VOT array, they were trained with a different category structure: for the F0 condition, category membership depended on values along F0, and for the VOT condition, category membership depended on values along VOT. The training categories were thus experimentally defined, and were designed to focus subject’s attention on a single category relevant dimension, depending on their training condition.

Figures 3.3 to 3.5 illustrate the category structures of the training stimuli for each of the three training conditions. In the VOT condition, subjects were trained with the F0-VOT and the VOT-VQ arrays. Subjects in this condition were “told” (by way of feedback) that VOT steps 1-3 of both training arrays corresponded to the strong stop. (In actuality, during training the feedback made reference to the Korean Hangul character/colored circle

![Figure 3.3: Category structure for training stimuli in the VOT training condition](image-url)
combination that represented the strong stop, i.e. ㄱ in a red circle, see figure 3.3. The VOT steps 4-6 corresponded to the weak stop (ㄱ in a blue circle), and VOT steps 7-9 corresponded to the aspirated stop (ㅋ in a yellow circle).

In the F0 condition, subjects were trained with the F0-VOT and the F0-VQ arrays. For this condition, F0 steps 1-3 of both training arrays corresponded to the weak stop, F0 steps 4-6 corresponded to the strong stop, and F0 steps 7-9 corresponded to the aspirated stop.

Subjects in the VQ condition were trained with the F0-VQ and VQ-VOT arrays. In this training condition, VQ steps 1-3 corresponded to the strong stop, VQ steps 4-6 corresponded to the weak stop, and steps 7-9 corresponded to the aspirated stop.

3.4.4 Creation of training stimuli

To make the training stimuli, a representative syllable with a consonant from each stop category was chosen from the test stimuli, based on three criteria: (1) they were all from the same recording set in order to control for any context effects during recording, (2) the VOT of
the aspirated syllable (/kʰa/) was long enough to allow for duration manipulations beyond the natural length that would not result in artifacts, and (3) the syllables showed good separation along the three training dimensions (VOT, F0, and voice quality, as indicated by spectral tilt). The recording set chosen for manipulation was set 1, where syllables were spoken in citation form in a carrier sentence. All training stimuli were created using signal processing techniques with the Praat software package (Boersma et al., 2002).

3.4.4.1 Voice quality dimension

The creation of the voice quality dimension presented two distinct challenges. First, it is unclear if the manipulation of spectral tilt, or any other single acoustic measure, will result in perception of different voice quality types. It is true that spectral tilt is often cited as an acoustic correlate to voice quality. In terms of spectral tilt, “breathy” voice has a steep drop-off in spectral energy for harmonic frequencies higher than the fundamental. Conversely, “creaky” voice shows a rising in spectral energy, and “modal” voice has fairly level spectral energy. Two potential reasons for using spectral tilt in acoustic descriptions of voice quality...
are that it is relatively straightforward to measure and it reliably demonstrates differences between voice quality types. However, the successful manipulation of spectral tilt may not necessarily result in the perception of different voice quality types. Instead, it is possible that the perception of voice quality may be due to a cluster of related acoustic features including, but not limited to, spectral tilt, harmonics-to-noise ratio, jitter, and formant frequency (Gordon and Ladefoged, 2001). Second, even if spectral tilt was the acoustic correlate solely responsible for the perception of voice quality differences, it is difficult to manipulate spectral tilt without distortion or the introduction of acoustic artifacts. In sum, the manipulation of a single acoustic feature associated with voice quality may not be enough to have the intended perceptual consequence for the listener, and may introduce distortion that interferes with perception in an unintended way.

In order to overcome these problems, a continuum consisting of blended vowel segments from the three syllables /kʰa/, /ka/, and /kʰa/ was created. This allowed for many of the acoustic features associated with voice quality to co-vary, resulting in a dimension with a redundancy of voice quality features. F0 of the vowel segments, one of the three training dimensions, was manipulated separately. Thus, the voice quality dimension consisted of 9 blended vowels with acoustic features that, crucially, were NOT F0 (or VOT, for that matter, as VOT is not an acoustic feature of vowels).

The voice quality dimension consisted of 9 steps, where each step was a vowel with the following specifications:

**Step 1** 100% /*a/ (the vowel from /kʰa/)
**Step 2** 80% /*a/, 20% /a/ (the vowel from /ka/)
**Step 3** 60% /*a/, 40% /a/
**Step 4** 40% /*a/, 60% /a/
**Step 5** 100% /a/
**Step 6** 80% /a/, 20% /ʰa/ (the vowel from /kʰa/)
**Step 7** 40% /a/, 60% /ʰa/
**Step 8** 20% /a/, 80% /ʰa/
**Step 9** 100% /ʰa/
Once the blended vowels were created, they were spliced onto a consonant segment. This splicing process also involved manipulation of VOT and F0, and is described in section 3.4.4.3 below.

To create these voice quality steps, the vowel segments of the three test stimuli were separated from the consonant segments at the onset of the vowel. The vowel segments were all normed to the same duration (344 ms, the duration of the vowel from /kʰa/₁) and intensity (70 dB). The pitch contour was copied from /kʰa/₁ to the other two vowels, resulting in three vowels with the same duration, intensity, and pitch contour. Then, the vowels were mixed in Praat using a method of amplitude addition. This resulted in a set of 9 vowel segments, which were the steps along the voice quality dimension. At one end of this dimension was the vowel segment /*a/₁, in the middle of the dimension was the vowel segment /a/₁, and at the other end of the dimension was the vowel segment /hᵃ/₁. The dimension was arranged in this way, with /a/₁ in the middle, because the vowel following a “weak” stop has spectral tilt values that fall between those of the vowels following weak and strong stops, and to the extent that spectral tilt can be used as a proxy for voice quality, this maintains the pattern for Korean stops where strong < weak < aspirated.

3.4.4.2 VOT and F0 steps

The VOT dimension consisted of 9 steps, where each step was a VOT value in milliseconds. The endpoints of this dimension were calculated according to the following procedure: the lowest VOT value of the test stimuli was subtracted from the highest VOT value, giving the full range of VOT of the test stimuli. Then, 15% of this full range was subtracted from the mean value of VOT for the strong category (the lowest mean VOT) and added to the mean value of VOT for the aspirated category (the highest mean VOT). This step was done to calculate an expanded VOT range based on the mean VOT of the stop categories and the range of individual values of VOT (rounded to the nearest millisecond). The endpoints of

---

3 Thanks to Mark Antoniou, for the use of his “Sound mixer, step maker” Praat script.
this dimension were used to specify VOT values when splicing consonant and vowel segments to create stimuli arrays (see section 3.4.4.3). The VOT steps of the training stimuli were:

**Step 1** 7 ms  
**Step 2** 23 ms  
**Step 3** 39 ms  
**Step 4** 55 ms  
**Step 5** 71 ms  
**Step 6** 87 ms  
**Step 7** 103 ms  
**Step 8** 119 ms  
**Step 9** 135 ms

The F0 dimension consisted of 9 steps, where each step was an F0 value in hertz. The endpoints of this dimension were calculated in the same way as the VOT continuum: 15% of the full F0 range of the test stimuli was subtracted from the mean F0 of the weak stops (the category with the lowest F0 values) and added to the mean F0 of aspirated stops (the category with the highest F0 values), resulting in an expanded F0 range. As in the case of the VOT dimension, the endpoints of the F0 dimension were used to specify F0 values when splicing consonant and vowel segments to create stimuli arrays. The F0 steps of the training stimuli were:

**Step 1** 97 Hz  
**Step 2** 108 Hz  
**Step 3** 118 Hz  
**Step 4** 129 Hz  
**Step 5** 140 Hz  
**Step 6** 151 Hz  
**Step 7** 162 Hz  
**Step 8** 172 Hz  
**Step 9** 183 Hz

Note that these F0 values were rounded to the nearest whole hertz, as required by the script used to create the stimuli arrays.
3.4.4.3 Splicing consonant and vowel segments to create stimuli arrays

The stimuli arrays were created with the use of a Praat script[^1] that splices consonants and vowels to create syllables with specific VOT and F0 values. These values lie along continua with endpoint values and number of steps specified by the user. The different VOT values are creating by adding or subtracting the aspiration portion of a voiceless stop consonant, and the F0 values are created by adjusting the pitch contour of a vowel to a target F0 value. The consonant segment is spliced with the vowel segment by blending the end of the aspiration portion from the consonant into the beginning of the vowel, a method that allows for a natural sounding transition between consonant and vowel.

The consonant segment chosen for splicing was from /kʰa/. This particular segment was chosen because it had a long enough period of aspiration to allow for the creation of VOT steps at the high end of the VOT range without introducing unwanted artifacts. Only the first portion of the consonant segment from /kʰa/ was used. This portion contained the stop burst and 60 ms of aspiration, about half of the total period of aspiration from /kʰa/, which had a VOT of 126 ms. The aspiration portion contained only noise with no periodic structure (as seen in both the waveform and spectrogram), and was used for splicing so that any vowel transition effects toward the end of the aspiration in /kʰa/ were not a part of the training stimuli. The splicing segment was also scaled to 60 dB. One investigation of Korean stops [Cho et al., 2002] showed that aspirated stops may have higher burst energy compared to strong and weak stops, which do not differ in terms of burst energy—though there is no study to date showing that burst energy functions as a perceptual cue to the perception of Korean stops. Decreasing the energy levels of the splicing segment mitigated possible effects of burst energy as a cue to stop type, as well as allowing for a more natural blend with vowel segments. These steps were taken to control for any potential bias in the training stimuli toward an aspirated stop. Furthermore, all training stimuli were created with

[^1]: Thanks to Matt Winn for the use of his “Make VOT/F0 Continuum” Praat script.
this same splicing segment, so acoustic features related to burst energy and vowel transitions did not vary in the training stimuli, and could not be used as contrastive cues in the training task.

For the F0-VOT array, the vowel segment from the middle of the voice quality dimension (Step 5–100% /a/) was combined with the consonant portion from /kʰa/. Using the VOT-F0 continuum Praat script, 81 stimuli were created by specifying the endpoints of VOT and F0 continua (steps 1 and 9 from the VOT and F0 dimensions, respectively), and the number of steps between these endpoints (8).

For the F0-VQ array, each of the 9 vowel segments from the voice quality dimension were combined with the consonant portion from /kʰa/. Using the VOT-F0 continuum script, 81 stimuli were created by specifying the endpoints of an F0 continuum (steps 1 and 9 from the F0 dimension), and the number of steps between these endpoints (8), along with a constant value for VOT (71 ms, the middle value from the VOT dimension).

For the VQ-VOT array, each of the 9 vowel segments from the voice quality dimension were combined with the consonant portion from /kʰa/. Using the VOT-F0 continuum script, 81 stimuli were created by specifying the endpoints of a VOT continuum (steps 1 and 9 from the VOT dimension), and the number of steps between these endpoints (8), along with a constant value for F0 (140 Hz, the middle value from the F0 dimension).

3.5 Results

Results from Experiment 2 are presented in this order: Korean category training (section 3.5.1, page 101), Korean post-training identification task (section 3.5.2, page 105), Korean pre- and post-training difference rating task (section 3.5.3, page 112). There is an overview of all of the results from Experiment 2 at the beginning of section 3.6, page 133.

Unless otherwise noted, all statistical analyses were done using R (version 3.2.3) (R Core Team, 2015).

As in Experiment 1, in cases where an analysis of the subject data included categorical
variables (e.g. in the analysis of training accuracy for each of the Korean training conditions), orthogonal contrast coding was used to specify mean comparisons between levels of the categorical variable(s). As in the case of the analyses for the subject data in Experiment 1, the details of coding schemes are described in the context of each model.

### 3.5.1 Korean category training

For the Korean category training task, subjects in each condition heard stimuli (Korean CV syllables, manipulated to highlight particular acoustic dimensions) from two of the three training arrays (depending on their training condition), and identified the stop category of the initial consonant.

Data from the category training tasks were analyzed using logistic mixed-effects regression models, which allowed for a binary dependent variable (because accuracy on a single trial was coded as 0 or 1), and random effects for subject and/or item, which can account for subject- and item-level variance in the case of a repeated-measures design. As in the previous chapter, unless otherwise noted, all logistic mixed-effects models were run using the R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2016), and all model comparisons to test for the significance of fixed effects were done using the "anova()" function in the R stats package. Contrast coded variables were used for specific mean comparisons, and coding schemes depended on the models (see descriptions below).

Figure 3.6 shows mean training accuracy as a function of block number, split by training condition.

To account for multiple testing with the Korean training data for the family of mixed models where accuracy was predicted by block number as a continuous fixed effect (the omnibus test, described in the subsequent paragraph, and a similar test that examined differences in accuracy between the training conditions in block 1 only), alpha levels were adjusted using the Holm-Bonferroni correction, which is more powerful than the classic Bonferroni correction (Holm, 1979). The correction is noted (e.g. “p<.001”, H-B corrected)
To evaluate if accuracy improved over the course of training, a logistic, mixed-effects model was fit to the Korean training data (see table 3.5). The dependent variable was accuracy on a single trial (coded as 1 or 0). Block number was considered as a continuous, fixed effect. Subjects and each subject’s block number effect were treated as random variables. The model also contained two categorical fixed effects variables that were coded for condition. These two variables allowed for comparisons in mean accuracy among the training conditions: one variable compared the mean accuracy for the VOT condition to the mean accuracy of the other two conditions (“VOT vs others”), and the other variable compared the mean accuracy of the F0 condition with that of the Voice quality condition (“F0 vs VQ”). Interactions between block number and each of the contrast coded predictors were included as well, to assess the extent to which the block number effect depended on training condition. The model also included interactions between block number and each of the contrast coded predictors, to assess the extent to which the block number effect depended on condition.

The VOT condition had higher accuracy overall, controlling for block number (signif-
icant contrast coded predictor in the omnibus model, “VOT vs others”, $\chi^2(1) = 22.47, p < .00001, \text{H-B corrected)}$. The interaction between block number and the coded variable “VOT vs others” was not significant, indicating that the block effect was not different for the VOT condition ($p = .09, \text{H-B corrected}$); in other words, although the VOT condition had higher accuracy overall, they did not improve at a faster rate.

Table 3.5: Fixed effects from Korean category training regression model, predicting accuracy with block number and contrast codes for training conditions. The coefficients in a logistic regression model represent the change in log odds as values increase along a continuous predictor, or from one level of a contrast coded predictor to another level. P values for $z$-statistics are not corrected. See text for F-statistics and corrected p values

|                          | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------------|----------|------------|---------|----------|
| (Intercept)              | 0.0005708| 0.0519738  | 0.011   | 0.9912   |
| Block number             | 0.0310418| 0.0057311  | 5.416   | 6.08e-08 *** |
| VOT vs. others           | 0.7673961| 0.1481937  | 5.178   | 2.24e-07 *** |
| F0 vs. VQ                | 0.1598901| 0.1272607  | 1.256   | 0.2090   |
| Block*VOT vs others      | 0.0324753| 0.0162900  | 1.994   | 0.0462   |
| Block*F0 vs VQ           | 0.0219956| 0.0139751  | 1.574   | 0.1155   |

The VOT subjects were higher in accuracy at the beginning of the training compared to the other training conditions. A post-hoc logistic mixed-effects model for block 1 only was run, predicting accuracy with two contrast coded variables for condition. These variables compared mean accuracy for the VOT condition compared to the other two conditions (“VOT vs others”) and the mean accuracy for the F0 condition vs the VQ condition (“F0 vs VQ”). The model also included mixed effects for subjects and each subject’s block effect. The results from this model indicated that in block 1, the VOT condition had significantly higher accuracy compared to the other two conditions (significant contrast-coded predictor, “VOT vs others”, $\chi^2(1) = 9.45, p = .006, \text{H-B corrected}$). There was no difference in mean accuracy in block one for the F0 vs VQ conditions (“F0 vs VQ”, $p = .09, \text{H-B corrected}$). Taken together, these results suggest that even at the beginning of training, the VOT condition had an advantage over the other training conditions, in terms of using VOT as a contrastive cue.
Post-hoc tests for individual training conditions were run, consisting of logistic mixed-effects models, where the dependant variable was accuracy in the Korean training task, and block number was treated as a continuous, fixed effect. In addition to the block number variable, these post-hoc tests also included a categorical variable for the stimulus array (because the set of training arrays was different for each condition). Subjects, each subject’s block number effect, and each subject’s stimulus array effect were treated as random variables.

The addition of the stimulus array variable meant that the hypotheses being tested in the post-hoc tests were slightly different from those being tested in the omnibus test described above: in the post-hoc tests reported below, the “block number” effect is over and above any effect of stimulus array. Therefore, because these are ancillary hypotheses and not considered as part of the “family” of tests where accuracy was predicted with block number only, no adjustment of p-values was done for these three models.

The post-hoc tests for individual training conditions indicated that subjects in both the VOT and the Voice Quality conditions did improve in accuracy over the course of training, but subjects in the F0 condition did not.

The model for the VOT condition included the continuous block number variable and a contrast coded variable that compared the mean accuracy of the two stimulus arrays for the VOT condition (“F0-VOT” vs. “VQ-VOT”), along with random effects for each subject and their block and array effects. Controlling for stimulus array, there was a significant block number effect in this model, whereby accuracy improved over the course of training ($\chi^2(1) = 14.46, p = .0001$). Controlling for block number, there was a marginally significant effect of stimulus array ($\chi^2(1) = 3.85, p = .05$), whereby mean accuracy was slightly higher for the F0-VOT array ($M = .67$) compared to the VQ-VOT array ($M = .64$).

The model for the VQ condition included the continuous block number variable and a contrast coded variable that compared the mean accuracy of the two stimulus arrays for the VQ condition (“F0-VQ” vs. “VQ-VOT”), along with random effects for each subject and their block and array effects. In this model, there was a significant block effect. Controlling
for stimulus array, accuracy improved over the course of training ($\chi^2(1) = 10.84, p = .001$). Controlling for block number, the effect of stimulus array was not significant ($p = .07$).

Finally, the model for the F0 condition included the continuous block number variable and a contrast coded variable that compared the mean accuracy of the two stimulus arrays for the VQ condition ("F0-VOT" vs. "F0-VQ"), along with random effects for each subject and their block and array effects. In this model, both block number and the contrast-coded predictor for stimulus array were not significant ($p = .23$ and $p = .47$, respectively), indicating that subjects in the F0 condition did not improve in accuracy over the course of training, and controlling for block number, there was no difference in accuracy for the two stimulus arrays.

These results suggest that subjects in both the VOT and the VQ conditions were learning to attend to their relevant dimensions in order to assign stimuli to the Korean stop categories. Subjects trained with F0, on the other hand, did not improve, which suggests that they were having difficulty using F0 as a cue to category membership.

### 3.5.2 Korean identification task

For the post-training identification task, subjects heard each of the 18 test stimuli (Korean CV syllables, unmanipulated except for loudness scaling) and identified the stop category of the initial consonant.

As in the analysis of the training data, data from the identification task were analyzed using logistic mixed-effects regression models, with a binary dependent variable coded for accuracy, and random effects for subject and/or item. Contrast coded variables were used for specific mean comparisons (variables and coding schemes depended on the models, see descriptions below). Figure 3.7 shows overall accuracy in the identification task as a function of training condition.

In the post-training identification task, the VOT training condition was higher in accuracy compared to the other two conditions. In order to investigate differences in accuracy
between the training conditions in the identification task, a logistic, mixed-effects model was fit to the post-training identification data. The dependent variable was accuracy on a single trial (coded as 1 or 0). The model contained two categorical condition variables, treated as fixed effects. The condition variables coded for comparisons in mean accuracy among training conditions ("VOT vs others" and "F0 vs VQ"). The model also contained random effects for subject (see table 3.6).

Table 3.6: Fixed effects from regression model for Korean identification task, predicting accuracy categorical variables for training condition. The regression coefficients in this model represent the change in log odds when moving from one level of a contrast coded predictor to another level.

|                | Estimate | Std. Error | z value | Pr(>|z|)  |
|----------------|----------|------------|---------|-----------|
| Intercept      | 0.08390  | 0.07791    | 1.077   | 0.282     |
| VOT vs. others | 1.79136  | 0.22644    | 7.911   | 2.55e-15  *** |
| F0 vs. VQ      | 0.12835  | 0.18665    | 0.688   | 0.49      |

Mean accuracy was significantly higher for the VOT condition compared to the other conditions ("VOT vs. others", $\chi^2(1) = 45.84, p < .00001$), but there was no difference in
accuracy for the F0 and VQ conditions \((p = .49)\).

Depending on their training condition, subjects were better at identifying certain stop types than others. Figure 3.8 shows mean accuracy in each training condition, split by consonant type.

![Figure 3.8: Korean identification training: Mean accuracy, split by training condition and consonant type. Dashed line represents “chance” (.33).](image)

Post-hoc tests were run for each condition to investigate how accuracy differed depending on the consonant type in the stimuli. No p-value adjustments were done for these post-hoc tests, because the hypotheses tested in the post-hoc tests were different from the model described above that compared mean accuracy in the identification task among the different training conditions. In these post-hoc tests, accuracy is modeled as a function of consonant type, and the mean comparisons between consonant types are different for the three training conditions.

For the VOT condition, a logistic mixed effects model was fitted to the post-training identification data, with accuracy on a single trial as the dependant variable. The independent variables were two categorical consonant type variables that were treated as fixed effects. These were coded for mean differences among consonant type (“Weak vs. others” and “Strong vs aspirated”). The model also contained random effects for each subject. For
the VOT condition, stimuli containing weak stops had significantly lower accuracy compared to stimuli containing the other stop types ("Weak vs. others", $\chi^2(1) = 10.97, p = .001$). Furthermore, stimuli containing aspirated stops were significantly lower in accuracy compared to stimuli containing strong stops ("Strong vs aspirated", $\chi^2(1) = 7.97, 0.005$). Taken together, these results indicate that subjects in the VOT condition have the highest accuracy when identifying stimuli containing strong stops, and are better at identifying strong and aspirated stops compared to weak stops. This suggests that by attending to VOT, subjects are able to separate the strong from the aspirated stops fairly well, which makes sense given that the strong and aspirated stops are on opposite ends of the VOT dimension, whereas weak stops lie in the middle of the VOT dimension. Identification accuracy for weak stops may be lower because subjects in the VOT condition are unsure, at least some of the time, how to categorize stimuli with VOT values that lie in the middle of the VOT range.

![Figure 3.9: Confusion plot for VOT condition, showing the proportion of subject responses as a function of stimulus consonant type.](image)

Figure 3.9: Confusion plot for VOT condition, showing the proportion of subject responses as a function of stimulus consonant type.

One way to visualize when and how subjects made mistakes when identifying stop categories is to plot the proportion of responses given by subjects as a function of the type of consonant in the test stimuli. Figure 3.9 is a "confusion plot" for the responses given by subjects in the VOT training condition, which reveals the responses given by subjects when
they are confused about the identity of stops.

Subjects in the VOT condition were fairly accurate when identifying both strong and aspirated stops: most of the time, subjects chose the aspirated category (i.e. ㅋ in a yellow circle) when they heard a syllable beginning with an aspirated stop, and they chose the strong category (i.e. ㄲ in a red circle) when they heard a syllable beginning with a strong stop. Stimuli beginning with a weak stop posed more of a problem for these subjects; though more than half of the time subjects did give the correct response for syllables with a weak stop, they also assigned weak stops to the aspirated category (and sometimes to the strong category).

For the F0 condition, a logistic mixed effects model was fitted to the post-training identification data, with accuracy on a single trial as the dependant variable. The model contained two categorical variables for consonant type as fixed effects. These variables coded for mean differences among consonant type ("Strong vs. others" and "Aspirated vs weak"). The model also included random effects for each subject. For the F0 condition, stimuli containing strong stops had significantly lower accuracy compared to stimuli containing the other stop types ("Strong vs. others", $\chi^2(1) = 4.92, p = .03$). Stimuli containing weak stops were slightly lower in accuracy compared to stimuli containing aspirated stops, but this difference was only marginally significant ("Aspirated vs weak", $\chi^2(1) = 3.88, p = .05$). These results indicate that subjects in the F0 condition had poor accuracy when identifying strong stops, compared to the other stop types. This makes sense if subjects are using F0 as cue: weak and aspirated stops are at opposite ends of the F0 dimension, so subjects may be able to identify them better compared to the strong stimuli that have middle F0 values.

Figure 3.10 is a confusion plot for the responses given by subjects in the F0 training condition. It is clear from this plot that subjects in F0 condition were confused about the identity of stops more often than the VOT subjects. When the stimulus had an aspirated stop, subjects gave the correct response about half the time. When they made a mistake, they were more likely to say that aspirated stops were "strong". However, they are equally
likely to assign stimuli with weak stops to the “weak” or the “aspirated” categories, which suggests that subjects in this condition may not be attending to F0 very well (or may be attending to other cues in addition to F0). If they were, we might expect to see mostly “weak” responses for weak stimuli (i.e. higher accuracy for the weak stops), and more “strong” mistakes when identifying weak stops (because strong stops are in the middle of the F0 dimension). Subjects were very confused when identifying the “strong” stops, for which their most common response was actually “weak” (i.e. ㄱ in a blue circle).

For the F0 condition, a logistic mixed effects model was fitted to the post-training identification data, with accuracy on a single trial as the dependent variable. The model contained two categorical variables for consonant type as fixed effects. These variables coded for mean differences among consonant type (“Weak vs. others” and “Aspirated vs strong”). The model also included random effects for each subject. For the VQ condition, stimuli containing weak stops had significantly lower accuracy compared to stimuli containing the other stop types (“Weak vs. others”, $\chi^2(1) = 5.81, p = .02$). There was no difference in accuracy for stimuli containing aspirated stops compared to stimuli containing strong stops (“Aspirated vs strong”, $p = .89$). These results indicate that subjects in the VQ condition
were better at identifying strong and aspirated stops compared to weak stops, which makes sense if they are attending to voice quality, because aspirated and strong stops are at the opposite ends of the voice quality dimension, whereas weak stops are in the middle. However, accuracy for both aspirated and strong stops was fairly low overall (especially compared to the VOT group), which suggests that though subjects may have been learning to attend to voice quality, they are not using their category-relevant dimension as efficiently as subjects in the VOT condition.

Figure 3.11 is a confusion plot for the responses given by subjects in the Voice Quality training condition. When identifying stimuli with aspirated stops, the majority of the time subjects gave the correct answer, and when they made mistakes were more likely to respond “weak”. Likewise, for stimuli with strong stops, subjects responded “strong” the majority of the time. However, they are at chance for stimuli with weak stops, because they are almost equally likely to respond with any of the three stop types. These results suggest that if these subjects are attending to voice quality primarily, they can use it to identify aspirated stops and strong stops (though with fairly low accuracy), but they are confused about the identity of the stops that fall in the middle of the voice quality continuum (i.e. weak stops).

![Confusion plot for Voice Quality condition](image)

Figure 3.11: “Confusion” plot for Voice Quality condition, showing the proportion of subject responses as a function of stimulus consonant type.
3.5.3 Korean difference ratings

For the difference rating task, subjects heard all pairwise combinations of the 18 test stimuli (Korean CV syllables, unmanipulated except for loudness scaling) and provided a difference rating for each pair on a scale from 1-9. These difference ratings were used in an INDSCAL analysis to obtain multi-dimensional scaling solutions for each training condition, pre- and post-training.

As in the previous chapter, the pre- and post-training difference ratings for the Korean test stimuli were analyzed with a combination of an INDSCAL and multiple regression models. Briefly, an INDSCAL analysis allows for subjects to have different weights for each of the dimensions in the scaling solution. The algorithm computes a group space, based on the average of values in dissimilarity matrices for each subject (pre- and post-training), along with individual weights for each subject. This allows for individual scaling configurations for each subject, pre- and post-training. These individual configurations can then be grouped together depending on training condition and training phase to visualize how difference ratings may change as a result of training, and how this change may depend on the training condition. As in the analysis for the Xorx and Bjaran data, the INDSCAL analysis for the Korean difference ratings was implemented in R, using the smacofIndDiff() function from the SMACOF package (De Leeuw and Mair, 2011), assuming an interval distance model.

3.5.3.1 Interpreting dimensions in the scaling solution

It was found that a two dimensional scaling solution was best for representing the Korean difference ratings, both due to the stress level and the interpretability of the scaling dimensions. The dimensions are called “D1” and “D2”. Note that these are simply labels for the dimensions in the scaling solution, and do not reflect any particular hierarchy of importance (they could just as easily be labelled “X” and “Y”).

The stimuli form three clusters, with the strong stimuli grouped together on the left
side of the plot, the aspirated stimuli at the top right of the plot, and the weak stimuli at the bottom right of the plot. The strong stimuli are further away from the other two categories in this group space, suggesting that overall (collapsing across all subjects, training phases, and training conditions) the strong stimuli, as a group, were heard as more different from the other two stop types, which were closer to each other in similarity space.

Figure 3.12 is a biplot of the group space from the INDSCAL analysis of Korean difference ratings. Each of the 18 test stimuli is represented with a colored dot that reflects the stop category of the initial consonant.

![Figure 3.12: Biplot of group space from INDSCAL of Korean difference ratings.](image)

Again, it is important to note that the dimensions in an MDS solution are NOT the physical dimensions of the stimuli per se. However, it is possible interpret MDS dimensions in light of physical dimensions through visual inspection of the scaling plot, and by modeling the extent to which the dimensions in the scaling plot correlate with the physical dimensions of the stimuli.

To evaluate the extent to which the scaling dimensions in the group space correlated
with the acoustic values of the stimuli (VOT, F0, and VQ, in this case a measure of spectral tilt, H1-F1) a set of 3 linear regression models were fit to the results of the Korean INDSCAL group space, using the biplotmds() function in the R “smacof” package (De Leeuw and Mair 2011). The dependent variable in these models were the acoustic values (F0, VOT, and VQ, z-scored) of the Korean stimuli, which were regressed on the scaling coordinates of the stimuli for each of the two dimensions from the group-space.

The vectors in figure 3.12 show how the acoustic dimensions of VOT, F0 and VQ “load” on each of the dimensions in the scaling solution. The endpoints of the vectors represent the regression coefficients for D1 and D2 from the individual models where a single acoustic property (z-scored) is regressed on the stimuli coordinates from the group scaling space. For example, in the linear model where \( VOT = \beta_1 D1 + \beta_2 D2 \), the endpoints of the VOT vector in figure 3.12 are \((\beta_1, \beta_2)\). See table 3.7 for a listing of dimension coefficients from each of the three regression models, obtained using the biplotmds() function.

Table 3.7: Coefficients from individual regression models where acoustic variables are regressed on D1 and D2 coordinates from the group space scaling solution.

<table>
<thead>
<tr>
<th>Acoustic variable</th>
<th>Coefficients from regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta_1 D1 )</td>
</tr>
<tr>
<td>VOT</td>
<td>1.46</td>
</tr>
<tr>
<td>F0</td>
<td>0.1</td>
</tr>
<tr>
<td>VQ</td>
<td>1.53</td>
</tr>
</tbody>
</table>

If the vectors are extended through the origin of the scaling solution, they can be thought of as acoustic “axes” along which individual stimuli lie. For example, as the F0 axis increases (in the direction of the arrow toward the top of the scaling plot), the F0 values of the stimuli also increase: weak stimuli (blue dots) are lowest on the F0 axis, strong stimuli (green dots) are in the middle of the F0 axis, and aspirated stimuli (red dots) are highest on the F0 axis, mirroring the F0 pattern for the Korean stops (weak < strong < aspirated).

The F0 axis is nearly overlapping with the axis for D2 in the scaling plot, which aids
in the interpretation of D2 (i.e., it is related to F0). In the regression model for F0, D2 was found to be significant \( F(1, 16) = 22.54, p = .0002 \), but D1 was not \( p = .73 \). In figure 3.12, the F0 vector is nearly aligned with D2, reflecting the higher coefficient for D2 compared to D1 in the regression model (see table 3.7). Furthermore, the distribution of stimuli along both the F0 axis and D2 in the scaling solution reflects the patterning of the stop types along the F0 continuum: vowels following weak stops have the lowest F0, vowels following strong stops are in the middle, and vowels following aspirated stops have the highest F0.

Voice Quality (H1-F1) seems to load more heavily on D1 than D2: in the regression model for VQ, D1 was found to be significant \( F(1, 16) = 50.12, p < .00001 \), but D2 was not \( p = .25 \).

VOT seems to be related to both dimensions: both dimensions were significant in the regression model (\( \text{D1: } F(1, 16) = 46.17, p < .00001 \); \( \text{D2: } F(1, 16) = 5.65, p = .03 \)). However, VOT loads more heavily on D1, as reflected in the higher coefficient for D1 compared to D2 (see table 3.7). Both of the vectors for VOT and VQ extend in the same direction (to the right of the scaling plot), and the distribution of stimuli along these vectors reflects the patterning of stop types along the two dimensions: for both the VQ and the VOT axes, strong stops have the lowest values, weak stops are in the middle, and aspirated stops have the highest values.

The group space biplot shows how stop categories separate along the individual acoustic dimensions that were used for training. The F0 axis provides decent separation between the the aspirated and weak stops, but the strong stops are dispersed in the middle of the F0 axis, overlapping considerably with the other two categories. The VOT axis provides good separation between the strong and the other stop types, but there is some overlap between the aspirated and the weak stops. The same is true for the VQ axis. Based on these scaling results, it appears as though an optimal separation of the three stop categories can be achieved by using two dimensions: both VOT and VQ separate the strong stops from the other types, and F0 separates the weak from the aspirated stops. This is an interesting result.
in and of itself, as it aligns with other research on how native Korean speakers perceive the stop categories (see section 3.1.2, page 78).

3.5.3.2 Scaling plots for the three training conditions

Figures 3.13, 3.14 and 3.15 show the INDSCAL results for each of the three training conditions. In these plots, points represent average stimulus coordinates across all subjects, pre-training. The coordinate values are collapsed across training condition, because it is assumed that training condition will not make a difference in pre-training difference ratings, since subjects were never informed that they were in a specific training condition. From each point extends a vector, the endpoint of which is the average stimulus coordinate, post-training, for subjects in that training condition. Therefore, each vector represents both the magnitude and direction of pre- to post-training shift for a single stimulus. These vectors give a global picture of how difference ratings may change as a result of both training and training condition.

In all three scaling plots, it is clear that even prior to training, there is quite a bit of separation between the strong and the aspirated/weak stops, and that this separation occurs more along D1 (or the VOT/VQ axes). This suggests that, even prior to training, there is a discontinuity along these dimensions for the English subjects. This likely reflects the English distinction between /k/ and /g/, a distinction that English speakers make by attending to VOT primarily. The question, then, is whether training on a single dimension will help English speakers “pull apart” the weak and the aspirated stops.

In the scaling plot for the VOT condition (3.13), the aspirated and weak stops moved toward each other, post-training. On average, subjects trained with VOT perceived the stops in these categories as more similar after training. At the same time, the strong stops moved away from the aspirated and weak stops, and toward each other. Training with VOT didn’t seem to result in aspirated and weak stops becoming less similar, but it did help to further separate the strong from the weak stops. The weak and aspirated stops moved toward each
other mainly along D2, which is related to F0. This suggests that as subjects learn to draw attention away from F0, they “compress” this dimension, so differences along F0 become less important. At the same time, the strong stops moved away from the other stops along D1, which is more related to VOT/VQ. As subjects learn to attend to VOT, they “expand” this dimension, so differences along VOT become more important. Despite this evidence to suggest that training with VOT resulted in subjects using VOT as the basis for difference ratings to a greater extent after training, they still had trouble separating the weak and
aspirated stops. This is perhaps due to the fact that aspirated and weak stimuli are closer in terms of VOT values.

![INDSCAL results for F0 (Familiar) condition. Points represent average stimulus coordinate, pre-training. Endpoints of vectors represent average stimulus coordinate, post-training.](image)

In the scaling plot for the F0 condition, the aspirated and weak stops moved away from each other, post-training. On average, subjects trained with F0 perceived the stops in these categories as less similar after training. At the same time, the strong stops are moved toward the other stops, and away from each other. Training with F0 did result in the aspirated and weak stops becoming less similar, but it caused the strong stops to be
perceived as more similar to the other two stop categories. The aspirated and weak stops moved away from each other mainly along D2, which is related to F0. As subjects learn to attend to F0, they are “expanding” this dimension, and pulling apart the categories that are at the endpoints of the F0 continuum. The strong stops moved toward the other stops mainly along D1, which is more related to VOT/voice quality. Subjects trained with F0 may be “compressing” this dimension as they draw attention away from acoustic properties that are not F0, so differences along this dimension become less noticeable. Some of the strong stops are moving toward the aspirated stops and some are moving toward the weak stops. As subjects learn to attend to F0, they are perceiving some of the strong stops as more similar to aspirated stops (likely those stops that are higher in F0) and some of the strong stops as weak stops (likely those stops that are lower in F0). They had difficulty in separating the strong stops from the other two stop types because the strong stops lie in the middle of the F0 continuum.

The scaling plot for the VQ condition (3.15) is similar in many ways to the F0 scaling plot. The strong stops moved toward the weak and aspirated stops, and the aspirated and weak stops moved toward each other. This is curious because if subjects are attending to VQ, we might expect their scaling solutions to mirror the VOT condition rather than the F0 condition, because the stops pattern similarly along VOT and VQ (strong < weak < aspirated). However, it appears that the subjects in the voice quality training condition may have been attending to F0, rather than VQ, as a basis for making difference judgements. Their results indicate similar expansion along the F0 axis, and compression along the V0T and VQ axes, which shouldn’t happen if they are attending to VQ only. One difference between the F0 and VQ plots is that the stimuli don’t seem to be shifting as much, pre- to post-training, for the VQ subjects. This suggests that training with VQ didn’t have as much of an effect on difference ratings compared to the other two conditions.
Figure 3.15: INDSCAL results for Voice Quality (New) condition. Points represent average stimulus coordinate, pre-training. Endpoints of vectors represent average stimulus coordinate, post-training.

3.5.4 Category learning in Korean: Acquired Similarity and Acquired Distinctiveness

The scaling solutions for the difference ratings were evaluated for evidence of category learning, specifically in terms of acquired similarity, where within-category distances become smaller as a result of category learning, and acquired distinctiveness, where between-category distances become larger as a result of category learning.
3.5.4.1 Acquired similarity

The comparison of pre- and post-training within-category distance can provide evidence of Acquired Similarity (AS): if post-training within-category distances are smaller, this suggests that categories are “tightening up”.

Within-category distances were obtained for each subject by calculating the mean Euclidean distance between the subject’s category centroids (the average spatial coordinates for all stimuli in a category) and each stimulus in that category. The overall within-category distances for each subject were found by taking the mean of all within-category distances for each stop category. See figure 3.16 for a schematic showing within-category distance.

![Figure 3.16: Schematic for within-category distance](image)

Each subject had six within-category distances: average within-category distance for aspirated stops, pre- and post-training, average within-category distance for weak stops, pre- and post-training, and average within-category distance for strong stops, pre- and post-training. Figure 3.17 shows the within-category distances (averaged across subjects and consonant types) for each of the three training conditions, pre- and post-training.

A measure of “within-category shift” was constructed by subtracting post-training distances from pre-training distances for each of the stop categories. The within-category
shift variable represents the degree and direction of change between post-training and pre-training within-category distance.

To evaluate the effect of training on judgments of similarity and in order to test for a significant change in within-category distance after training, a linear model was fit to the results of the pre- and post-training Korean INDSCAL scaling configurations using the R Stats package. The dependent variable was the “within-category shift” variable (the difference between the pre- and post-training within-category distance, which was the Euclidean distance between each subject’s category centroid and the positions of each stimulus in each subjects individual configuration, see above for full description of how this variable was computed). The independent variables were two categorical variables for training condition, which were contrast coded for specific mean comparisons. The two condition variables compared the mean within-category shift for different training conditions (“VOT vs others”, “F0 vs. VQ”).

The mean within-category shift was larger for the VOT condition compared to the other two conditions, as indicated by a significant condition variable (“VOT vs others”,

Figure 3.17: Within-category distance, pre- vs. post-training, split by training condition. Error bars represent $SEM_{betw}$. 

![Graph showing within-category distance pre- and post-training, split by training condition.](image-url)
\[ F(1, 177) = 28.71, p < .00001 \]. In other words, the measure of within-category distance changed more for the VOT group in post-training, compared to the other conditions. There was no difference in the mean within-category shift for the F0 and VQ conditions (“F0 vs. VQ”, \( p = .65 \)).

To investigate the effect of consonant type on within-category shift in each condition, post-hoc linear models were fit to the results of the pre- and post-training Korean INDSCAL configurations for each of the three consonant types, within each of the three training conditions (a total of 9 linear models). The dependent variable in each of these post-hoc models was the “within-category shift” variable. These models were simple, intercept-only models, which were used to test if the within-category shift was significantly different from zero for each condition (\( \text{Within-category Shift} = \beta_0 \)). In other words, these models tested if there was a significant shift in within-category distance, from pre- to post-training.

As in the AS analysis of the Xorx and Bjaran data in chapter 2, in order to obtain an F-statistic to test for the significance of the intercept in these post-hoc models, the full model was compared with a reduced model where the variable of interest was left out (in this case, the intercept). The full and reduced models were compared using the \texttt{anova()} function in the R Stats package.

Figures 3.18 to 3.20 show the within-category distances (averaged across subjects) for each of the training conditions, split by consonant type. Asterisks indicate significant differences between pre- and post-training in the post-hoc tests. “AS” indicates an Acquired Similarity effect (where within-category distances are smaller, post-training).

Post-hoc models for the VOT condition indicate that there were significant shifts for all three consonant types from pre- to post-training, in the predicted direction for an AS effect. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated within-category shift of each consonant type in the VOT training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “\( p < .001 \)”, H-B corrected) when applicable. For aspirated stops,
Table 3.8: Intercepts for post-hoc models of within-category shift for each training condition and stop type.

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT-aspirated</td>
<td>0.005431</td>
</tr>
<tr>
<td>VOT-strong</td>
<td>0.028860</td>
</tr>
<tr>
<td>VOT-weak</td>
<td>0.0023128</td>
</tr>
<tr>
<td>F0-aspirated</td>
<td>-0.003573</td>
</tr>
<tr>
<td>F0-strong</td>
<td>-0.023055</td>
</tr>
<tr>
<td>F0-weak</td>
<td>-1.936e-05</td>
</tr>
<tr>
<td>VQ-aspirated</td>
<td>-0.005173</td>
</tr>
<tr>
<td>VQ-strong</td>
<td>-0.02672</td>
</tr>
<tr>
<td>VQ-weak</td>
<td>-0.0013624</td>
</tr>
</tbody>
</table>

the intercept in the simple model was significant \( F(1, 19) = 16.45, p = .002, \) H-B corrected), indicating that the mean within-category shift was non-zero. Furthermore, the intercept was positive (see table 3.8 for a list of the intercepts in the 9 post-hoc AS models). A positive intercept in this case shows that the within-category shift was in the predicted direction for an AS effect, i.e. post-training within-category distances were smaller after training. A significant positive intercept was found for the other two stop types as well (strong stops: \( F(1, 19) = 10.52, p = .004, \) H-B corrected; weak stops: \( F(1, 19) = 15.45, p = .002, \) H-B corrected).

Post-hoc models for the F0 condition indicate that there were significant shifts for some of the consonant types from pre- to post-training, but not in the predicted direction for an AS effect. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated within-category shift of each consonant type in the F0 training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “\( p<.001\)”, H-B corrected) when applicable. In the model for aspirated stops, the intercept was significant, but negative \( F(1, 19) = 10.48, p = .008, \) H-B corrected), and the same was true for strong stops \( F(1, 19) = 14.09, p = .004, \) H-B corrected). The intercept in the simple model for weak stops was not significant \( p = .97, \) H-B corrected), indicating that there was no shift between pre- and post-training within-
Figure 3.18: **VOT training condition**: within-category distance, pre- vs. post-training, split by consonant type. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AS” indicates an Acquired Similarity effect (within-category distances are smaller post-training). Error bars represent $SEM_{betw}$.

category distances. These results demonstrate that there were no AS effects for any of the consonant types in the F0 condition; in fact, within-category distances actually got bigger for the aspirated and strong categories after training.

Post-hoc models for the VQ condition indicate that there were no significant shifts for any of the consonant types from pre- to post-training. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated within-category shift of each consonant type in the VQ training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “$p<.001$”, H-B corrected) when applicable. For the VQ condition, the within-category shift for all three stop types was marginally significant, but the intercept was negative ($F(1,19) = 7, p = .05$, H-B corrected). There was no significant within-category shift for the other consonant types (strong:$p = .07$, H-B corrected; weak:$p = .07$, H-B corrected). There were no AS effects for any of the consonant types in the VQ condition.
Figure 3.19: **F0 training condition**: within-category distance, pre- vs. post-training, split by consonant type. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. Error bars represent $SEM_{betw}$.

Figure 3.20: **Voice Quality training condition**: within-category distance, pre- vs. post-training, split by consonant type. Periods indicate marginal difference between pre- and post-training in post-hoc tests. Error bars represent $SEM_{betw}$. 
3.5.4.2 Acquired distinctiveness

Acquired distinctiveness is the process by which between-category distances become larger as a result of category learning. The comparison of pre- and post-training between-category distance can provide evidence of Acquired Distinctiveness (AD): if post-training between-category distances are larger, this suggests that categories are “moving apart” in psychological space, and becoming less similar to each other.

Between-category distances were obtained for each subject by calculating the mean Euclidean distance between each category centroid (aspirated to weak, aspirated to strong, and strong to weak). Figure 3.21 shows a schematic of the between-category distances.

To express the between-category distance between all centroids, the area of the triangle formed by the three between-category distances was calculated using Heron’s formula for the area of a triangle with sides $a$, $b$, and $c$:

$$A = \frac{\sqrt{(a^2 + b^2 + c^2)^2 - 2(a^4 + b^4 + c^4)}}{4}$$

![Figure 3.21: Schematic for between-category distance, where $d_x$ = the Euclidean distance between two category centroids Overall between-category distance was found by calculating the area of the triangle formed by the three pair-wise distances. Error bars represent SEMbetw.](image)

Each subject had one between-category area, and three between-category distances for each category centroid pair, to be used in post-hoc tests (Aspirated/Strong, Aspirated/Weak, and Strong/Weak). Figure 3.22 shows the between-category area (averaged across subjects) for each of the three training conditions.
Figure 3.22: Area between category centroids, pre- vs. post-training, split by training condition.

A measure of “between-category area shift” was computed by subtracting pre-training centroid area from post-training centroid area. This between-category shift variable represents the degree and direction of change between post-training and pre-training between-category distance.

In order to test for a significant change in between-category distance after training, and to assess the extent to which that change depended on training condition, a linear model was fit to the results of the pre- and post-training Korean INDSCAL scaling configurations using the R Stats package. The dependent variable was the “between-category shift” variable (described above). The independent variables were two categorical variables for training condition, which were contrast coded for specific mean comparisons. The two condition variables compared the mean between-category shift for different training conditions (“VOT vs others”, “F0 vs VQ”).

The mean between-category shift was significantly larger for the VOT condition compared to the other training conditions (significant condition variable, “VOT vs others”: $F(1, 57) = 24.76, p < .00001$). There was no difference in the mean between-category area
shift for the F0 and the VQ condition \( (p = .36) \).

To investigate the effect of between-category shift for specific centroid pairs (i.e. distance between aspirated and weak category centroids), post-hoc linear models were fit to the results of the pre- and post-training Korean INDSCAL configurations for each of the category pairs, within each of the three training conditions (a total of 9 linear models). The dependent variable in each of these post-hoc models was the “between category shift” variable for specific centroid pairs. For example, the between-category shift for the Aspirated-Strong pair was found by subtracting the pre-training distance between aspirated and strong centroids from the post-training distance between these centroids. The models were simple, intercept-only models, which were used to test if the between-category shift for specific category pairs was significantly different from zero \( (\text{Between - categoryShift} = \beta_0) \).

As in the AS analysis above, in order to obtain an F-statistic to test for the significance of the intercept in these post-hoc models, the full model was compared with a reduced model where the variable of interest was left out (in this case, the intercept). The full and reduced models were compared using the \texttt{anova()} function in the R Stats package.

Figures 3.23 to 3.25 show the pre- and post-training between-category distances (averaged across subjects) for each training condition, split by pair. Asterisks indicate significant differences between pre- and post-training in the post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training).

Post-hoc models for the VOT condition indicate that there were significant between-category shifts for all three centroid pairs, and two were in the correct direction for an AD effect. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated between-category shift for each centroid pair in the VOT training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “\( p < .001 \)”, H-B corrected) when applicable. For the Aspirated/Strong centroid pair, the intercept in the simple model was significant \( (F(1,19) = 6.75, p = .04, \text{H-B corrected}) \), indicating that the mean between-category shift for this pair was non-zero.
Table 3.9: Intercepts from post-hoc models of between-category shift for each training condition and centroid pair

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT-Aspirated/Strong</td>
<td>0.04269</td>
</tr>
<tr>
<td>VOT-Aspirated/Weak</td>
<td>-0.07189</td>
</tr>
<tr>
<td>VOT-Strong/Weak</td>
<td>0.02797</td>
</tr>
<tr>
<td>F0-Aspirated/Strong</td>
<td>-0.05748</td>
</tr>
<tr>
<td>F0-Aspirated/Weak</td>
<td>0.05841</td>
</tr>
<tr>
<td>F0-Strong/Weak</td>
<td>-0.03819</td>
</tr>
<tr>
<td>VQ-Aspirated/Strong</td>
<td>-0.05304</td>
</tr>
<tr>
<td>VQ-Aspirated/Weak</td>
<td>0.06736</td>
</tr>
<tr>
<td>VQ-Strong/Weak</td>
<td>-0.03425</td>
</tr>
</tbody>
</table>

Furthermore, the intercept was positive (see Table 3.9) a list of the intercepts in the post-hoc between-category shift models. A positive intercept in this case shows that the between-category shift was in the predicted direction for an AD effect (i.e. post-training between-category distances are larger after training). A significant, positive intercept was found for the Strong/Weak pair as well ($F(1, 19) = 6.06, p = .01$, H-B corrected). For the Aspirated-Weak pair, there was a significant difference between pre- and post-training distances, but the intercept in this model was negative, indicating that between-category distances actually got smaller after training ($F(1, 19) = 10.6, p = .04$, H-B corrected). Taken together, these results demonstrate that for the VOT condition, there was an AD effect for two of the category pairs, Aspirated-Strong and Strong-Weak. For the aspirated-weak pair, the centroids moved closer together.

Post-hoc models for the F0 condition indicate that the between-category shift was significant for all three centroid pairs, but only one of the pairs showed an AD effect. These results show that there was an AD effect for one of the pairs (Aspirated/Weak), but for the other pairs, the category centroids moved closer together. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated between-category shift for each centroid pair in the F0 training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “p<.001”, H-B
Figure 3.23: **VOT training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent $SEM_{betw}$.

Figure 3.24: **F0 training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Asterisks indicate significant difference between pre- and post-training in post-hoc tests. “AD” indicates an Acquired Distinctiveness effect (between-category distances are larger, post-training). Error bars represent $SEM_{betw}$.

corrected) when applicable.
The intercept in the simple model for the Aspirated/Weak pair was positive and significant \(F(1, 19) = 14.18, p = .003, \) H-B corrected), indicating an AD effect. The Aspirated/Strong pair did have a significant shift measure, but the intercept was negative, indicating that the distance between centroids was smaller post-training \(F(1, 19) = 15.04, p = .003, \) H-B corrected). The same was true for the Strong/Weak pair \(F(1, 19) = 14.78, p = .003, \) H-B corrected).

Figure 3.25: **Voice Quality training condition**: between-category distance, pre- vs. post-training, split by consonant type pair. Error bars represent \(SEM_{\text{betw}}\).

Post-hoc models for the VQ condition showed that no pairs had a significant between-category shift. To account for multiple testing with the Korean difference rating data for the family of 3 post-hoc models that investigated between-category shift for each centroid pair in the F0 training condition, alpha levels were adjusted using the Holm-Bonferroni correction. The correction is noted (e.g. “\(p < .001\)”, H-B corrected) when applicable. None of the intercepts in the models for the VQ condition were significant (Aspirated/Strong, \(p = .1, \) H-B corrected; Aspirated/Weak, \(p = .1, \) H-B corrected; Strong/Weak, \(p = .1, \) H-B corrected).
3.6 Discussion

Experiment 2 was designed to train English speakers with no experience in Korean to hear a Korean stop contrast. This was done using a method of high-variability perceptual training with a single acoustic training dimension. There were three training conditions, where subjects in each condition were trained with a different acoustic dimension that was related to English in a specific way. Results from pre- and post-training difference ratings and a post-training identification task showed that phonetic training affected subjects' ability to distinguish between the types of stops. A comparison of different training dimensions indicated that some training dimensions were more effective than others for learning the Korean stop contrast.

3.6.1 Overview of results

3.6.1.1 Category training

For the Korean category training, subjects in each condition heard stimuli (Korean CV syllables, manipulated to highlight particular acoustic dimensions, see section 3.4.2, page 92) from two of the three training arrays (depending on their training condition), and identified the stop category of the initial consonant.

Subjects in the VOT condition had higher accuracy, overall, compared to the other conditions. Models for individual training conditions showed that accuracy increased over the course of training for both the VOT and VQ conditions. Accuracy in the F0 condition did not increase, which suggests that while subjects in both the VOT and the VQ conditions did utilize VOT and VQ to some extent to identify the Korean stops, the subjects in the F0 condition were not as successful in utilizing F0 as a contrastive cue.

The VOT condition had an advantage in terms of using VOT as a contrastive cue at the beginning of training, as evidenced by higher accuracy in block 1 for the VOT condition.
3.6.1.2 Identification task

For the post-training identification task, subjects heard each of the 18 test stimuli (Korean CV syllables, unmanipulated except for loudness scaling, see section 3.4.1, page 87) and identified the stop category of the initial consonant.

Subjects in the VOT condition had the highest accuracy overall compared to the other training conditions, and there was no difference in overall accuracy between the F0 and VQ conditions. The type of consonant in the test stimuli also made a difference in accuracy for each of the training conditions.

Subjects in the VOT condition were more accurate when identifying aspirated and strong stops, which are at opposite ends of the VOT dimension. Over half the time, they also responded correctly when identifying weak stops. When they did make mistakes with weak stops, they were more likely to assign them to the aspirated category. Subjects trained with VOT seem to have used VOT successfully to identify the stop categories, particularly strong and aspirated. It appears as if subjects had more difficulty determining which VOT values corresponded to weak stops.

Subjects in the F0 condition were better at identifying aspirated and weak stops, which lie at opposite ends of the F0 dimension. When the stimulus had a weak stop, they were equally likely to say the stop was “aspirated” or “weak”. This suggests that while they may have recognized that higher pitched syllables were part of the “aspirated” category, for the other stop categories they may not have been utilizing F0 efficiently (or in isolation). If they were, we might expect higher accuracy for the weak stops, and more confusion between weak and strong stops, which are closer together along the F0 dimension.

Subjects trained with voice quality were best at identifying strong and aspirated stops, which lie on opposite ends of the voice quality dimension. They were at chance for the weak stops. Overall accuracy is low in this condition, which suggests that if subjects used voice quality to identify the aspirated and strong stops, it is not a very effective cue, particularly
for identifying strong and weak stops. Furthermore, when they had to identify a strong stop, they more often mistook it for an aspirated stop. This calls into question if the subjects are attending to the VQ dimension effectively, or in isolation. If they were, we might expect to see more “weak” mistakes for strong stops, because weak stops lie in the middle of the voice quality dimension.

### 3.6.1.3 Difference ratings

For the difference ratings, subjects heard all pairwise combinations of the 18 test stimuli (Korean CV syllables, unmanipulated except for loudness scaling, see section 3.4.1, page 87) and provided a difference rating for each pair on a scale from 1-9. These difference ratings were used in an INDSCAL analysis to obtain multi-dimensional scaling solutions for each training condition, pre- and post-training.

The scaling results showed that the test stimuli formed three coherent clusters, where each cluster was made up of stimuli with the same consonant type. The acoustic values of the stimuli formed axes along which the stimuli clusters could be separated. The VOT axis loaded more heavily along D1 in the scaling results, which separated the strong cluster from the other two. The VQ axis loaded more heavily along D1 as well. The F0 axis loaded heavily along D2, which separated the weak and aspirated clusters. It appears that two dimensions, rather than a single dimension, provides optimal separation between the three stop categories.

Prior to training, there was separation between the strong stops and the other stop types, for all three training conditions. This separation occurred more along the VOT/VQ axes, which suggests that there is a discontinuity along one or both of these acoustic dimensions for English subjects. This is likely due to the distinction between English /k/ and /g/: the VOT boundary separating the two English categories lies between the VOT values for Korean strong stops and other stop types.

In the scaling plot for the VOT condition (figure 3.13, page 117), the aspirated and
weak stops moved toward each other post-training, while the strong stops moved away from the others. This means that after training with VOT, subjects rated strong stops as less similar compared to the other two stop types, but aspirated and weak stops were rated as more similar. In a sense, VOT differences between the strong stops and the other stop types were enhanced as a result of training with VOT.

For the F0 condition, the scaling plot (figure 3.14, page 118) showed the aspirated and weak stops moving away from each other post-training, while the strong stops moved toward the aspirated and weak stops. This means that after training with F0, subjects rated aspirated and weak stops as less similar. At the same time, some of the strong stops were rated as similar to aspirated, and some of the strong stops were rated as similar to weak stops, which suggests that while training helped these subjects to separate weak and aspirated stops, it decreased their ability to separate strong stops from the other two categories.

The scaling plot for the VQ condition was similar to the F0 condition (figure 3.15, page 120), where aspirated and weak stops moved away from each other, and strong stops moved toward the other two categories. This is an unexpected result if VQ subjects are attending to VQ in isolation. If they were, their scaling solutions would be more similar to the VOT condition rather than the F0 condition, due to the fact that the stops pattern similarly along both VOT and VQ (strong < weak < aspirated). It may be that subjects in the VQ condition are attending to F0 as a basis for making difference ratings, which may suggest something about the relative salience of these two training dimensions (see discussion below).

There was evidence of acquired similarity (AS) in the VOT condition for all three stop types: training with VOT resulted in within-category distance becoming smaller for aspirated, strong, and weak stops (but to a larger extent for the strong stops). There was no evidence of AS in the F0 condition; in fact, the aspirated and strong clusters became more dispersed after training (within-category distances were larger), and there was no change in the within-category distances for the weak stops. For the VQ condition there was also no evidence of an AS effect.
An acquired distinctiveness (AD) effect was found for some category pairs, but this depended on the training condition. There was evidence of AD in the VOT condition for the Aspirated/Strong and Strong/Weak pairs, but between-category distance got smaller for the Aspirated/Weak pair. In the F0 condition, there was evidence of an AD effect for the Aspirated/Weak pair, but between-category distance got smaller for the Strong/Weak pair and the Strong/Aspirated pair. In the VQ condition, there was no shift in between-category distance for any of the centroid pairs, and no AD effect.

3.6.1.4 Exit surveys

Though a systematic analysis of the exit surveys was not done, the surveys do provide some insight into the perceptual experiences of subjects. See Appendix C for a list of the questions in the exit survey.

When asked to characterize the Korean syllables in the experiment (question #8, see Appendix C), many subjects said that the consonants sounded like /k/ and /g/, or like /g/ and two /k/ sounds. Many subjects also noted that the three sounds were very similar and hard to distinguish. These answers support the claim that English speakers are mapping the Korean consonants to English sounds, at least to some extent.

There were subjects who gave more detailed characterizations of the sounds, which may reflect the features that subjects utilizing as contrastive cues. In some cases, subjects’ descriptions aligned well with their training dimension.

For instance, one subject in the F0 training group mentioned that “the blue one” (the weak stop) “was a lower pitch”. One subject in the VOT condition said that “…perhaps some were voiced and others voiceless” (which seems to imply that this subject was attending to VOT in part, and also that they had a working knowledge of phonetic terminology). Several people in the VQ group used language that could be construed as naive descriptions of voice quality. For instance, one VQ subject said that “some had softer sounds and others were sharper, [and] I had an easier time recognizing yellow [aspirated] because it was the most
distinguished of the 3 categories”. Another VQ subject said that “The yellow Korean syllable [aspirated] was softer and had a lighter ending sound. The blue Korean syllable [weak] was harsher with its ending sound.” Taken together, these detailed descriptions suggest that in some cases, subjects were paying attention to a specific dimension, and in some cases, those dimensions aligned with their training condition.

Subjects were also asked to describe any particular learning strategies they used over the course of the training (question #10). Again, some subjects described strategies that aligned well with their training condition. Several subjects in the F0 condition indicated that they focused on “pitch”, “tones” or even “stress” (which could be a description of pitch, as pitch contributes to stress in English). One F0 subject said “Yellow [aspirated] seemed to have a rising sound. Blue sounded deeper & red […] sounded in between”, and another said that they “tried to identify the way that certain words were pronounced, like using an arch in the voice”, which could be interpreted as a description of pitch or perhaps a tone contour. One subject in the VOT condition mentioned “breath before sound” as a strategy, which could be a naive description of VOT. Subjects in the VQ condition described some strategies that could be related to voice quality, for instance, “I designated strengths to each consonant, the blue was most harsh. The yellow was the most calm/soothing and red was in between.” Another VQ subject said that they “listened for the airy sounds and the more pure sounds and the categorized them form there”. One VQ subject mentioned “the sound the vowel made after”, which at least indicates that this subject was listening for a vocalic cue, rather than a cue carried on the consonant portion of the syllable.

3.6.2 Which training dimensions are most effective?

Based on the results from all three tasks in Experiment 2, subjects trained with VOT were most successful. They had the highest accuracy and improved the most over the course of training. Even though they were already good at using VOT as a contrastive cue at the beginning of training, they got better at ignoring other acoustic variation and attending only
to VOT. The VOT condition also had the highest accuracy in the identification task. Though they had trouble identifying weak stops some of the time, confusing them for aspirated stops, they did identify the weak stops correctly about 60% of the time. Training with VOT resulted in acquired similarity for all three stop categories, where within-category distances became smaller post-training, and acquired distinctiveness for the strong stops, where between-category distance from the other two categories became larger post-training.

In general, subjects in the F0 condition did not improve over the course of training, meaning that they were not able to attend to F0 while ignoring other potential cues. Accuracy in the identification task was quite low for this condition. The mistakes made in identifying stops suggests that they weren’t using F0 effectively or in isolation to identify the stops. If they were, we would expect them to confuse weak stops and strong stops more often, because strong stops are in the middle of the F0 dimension. However, when identifying weak stops, they often mistook them for aspirated stops. The scaling results do suggest that subjects in this condition are attending to F0 to make difference ratings, because the stimuli at the endpoints of the continuum (weak and aspirated) move away from each other, and there was an acquired distinctiveness effect for the aspirated-weak pair. However, the aspirated and strong categories are becoming more dispersed as a result of training, and the strong stops are moving closer to both the weak and aspirated stops. Training with F0 seems to cause subjects to suppress their native English-speaker tendencies to focus on VOT as a cue for stop contrasts.

Although the subjects in the voice quality condition did improve over the course of training, which suggests that they were able to ignore VOT and F0 somewhat and focus on the VQ dimension, there is little else to suggest that attending to VQ helped them to learn the Korean stop categories. In fact, trying to focus on VQ may have confused these subjects. The Korean stops pattern similarly along both the VQ and VOT dimension, but in the identification task, subjects in VQ had much lower accuracy compared to the VOT group for the strong and aspirated stops, the endpoints of both dimensions. Furthermore,
the mistakes made when identifying strong stops suggests that subjects may not have been utilizing VQ to a great extent in the identification task. As a result of training, all three categories of test stimuli became more dispersed for this condition. The scaling results for the VQ condition are more similar to the F0 condition, which suggests that whatever subjects in this condition are basing their difference ratings on, it isn’t strictly the VQ dimension. If they were attending to VQ to make the difference ratings, their results would be more similar to the VOT condition.

F0 and voice quality were not very effective training dimensions, taking all results into account. In fact, training with these dimensions seemed to work against the natural advantage that English speakers have in using VOT to make a distinction between stop consonants at the same place of articulation. The strong stops were already separated from the weak stops prior to training, but training with VQ and F0 caused the strong stops to become more similar to the other two categories. It is worth noting, however, that while F0 may not have been an effective training dimension for all three stop types, it may be quite effective as a training dimension to separate aspirated and weak stops.

3.6.3 Phonetic training for Korean stops

Phonetic training, as implemented in this experiment, did affect perception. There were differences between difference ratings when comparing pre- to post-training results, and differences among the three training conditions in post-training tasks. Furthermore, phonetic training seems to be working in the expected way for at least two of the training conditions (VOT and F0). In both cases, subjects shifted attention to their relevant dimensions (VOT or F0, respectively), a process that is especially evident in the scaling results. These results show that subjects are “pulling apart” the stop categories that lie at the endpoints of their respective training dimensions. The perception effects due to training are transferable, because the differences between pre- and post-training tasks and training conditions were seen in the identification and difference ratings, which used different stimuli from those used in
training.

The VOT condition was the most successful based on the results of all experimental tasks, but are subjects really learning the three-way Korean stop contrast? There is evidence to suggest that they are perceiving at least two categories, those at the ends of the VOT dimension... but this could just be a demonstration of their knowledge of English /k/ and /g/. The question is whether training with the VOT dimension helped English speakers to “pull apart” the weak and the aspirated stops in order to make the three-way Korean stop distinction. There are a few pieces of evidence to suggest that subjects trained with VOT might be on their way to making the three-way distinction (and separating weak from aspirated stops). First, results from the identification task showed that around 60% of the time, subjects identified the weak stops correctly. Second, there was an acquired similarity effect for the weak stops, so subjects did perceive weak stimuli as more similar to each other, and they did this to the same degree as they did with aspirated stops.

However, even with some evidence to suggest that subjects in the VOT condition may be learning to separate the weak and aspirated stops, it seems that one dimensional training doesn’t work very well in the particular case of Korean stops, even when the single cue is an easy one for subjects to use (based on experience in English). Using the single dimension of VOT, subjects can do fairly well because they can map two of the Korean stops to English stop categories, but they are not very successful in separating the weak and aspirated stops. This in itself is not so surprising: English speakers have learned not to perceive the VOT boundary between aspirated and weak stops, because this boundary lies in the middle of their native /k/ category.

Training with F0 in isolation seemed to weaken the VOT advantage for separating strong and aspirated stops, because it caused subjects to work against their “native” intuitions about stop categories. By focusing on F0, subjects were able to “pull apart” the stop categories at the two ends of the F0 dimension to some extent. However, they were confused about the identity of strong stops: even though they may have perceived the strong stops
as a single category (based on VOT), training “told” them to ignore the “/g/-ness” of the strong stops and focus on pitch.

Training with the two dimensions of VOT and F0, rather than one, may result in the best performance. Subjects are already good at making stop distinctions using VOT because of their experience with English stop categories, and this knowledge can be used to bootstrap identification of strong stops, probably without much training at all (note that in the pre-training scaling results all subjects hear the strong stops as less similar than the other stop types). Training with F0 can help subjects separate the weak from the aspirated stops. In fact, in a way, F0 is the “best” training dimension for English speakers, because it is training subjects to attend to F0 separately from VOT that allows subjects to pull apart the aspirated and weak categories. This is difficult when they are only attending to VOT. Similarly, difficulties arise when subjects are only attending to F0 as well: training with F0 in isolation seems to have a deleterious effect for the identification of strong stops.

To avoid the negative effect of training with F0 on strong stops, perhaps the best approach for training English speakers to hear the Korean stops would be to focus on the contrast between aspirated and weak stops, using F0 as a training dimension. English speakers are already able to identify Korean strong stops with high proficiency (because they map well to English /g/), but some training with VOT on the contrast between strong and weak stops might help subjects to define a more Korean-like VOT boundary. Essentially, the three-way contrast between stops would be cast as two two-way contrasts between aspirated and weak stops (trained with F0), and between strong and weak stops (trained with VOT). Subjects trained with the two dimensions of VOT and F0 could then have a framework for contrasting the three Korean stops, something along the lines of “if it sounds like /g/, it’s the strong stop. If it sounds like a high pitched /k/, it’s aspirated, and a low pitched /k/ is weak”.

It is important to note that a two-dimensional strategy is probably what Korean speakers do when making stop distinctions. Francis et al. (Francis and Nusbaum, 2002), in a study
that used MDS to investigate how Korean and English speakers perceive the Korean stops, found that Koreans used a two dimensional solution to separate the three stop categories, where each dimension may itself be a bundle of acoustic features (e.g. VOT and VQ may combine to result in one unitary perceptual cue). English speakers with no previous exposure to Korean stops used a single dimension (likely VOT), but, as in the present experiment, were unable to separate weak and aspirated stops. After training, which consisted of category identification with feedback of natural Korean syllables containing the stops, with no focus on specific acoustic dimensions, there was evidence that English speakers were beginning to use a two-dimensional solution to separate the stop categories. However, there is still some doubt that these learners were proficient at separating the weak and aspirated stops, given the scaling results in this study (weak and aspirated stops were much closer together compared to the Korean results).

For English speakers learning Korean, there is certainly some benefit to phonetic training with F0, which can help English learners focus on a useful cue to separate the weak and aspirated stops. It may help that this cue is already perceptually “available” to English speakers as a potentially useful acoustic cue in speech perception. Even though F0 is not the primary contrastive cue for any segmental contrasts in English, it does co-vary with other contrastive cues. Furthermore, F0 is relatively salient and can be explained to novices in a clear way.

3.6.4 Voice quality as a training dimension

The results from Experiment 2 for the VQ training condition suggest that this dimension was not very effective as a contrastive cue for Korean stops. This might be due to several factors.

First, for the test stimuli, the stops overlap more in voice quality compared to the other dimensions (particularly weak and aspirated stops, see figure 3.2). This suggests that voice quality is not a particularly strong contrastive cue for separating the three stop types. An
effective training dimension should provide good separation between categories.

Furthermore, results from the difference ratings suggest that subjects may not have been attending to voice quality much when judging the test stimuli. Voice quality may be a poor training dimension because it is just not very salient to listeners as a contrastive cue, particularly compared to the other dimensions of VOT and F0. The relative salience of a training dimension compared to other potential acoustic cues is an important consideration when deciding on a training strategy. If listeners can’t “hear” a dimension to begin with, it’s going to be fairly useless as a training dimension. If other cues are more salient, then they will be more effective as training dimensions. This salience could be due to psychoacoustic properties of the dimension itself (e.g. it is more robust or small differences along the dimension are easier to hear) or it could be that native language experience functions to highlight some cues over others, effectively making them more salient.

It should be noted that the higher relative salience of VOT and F0 compared to VQ is not likely due to any issues of psycho-acoustic scaling in the training stimuli. Pilot work with the training stimuli (reported in Appendix B) indicated that, in terms of psycho-acoustic scaling, the dimensions in the training arrays were fairly equal. For the F0-VOT array, mean discrimination along the F0 dimension was not statistically different from mean discrimination along the VOT dimension. Likewise, in the VQ-VOT array, mean discrimination along VQ was not statistically different from mean discrimination along VOT. In the F0-VQ array, the VQ dimension actually had higher discriminability compared to F0. This was due to a high peak in d-prime along the VQ dimension between steps 6 and 7. However, when leaving this specific pair out, there was no difference in average discriminability for the dimensions in the F0-VQ array.

Relative salience of training dimensions is discussed in more detail in the following chapter, in the context of both Experiment 1 and 2.

Finally, the manipulation of voice quality dimension, as done in Experiment 2, may not have been as successful as the manipulation of F0 and VOT. Manipulating voice quality is
difficult because the acoustic correlates of voice quality are somewhat unclear. As discussed on page 95, the perception of voice quality may be due to a bundle of related features. It is possible the vowel continuum which constituted the VQ dimension did not contain the sufficient acoustic cues for the perception of different voice qualities, whatever they may be. At the same time, the acoustic properties of F0 and VOT are more directly linked to the perceptions of pitch and voicing, so manipulating those properties may have had a more direct affect on perception. Furthermore, the manipulation for the VQ dimension was more cumbersome compared to the manipulation of either F0 or VOT, which was fairly straightforward. Given the difficulty of this manipulation, it is possible that the manipulation itself was less successful.

### 3.6.5 L1-L2 relationships and phonetic training

The main prediction from Experiment 1 was that training with the “same” dimension (i.e. a dimension that is used as a primary cue for target and native contrasts, with different boundaries) would be the least effective. This prediction was not borne out in Experiment 2 to a large degree. VOT had a similar relationship in Korean and English as pitch did in regards to Xorx and Bjaran, in that both dimensions were the “category relevant” dimensions in L1 and L2. In Experiment 1, pitch was the least effective training dimension. However, in the context of training with Korean stops, VOT was the most effective training dimension. At the same time, it was predicted that the “new” dimension of VQ would be the most effective training dimension because of a lack of L1 interference with voice quality. Results from Experiment 2, however, suggest that the “new” dimension was not effective.

Training with a “familiar” dimension, F0, was effective to some extent because it helped subjects separate two of the categories (weak and aspirated) that were likely both heard as exemplars of English /k/ prior to training. In this way, F0 can perhaps be thought of as the “best” training dimension, because it helped English speakers to overcome their “native” habits in regards to collapsing the weak and aspirated categories together. However,
in the context of Korean stops, which do seem to benefit from a two dimensional solution to optimally separate the three categories, training with F0 alone was not enough to distinguish the three categories, and poor performance on the identification task for this condition indicates that trying to use F0 alone for the three-way stop contrast ended up confusing the subjects rather than helping them.

The next chapter will revisit the results from both Experiment 1 and 2 with respect to the main research questions. This is followed by a discussion of the theoretical implications of this work, with a focus on how how understanding the relationship between a learner’s first language and the target language can inform strategies for phonetic training.
The goal of this dissertation research was to find effective methods for training adult learners to hear sounds in a new language. The experiments presented here explore how phonetic training can improve non-native speech perception. This method helps learners focus on underlying acoustic cues through a process of high variability feedback training. The approach taken is informed by several models of non-native speech perception in addition to the attention-to-dimensions (A2D) model of category learning, which conceptualizes non-native speech perception as a process of category learning via attention to acoustic dimensions that underlie phonemic contrasts. Phonetic training can help a learner to focus on the relevant acoustic dimensions for learning a new contrast. However, a crucial question raised by the A2D approach is which acoustic dimensions are most relevant for training adults to hear sounds in a new language, given the first language of the learner?

Unlike other phonetic training studies, the research presented here did not begin with the assumption that phonetic training strategies should mirror the perceptual strategies of native speakers. Rather, a main focus of the study was to consider native language experience and the relationship between the native language of the learner and the target language in order to choose effective training dimensions.
4.1 Research questions, revisited

Previously, three research questions were posed which were addressed by two experiments:

**Research Question 1** Will phonetic training result in better perception of non-native sounds?

**Research Question 2**

2A Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others?

2B Given this L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language?

In the following sections, these research questions are revisited with respect to the results from Experiments 1 & 2.

4.2 Will perceptual training result in better perception of non-native sounds?

The type of perceptual training implemented in both experiments has the potential to improve perception of non-native speech. The “alien words” from Experiment 1 aren’t non-native speech sounds *per se*, but to the extent that subjects learned to distinguish novel categories of pseudo-speech stimuli, the perceptual training was successful. In both experiments, subjects learned to utilize specific acoustic dimensions to distinguish between categories of sounds.

The method of perceptual training implemented in the two experiments did not involve explicit instructions about which acoustic dimensions were important for categorization.
Rather, subjects learned implicitly which acoustic dimensions were category relevant and which were irrelevant, through a process of category identification with feedback. The results from Experiment 2 indicate that trained knowledge was transferable to some degree: subjects used what they learned during training as a basis for making similarity ratings and to identify Korean stops in the post-training tasks, which involved a different set of stimuli from those used in training.

4.3 Does experience with the acoustic dimensions in an L1 result in some training dimensions being more effective than others? Given a learner’s L1 experience, which training dimensions are most effective for learning to discriminate sounds in a new language?

In experiment 2, there were differences among the training conditions in terms of the ability of subjects to identify and distinguish the Korean stop categories. This result suggests that L1 experience does result in some training dimensions being more effective than others, because the main difference between the experimental conditions in Experiment 2 was the particular relationship between the training dimension and English.

Based on the results from Experiment 2, given the L1 experience of a learner, effective training dimensions are:

(1) **The acoustic dimensions learners are already familiar with from their native language.**

   1A Salient dimensions

   1B Verbalizable dimensions

(2) **The acoustic dimensions that native speakers of the target language use.**

   2A Strongly contrastive dimensions

It is important to note here that the ancillary features listed above (strongly contrastive, salient, verbalizable) were not strictly tested for in experiments 1 and 2. Rather,
they are hypotheses, consistent with the data from the experiments, about why (1) and 
(2) are effective. In a sense, (1A), (1B), and (2A) are *outcomes* of (1) and (2). In other 
words, acoustic dimensions that are familiar because of native language experience will also 
be salient and verbalizable. A dimension that is strongly contrastive for the target sounds 
will likely be one that native speakers use.

The features of effective training dimensions are considered in further detail below. 
While (1) and (2) may, at the outset, appear to be in conflict, they need not be. For instance, 
it is possible that within the set of dimensions that native speakers use to distinguish a target 
contrast, some of those dimensions are also familiar to non-native learners.

### 4.3.1 The acoustic dimensions learners are already familiar with from their 
L1

It was hypothesized that phonetic dimensions already used as a primary cue to a 
similar native contrast might cause too much L1 interference, and thus be less effective 
when used for phonetic training. Contrary to this hypothesis, results from Experiment 2 
suggest that this kind of cue can be quite effective. In Experiment 2, subjects trained with 
a dimension that they already used to make stop distinctions in their own language (VOT) 
was the most successful in terms of the subjects’ accuracy during training and in the post-
training identification task. Because English speakers are accustomed to relying on VOT as 
a contrastive cue for stops, they can learn to co-opt it for Korean stops fairly effectively. 
The costs associated with moving a boundary along a learner’s “L1 dimension”, or creating 
a new boundary (as in the case of learning a three-way Korean stop contrast), did not seem 
to be as problematic for subjects compared to the difficulties faced by subjects in the other 
two conditions, particularly the VQ condition, where subjects had to learn a completely new 
contrastive cue.

It seems that completely novel cues might be useful in cases of learning artificial sound 
categories, as in Experiment 1. In the case of artificial linguistic categories, learners don’t
have strong perceptual “habits” from native-language experience. It may be easier to learn an entirely new dimension of contrast rather than shift a boundary along a just-learned dimension. However, in the context of actual language learning, it is beneficial to leverage what you already know. Experience with a dimension that is common in both L1 and L2 has a bootstrapping effect, particularly for short term phonetic training as implemented in Experiment 2. Learners may experience some L1 interference, but they can learn to use old dimensions in new ways. In the context of language learning, a flexibility with “known” dimensions wins out over the cost associated with having to learn an entirely new dimension of contrast.

An effective training dimension can also be one that learners have some prior experience with from their native language, though not necessarily as a primary contrastive cue for L1 sound categories. In Experiment 2, there was some evidence that training with F0, a “familiar” dimension, was helping subjects to distinguish weak and aspirated stops. Because F0 is a cue that is familiar to learners, even potentially used in the same contrastive context (i.e. as a secondary cue to stop voicing in English), it was somewhat effective to use as a training dimension (though training with F0 in isolation was probably harmful in the context of the “three way” Korean stop contrast).

4.3.1.1 Salient dimensions

Though Experiments 1 and 2 were not designed to directly test salience as a feature of effective training dimensions, the ability of a dimension to “stand out” compared to others may determine its effectiveness as a contrastive cue.

A dimension can be “salient” in two ways. First, it can be acoustically salient: robust, loud, or resilient to noise or degradation. In short, it can be “easy to hear”, from an acoustic/auditory standpoint. “Acoustic salience” is important for effective speech perception and may determine patterns of phonological acquisition, both for infants acquiring a first language (Narayan et al., 2010; Aslin and Pisoni, 1980) and for adults learning non-native
contrasts (Polka, 1991; Pruitt, 2004).

Acoustic salience may have played a role in the outcomes of both experiments. In Experiment 1, subjects in one Bjaran condition were trained with duration, which may not have been as acoustically salient as the relevant amplitude dimension that subjects were trained to ignore. The imbalance in acoustic salience could have affected the way subjects learned the Bjaran categories in this condition. Indeed, the scaling results from Experiment 1 indicated that subjects may have been attending to relative amplitude rather than duration. Also, in Experiment 1, the “pitch rise” dimension could have been more acoustically salient than the other Bjaran training dimensions. The acoustic salience of pitch rise could make it an effective training dimension, over and above its relationship to the dimensions in the “alien-L1”. In Experiment 2, it was difficult for subjects to focus on voice quality while ignoring the other varying dimensions of F0 and VOT: perhaps voice quality was not as acoustically salient as the other dimensions.

Another way in which a dimension can be salient is that it is “easy to perceive”, by virtue of native language experience. Several models of non-native speech perception, including the Attentions-to-Dimensions and Native Language Magnet theories, propose that “perceptual warping”, whereby some acoustic dimensions become more salient than others, is a consequence of experience with native language sound categories. This type of “perceptual salience” differs from “acoustic salience”, and may affect learning in different ways (Narayan et al., 2010). In Experiment 2, VOT was a salient dimension precisely because English speakers have so much experience using it as a contrastive cue for stop distinctions.

It seems that “perceptual salience” is quite important in the context of phonetic training with human languages. It is possible that any unfamiliar dimension, even a very “acoustically salient” one, will be less effective as a contrastive cue compared to a dimension that learners have some experience with. However, the experiments presented here do not directly address how acoustic salience interacts with perceptual salience due to native language experience. More work should be done to explore the interaction of acoustic and perceptual
salience in the context of phonetic training.

4.3.1.2 Verbalizable dimensions

One reason why native language dimensions may be helpful, particularly for the kind of short-term phonetic training implemented here, is that they are dimensions that learners can easily characterize. In other words, they are verbalizable. In cases of implicit learning, learners can internalize “rules” for a certain dimension so long as they can characterize that dimension. A dimension that is not verbalizable may be harder to use, especially in the short term. Over longer periods of training, learners may develop a sufficient characterization of a dimension which will allow them to use it as an effective contrastive cue. However, in cases where a more verbalizable dimension co-exists with a less verbalizable one, learners might gravitate toward the former.

In Experiment 2, subjects may have been able to use F0 effectively to some extent because it was a cue that learners could characterize. They recognized the fact that “pitch” was category-relevant, so were able to attend to it more easily compared to, e.g. voice quality, which might be less verbalizable. In Experiment 2, there were several subjects in the VQ condition, who, in the exit survey, described the stimuli with terminology that could be interpreted as naive descriptions of voice quality. However, relatively few subjects made these observations, and there was no overlap among the “naive voice quality descriptions”, suggesting that even if subjects were attending to voice quality, they had a hard time characterizing it. As in the case of acoustic vs. perceptual salience, the idea that verbalizable dimensions may be more effective in phonetic training was not directly tested, and is a promising direction for future work.

4.3.2 The acoustic dimensions that native speakers use.

The results from Experiment 2 demonstrated that a two-dimensional solution was optimal for separating the three Korean stop categories. There is evidence from previous
work (e.g. Francis and Nusbaum, 2002) that a two-dimensional solution is probably the perceptual strategy that Korean speakers use to distinguish the three stop types. There is not reason to assume that English speakers wouldn’t benefit from using a similar strategy, so it is useful to approach the particular problem of Korean stops with the “native strategy” in mind.

Of course, it is not the case that English speakers are going to learn Korean stops in the exact same way that Korean speakers do. Because English speakers have strongly ingrained perceptual categories based on English sounds, non-native phonemes have to “fit” into this system, at least at first. Knowing which contrastive dimensions a Korean speaker uses in order to perceive the Korean stop contrast is useful, but it is also important to consider that native English speakers might use the same dimensions in different ways. For instance, it is clear that Koreans use F0 to help them separate weak and aspirated stops. Nonetheless, English speakers will likely use VOT to distinguish all three Korean stops at first, mapping them to English /k/ and /g/. Therefore, the phonetic training strategy should focus on training English speakers to use F0 as a separate cue. This strategy will help them “break apart” the English /k/ category to accommodate Korean weak and aspirated stops. In this way, the “native strategy” can inform, rather than determine the phonetic training strategy.

4.3.2.1 Dimensions that are strongly contrastive

In order to be effective, a training dimension needs to be strongly contrastive for the target sounds. In other words, sound categories should not overlap in values along a contrastive dimension, because this effectively weakens the contrast between them. For the test stimuli in Experiment 2, there was some overlap between aspirated and weak stops in terms of voice quality/spectral tilt. Even if subjects had been completely successful in learning to use VQ as a contrastive cue to distinguish the training stimuli (and there isn’t much evidence to support this claim), it is unclear if they would have been able to use VQ to separate weak and aspirated stops in the test stimuli.
The dimensions that native speakers use as contrastive cues tend to be strongly contrastive. For instance, in the context of Korean, we know that native Korean speakers likely attend to VOT and F0, and though they may attend to voice quality somewhat (either as a part of a combined cue with VOT or to different degrees in certain dialects), VOT and F0 both provide a strong contrastive cue for the Korean stops, particularly when used in conjunction with each other. One of the reasons why “native speaker dimensions” make for effective training is because they provide “good” contrast.

4.3.3 Are the predictions of Experiment 1, based on artificial languages, borne out in cases of training with real languages? (RQ 3)

The predictions of Experiment 1 were that “new” and “familiar” dimensions might be more effective as training dimensions, because they don’t interfere with native language categories in the way that shifting boundaries along a “same” dimension does. These predictions were based on results from Experiment 1, which showed that there was a cost associated with having to create new boundaries along a dimension that was previously used as a category relevant dimension in the “alien-L1”. When a “new dimension” was available, learners were able to utilize it more successfully as a contrastive cue.

In Experiment 2, the “new” dimension of voice quality was not successful. In fact, the “same” dimension in Experiment 2 was actually the most successful. Taken together, these results show that the predictions from Experiment 1 were not borne out in cases of training with real languages.

However, the effect from Experiment 1 regarding interference with a “same” dimension was not entirely absent from the results of Experiment 2. Results from the scaling analysis suggest that the two non-native categories of weak and aspirated Korean stops were at first mapped to a single English category (/k/). Training with the “same” dimension of VOT caused subjects to further collapse the weak and aspirated categories together. These learners, in heavily weighting a “native” dimension, ended up mapping Korean sounds to
native categories, as predicted by several models of second language acquisition (PAM, SLM, and NLM). This result suggests that in the case of Korean, training with other dimensions in addition to VOT might be useful. As in Experiment 1, the “familiar” dimension was somewhat successful, in that it helped learners pull apart weak and aspirated stops.

One of the reasons why the predictions from Experiment 1 weren’t upheld to a strong degree in Experiment 2 is because the learning context for the two experiments is very different. In Experiment 1, subjects learned artificial sound categories. Though the stimuli were pseudo-speech with acoustic dimensions based on actual linguistic features, the learning context was qualitatively different from real language learning. In Experiment 1, the dimensions learned in the “alien L1” were recently learned (i.e. in the very same experimental session as the “alien L2”). In Experiment 2, some of the dimensions (particularly VOT) were deeply ingrained through 20+ years of speaking and hearing English sounds. In short, the “L1 experience” was much different in Experiment 1 versus Experiment 2. Perhaps it is the case that in cases of artificial category learning, learners have no strong “perceptual habits”, so learning a new dimension of contrast is easier compared to reconfiguring boundaries along a recently learned dimension. But in the case of language learning, native language experience strongly determines which acoustic dimensions are “important” for making phonetic contrasts, so it is harder to learn to attend to something new then to attend to a known dimension in a slightly different way.

The L1/L2 relationship for specific training dimensions was slightly different in Experiments 1 and 2. In Experiment 1, the “familiar” training condition was duration, which was present in the “alien L1” (as a category-irrelevant “noise” dimension that subjects were trained to ignore). In Experiment 2, the “familiar” dimension was one that subjects probably use in their first language, even in the context of making stop distinctions (i.e. as a secondary cue). In other words, F0 wasn’t a cue that listeners actively ignored in their L1, which made it perceptually “available” to them in a way that duration perhaps wasn’t for learners of Bjaran.
Finally, one reason that training with the “new” dimension of voice quality was unsuccessful may be that the manipulation of voice quality, as implemented in experiment 2, was not as directly related to to the contrastive cues for voice quality types.

Ultimately, the specific context of language learning has specific requirements in terms of phonetic training strategies. The difference in experimental contexts is one clear reason why the predictions from Experiment 1 weren’t borne out in Experiment 2. Despite this fact, Experiment 1 was useful, particularly for piloting training methodologies as well as developing the analysis techniques using MDS to explore specific category learning processes.

### 4.4 Theoretical implications

The research presented here adds to the body of work showing how native language experience affects the perception of non-native sounds. This dissertation research adds the idea that the details of L1 phonetic experience should be carefully considered when deciding on a phonetic training strategy for learning a non-native contrast. It’s important to understand what learners already know, and the way in which that knowledge will affect their perception of the target sounds. If learners are likely to have L1 interference, and the patterns of this interference are predictable in terms of assimilation to native categories, it may be possible to find specific cues that can help learners break their native language “habits”. Training cues are most effective if they are familiar (and perhaps acoustically salient and verbalizable). A good training strategy should also consider how the native speakers distinguish sound categories in the target language, because the “native” contrastive cues are likely strongly contrastive—but the “native strategy” shouldn’t be the entire focus of a training program. It is important to take into account both L1 experience and L2 native strategies.

#### 4.4.1 A “dimensional approach” to the development of new sound categories

The idea that L1 experience will affect a learner’s ability to utilize a training dimension, which motivated the experiments presented here, was inspired by three models of non-native
speech perception: the Perceptual Assimilation Model (PAM), the Native-Language Magnet Model (NLM), and the Speech Learning Model (SLM). Several of the results from Experiment 2 fit with the explicit predictions of these models. All three models predict that novice learners will at first map new sound categories to native phoneme categories. This mapping process can cause interference when trying to hear new sounds, because often learners don’t create new sound categories. Instead, they assimilate non-native sounds to native categories. All three models suggest that the strength of this assimilation depends on the degree of similarity between non-native sounds and native sounds. The results from Experiment 2 align well with these predictions. The results of Experiment 2 also adds an important consideration to theories of non-native speech perception based on similarity, which is that the equivalence of L1 and L2 sound categories, i.e. their degree of similarity, can depend on which acoustic dimension is being used as a basis for comparison.

The results from Experiment 2 are particularly well-suited to an analysis based on the PAM model. According to PAM, if two non-native sounds span a native boundary, they will be fairly easy to perceive by non-native speakers; PAM calls this kind of contrast a “Two Category” (TC) contrast. In a TC contrast, the non-native sounds map well onto two distinct native phonemes, so non-native speakers can easily discriminate them. Based on the results of Experiment 2, the contrast between Korean strong and aspirated stops seems to function as a TC contrast: /k*/ mapped well to /g/ for English speakers while /kʰ/ mapped well to English /k/. Strong and aspirated stops span a native English VOT boundary that separates English /k/ and /g/, so when subjects focused on VOT, they were quite good at identifying these stops and detecting the contrast between them. Weak and aspirated Korean stops, however, seem to fit a different contrastive pattern for English speakers, where /kʰ/ is heard as a fairly “good” version of English /k/ (because in initial position in English, /k/ is allophonically aspirated), while /k/ (weak) is heard as a “poor” exemplar of the English /k/ category. This kind of contrast, in the PAM model, is known as a “Category Goodness” (CG) contrast. In a CG contrast, one member of the contrastive pair is more similar to the
prototype for the native category. PAM predicts that discrimination and identification for a CG contrast is not as good as for a TC contrast. The post-training identification results from Experiment 2, particularly for the VOT training condition, suggest that strong and aspirated stops (a TC contrast) were easier for learners to identify compared to the weak stops. Furthermore, because the weak stop was similar to the aspirated stop in terms of VOT (and they both lie within the range of the /k/ phoneme for English speakers), when learners were focused on VOT, the contrast between weak and aspirated was weakened, drawing the two categories even closer together.

A compelling outcome from Experiment 2 in the context of PAM, and other models that predict assimilation patterns as a function of similarity, is the idea that the degree of similarity between sound categories can depend on which acoustic dimension underlies the equivalence. Shifting attention from a “native dimension” to a different acoustic dimension can actually change the nature of the non-native contrast, which can impact the ability of non-native speakers to detect it. For instance, while Korean weak and aspirated stops are more similar to each other in terms of VOT, they are more different in terms of F0. So when subjects in Experiment 2 focused on F0, they could detect the contrast between weak and aspirated stops better (as evidenced by AD between weak and aspirated stops in similarity ratings, as well as higher identification accuracy for these two stops compared to the strong stops). In the terminology of the PAM model, weak and aspirated stops form a CG contrast, as long as the basis for similarity is VOT. The model predicts that such a contrast will be more difficult for non-native speakers to detect, compared to a TC contrast where non-native sounds are mapped onto native categories. If the basis for similarity is instead F0, this makes the contrast between weak and aspirated stops more detectable. A “newly detectable” contrast is one that the current PAM model doesn’t account for. To the extent that learners can shift their attention to a “more contrastive” dimension, if it exists, non-native sounds that were once heard as better or worse exemplars of a single native sound (a CG contrast) can be “broken apart” into two new categories. The closest thing in the PAM model to this
process is the so-called “NA” contrast, where a contrast bears no “detectable similarity to any native phoneme” and are perceived as “non-speech”. However, it is not the case that a “newly detected” contrast is heard as “non-speech”; rather, the process of focusing on a different dimension of similarity to help detect contrast essentially results in the creation of a new contrast, and two new speech categories.

This “dimensional approach” to contrast detection bears some resemblance to the native language magnet model, which claims that native-language assimilation is based on the shared acoustic dimensions that underlie both the target non-native sounds and the native sounds to which non-native sounds are mapped. The main focus of NLM is on how well a novel sound category will be assimilated, rather than the creation of a new sound category. Like PAM, NLM doesn’t account for the creation of new sound categories through a process of shifting attention to a dimension that can provide a contrast between sounds that were, at first, assimilated to the same native category.

PAM, NLM and SLM (though it is focused on production) all agree that “dissimilarity” is the key to forming novel sound categories. What these models are lacking is the idea that similarity can depend on which dimensions (or features or cues) used as a basis for comparison. The idea of a contrast being more or less detectable depending on the dimension of contrast has important implications for the development of new sound categories and the degree to which a novel contrast is learnable.

The extension of the A2D model to speech perception comes closest to integrating a dimensional approach with approaches based on similarity and assimilation. The A2D approach focuses on training learners with the “right” dimensions to allow for the development of new sound categories. However, the development of a new sound category (or detection of a non-native contrast) doesn’t always have to follow the trajectory of native speakers who learn the target language from birth, which is an assumption (though perhaps implicit) of the A2D approach to speech perception. The “right” dimensions for learning a new contrast are whatever dimensions help people learn a contrast. What those helpful dimensions are should
be determined empirically, based on the L1 experience of the learner and the characteristics of the L2 sounds.

4.5 Methodological contributions for phonetic training studies

A novel methodological contribution of Experiments 1 and 2 was the addition of a between-groups manipulation involving training dimensions that each had a different status in the L1 (or quasi-L1) of the learners. Some phonetic training studies have compared different types of training techniques, e.g. high variability training with feedback or explicit vs implicit training but few (if any) thus far have considered how training with different types of dimensions will affect learning.

Another contribution of this research is the use of INDSCAL as a technique to understand the nature of category learning in the context of phonetic training. Some phonetic training studies have used MDS and/or INDSCAL (Francis and Nusbaum, 2002; Guion and Pederson, 2007; Chandrasekaran et al., 2010), but often these analyses end with a qualitative interpretation of the scaling plots (and perhaps an attempt at rigorous interpretation of the scaling dimensions). To the author’s knowledge, the analysis presented here is the first time INDSCAL results have been analyzed in terms of the category-learning processes of acquired similarity and acquired distinctiveness. Seen through the framework of the A2D model of category learning, the analysis of INDSCAL results can yield insights into how people learn perceptual categories, and how those categories may be related to the physical properties of sensory input.

4.6 Future work and applications

4.6.1 Subjects with different native languages

To further test the finding from Experiment 2 that an effective contrastive cue is one that is familiar to learners would be to repeat Experiment 2 with subjects who have an L1
that is NOT English. For instance, if subjects do have phonemic contrasts based on voice quality in their native language, they might be better at using it as a contrastive cue for Korean stops. Training subjects who have a voice quality contrast in their L1 could also indicate whether or not the voice quality manipulation in Experiment 2 served to highlight the acoustic properties that underlie voice quality contrasts. Finally, training subjects with other L1s could also help to ascertain how much the “native strategy” actually matters. In other words, does the appropriate training cue depend *more* on the L1 or the L2? If, in several cases of training with different L1s, it became apparent that VOT and F0 were *always* the most effective training dimensions for Korean stops, this would suggest that the “native strategy” is most important.

### 4.6.2 Time course of training

The time course of training in the experiments presented here was quite short, around two hours for all components (including pre- and post-testing). Perhaps familiar dimensions are effective in short-term training contexts, because of their bootstrapping effect: people can effectively use a dimension that they already know in a new contrastive context. However, long term training might show that new dimensions *can* be helpful, but it may take time for subjects to “tune in” to them.

A longer term study would also help mitigate training fatigue. Even though both Experiment 1 and 2 involved a short training period overall, and the subjects were allowed breaks throughout training, the two hour phonetic training experiments may have been perceptually taxing.

### 4.6.3 Motivation and learning context

In Experiment 2, subjects were, on the whole, not very motivated to learn Korean. Of course, all studies which make use of university subject pools run the risk of having subjects that are more or less motivated to “do well” in a study. However, motivation to “succeed”
at the experimental task may not influence outcomes to the a large degree in some social science experiments, compared to a language learning experiment. When it comes to learning a language, motivation can greatly affect outcomes. Results from Experiment 2 might have differed if subjects were already motivated to learn the Korean stops, for instance if they were students in a Korean language class. Training with more motivated learners would also provide an opportunity to compare how training a true novice (someone with no experience in the target language) might be different than training a “beginner” who has some experience with the target language.

4.6.4 Type of training

In Experiment 1 and 2, the training methodology involved implicit instruction. Subjects were not told to “pay attention to $X$”. Rather, through a process of feedback, they were implicitly “told” which dimensions were category relevant. One potential follow-up experiment would be to compare implicit training with a type of training that involved explicit instruction. It may be that if subjects are explicitly told what to listen for, they would be very good at perceiving a novel contrast. Of course, the outcome might depend on the type of dimension learners should listen for, and how “verbalizable” it is. F0, for instance, is easy to explain to English speakers (we can even use a more common word, “pitch”). VOT and voice quality, however, are harder to explain to someone with no background in phonetics—especially voice quality, the characterization of which is difficult even for trained phoneticians!

Again, L1 experience plays a role in characterizing training dimensions for explicit instruction. VOT can be discussed with English speakers indirectly, by making appeal to the native categories where VOT plays a contrastive role (e.g. “did you notice that the strong stops sounds like /g/?”). There is no such opportunity to refer to native categories in order to characterize a dimension such as voice quality, which plays little to no role in distinguishing native contrasts.
Another training strategy to explore would involve a combination of explicit and implicit training. For instance, at the beginning of training, subjects would undergo category identification with feedback (as in the training portions of Experiments 1 and 2). Then, after the subjects have some experience with implicit training, the experimenter could instruct subjects to listen for a specific cue. This kind of combination feedback-plus-explicit instruction has been shown to have promising results for learning difficult perceptual categories (see, e.g., Biederman and Shiffrar [1987]). In the specific context of training the Korean stops, a combination training strategy could proceed in this way: subjects first undergo category identification with feedback. Keeping in mind that subjects are likely to map the Korean stops onto English /k/ and /g/, after subjects have had some experience with the sounds, the experimenter could tell them to focus on pitch, specifically for the stimuli that “sound like /k/” to English ears.

4.6.5 Production

Perception and production are certainly related, and though speech perception precedes production in first language acquisition, the two reinforce each other during the process of language learning. A potential followup to this dissertation research would be to explore the effects of phonetic training on non-native production. The ability to attend to specific dimensions of contrast might enhance the production abilities of learners. If learners can hear contrasts better, it is reasonable to assume that they will produce the sounds better as well, because they can monitor their own productions and adjust them to match perception. Furthermore, a focus on specific acoustic features could help learners develop more robust representations of the articulatory gestures associated with a new speech sound.

4.6.6 Applications

Educational programs for second-language learners are sorely lacking in any kind of targeted phonetic training. The ability to effectively and efficiently hear small distinctions in
a language increases comprehension and likely has positive benefits for production as well. Effective comprehension and production both serve to increase a learner’s confidence, which leads to fuller engagement when learning a second language.

Furthermore, second language pedagogy, both in the classroom and in consumer-facing language learning programs, largely focuses on the target language without considering L1 influences or the contexts wherein learners might experience L1 interference. The research presented here shows that an effective training strategy should take the specific L1 experience of the learner into account.

Though one obvious application for the research presented here is in second language education, some of the training methodologies in this research could be adapted for use in clinical settings. For instance, the phonetic training methodologies developed in this dissertation research may be helpful for children who have phonological impairments. There is evidence to suggest that perceptual training, in some form, is beneficial as part of a treatment plan for phonological impairment (Rvachew, 1994; Rvachew et al., 2004). The kind of high variability training implemented here, which helps learners to focus on a specific cue, might prove highly effective as part of an intervention program for phonological impairment.

4.7 Conclusion

The research presented in this dissertation contributes to the field of non-native speech perception in several ways. First, we stand to gain a fuller understanding of how the relationship between L1 and a new language affects learning and perception. Second, the research adds to theories of non-native speech perception, by showing how the development of new sound categories can be based on different dimensions of similarity. The methodological techniques in this research can be extended to other studies that use phonetic training in order to assess the effectiveness of different training dimensions and techniques. The use of multi-dimensional scaling in this study was a novel approach to exploring several category learning processes as conceptualized in the Attention-to-dimensions model of category learning. One
of the strengths of the A2D model, as applied to speech perception, is that the learning mechanisms are thought to be domain-general. Therefore, this research adds to the body of work which investigates the relationship between category learning and perception. In terms of applications, this dissertation research could lead to the implementation of effective phonetic training programs to be utilized in educational and clinical contexts.

This research showed that to effectively learn to hear sounds in a new language, it is important to consider both where you’re going (that is, the “native” perceptual strategies for target sounds) and where you’re coming from (L1 experience). Most phonetic training studies to date focus on the former, but not the latter. This work suggests that it’s particularly important to know where a learner is “coming from” in order to develop effective training strategies, because L1 experience has such a strong effect on the perceptual abilities of learners.
Bibliography


Kim, M. (2004). Correlation between vot and f0 in the perception of korean stops and affricates. In INTERSPEECH.


Test stimuli sets and recording contexts

Test stimuli for Experiment 2 (used in the pre- and post-training difference rating tasks and in the post-training identification task) were recorded in a number of different contexts. The carrier phrase “ige [blank] hago kata” translates to *the word is [blank]*. Citation form refers to slow, careful articulation, “as if speaking to someone learning Korean”. Conversation form is speaking quickly, “as if speaking to a friend”.

<table>
<thead>
<tr>
<th>Carrier Phrase</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>kⁿa¹</td>
<td>ige _____ hago kata</td>
</tr>
<tr>
<td>k* a¹</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ¹</td>
<td>___________</td>
</tr>
<tr>
<td>kʰa²</td>
<td>ige _____ hago kata</td>
</tr>
<tr>
<td>k* a²</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ²</td>
<td>___________</td>
</tr>
<tr>
<td>kʰa³</td>
<td>ige _____ hago kata</td>
</tr>
<tr>
<td>k* a³</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ³</td>
<td>___________</td>
</tr>
<tr>
<td>kʰa⁴</td>
<td>ige _____ hago kata</td>
</tr>
<tr>
<td>k* a⁴</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ⁴</td>
<td>___________</td>
</tr>
<tr>
<td>kʰa⁵</td>
<td>(none)</td>
</tr>
<tr>
<td>k* a⁵</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ⁵</td>
<td>___________</td>
</tr>
<tr>
<td>kʰa⁶</td>
<td>(none)</td>
</tr>
<tr>
<td>k* a⁶</td>
<td>___________</td>
</tr>
<tr>
<td>kᵃ⁶</td>
<td>___________</td>
</tr>
</tbody>
</table>
Appendix B

Training stimuli: discrimination pilot experiment

A pilot experiment was run prior to Experiment 2, with a different set of subjects, in order to better understand the nature of the stimuli used for training. The pilot experiment involved a discrimination (AX) task, and was designed to investigate the perceptual scaling of the acoustic dimensions in the training stimuli.

B.1 Perceptual distance

The pilot experiment was designed to determine if the perceptual distance between steps along each dimension were roughly comparable. One way to assess if this is the case is with a discrimination task, where subjects discriminate between pairs of stimuli. A measure of discrimination (d-prime) can be derived for pairs of stimuli along a dimension. Peaks in d-prime along a single dimension suggest that there is a perceptual discontinuity along the dimension, which could indicate a boundary between perceptual categories (because discrimination is generally higher at boundaries between phonemes). One hypothesized effect of perceptual training, according to A2D theories, is to change the perceptual distance between categories along a particular dimension. Therefore, it’s important to know if there are any perceptual discontinuities along a dimension prior to training. These discontinuities could be due to prior language experience or natural discontinuities along a particular dimension, as these have been shown to exist for perceptual continua such as VOT (see, e.g. [Kuhl and Miller 1975] for an investigation of VOT discrimination by chinchillas) and color [Witzel and ...]
It is important to see if there are discontinuities prior to training so we can see if and how these pre-existing discontinuities may affect training.

Another goal of the pilot study was to determine if the two dimensions in a single training array (e.g. F0 and VOT in the F0-VOT array) were equally discriminable, i.e. if one dimension “stood out” more than the other dimension. If one dimension is more discriminable, overall, this could affect the ability of subjects to attend to this dimension. This could result in an advantage for subjects who are trained with this dimension, and a disadvantage for subjects who are supposed to ignore this dimension. To investigate overall discriminability of dimensions in a single stimulus array, the average d-prime for a dimension can be compared to the average d-prime for the other dimension. If one dimension has a higher overall d-prime, this might make variations along that dimension harder to ignore.

### B.1.1 Stimuli

Stimuli for the pilot experiment were subsets of the training arrays created for experiment 2 (F0-VOT, F0-VQ, VQ-VOT), and consisted of the middle row and column from each array. For example, for testing the F0 dimension in the F0-VOT array, pilot stimuli consisted of F0 steps 1-9 at VOT step 5. For testing the VOT dimension in this same array, stimuli consisted of VOT steps 1-9 at F0 step 5. See figure B.1 for a schematic showing the subset of the F0-VOT array used in the pilot experiment. The same subsets (middle row and middle column) from the other arrays were also tested in the pilot experiment. Each array was tested by a separate set of subjects.

### B.1.2 Design

The discrimination task followed an “AX” design, where subjects heard pairs of adjacent stimuli along a single dimension, and indicated if those stimuli were the same or different. For each dimension, there were 8 adjacent pairs (Step 1 vs Step 2, Step 2 vs Step 3, Step 3 vs Step 4, etc.).
Figure B.1: Stimuli used in the pilot experiment (red boxes). Figure shows F0-VOT array; the same middle column and row tested for the other two stimuli arrays (F0-VQ, VQ-VOT).

For each stimulus pair, there were four trial types: AA, AB, BA, and BB. In an AA trial type, subjects heard the same stimulus played twice (e.g. F0-Step 1, F0-Step 1). In an AB trial type, subjects heard one stimulus followed by the next adjacent stimulus (e.g. F0-Step 1, F0-Step 2). In a BA trial type, the stimulus presentation was reversed (F0-Step 2, F0-Step 1). Finally, a BB trial type consisted of two presentations of the second stimulus (F0-Step 2, F0-Step 2). All four trial types were repeated 8 times for each stimulus pair. Thus, for a single dimension, there were (8 pairs) x (4 trial types) x (8 repetitions) = 256 trials.

Subjects were tested on both dimensions of a stimulus array. The order of dimensions was counterbalanced such that half of the subjects were tested on one dimension first, while half the subjects were tested on the other dimension first. The entire discrimination task consisted of 512 trials (256 trials x 2 dimensions).

Each stimulus array was tested separately, i.e. the F0-VQ array was tested with one group of subjects, the F0-VOT array was tested with another group, and the VQ-VOT array was tested with a third group. There were 16 subjects in each group.
B.1.3 Procedure

The experiment was presented using Psychopy. Subjects used an ioLabs button box to enter their response. Stimuli were presented through headphones at a comfortable listening volume, in the sound booth of the CU phonetics lab.

The instructions for the perceptual discrimination experiment were as follows:

*On each trial, you will hear two sounds play. After the sounds play, decide if the sounds are the SAME or DIFFERENT. Use your DOMINANT HAND to press the YELLOW button ("SAME") or the green button ("DIFFERENT").*

Before beginning the trials, subjects were told that there was an equal number of SAME and DIFFERENT trials in the experiment.

On each test trial, subjects heard a pair of stimuli play. Then, subjects decided if the stimuli were the same or different, and indicated their choice by pressing the yellow or green button, respectively. After a choice was made, there was an ISI of 500 ms before the next pair played. Feedback was provided in the form of an accuracy tracker at the top of the screen, which displayed a running total of the percent correct for that block of trials (i.e. the total percent correct for a single dimension). When discrimination is difficult, feedback can mitigate the tendencies for subjects to switch perceptual strategies over the course of a discrimination task (McGuire, 2010).

Before beginning the test trials, subjects were presented with a block of practice trials. Practice trials consisted of a single repetition of all 4 trial types (AA, AB, BA, BB) for three pairs of the first dimension being tested (pair 1, pair 4, and pair 8), for a total of 12 practice trials.

B.1.4 Subjects

For each of the three stimulus arrays, 16 subjects were tested, for a total of 45 subjects. All subjects were undergraduate students at CU Boulder, and were drawn from the
psychology and linguistics subjects pools. Subjects were native English speakers, and most were monolinguals with no significant experience in another language. One subject had learned French in school from the age of 6, and the other had learned Russian at home from birth, but was not fluent in Russian anymore. All subjects had normal hearing (by self report).

B.1.5 Results

For the pilot study, stimuli consisted of a single row and column from a training array. Each array was tested separately by a different group of subjects. Each subject was tested with two dimensions (e.g. F0 and VOT from the F0-VOT array), and there were eight stimuli pairs along each dimension (e.g. F0 step 1 vs F0 step 2).

A measure of adjusted d-prime was calculated for each stimulus pair. This measure of adjusted d-prime allows for data from an AX discrimination paradigm (which involves four trial types, AA, AB, BA, and BA) to be interpreted as data from a “same-different” paradigm (A vs B), allowing for an interpretation of d-prime as a measure of discrimination between A and B (Macmillan and Creelman, 2004). Adjusted d-prime is calculated using the following method (from (Macmillan and Creelman, 2004), p. 214-216): Where $HR =$ the subject’s hit rate for a stimulus pair, and $FAR =$ the subject’s false alarm rate for a stimulus pair:

$$d' = z(HR) - z(FAR)$$

Percent correct (PC) is given by:

$$PC = P(d'/2)^2 - 2 \ast P(d'/2) + 1$$

Finally, adjusted d-prime is given by:

$$d'(adjusted) = 2 \ast z(.5(1 + (2PC - 1)^5))$$
For simplicity, hereafter this adjusted d-prime will be referred to as simply “d-prime”.

Data were analyzed using a combination of repeated measures ANOVA models (using the R stats package) and mixed-effects regression models (using the R packages lme4 \cite{Bates2015} and lmerTest \cite{Kuznetsova2016}).

B.1.5.1 F0-VOT array

Figure B.2 shows the average d-prime (across subjects) for the F0 and VOT dimensions from the subset of the F0-VOT training array (the middle row and column of the array, respectively).

![D-prime for each dimension in F0-VOT array](image)

Figure B.2: D-prime for each dimension in F0-VOT array, averaged across subjects

There are no noticeable peaks in d-prime for any of the pair numbers along the F0 dimension. A repeated measures ANOVA was run for the F0 dimension, comparing the effect of pair number on d-prime, with pair number as a within-subjects factor. There was no effect of pair number ($p = .62$). In other words, there were no statistically significant “peaks” in d-prime along the F0 dimension for this array. This suggests that discriminability along this dimension is fairly even: there is no particular pair that elicits a higher d-prime,
compared to the other pairs.

For the VOT dimension, there was a significant effect of pair number. In a repeated measures ANOVA, comparing the effect of pair number on d-prime for the VOT dimension, with pair number as a within-subjects factor, pair number was significant $F(7, 105) = 2.84, p = .009$). In figure B.2 there is a peak in d-prime at pair two. To see if the mean d-prime of pair 2 was significantly higher than other pairs, a mixed effects regression model was run, predicting VOT d-prime with dummy coded categorical variables for pair number, treated as fixed effects, and random effects for subject. In this model, pair 2 was set as the reference level to be compared against every other pair. The d-prime of pair 2 was found to be significantly higher in every mean comparison, except when compared to pair 1 and pair 7 (where the mean comparison was only marginally significant). See table B.1 for the individual comparisons with pair 2.

Table B.1: Fixed effects from mixed-effects regression model for VOT, with pair 2 as the reference level

|           | Estimate | Std. Error | t value | Pr(>|t|)    |
|-----------|----------|------------|---------|-------------|
| (Intercept) | 0.69211  | 0.08035    | 8.614   | 2.38e-14 ***|
| pair 1    | -0.17305 | 0.11149    | -1.552  | 0.123457    |
| pair 3    | -0.23782 | 0.11149    | -2.133  | 0.035103 *  |
| pair 4    | -0.38818 | 0.11149    | -3.482  | 0.000712 ***|
| pair 5    | -0.33148 | 0.11149    | -2.973  | 0.003610 **  |
| pair 6    | -0.32702 | 0.11149    | -2.933  | 0.004071 **  |
| pair 7    | -0.20611 | 0.11149    | -1.849  | 0.067149 .   |
| pair 8    | -0.42538 | 0.11149    | -3.815  | 0.000223 ***|

Another way to visualize subjects’ sensitivity to pairs of stimuli is to plot the distribution of maximum d-primes as a function of pair number, which can show if there are certain pairs that consistently elicit the highest d-prime. Figure B.3 shows these distributions for the F0 and VOT dimensions in the F0-VOT array.

For the F0 dimension, the highest d-prime occurred most often for pairs 3 and 8. However, in each case, only 4 subjects had the highest d-prime for these pairs—less than a
third of the 15 total subjects, and by no means a majority. In the case of VOT, 5 people (a third of subjects) had the highest d-prime for pair number 2 (which is reflected in the peak in mean d-prime at pair 2). Again, by no means did a majority of people have the highest d-prime for pair 2, or any other pair.

Taken together, these results suggest that discriminability is fairly uniform for the F0 dimension, but that there is a peak in discriminability for the VOT dimension around pairs 1 and 2.

In order to test whether one dimension was more discriminable, overall, compared to the other dimension, the mean d-primes for each dimension were compared. Figure B.4 shows the mean d-prime for each dimension. A linear model was run on the d-prime data for the VOT and F0 dimensions, where the dependent variable was d-prime, and the independent variable was a categorical variable dimension variable (“F0 vs VOT”), which was contrast coded to compare the mean d-prime for F0 with that of VOT. In this model, the dimension variable was not significant ($p = .23$), meaning that there was no significant difference in mean d-prime for the F0 and VOT dimensions. This suggests that, overall (collapsing across
pair number), the VOT dimension is not more discriminable than that F0 dimension.

Figure B.4: Mean d-prime for F0 and VOT dimensions, F0-VOT array. Error bars represent $SEM_{betw}$.

B.1.5.2 F0-VQ array

Figure B.5 shows the average d-prime (across subjects) for the F0 and VQ dimensions in the F0-VOT training array (the middle row and column of the array, respectively.)

A repeated measures ANOVA was run on the d-prime data from the F0-VQ array, comparing the effect of pair number on d-prime for the F0 dimension, with pair number as a within-subjects factor. There was no effect of pair number ($p = .84$). In other words, there were no statistically significant “peaks” in d-prime along the F0 dimension for this array. As in the case of the F0 dimension in the F0-VOT array, discriminability along F0 in the F0-VQ dimension is similarly even: there is no particular pair that elicits a higher d-prime, compared to the other pairs.

There is, however, a noticeable peak in d-prime for the VQ dimension, at pair 6 (steps 6 and 7 along the VQ dimension). In a repeated measures ANOVA comparing the effect of
pair number on d-prime in the VQ dimension, with pair number as a within-subjects factor, there was a significant effect of pair number \((F(7, 105) = 9.42, p < .00001)\). To see if the mean d-prime of pair 6 was significantly higher than other pairs, a mixed effects regression model was run, predicting VQ d-prime with dummy coded categorical variables for pair number, treated as fixed effects, and random effects for subject. In this model, pair 6 was set as the reference level to be compared against every other pair. There was a significant effect for each mean comparison; in other words, d-prime for pair 6 was significantly higher in every pair-wise comparison (see table B.2).

Figure B.5 shows the distributions of the highest d-prime for each subject as a function of pair number. For the F0 dimension, this distribution is fairly flat: the highest number of maximum d-primes was for pair four, but this was only four subjects out of 15. For the VQ dimension, the majority of subjects (10/15) had the highest d-prime for pair 6.

Taken together, these results suggest that discriminability is fairly uniform for the F0 dimension. For the VQ dimension, there is a peak in discriminability for pair 6 for the majority of the subjects.
Table B.2: Fixed effects from mixed-effects regression model for VQ (F0-VQ array), with pair 6 as the reference level

|        | Estimate | Std. Error | t value | Pr(>|t|) |
|--------|----------|------------|---------|----------|
| (Intercept) | 1.14715  | 0.08617    | 13.313  | <2e-16 *** |
| pair 1   | -0.69819 | 0.12099    | -5.771  | 7.14e-08 *** |
| pair 2   | -0.74355 | 0.12099    | -6.145  | 1.25e-08 *** |
| pair 3   | -0.81394 | 0.12099    | -6.727  | 7.66e-10 *** |
| pair 4   | -0.63413 | 0.12099    | -5.241  | 7.59e-07 *** |
| pair 5   | -0.77258 | 0.12099    | -6.385  | 4.01e-09 *** |
| pair 7   | -0.81041 | 0.12099    | -6.698  | 8.83e-10 *** |
| pair 8   | -0.76605 | 0.12099    | -6.331  | 5.19e-09 *** |

Figure B.6: Maximum D-prime by pair (counts), F0-VQ array

Figure B.7 shows the mean d-prime for each dimension.

In order to test whether one dimension was more discriminable, overall, compared to the other dimension, the mean d-primes for each dimension were compared. A linear model was run on the d-prime data for the F0 and VQ dimensions from the F0-VQ array, where the dependent variable was d-prime, and the independent variable was a categorical variable dimension variable ("F0 vs VQ"), which was contrast coded to compare the mean d-prime for F0 with that of VQ. In this model, the dimension variable was significant, meaning that
the mean d-prime for the VQ dimension was significantly higher than for the F0 dimension ($F(1, 30) = 11.87, p = .002$). Overall (collapsing across pair number), the VQ dimension is more discriminable than that F0 dimension. This effect is most likely driven by the large peak in d-prime at pair 6. In fact, when this pair is left out of the analysis, there is no longer a significant difference between mean d-prime of F0 compared to VQ ($p = .17$).

Figure B.7: Mean d-prime for F0 and VQ dimensions, F0-VQ array. Error bars represent $SEM_{betw}$.

B.1.5.3 VQ-VOT array

Figure B.8 shows the average d-prime (across subjects) for the VQ and VOT dimensions in the VQ-VOT training array (the middle row and column of the array, respectively.)

A repeated measures ANOVA was run on the d-prime data from the VQ-VOT array, comparing the effect of pair number on d-prime for the VQ dimension, with pair number as a within-subjects factor. There was a significant pair number effect ($F(7, 105) = 8.36, p < .00001$). As in the F0-VQ array, the highest mean d-prime was for pair 6. To see if the mean d-prime of pair 6 was significantly higher than other pairs, a mixed effects regression model
A repeated measures ANOVA was run on the d-prime data from the VQ-VOT array, comparing the effect of pair number on d-prime for the VOT dimension, with pair number
as a within-subjects factor. In this model, there was no significant effect of pair number (repeated measures ANOVA, \( p = .15 \)), meaning there was no significant peak in d-prime for the VOT dimension.

Figure B.9 shows the distributions of the highest d-prime for each subject as a function of pair number. For the VOT dimension, a third of the subjects had the highest d-prime for pair 1. For the VQ dimension, the majority of subjects (11/15) had the highest d-prime for pair 6.

Taken together, these results suggest that discriminability is fairly uniform for the VOT dimension, but that there is a peak in discriminability for the VQ dimension around pair 6. Again, this is a strong effect; the majority of subjects had the highest d-prime for pair 6.

Figure B.9: Maximum D-prime by pair (counts), VQ-VOT array. Note that pair 7 is missing in both dimensions because in neither case did any subject have pair 7 as their maximum dprime.

Figure B.10 shows the mean d-prime for each dimension. A linear model was run on the d-prime data for the VQ and VOT dimensions from the VQ-VOT array, where the dependent variable was d-prime, and the independent variable was a categorical variable dimension variable (“VQ vs VOT”), which was contrast coded to compare the mean d-prime
for VQ with that of VOT. In this model, mean d-prime (collapsing across pair number) for the VQ dimension was not significantly different from the mean d-prime for the VOT dimension (contrast coded variable for dimension was N.S., \( p = .98 \)).

Figure B.10: Mean d-prime for VQ and VOT dimensions, VQ-VOT array. Error bars represent SEM_{betw}.

B.2 Discussion

This pilot study explored the perceptual scaling of the dimensions in the three stimulus arrays used for training in Experiment 2. A subset of each array (one row and one column) was tested. The three subsets were tested separately (16 subjects each).

B.2.1 Discontinuities along training dimensions

Discontinuities along particular training dimensions were observed for some of the dimensions (VOT in the subset of the F0-VOT array and VQ in both subsets). These discontinuities were indicated by significant peaks in d-prime at single pair numbers.
B.2.1.1  F0

There was no significant effect of pair number for the F0 dimension in either tested subset, and there was no pair number that for the majority of subjects elicited the highest d-prime. In both the subset of the F0-VOT array and the subset of the F0-VQ array, the perceptual distance between steps along the F0 dimension are, on average, fairly even.

B.2.1.2  VOT

For the VOT dimension, there was no significant effect of pair number in the subset of the VQ-VOT array. However, there was a significant effect of pair number for VOT in the F0-VOT array, with a peak in d-prime at pair 2. Pair 2 along the VOT dimension is comprised of stimuli with VOT values of 23 and 39 ms (see description of VOT steps below).

Step 1  7 ms
Step 2  23 ms
Step 3  39 ms
Step 4  55 ms
Step 5  71 ms
Step 6  87 ms
Step 7  103 ms
Step 8  119 ms
Step 9  135 ms

It is possible that the higher d-prime for between steps 2 and 3 along the VOT training dimension reflects a the natural discontinuity in VOT for English speakers between /g/ and /k/. However, the fact that only a third of subjects had the highest d-prime for this pair suggests that if there is a VOT boundary here for English speakers, this effect is not very consistent. Furthermore, there is no evidence that the same boundary exists along the VOT dimension for the subset of the other array.

Regarding the use of VOT as a training dimension in Experiment 2, steps 2 and 3 were both in the “strong” category for the VOT subjects. There is no discrimination data
for the training stimuli in Experiment 2. However, even if the subjects in the VOT group had shown evidence of having a perceptual discontinuity between these two steps in the F0-VOT array, there was no difference in accuracy between the two training arrays for the VOT condition. Thus, there is no evidence that a potential discontinuity in VOT affected the ability of subjects to learn the Korean stop categories during training in Experiment 2.

B.2.1.3 VQ

For the VQ dimension, there was a significant effect of pair number for both subsets of the two training arrays (F0-VQ and VQ-VOT). In both cases, the peak in d-prime occurred at pair 6. This was a fairly strong effect, because for both stimuli subsets, the majority of subjects had the highest d-prime for pair 6.

Pair 6 along the VQ dimension is comprised of two stimuli with blended vowels, steps 6 and 7 along the VQ array (see description of VQ steps below).

<table>
<thead>
<tr>
<th>Step</th>
<th>Percentage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100%</td>
<td>/<em>a/ /k</em>a/ / (the vowel from /k*a/)</td>
</tr>
<tr>
<td>2</td>
<td>80%</td>
<td>/*a/ , 20% /a/</td>
</tr>
<tr>
<td>3</td>
<td>60%</td>
<td>/*a/ , 40% /a/</td>
</tr>
<tr>
<td>4</td>
<td>40%</td>
<td>/*a/ , 60% /a/</td>
</tr>
<tr>
<td>5</td>
<td>100%</td>
<td>/a/</td>
</tr>
<tr>
<td>6</td>
<td>80%</td>
<td>/a/ , 20% /ʰa/</td>
</tr>
<tr>
<td>7</td>
<td>40%</td>
<td>/a/ , 60% /ʰa/</td>
</tr>
<tr>
<td>8</td>
<td>20%</td>
<td>/a/ , 80% /ʰa/</td>
</tr>
<tr>
<td>9</td>
<td>100%</td>
<td>/ʰa/</td>
</tr>
</tbody>
</table>

Step 6 had a vowel that was 80% “weak” and 20% “aspirated”. Step 7 had a vowel that was 40% “weak vowel” and 60% “aspirated vowel”. For both portions of the blended vowel, there is a 40% change from step 6 to step 7 (“weak vowel” goes from 80% to 40%, and “aspirated vowel” goes from 20% to 60%). It could be the 40% change for each portion, moving from step 6 to 7, is simply quite noticeable. However, pair 4 (steps 4 and 5 along the VQ continuum) also have a 40% change for each portion of the vowel, moving from one
step to the next (the “strong vowel” goes from 40% in step 4 to 0% in step 5, and the “weak vowel” goes from 60% in step 4 to 100% in step 5. But, there is no peak in d-prime for this pair. So, it is not necessarily the proportions of vowels that lead to this effect, but rather the addition of 40% more aspirated vowel that leads to the peak in d-prime.

Regarding the use of VQ as a training dimension in Experiment 2, steps 6 and 7 were in different training categories for the VQ subjects (“weak” and “aspirated”, respectively). Again, there is no discrimination data for the training stimuli in Experiment 2. However, even if subjects in Experiment 2 noticed a large discontinuity along the VQ dimension in the training stimuli, this did not seem to help them to attend to vocalic cues related to voice quality for the testing stimuli. Rather, they seemed to attend to F0.

B.2.2 Comparing discriminability of dimensions within a single array

In both the subsets containing the VOT dimension (F0-VOT and VQ-VOT), the overall d-prime of the two dimensions (average d-prime across pair numbers) was not statistically different. These results suggest that for subjects trained with these arrays, one dimension is not more discriminable than the other. In other words, when training with the F0-VOT array, neither the F0 or the VOT group has any “advantage” due to better overall discriminability along their training dimension. Likewise, when training with the VQ-VOT array, neither the VQ condition nor the VOT condition has an advantage.

In the F0-VQ array, overall d-prime of the VQ dimension was higher. This could lead to an unfair advantage for the VQ training condition for this array. However, the VQ subjects were also trained with the VQ-VOT array, so half the time they wouldn’t have had a training advantage. Furthermore, if the discontinuity along VQ had resulted in an advantage for the VQ subjects for identifying categories, then accuracy should have been higher for the F0-VQ array. However, in the post-hoc test for the VQ condition comparing the mean accuracy for each training array (controlling for block number), there was no significant difference between the arrays. Thus, the relative boost in discriminability for the VQ dimension compared to
the F0 dimension didn’t ultimately matter in the context of training.
Appendix C

Exit Survey for Experiment 2

(1) To the best of your knowledge, do you have any hearing loss or other hearing problems?

(2) What is your major?

(3) Have you ever taken a linguistics or phonetics course? If so, which one(s) and where?

(4) What, if any, other languages do you speak fluently? How long and how did you learn them?

(5) What other languages have you studied? How long and how did you learn them?

(6) Have you had any musical training? If so, what kind and for how long?

(7) What do you think this study was about?

(8) How would you characterize the Korean syllables you heard?

(9) How difficult was your training, overall? (scale of 1-6, 1 = very easy, 6 = very difficult)

(10) Did you use any particular strategy in order to learn the Korean consonants?

(11) Do you have any other comments about your experience in this experiment?