Aspects of articulatory and perceptual learning in novel phoneme acquisition

by

Emily Suzanne Cibelli

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Keith A. Johnson, Chair
Professor Susanne Gahl
Professor Robert T. Knight

Summer 2015

Aspects of articulatory and perceptual learning in novel phoneme acquisition

Abstract

Aspects of articulatory and perceptual learning in novel phoneme acquisition

by

Emily Suzanne Cibelli

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor Keith A. Johnson, Chair

In this dissertation, I describe three related experiments which explore the relationship between perceptual and articulatory learning during the process of second language phoneme acquisition. Novel phoneme acquisition is a well-documented challenge for adult learners which can persist even after extensive experience with the target language. The perceptual challenge typically manifests as an inability to distinguish between two or more target categories. Articulation performance often reveals a significant effect of the native language on pronunciation. While these problems have been extensively studied as independent phenomena, there is less work relating the joint development of articulatory and perceptual categories in second language acquisition. As a result, questions remain about the effects of cross-modal training - the extent to which learning in one domain can support development in the other.

This dissertation contributes to that body of literature with experiments that compare the effects of perceptual and articulatory training on the perception and production of Hindi coronal stop consonants by native English speakers. It focuses on adult learners who have no prior exposure to Hindi to explore patterns of learning at the earliest stages, before stable second language targets have formed. Of particular interest is the transfer of articulatory learning to perceptual categorization, a trajectory which has been explored in only a small handful of studies.

In experiment 1 (chapter 2), the joint contributions of perceptual and articulatory learning on category acquisition were assessed with a multi-day training experiment. Benefits were found for within-mode learning - perceptual training aided discrimination, and pronunciation was improved during articulatory training. However, cross-modal learning did not have an effect - pronunciation was not significantly improved by perceptual learning, and articulatory learning did not have an additional benefit on discrimination performance.

Experiment 2 (chapter 3) sought an explanation for the lack of a cross-modal effect in experiment 1. New learners who received only a single session of training showed discrimination improvement whether they received articulatory or perceptual training. This finding suggests that learners in experiment 1 failed to show an improvement from cross-modal learn-

ing not because of general inefficacy, but because they had already received prior training during perceptual learning. This result is taken as evidence that completely novice learners benefit from instruction about category targets regardless of the mode of training, but that within-mode learning may be more beneficial once some experience is gained.

In experiment 3 (chapter 4), the neural correlates of category learning were tested. The mismatch negativity response, a component of the electrophysiological response to auditory stimuli, was used to test pre-attentive categorization of selected target categories. The findings from this study indicate that learners are able to detect category differences both before and after training - a contrast to the behavioral results reported in experiments 1 and 2. This study suggests that pre-attentive auditory detection of non-native contrasts is possible even when behavior indicates an inability to categorize targets.

Taken together, these results provide some evidence for the efficacy of cross-modal training, but the effect is restricted to certain conditions. I argue that cross-modal learning may be most effective for purely novice learners; when a learner has not been introduced to a category paradigm before, it may be that any cue to category identity can be leveraged to begin to detect contrasts. This is provided that detection is undertaken as a conscious categorical task, and not a pre-attentive response. After some learning has taken place, within-mode training becomes more important to continued development of category representations. This work supports arguments that acoustic and perceptual processes may be the primary source of robust perceptual phonetic categories, while simultaneously supporting the involvement of motor representations as a secondary mechanism to support those categories. The dissertation concludes with predictions about the trajectory of learning beyond the novice stage, and suggestions for future avenues of behavioral and neural studies of non-native phonetic category representations.

To generosity and perseverance.

Contents

List of Figures

List of Tables

Acknowledgments

Finishing a dissertation always requires support from a big community, and I am no exception to this rule. Still, it is humbling to come to the end of this project and reflect on just how many people had a role in making it possible. It's safe to say that I would not be writing this page without the guidance, kindness, and friendship of the community that made me feel at home in Berkeley for six years.

First to be thanked are the members of my committee, who have utterly deprived me of the chance to commiserate with the downtrodden grad students of PhD Comics. Susanne Gahl has been my model for rigorous argumentation, thoughtful criticism, and clear organization in scientific writing. She also taught me how to love R, and shared an excellent and storied cookie recipe; both are gifts I treasure. Bob Knight gave me full access to his lab equipment and his students' expertise so that I could fast-track an EEG experiment in my last full semester, and was always available to sit down and help me make sense of my data. (He also shares my wonderful home state of New Jersey!) Keith Johnson, my advisor, has been a joy to work with. He recruited me to Berkeley and has let me have free reign of my direction ever since, encouraging me to take risks in my work while simultaneously providing gentle guidance back to structure when I needed it. For the numerous debugging sessions, career counseling talks, stories of legendary phoneticians, and unrelenting encouragement, I will always be grateful.

Not every faculty member who had a profound influence on me is listed on the first page of this document. Terry Regier gave me one of my first research projects in grad school, taught me how to break an argument down and build it back up again, and trained me to reason and question - everything - as a scientist. Eddie Chang welcomed me into his lab and allowed me access to the rare and exciting ECoG research going on at UCSF. Both Terry and Eddie advised projects that spanned much of my grad school career, and while the results of those studies are not included here, those experiences nevertheless helped shape the quality and content of this work.

Thanks are also due to the fantastic support staff in the linguistics department at Berkeley. Paula Floro kept track of an amazing number of things for our department, and still had time to help me manage the small details of my own administrative needs. I feel certain that I would have missed an embarrassing number of deadlines and forms without Belen Flores keeping an eye on me. And Ron Sprouse spent hours helping me debug code and optimize analyses, and was so kind about the two (two!) times that I brought down the department website by running huge models on the server.

Even with excellent mentors, grad school can be stressful. So I'm grateful to have chosen a department with grad students and postdocs who acted as a critical support system when things got tough. My cohort, lab mates, and collaborators (in particular, Elise Stickles, Shinae Kang, Greg Finley, Stephanie Farmer, Sarah Bakst, Clarah Cohen, Andrea Davis, Jevon Heath, Melinda Fricke, Elisabeth Wehling, Andrew Cheng, Hannah Sande, Nico Baier, Matt Goss, Chundra Cathcart, Roslyn Burns, I-Hsuan Chen, Clare Sandy, Eric Prendergast, Matt Faytak, Stephanie Shih, Yang Xu, Geoff Bacon, Joe Austerweil, Francine Foo, and

Matt Leonard) are brilliant, supportive, clever, and the model for collaborative colleagues. They're pretty fun people, too.

Support also came from outside of the walls of Dwinelle Hall. The Hucking Wugs (especially Greg Finley, Stephanie Farmer, Melinda Fricke, Jevon Heath, Hannah Sande, Marcus Ewert, Alex Dougal, Vivan Wauters, Marc Juberg, Geoff Bacon, Andrew Cheng, and Randy O'Connor) were key to my sanity - sometimes you need to get out of the lab, take advantage of the eternal California spring, and play some frisbee. Jen Boyko McGinniss, Devin Toohey, and Nick Principe have been rooting for my writing since 11th grade English class. Jen, along with Benjamin Baxter and Ronnie Phillips, took the time to read through every page of this dissertation and flag my typos and LaTeX fails. Nicole Cibelli, my sister, sent me cheery packages and mountains of photos of my nieces so their absent aunt could get to know them while away in an ivory tower across the country. My parents, Diane and Ken Cibelli, bolstered me with unwavering confidence in my abilities, and kept me fed with regular shipments of candy and contraband Taylor ham.

Finally, my partner Randy O'Connor has been my constant support for eight years. (He is making me dinner as I type this!) He taught me to appreciate my critical reasoning skills, gave me confidence when I had doubts about my ability to succeed, and is my role model for how to pursue work with passion and creativity.

Chapter 1

Introduction

1.1 Acquiring novel speech sounds

The acquisition of speech sounds in a new language is a well-known challenge for adult learners. Phonetic instruction is often not emphasized in the second language classroom (Arteaga, 2000), and yet the phonetic categories of the native language can be one of the most pervasive sources of interference when adults attempt to acquire a second language (Best, McRoberts, & Goodell, 2001; Iverson et al., 2003; Flege, Bohn, & Jang, 1997). This challenge has practical implications for the adult language learner. It also has theoretical implications for models of phonetic representations that try to account for the interaction of old and new category systems.

A good deal of the literature on adult phoneme acquisition is concerned with the ongoing difficulties that speakers of some languages encounter with particular non-native contrasts (Akahane-Yamada, Tohkura, Bradlow, & Pisoni, 1996; Flege et al., 1997; Bradlow, Pisoni, Akahane-Yamada, & Tohikura, 1997; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Hattori & Iverson, 2009; Lai, 2009; Diaz, Baus, Escera, Costa, & Sebastián-Gallés, 2008), even after many years of experience with the language and proficiency at other levels of linguistic representation. A second line of work is concerned with novice learners, and how the native language shapes representations of unfamiliar targets when learners are exposed to them for the first time (Best et al., 2001; Best & Avery, 2007; Best, Goldstein, Tyler, & Nam, 2009; Golestani & Zatorre, 2009; Pruitt, Jenkins, & Strange, 2006; Lim & Holt, 2011; Song, Skoe, Wong, & Kraus, 2008). Often, both types of studies employ training paradigms to understand more about the types of information that are helpful for novice or experienced learners to improve their perceptual and articulatory targets.

It is in the tradition of the second line of work that this dissertation is situated. The work presented here is concerned with the establishment of early categories, at the beginning stages of second language acquisition in adulthood. For learners who do not have significant exposure to the language, and have not acquired it meaningfully at other levels of representation, what are the best procedures to start the process of phoneme category acquisition?

What types of information are most useful in beginning to overcome native language biases when learners are first exposed to new targets? In learning tasks such as the ones described in this dissertation, the use of novice learners with no previous exposure to the target language may be somewhat artificial from an applied linguistics perspective, as they do not come into the task with an inherent desire to learn the language or a motivation for acquiring it, and they are not learning other aspects of the language at the same time. However, they provide a "clean slate" in terms of duration and extent of exposure, as well as motivation to acquire the language. This makes them particularly suited to test theoretical questions about the nature of category representations at a cognitive level.

A particular interest of this dissertation concerns the interaction of perceptual and articulatory representations during the acquisition of novel categories. If category invariance exists at some level, then these representations must be linked in some way (Flege et al., 1997). Nevertheless, the two representations do not always develop simultaneously in second-language acquisition (Sheldon & Strange, 1982), and the trajectory of improvement in the two domains is not always consistent (Bradlow et al., 1997; Sheldon & Strange, 1982; Baese-Berk, 2010). The transfer of perceptual learning to pronunciation of second-language targets has received some attention in the literature (Pimsleur, 1963; Akahane-Yamada et al., 1996; Bradlow et al., 1997, 1999; Wang, Jongman, & Sereno, 2003; C. Wilson, Davidson, & Martin, 2014), with generally facilitatory effects. Less work has been conducted on the reverse direction - the influence of articulatory learning on perceptual targets - though the studies which do exist report generally favorable, if somewhat mixed, effects (Catford & Pisoni, 1970; Schneiderman, Bourdages, & Champagne, 1988; Lacabex, García Lecumberri, & Cooke, 2008; Gómez Lacabex, 2009; Hirata, 2004; Herd, Jongman, & Sereno, 2013; Baese-Berk, 2010).

The link between perceptual and articulatory information is important from the classroom perspective; a better understanding of the links between the two will inform the amount of crossover that can be expected from the instruction of one domain or the other. From a theoretical perspective, understanding of this link will contribute to models which posit a link between perception and production as a mutual support system (e.g. Hickok & Poeppel, 2007) or even as compulsory ties which enable perception altogether (e.g Liberman, Delattre, Cooper, & Gerstman, 1954; Fowler, 2008).

1.2 Theories of novel phoneme acquisition

There are three prominent theories in the literature on novel phoneme acquisition that have informed experimental work on the topic. Each approaches the topic from a different perspective, and while they differ in their assumptions and mechanistic explanations, all three have enjoyed some explanatory power in the literature.

This dissertation is not explicitly designed to discriminate between contradictory predictions of these theories (and indeed, not all of these predictions are mutually exclusive). This is primarily because none focuses in much detail on the central concern of this work,

articulatory learning and its connection to perceptual learning. However, all three theories have influenced the development of the studies described here, and an overview of them is essential to contextualizing the work that follows. The extent to which its findings corroborate or clash with the theoretical predictions of each model will be explored in detail in chapter 5.

1.2.1 The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM), advocated by Best and colleagues (Best & Strange, 1992; Best et al., 2001; Best & Avery, 2007; Best et al., 2009) is designed to account for the difficulties that adult listeners have upon exposure to a contrast in a second language that is not present in their native language. It recognizes variability in the difficulty of this task, and predicts the discrimination abilities of the listener as a function of the degree to which the target sounds in the non-native contrast assimilate to categories in the native phonology. A contrast will be challenging if the two target sounds assimilate to the same native category, relatively simple if they assimilate to different categories, and somewhat challenging if there is variation in the goodness of fit of the non-native sounds to native categories.

A key tenet of PAM is that assimilation is a function of articulatory similarity. A corollary of the theory, the *articulatory organ hypothesis* (Studdert-Kennedy & Goldstein, 2003), holds that place of articulation is the primary determinant of assimilation. This predicts that two non-native sounds which use the same primary articulator (e.g. lips, velum) should be harder to discriminate than sounds which use different primary articulators, even if they share their manner of articulation (Levitt, Best, Goldstein, & Carpenter, 2006; Best & McRoberts, 2003; Best et al., 2009).

1.2.2 The Speech Learning Model

In contrast to PAM, Flege and colleagues' Speech Learning Model (SLM) (Flege, 1995; Guion, Flege, Akahane-Yamada, & Pruitt, 2000; Baker, Trofimovich, Mack, & Flege, 2002) is explicitly concerned with second language learners, rather than novice adult listeners who may not intend to acquire the language. It shares the concept of assimilation of second-language phonemes to native categories. However, it calculates similarity on the basis of what Guion et al. (2000) call phonetic distance, rather than shared articulatory gestures. More phonetically distinct sounds are less likely to assimilate to native categories. The details of what constitutes phonetic distance are underspecified in the theory's original formulation, which makes its predictions somewhat challenging to tease apart. For example, if phonetic distance were primarily a function of acoustics, then SLM would have different predictions than PAM. Contrasts which differ on manner of articulation are more acoustically distinct than contrasts which differ on place of articulation; as a result, SLM would predict better discriminability for the former, and PAM for the latter.

SLM does lay out specific predictions for the process of category formation. Its proponents argue that detection of a contrast is a necessary precursor to category acquisition. SLM is also sensitive to bidirectional influences between a speaker's two languages: second-language (L2) categories are argued to influence first-language (L1) representations, just as L1 categories bias the perception of L2 tokens (although the relative weight of each system will vary as a function of experience and age of acquisition).

1.2.3 The Native Language Magnet Theory

The Native Language Magnet Theory (NLM and later NLM-e, Kuhl, 2000; Iverson et al., 2003; Kuhl, Conboy, & Coffey-Corina, 2008) differs from PAM and SLM in two ways. First, it contends that acoustic and auditory information form the basis of category representations. Second, it focuses on the forces that govern first language acquisition, and uses that as a foundation to understand bias patterns observed in adult learners. According to NLM, infants begin to learn the categories of their native language in part because of the enhancement of critical perceptual cues in child-directed speech. This leads to categories which are built around strongly-weighted canonical exemplars. As these categories strengthen through repeat exposure, they warp the perceptual space (the *perceptual magnet effect*) by drawing incoming tokens to be perceived as more similar to these existing categories. This has the effect of concentrating perceptual representations towards category centers, and away from category boundaries, giving rise to categorical perception. This has a beneficial effect in first language acquisition, where reinforcement of L1 categories is the goal. However, this same effect can hamper the efforts of adults to acquire a language, as incoming L2 tokens will likewise be perceptually warped towards existing L1 category centers.

Recently, parallels have been drawn between NLM and Bayesian models of cognitive processes. A Bayesian model of the perceptual magnet effect (Feldman, Griffiths, & Morgan, 2009) effectively modeled category bias as a function of uncertainty about the speech signal: memory demands or acoustic noise increased reliance on existing categories (in Bayesian terms, the prior) at the expense of veridical representation of the speech signal. Such a system builds strong category priors over time, which can help explain why L1 phonetic categories are strongly ingrained by adulthood. Köver and Bao (2010) used a similar modeling approach to explain tuning in the auditory cortex, providing a neural account consistent with the computational and cognitive approaches of Feldman et al. (2009) and Kuhl et al. (2008), respectively.

1.3 Structure of the dissertation

This dissertation contains three sets of experiments designed to test the nature of early phonetic category learning. It investigates the very beginning stages (chapters 2 and 3) of acquisition and acquisition after a small amount of training (chapter 2), as well as pre-attentive responses to training (chapter 4).

In chapter 2, the findings of a multi-day training study (experiment 1) are described. As an introduction, past experimental literature on novel phoneme acquisition and various methodological approaches to the problem are discussed. This study explicitly tests perceptual learning in novice learners, and investigates the additive effects of articulatory training on perceptual representations after perceptual training has already occurred. It also tests production accuracy as a function of both articulatory and perceptual training. Control studies are included to test the effects of exposure without training, as well as the nature of retention after a month-long period of no training.

The study in chapter 3 (experiment 2) was designed to directly compare the efficacy of perceptual and articulatory training on the development of perceptual categories. All training in chapter 2 begins with perceptual training, followed by articulatory training. In the study designed in chapter 3, groups receive either perceptual or articulatory training from the start, to test their relative efficacy on completely novice learners. The effects of different types of articulatory training are also compared.

In chapter 4, the pre-attentive categorization of novel categories is tested. There is some debate about the extent to which conscious decision-making aids or interferes with the acquisition of target contrasts which cross-cut native categories. Study 3, which is described in this chapter, makes use of the mismatch negativity response, a pre-attentive component of the ERP signal which indicates detection of a change in a signal (such as a change from one category to another). This paradigm allows for the detection of category discrimination without the need for an overt behavioral response, and may be sensitive to detection earlier than a behavioral response.

Chapter 5 summarizes the results of all three studies, and discusses the points of agreement and contradiction between them. The nature of the link between perception and production is discussed in light of the findings from each study. In addition, a proposal is put forth about the nature of phonemic representations in the novice learner, and the conditions in which cross-modal information may be more or less facilitatory for their development. Finally, methodological considerations and future directions are discussed.

Chapter 2

Multiple cues to establish fledgling novel categories

2.1 Introduction

In this chapter, I present the results of three related behavioral studies. These experiments were designed to examine how multiple sources of information help learners take the first steps towards the establishment of a new phonetic category in the target language. Of specific interest are the types of procedures that lead to successful learning in a multi-day short-term laboratory training paradigm, the properties of the acoustic signal that facilitate learning, and the nature of cross-modal learning in the perceptual and articulatory domains.

2.1.1 Study aims

Biases from phonetic categories in the native language (e.g. Best et al., 2001; Flege et al., 1997; Kuhl et al., 2008) often impact the ability of adult listeners to accurately discriminate between contrasts in a second language, and to produce them accurately. It has been proposed that some catalyst must provide the means to overcome the biases of the native language in order for accurate acquisition of second-language targets to take place (McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; Vallabha & McClelland, 2007). Because these difficulties can sometimes persist even with substantial experience with the language (e.g. Bradlow et al., 1997; Diaz et al., 2008), it is clear that exposure alone is not a strong predictor of success in articulation or perceptual discrimination. Therefore, it is important to explore different active training interventions to better understand which procedures - and ultimately, which sources of information - are effective in training second-language phonetic categories.

The experiments reported in this chapter have two primary goals. The first goal is to replicate the findings of previous studies which have demonstrated better discrimination of a non-native contrast after perceptual training in a short-term laboratory setting. In the main study described in this chapter (experiment 1A), I make use of three perceptual

interventions: repeated exposure, explicit performance feedback, and adaptive fading. These and other common methodologies used to study this topic are discussed in detail in section 2.2. This procedure is designed to test the hypothesis that perceptual biases inherited from the native language can be overcome to some degree even in a short time frame. The training paradigm is designed to test the combined effects of these perceptual interventions, rather than separate tests of their efficacy as individual training approaches.

The second goal is to incorporate the interaction of perceptual and articulatory learning into such a paradigm, in order to measure whether cross-domain (that is, perceptual-to-articulatory, and vice-versa) category information can contribute to the types of information necessary to overcome a native-language category bias. This goal is informed by Hebbian learning accounts of phoneme learning (McCandliss et al., 2002; Vallabha & McClelland, 2007), which suggest that multiple sources of information can have an influence on novel category formation. While the findings of Catford and Pisoni (1970) suggest that there should be crossover between the two domains, learning does not always go hand-in-hand in a predictable manner (Bradlow et al., 1997). As a result, the efficacy of this cross-domain information in such a paradigm remains an open question. The studies described in this chapter test the hypothesis that cross-modal information can contribute to the development of new phonetic categories in novice learners, even when within-mode information is also being used to facilitate learning.

2.2 Background: Past findings on perceptual learning

This section surveys experimental methodologies that have been used as training paradigms to aid the development of phonetic representations in adult second-language learners. A useful introduction to the topic can be found in Bradlow (2008), which surveyed more than 20 years of training studies to summarize general trends and examine the limits of various approaches. The review focuses on studies with particularly challenging contrasts - most notably, Japanese speakers' perception of English /ɹ/-/l/, but also the Hindi dental-retroflex contrast for English speakers - on the logic that success in the hardest cases will generalize to a broader set of learned contrasts.

In her review, effective training strategies included explicit cue training of the critical acoustic feature in the contrast (Strange & Dittmann, 1984; McCandliss et al., 2002), adaptive training which reduced the dispersion between tokens as learning progressed (McCandliss et al., 2002; Pruitt, 1995), a broad stimulus set and/or high variability training (Logan, Lively, & Pisoni, 1991), and both long- (Flege, 1995) and short-term (Logan et al., 1991) paradigms. Bradlow concludes that training in the laboratory *can* be effective under optimal conditions, that high-variability stimulus sets are ultimately preferable for long-term storage and robust maintenance of categories, and that expanding beyond syllables to words and sentences may be most effective. The review also emphasized the need for a reliable pre-test measure of competence in order to accurately assess change as a function of training.

In recent years, these specific aspects of training paradigms have been examined in detail,

in an attempt to hone in on the most promising strategies to promote learning and retention of new phonemic categories. Several factors that have received particular attention in recent research are reviewed below. The first two sections relate to the first goal of this study - effective techniques for perceptual learning - while the third section addresses the second goal, cross-modal learning during phoneme acquisition.

2.2.1 Feedback and explicit attention

The extent of the benefit of explicit feedback given to the learner is somewhat controversial. On the positive side, feedback in the laboratory is an aid to keep subjects on track and progressing towards accurate representations without the long periods of exposure available to individuals learning in a naturalistic environment. A number of studies have found that certain training paradigms with feedback can support learning (e.g. Goudbeek, Cutler, & Smits, 2008; McCandliss et al., 2002) even for infamously difficult contrasts.

Importantly, explicit feedback on perceptual judgments in the lab is not entirely divorced from naturalistic language learning. While adult learners may not receive explicit information about perceptual cues in the classroom or the ambient language environment, it is argued that infants may benefit from such information when acquiring their first language. Goldstein and Schwade (2008) instructed caregivers to respond to their prelinguistic infants' babbling with utterances that reflected these productions; this exposure in turn shaped the structure of further babbling, suggesting that the infants had generalized the phonological patterns of the adults' speech and applied it to their own productions.

In addition to feedback, subjects can be directed to explicitly attend to a particular feature or contrast as a means of manipulating training variables. Pederson and Guion-Anderson (2010) examined how well English speakers would learn challenging Hindi contrasts (VOT contrasts, vowel length, and dental-retroflex stop place contrasts) when told to only attend to the consonants or to the vowels. They found that attention modulated performance on consonant contrasts: those participants who were explicitly instructed to pay attention to consonants showed discrimination improvement after training, while those who attended explicitly to vowels showed no improvement on consonants.

However, not every study demonstrates an advantage for explicit feedback or attention. Several recent studies (Vlahou, Protopapas, & Seitz, 2011; Seitz et al., 2010; Lim & Holt, 2011; Gulian, Escudero, & Boersma, 2007) compare explicit and implicit training - the latter involving no explicit feedback, or else a redundant cue for subjects to attend to - and find an advantage for the implicit approach. Vlahou et al. (2011) propose that learners presented with explicit feedback may generate incorrect hypotheses about the generalizations to be learned, impeding accurate perception of the true targets. In a related vein, Kondaurova and Francis (2010) found that inhibition of *native* cues which conflict with a target contrast to be learned can be as effective as explicit cuing of that target contrast's cues.

One consideration to keep in mind is that in some (but not all) studies which show a benefit for implicit learning, the target contrast is still cued with a redundant cue that learners must attend to. This may artificially enhance the naturalness of the target contrast,

making the task easier than it would otherwise be. Given the mixed evidence regarding explicit feedback, it may be that language-specific factors (do native-language categories have conflicting cues?) and perhaps even individual differences (is a learner likely to form an incorrect hypothesis from feedback?) may determine whether or not explicit attention and feedback are beneficial or detrimental in a learning study.

2.2.2 Stimulus properties

Equally as important as study structure in a laboratory learning study is the design of stimuli. No study can encompass all of the variation and context found in a naturalistic speech setting, so choosing the subset used in a learning study is an important consideration, as it determines which features learners will be expected to generalize over.

Cue enhancement and dispersion

One particularly effective way to leverage stimulus variation is to use an adaptive fading or gradient stimulus paradigm (Jamieson & Morosan, 1986; Pruitt, 1995; Escudero, Benders, & Wanrooij, 2011), a technique that has been found to be robust throughout a number of psychological and cognitive domains (Terrace, 1963). In this paradigm, listeners are first presented with distinctive, well-separated tokens along some acoustic dimension(s). As subjects demonstrate improvement, less prototypical or more varied tokens are introduced. This allows the learner to begin to generalize from canonical tokens to less-central category members. In some cases, the most disperse contrasts are artificially exaggerated beyond the cues found in natural speech, in order to increase the likelihood that listeners will pick up on the contrast (Protopapas & Calhoun, 2000; McCandliss et al., 2002).

McCandliss and colleagues (2002) were able to demonstrate robust learning of /ɹ/-/l/ by Japanese speakers in a relatively short time frame (three 20-minute sessions) with an adaptive fading structure. They frame this finding in terms of a Hebbian account of learning, which argues that - because of the reinforcement of jointly-firing neurons in response to a frequently-encountered stimulus - non-native tokens that are not sufficiently distinct from native categories will not evoke different neural patterns of activity, but will instead reinforce established categories. Thus, for particularly challenging contrasts, exaggerated tokens are necessary to break this association with native categories and establish novel representations. In follow-up work (Vallabha & McClelland, 2007), they expand this account to incorporate findings that feedback can work in tandem with cue enhancement to create new representations, suggesting that the model of Hebbian learning is sensitive to multiple sources of external information.

While McCandliss and colleagues (2002) question the ecological validity of cue enhancement, Escudero et al. (2011) interpret their participants' successful learning with adaptive training as analogous to first language acquisition. Furthermore, they suggest that similar processes may be at work for second language acquisition, particularly in challenging cases where ambient exposure to the language does not provide sufficiently distinct input for the

learner to generalize from. This is also consistent with the predictions of the Native Language Magnet Theory (Kuhl et al., 2008), which argues that exaggerated infant-directed speech is the basis for the formation of strong category representations, a premise which is supported by recent ERP work on the enhancement of neural activity for formant-exaggerated speech in the infants (Zhang et al., 2011) and adults (Zhang, Kuhl, & Imada, 2009).

Variation and generalization

Training studies take different approaches to the amount of variability present in training stimuli. Some studies restrict exposure to a limited number of speakers or tokens, while others use more high-variability training. Limiting variability during training provides a space to test generalization to novel speakers, tokens, or context during post-training test phases. In her survey of training paradigms, Bradlow (2008) found that variation in both speakers and words or contexts seems to aid in the generation of robust perceptual traces for speech sounds. Similar results have been found in recent studies (Lim & Holt, 2011; Perrachione, Lee, Ha, & Wong, 2011; Kondaurova & Francis, 2010). Sadakata and McQueen (2011) found that listeners in high-variability training on a single-geminate stop contrast generalized not only to novel speakers and tokens, but also to new segments with different place and manner features which made use of the same durational contrast. This suggests that high-variability training may also facilitate the acquisition of more abstract phonological paradigms.

The benefits of high-variability training are intuitive, as the input more accurately reflects the nature of ambient language in natural speech contexts. Assuming that learning from such paradigms does occur, it also seems that generalization to novel forms would follow more easily, because listeners have already successfully synthesized a wide range of phonetic variation to arrive at a stable percept. In fact, there is some evidence that enhanced variability in lab training can be beneficial even when learners are already getting natural input in their linguistic environment (Iverson, Pinet, & Evans, 2012). However, naturalistic long-term learning is not the goal of every training study. As such, some paradigms may limit variability in order to demonstrate significant effects within the constraints of short-term laboratory training.

2.2.3 Cross-modal learning

The interaction between perceptual and articulatory learning is the second major aim of the current study. With respect to the effects of articulatory learning on perception, Catford and Pisoni (1970) introduced a series of unfamiliar non-native phonemes to English speakers, and measured their ability to discriminate and produce the sounds after either (1) purely acoustic training or (2) purely articulatory training. In the latter case, explicit information was given about vocal tract anatomy and articulatory postures of the target sounds. Unsurprisingly, learners in group two were more successful than those in group one at producing target

sounds after training; more remarkably, the advantage held for post-training discrimination performance as well.

Since this work, there have been a small handful of studies investigating the facilitatory effects of articulatory knowledge on perceptual learning in a second language through either audiovisual cues or explicit articulatory instruction (Hazan & Sennema, 2007; Lacabex et al., 2008; Hirata, 2004; Herd et al., 2013). In addition, a larger literature has generally found successful crossover from perceptual training to production performance (Pimsleur, 1963; Flege et al., 1997; Akahane-Yamada et al., 1996; Wang et al., 2003; Baese-Berk, 2010), albeit with some variation in how well perception and pronunciation seem to be linked at the individual level (Iverson et al., 2012; Bradlow et al., 1997).

Questions remain regarding the nature of the link between perception and production skills. In part this is because of variation in articulatory training approaches (Hirata, 2004). It may also be the case that articulatory and acoustic representations in the learner may not be clearly or consciously linked, and in some cases may even be disruptive (Schneiderman et al., 1988; Baese-Berk, 2010). As a result, the optimal way to integrate articulatory and acoustic learning in the laboratory or the classroom remains to be discovered. Questions about the nature of the representation in the learner also link to a broader debate in the phonetics literature regarding the status of acoustic or motor representations as the ultimate source of phonemic categories (Diehl, Lotto, & Holt, 2004), and the way that learners are able to - or fail to - capitalize on cross-modal information may inform these theories.

2.3 The current study

2.3.1 Research questions

1. Can adult novice learners of a new target language overcome their native language phoneme category biases with a combination of perceptual interventions in a short-term laboratory training paradigm?

2. Do non-native contrasts differ in learnability as a result of acoustic properties of those contrasts, and/or as a function of their similar to native language categories?

3. Is there an interaction between perceptual learning and articulatory learning such that each supports the development of categories in the other?

    a) Can knowledge about articulation of a target sound serve as an anchor for the establishment of a perceptual category?

    b) Can enhanced perceptual discrimination as a function of learning support a learner's ability to adjust their articulatory targets when attempting to produce new target phonemes?

2.3.2 Experiments

To address these questions, three related experiments were conducted. In experiment 1A, subjects participated in a eight-day training study designed to test the interacting effects of repeated exposure, performance feedback, adaptive fading, and articulatory training on the discrimination of Hindi coronal stop consonants. Experiment 1B served as a control experiment, which tested discrimination of the same stimuli exclusively as a function of repeated exposure. In experiment 1C, a small group of subjects who completed experiment 1A returned to the lab for re-testing approximately one month after the end of the final regular testing session.

2.3.3 Theoretical implications

While not explicitly designed to discriminate between prominent theories of second-language phoneme acquisition - the Speech Learning Model (SLM, Flege, 1995), the Perceptual Assimilation Model (PAM, Best et al., 2001), and the Native Language Magnet Theory (NLM, Kuhl et al., 2008) - there are theoretical implications related to each for the possible outcomes of the second and third research questions described above.

Question 2 asks whether learnability varies between different target contrasts. The theory which has discussed this in most detail is PAM. This model outlines a gradient of difficulty for non-native contrasts on the basis of (a) the number of native categories that cause assimilation, and (b) the goodness of fit to those native categories. If two targets in a non-native contrast assimilate to different native categories, they are expected to be well-discriminated. If the two targets assimilate to a single native category, but differ in how well they match the native category, then discrimination may be somewhat difficult, but not impossible. If both assimilate equally well to a single native category, then discrimination will be very difficult.

In the present study, many different contrasts are being tested. It is likely that they will differ in the extent to which they are discriminable, even prior to training, and these differences may remain to some extent even after training. If either pattern is observed, this would be consistent with the general predictions of PAM that non-native phoneme perception is a phenomenon with gradient difficulty.

Question 3 asks about the interaction of articulatory and acoustic representations during learning. Different theories make different underlying assumptions about the underlying representations of phonetic categories. Work on SLM has generally been consistent with the assumption that phonetic categories are represented by acoustic information. PAM proposes a larger role for articulatory features underlying phonetic representations. NLM (particularly its more recent formulations) emphasizes the connection between both representational systems as a mechanism for first language category acquisition. While a positive result for the role of cross-modal training in perceptual learning would not specifically refute any of the theories, it would be particularly parsimonious for PAM and NLM, which place more emphasis on the role of articulatory features or articulatory learning in perception.

2.4 Experiment 1A: Training study

2.4.1 Methods

Stimuli

Hindi phonology The stimuli for this study were chosen from the coronal stop series in Hindi. Hindi has a dental-retroflex place of articulation contrast. This contrast is well-studied in the literature on second language phoneme acquisition, as it presents a particular challenge to native English speakers (Pruitt et al., 2006; Pederson & Guion-Anderson, 2010; Golestani, 2014; Tees & Werker, 1984; Vlahou et al., 2011), who tend to map both onto the English alveolar stop category. Hindi also has a four-way voicing distinction which contrasts voiceless unaspirated, voiceless aspirated, voiced, and breathy voiced stops.

It also has a four-way voicing distinction which contrasts voiceless unaspirated /t/, voiceless aspirated/t^h/, voiced /d/, and breathy voiced /$d^{\text{ɦ}}$/ stops, which have also been studied with adult learners (Tees & Werker, 1984). These contrasts have a somewhat complex mapping to English, with a good match of Hindi /t/ to English [t] and Hindi /t^h/ to English [t^h]. The phonetics of the Hindi /d/ map onto the voiced allophone of English /t/; under Best's account (Best et al., 2001; Best & McRoberts, 2003; Best & Avery, 2007) this causes Hindi /t/ and /d/ to be assimilated into a single category, while Hindi /$d^{\text{ɦ}}$/ is more acoustically distinct.

This variation in mapping voicing categories to the English phonology, as well as differences in acoustic salience between the place contrast and the voicing contrasts, means that some contrasts are easier and some ultimately harder for native English speakers. Crossing place of articulation and voicing in the coronal stop series of Hindi, therefore, provides a rich set of stimuli with a variation in difficulty for English learners.

Stimulus construction Stimuli consisted of consonant-vowel (CV) and vowel-consonant-vowel (VCV) syllables, with one of three vowels, /a/, /i/, or /u/, combined with one of eight consonants. The consonants represent the coronal stop series of Hindi, with two places of articulation - dental and retroflex - crossed with four voicing types: voiceless unaspirated, voiceless aspirated, plain voiced, and breathy voiced. An example set of stimuli is shown in table 2.1.

A native speaker of Hindi was recruited to record all stimuli for the study. The speaker was a 20-year-old female who was born in India and lived in both India and the United States throughout her childhood. At the time of recording, she was a college student at the University of California, Berkeley.

Stimuli were recorded in blocks of 80. Each block consisted of one vowel/syllable-structure (VCV or CV) pairing, combined with each of the 8 unique consonants, at 10 repetitions each. This ensured that contrast between consonants was emphasized. However, this recording method tended to lead to contrastive stress between successive syllable types, introducing an potential extra cue for the identification of phonemic contrasts. Because this cue was

		CV stimuli		
	voiceless unaspirated	voiceless aspirated	voiced	breathy voiced
dental	t̪a	t̪ʰa	d̪a	d̪ɦa
retroflex	ʈa	ʈʰa	ɖa	ɖɦa
		VCV stimuli		
	voiceless unaspirated	voiceless aspirated	voiced	breathy voiced
dental	at̪a	at̪ʰa	ad̪a	ad̪ɦa
retroflex	aʈa	aʈʰa	aɖa	aɖɦa

Table 2.1: Example of stimulus types with the vowel /a/.

an artifact of the recording rather than a feature of the language, it is not desirable in the current study. Therefore, pitch of all stimuli was flattened to the mean of the pitch across the whole stimulus set[1].

The speaker recorded all blocks of stimuli in two styles. One was a "careful" style, where she was told to speak clearly and enunciate as if she were teaching the consonants to a class of learners. The second was a "natural" style, where she was asked to speak more naturally, with less focus on meticulous pronunciation. The crossing of the syllable structure (VCV or CV) and speaking style (careful or natural) provided four sets of stimuli with increasing difficulty, from VCV careful stimuli (most salient acoustic cues) to CV natural stimuli (least salient acoustic cues), which were used in an adaptive fading paradigm (see section 2.4.1 for a full explanation of the paradigm).

In total, 960 stimuli were initially recorded. A subset of tokens were re-recorded at a later date because of poor identification of 13 syllable types in the initial set by native speakers (see section 2.4.1 for details).

Stimulus norming Each recorded token was hand-segmented at the acoustic onset and offset of speech to create a set of 960 stimuli. From this set, a subset of 384 training tokens were selected for inclusion in the training study. This subset consisted of four tokens of each syllable type listed above.

Syllables were selected for inclusion in the training study with a norming study conducted with native Hindi speakers. The goal of the norming study was to ensure that selected tokens were clearly identifiable to native speakers. This ensured that learners were not required to learn tokens that were not good exemplars of Hindi categories.

The study consisted of an identification task. Participants listened to all recorded sylla-

[1]This approach has the disadvantage of eliminating naturally-occurring cues to consonant identity that are present in the fundamental frequency. For example, it is known that the breathy-voiced stop causes significantly-pronounced lowering of the onset F_0 of the following vowel compared to the other stops (Hombert, Ohala, & Ewan, 1979). However, the avoidance of extraneous pitch cues, which listeners might rely on to the exclusion of naturally-occurring cues, was considered a priority for the present study.

Stimulus class	Accuracy (careful)	Accuracy (natural)
d̪	79.4%	61.3%
d̪ʱ	98.1%	87.4%
t̪	90.4%	92.2%
t̪ʰ	96.4%	97.8%
ɖ	44.7%	79.0%
ɖʱ	20.4%	57.4%
ʈ	88.2%	76.1%
ʈʰ	88.6%	84.1%

Table 2.2: Norming study performance by consonant type and style, averaged over subjects 1 and 4.

bles presented over headphones. Each stimulus was accompanied by a set of eight competitor syllables spelled out in Devanagari and presented on a computer monitor. One of the eight syllables matched the auditory stimulus. Participants were asked to indicate which syllable they heard with a keypress corresponding to one of the eight syllables. Competitor syllables always contrasted with the target syllable in consonant voicing and/or consonant place, but not in vowel or in syllable structure (CV vs. VCV).

Because of the large number of stimuli, the norming task was split into two 60-minute sessions. Only one style (careful or natural) was presented per session. This was done to encourage identification judgments to be based on the acoustic space available within that stimulus set.

Four native speakers of Hindi were initially recruited to participate in the preliminary norming study. The data from subjects 2 and 3 was excluded from analysis, as their identification accuracy was relatively low (32.4 % and 16.8%). In addition, initial analysis showed that these two subjects had a high number of trials with reaction times below 150 ms (13% and 18%), faster than is reasonable for an 8-choice identification task, suggesting that they were not sufficiently engaged in the task to give accurate data.

For the two subjects who remained, accuracy was 76.2% (subject 1) and 78.9% (subject 4) across all tokens. Accuracy rates were similar for careful (subject 1: 71.8 % correct, subject 2: 78.7% correct) and natural (subject 1: 80.5% correct, subject 2: 77.8 % correct) tokens. The four tokens which received the highest rate of accurate identification within each syllable type were selected for inclusion in the training study. Accuracy rates by syllable type are reported in table 2.2.

As noted above, because 13 syllable types had low accuracy, a set of stimuli were re-recorded by the original speaker. In addition to recording syllable types that were poorly-identified in the norming study, the speaker also recorded examples of the syllable types that were most confusable with the poorly-identified stimuli, which was done to aid the speaker in emphasizing the contrast between the syllable types. To test whether these stimuli were better exemplars of the target categories than the original recordings, one additional native Hindi speaker was recruited to identify these stimuli in a single session, using the same

identification paradigm as for the original set. Accuracy rates for the 13 target re-recorded syllable types were still low (maximum performance: 9.7 % correct by consonant, 5.1% by syllable type).

In addition to these identification data, the native speaker recruited to evaluate the re-recorded stimuli also participated in an AX discrimination task, which contrasted the re-recorded syllable types which were challenging with those types that were re-recorded to emphasize contrast. Accuracy on this task across all trials was 58.9 %. This data was used to supplement the norming data in selecting re-recorded tokens. Taken together, this data was used to compare re-recorded stimuli to the original stimuli, to see if any tokens appeared to be stronger exemplars of the target categories. From the re-recorded set, 28 tokens were selected to replace tokens in the original stimulus set.

Acoustic properties of stimuli To quantify the acoustic properties of the stimuli along the dimensions (voicing and place of articulation) that distinguish them from their closest analogues in English, measurements of voice onset time (VOT), formant frequencies, and burst frequency of each stimulus were made. To do so, stimuli were hand-aligned for the onset and offset of each syllable, phoneme, and VOT properties of each consonant. These and subsequent transcriptions were conducted in Praat (Boersma & Weenink, 2014). These descriptions of the acoustic correlates of voicing and place of articulation provide a catalog of the acoustic cues that listeners may leverage in acquiring new perceptual contrasts; in addition, they provide a benchmark for evaluating learners' production accuracy as they progress through the training.

Durational metrics (voicing): Three durational measures were used to define and measure VOT in the stimuli, with different features applied to different phonemic voicing types. Negative VOT (i.e. voicing during the stop closure) was measured in all syllables with a voiced stop (/d/, /d$^{\text{fi}}$/), and was defined as the interval from the onset of prevoicing until the onset of the stop release burst. Closure duration was measured for syllables with voiceless consonants (/t/, /t$^{\text{fi}}$/) only in vowel-consonant-vowel (VCV) syllables (as the onset of the closure could not be determined in CV syllables), and was defined as the time from the offset of the first vowel until the stop release burst. Positive VOT was defined in all syllables as the time from the onset of the burst to the onset of the vowel. (The burst was also segmented within the positive VOT interval for analysis of its spectral properties; these properties are discussed below.)

A major motivator of comparing these durational properties is to test which properties signal the difference between phonemic voicing types (e.g. voiced vs. voiceless, breathy vs. aspirated). A secondary goal is to examine whether durational differences were affected by speaking style (careful vs. natural stimuli). To that end, a linear model was fit for each VOT metric (closure duration, positive VOT, and negative VOT), with duration as the dependent variable in each model, and the interaction of voicing type and speaking style as predictors. VOT differences across stimuli can be seen in figure 2.1.

For the closure duration model, the simple effect of style was significant (β = -13.2927, p = 0.017), indicating that closure durations were shorter in natural stimuli than careful stimuli. But importantly, pairwise tests (with post-hoc corrections for multiple comparisons) revealed that the difference between aspirated and voiceless stops was cued in part by closure duration, and that this difference was preserved in both careful (adj. $p < 0.001$) and natural (adj. $p < 0.001$) stimuli, despite the overall duration differences attributable to style.

The presence or absence of negative VOT is a major cue to the distinction between the pre-voiced (voiced and breathy) and the non-pre-voiced (voiceless and aspirated) stops. However, within pre-voiced segments, no differences were found in negative VOT as a function of style, voicing type, or the interaction of style and voicing type.

Positive VOT is an important cue to the distinction between the aspirates (aspirated and breathy stops) and the non-aspirates (voiced and voiceless) in Hindi. The positive VOT model recovered those differences; the duration of positive VOT in aspirated and breathy stops were significantly different from voiceless and voiced stops in both careful and natural stimuli (adj. $p < 0.001$ for all comparisons), indicating that these feature distinctions were preserved in both sets of stimuli. Within aspirate category (breathy/aspirated, and voiced/voiceless) there were no significant differences, as predicted. The positive VOT model also had a significant simple effect of style (β = -10.791, p = 0.016); as with closure duration, positive VOT duration was shorter in natural stimuli, even though critical phonemic differences between aspirates and non-aspirates were maintained.

Taken together, the duration metrics reveal that positive VOT and closure duration, but not negative VOT, provided important cues to the voicing features of the stimuli. In addition, these critical cues were affected (but not eliminated) by speaking style, indicating that the careful stimuli are longer, clearer, and "easier" tokens than their natural counterparts, a critical feature of the adaptive fading paradigm used in the current study.

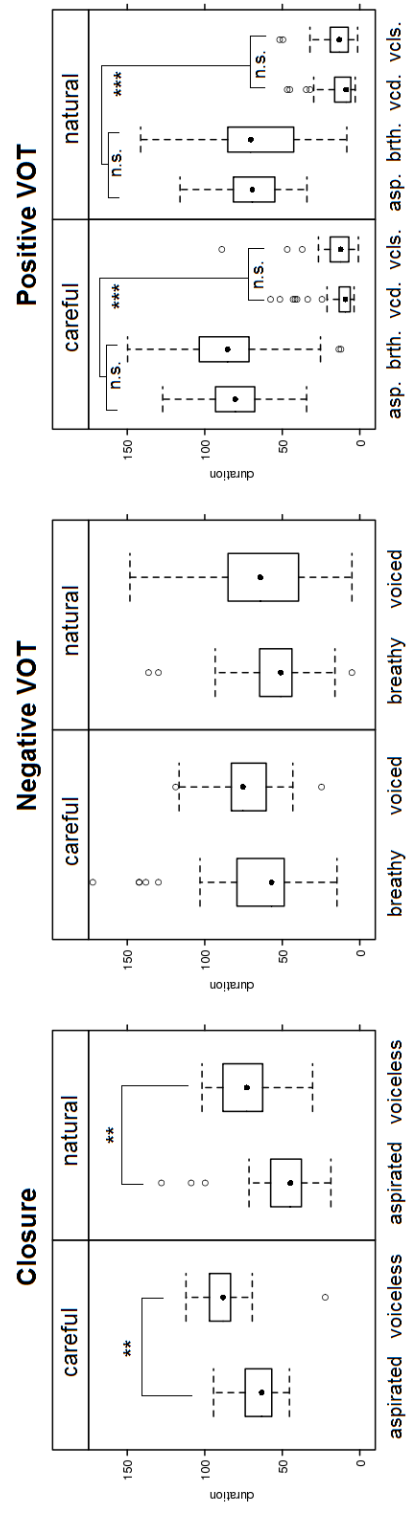


Figure 2.1: Durational VOT properties of stimuli. Significant contrasts between voicing types (e.g. aspirated vs. breathy) within speaking style (careful or natural) are signaled with significance stars (. $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$) as determined by pairwise comparisons (adjustment with Tukey's Honest Significant Difference Test).

Formant analysis: Formant frequencies have been identified as an important (if not independently sufficient) cue for place of articulation information, both at the vowel onset (Liberman et al., 1954; Delattre, Liberman, & Cooper, 1955; Kewley-Port, 1982) and potentially across the vowel duration (Sussman, McCaffrey, & Matthews, 1991; Sussman, Hoemeke, & Ahmed, 1993). The second formant (F2) is usually identified as key to place cues, but is unlikely to be sufficient for a dental-retroflex place contrast (Sussman et al., 1993) as retroflexion is often argued to lower the third formant (F3) (Stevens & Blumstein, 1975; Werker & Tees, 1984; Werker, Gilbert, Humphrey, & Tees, 1981) compared to the dental variant.

Measurements of F2 and F3 at vowel onset and vowel midpoint were taken from stimuli with voiced consonants ([d̪] and [d̪]). As a preliminary step, several logistic regression models were constructed of consonant-vowel (CV) stimuli only, to assess which combination of features (F2 onset, F2 midpoint, F3 onset, and/or F3 midpoint) most accurately categorized stimuli as dental or retroflex. The best model fit predicted place of articulation as a function of F2 onset, F3 onset, and F3 midpoint (AIC = 53.89214; range of other model AICs = 55.20171 - 71.07171; the worst model represented the "classic" locus equation, with independent variables of F2 onset and F2 midpoint only). Taking speaking style (careful vs. natural) into account did not further improve the model, suggesting that these formant properties were not strongly affected by this manipulation.

To further explore the classifying power of these features, linear discriminant analysis (LDA) with jacknifed (leave-one-out) cross-validation was used to classify the 48 CV stimuli with voiced segments as either dental or retroflex, again using combinations of the four features above. The best overall classification accuracy was again obtained with information about F2 onset, F3 onset, and F3 midpoint (overall classification accuracy: 83.3%; dental accuracy: 79.2%; retroflex accuracy: 87.5%.) This indicates, perhaps unsurprisingly, that the consonant-vowel transition is a key timepoint for formant-cued place of articulation information. But in addition, information contained in F3 at the vowel midpoint may also be an important cue to place identity, suggesting - in line with the concept of locus equations (Sussman et al., 1991), if borne out in a different formant - that consonant place information is maintained to some degree throughout a long portion of the duration of the following vowel.

Because vowel-consonant-vowel (VCV) stimuli contain transitional information both from the first vowel to the consonant, and then out of the consonant into the second vowel, it was hypothesized that these stimuli would be more robust carriers of place of articulation information. However, when comparable logistic regression and LDA models were run on VCV stimuli with voiced consonants (48 stimuli), the regression model fits and classification accuracy were not improved by midpoint or offset F2 and F3 from the first vowel. In fact, the best VCV models contained F2 onset, F3 onset, and F2 midpoint information from the second vowel only, and these models slightly underperformed compared to the CV models (logistic regresion AIC = 54.3601; overall LDA classifciation accuracy = 75.0%).

Stop burst spectra: While the formant analysis yielded good classification of place of articu-

lation, its success is limited to unaspirated segments (specifically voiced tokens in the present analysis). To find a metric that could describe all stimuli, a second component was considered: the spectral properties of the stop burst, which is also known to be a predictor of place of articulation information (Blumstein & Stevens, 1979; Kewley-Port, Pisoni, & Studdert-Kennedy, 1983). Following the analysis in Forrest, Wiesmer, Milenkovic, and Dougall (1988), eight spectral properties were measured from a spectrum extracted at center of each stop burst. The first four of these are the linear frequency scale spectral moments: centroid, standard deviation (variance), skewness, and kurtosis (a measure the relative peakedness or flatness of the spectrum). The other four were Bark-transformed equivalents of the linear scale spectral moments. Stimuli were excluded from this analysis if there was no visible stop burst or when the burst duration was shorter than 2 milliseconds; as a result, 338 of the 384 stimuli were included in this analysis.

Figure 2.2 shows the distribution of each spectral moment across stimuli in the linear and Bark scales. All metrics showed a significant difference between dental and retroflex stimuli (Wilcoxon rank-sum test, $p < 0.001$). The centroid was slightly lower for retroflex bursts than dental bursts, but had slightly higher standard deviation. Skewness was more negative for dental than retroflex bursts. Kurtosis was lower for retroflex bursts, indicating that they had relatively more flat/diffuse spectra compared to dental bursts.

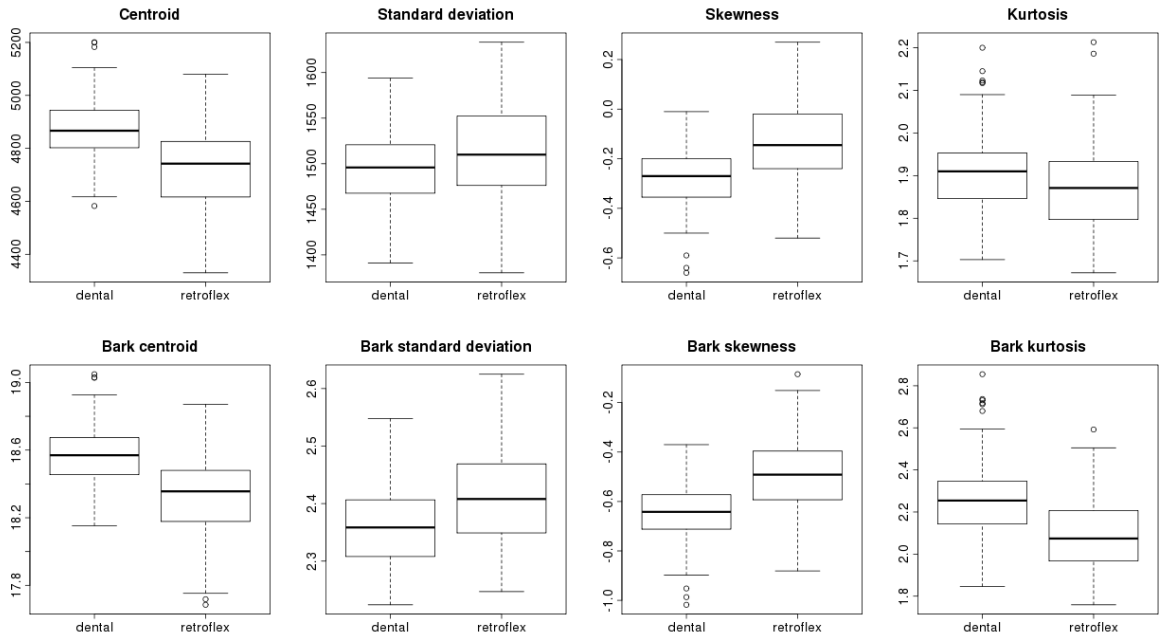


Figure 2.2: Linear scale and Bark scale spectral moments of stimuli burst spectra. Contrasts between dental and retroflex stimuli were significant (Wilcoxon rank-sum test, $p < 0.001$) for all metrics.

A linear discriminant analysis with cross-validation was run separately for the linear scale

Session code	Session	Tasks	Stimuli	Feedback?
A	Pre-test	discrimination; repetition	CV natural	no
B	Perception training 1	discrimination	VCV careful	yes
C	Perception training 2	discrimination	VCV natural	yes
D	Perception training 3	discrimination	CV careful	yes
E	Perception training 4	discrimination	CV natural	yes
F	Post-test	discrimination; repetition	CV natural	no
G	Production training	production training; repetition	CV natural	no
H	Re-test	discrimination; repetition	CV natural	no

Table 2.3: Structure of experiment 1A.

and Bark scale moments. In both cases, the best classification accuracy was achieved with all four spectral moments. The linear model (overall accuracy = 75.1%, dental accuracy = 84.1%, retroflex accuracy = 66.7%) slightly outperformed the Bark model (overall accuracy = 74.3%, dental accuracy = 80.5%, retroflex accuracy = 68.4%). In both cases, the dental stimuli were better classified than the retroflex stimuli. The inclusion of speaking style (careful vs. natural) did not improve the classification accuracy of either model. As a whole, the results indicate that burst spectral properties are also an important cue to place of articulation of the stimuli.

Experiment procedure

Experiment 1A consisted of eight sessions of testing and training, designed to teach subjects the novel contrasts and gauge their progress at several points along the way. The study design involved four major components which are hypothesized to improve learning of a non-native contrast: repeated exposure (through multiple days of training), explicit feedback on discrimination performance, an adaptive fading paradigm during training, and explicit training on articulation of novel phonemes. Table 2.3 lays out the content of each session with respect to task and stimulus type.

All sessions consisted of one or more of three independent tasks. The perception task was an AX discrimination task which gauged how well subjects could discriminate between pairs of Hindi consonants. Subjects heard two syllables and had to judge whether the consonant in each was the same or different. "Same" trials matched two syllables of the same type, but never the same token. Vowels were always the same in both syllables in a trial pair. The discrimination task contained 432 trials, which were always presented in two blocks with a break in-between each block. Participants received feedback about their performance after each trial during sessions B-E.

The production task was a repetition task, and was used to gauge how accurately subjects could articulate each critical consonant. For each trial, subjects heard a single token and were asked to repeat it in a clear voice. Subjects heard each of the 96 unique tokens in the CV natural stimulus set once, in two blocks of 48 trials.

The production training task gave subjects explicit information about the articulation of the Hindi phonemes they had been exposed to throughout the experiment. Training began with an explanation of place of articulation, including instructions on tongue placement for the dental and retroflex consonants, and the introduction of sagittal sections which were color-coded (red for retroflex, green for dental). The colors and sagittal sections were used throughout production training to cue subjects about the place of articulation of the syllable they were hearing. Subjects practiced saying dental and retroflex consonants several times.

After this, voicing was introduced. The English /t/-/d/ (Hindi /t^h/ - /t/) distinction was described first, to introduce subjects to the concept of aspiration. Subjects were instructed to hold their hands in front of their faces while articulating "t" and "d", with attention to the presence or absence of a "puff of air", to further clarify the distinction. Then prevoicing was introduced for the Hindi /d/ consonant, by asking subjects to hold their hand to their throat while making a buzzing sound. Hindi /d^h/ was introduced as a combination of the buzzing and puff of air. To help cement these distinctions, visual cues were introduced alongside the introduction of the voicing types. Finally, subjects were taught to combine the place and voicing features with several practice trials that included both sets of visual cues. In this way, visual cues acted as a kind of external label that learners could use to attach to an articulatory category.

The production training was followed by an enhanced repetition task. This followed the same procedure as the repetition tasks in the pre- and post-test sessions, but also included the visual cues used during the production training, to assist subjects in linking the perceptual information to the articulatory routines they had just learned. Examples of training visuals and enhanced repetition are shown in figure 2.3; the full set of training materials can be found in Appendix A.

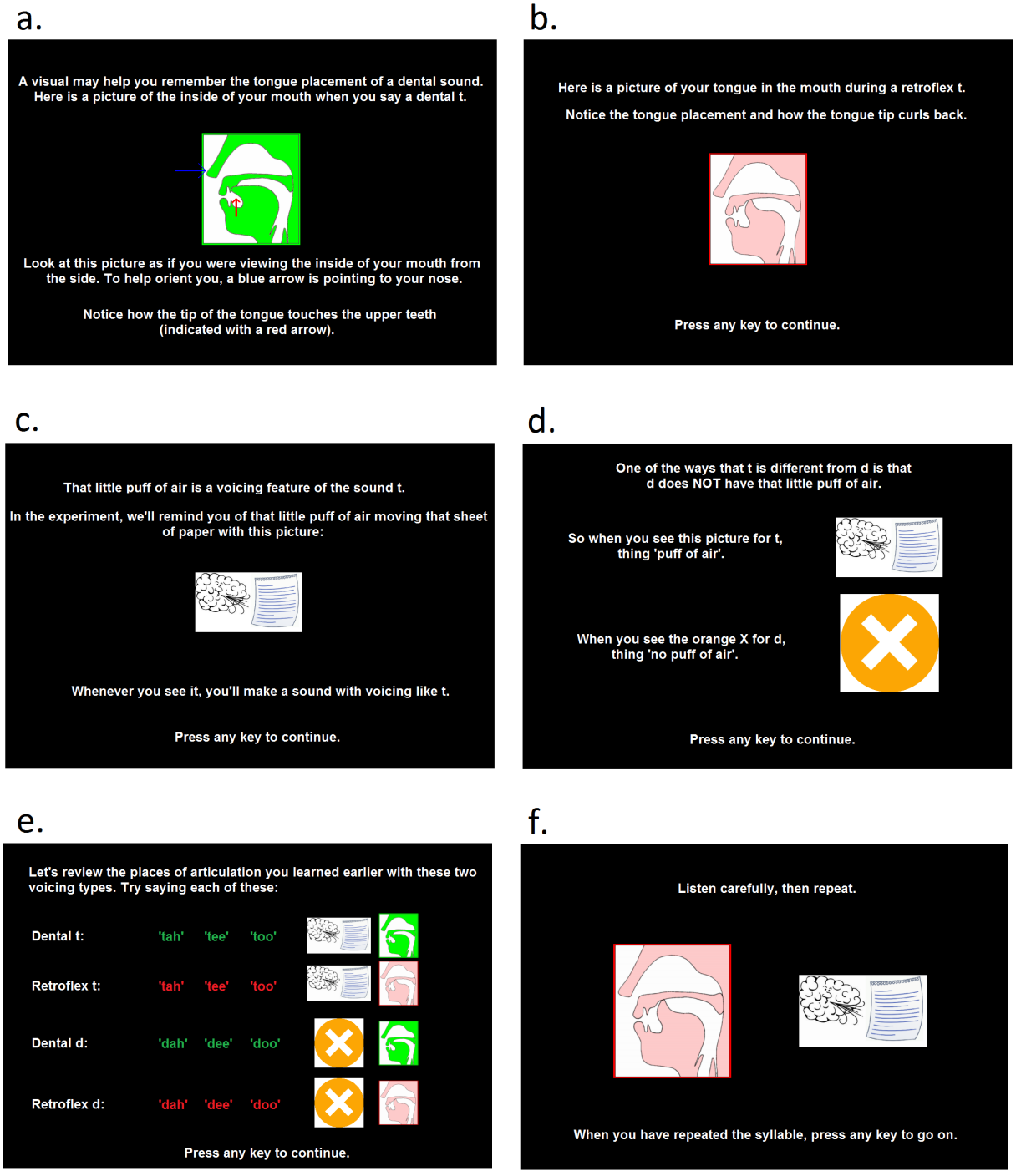


Figure 2.3: Examples from production training. Figure (a) and (b) show training of the dental-retroflex place contrast by introducing subjects to major articulatory landmarks using sagittal sections. The color-coding will consistently follow dental (green) and retroflex (red) tokens throughout training. Figure (c) introduces the concept of aspiration, and a picture to associate with the concept. Figure (d) compares the aspirated "t" to the unaspirated "d" (English orthography). Figure (e) asks subjects to practice combining place and voicing with visual cues. Figure (f) demonstrates a repetition trial, with visual cues.

In each session, subjects interacted with one of the four stimulus types: VCV careful, VCV natural, CV careful, or CV natural. All tests and production training used the CV natural stimuli; the other stimuli types were used during sessions B-E in an adaptive fading paradigm to create a progression from easiest/clearest tokens (VCV careful) to most challenging (CV natural) tokens.

Up to two sessions were run per day, with the restriction that post-test sessions (sessions F and H) could not follow a training session on the same day as that training. This was done to ensure that performance on the post-tests reflected some consolidation of the learned information (Earle & Myers, 2013), rather than simply a boost due to recent exposure to stimuli in the preceding training session. The median number of days to complete the eight-session study was 16 days (range: 7 - 29 days; mean = 15.42 days.)

Experimental sessions were run in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012). All stimuli were presented over headphones. Responses during the discrimination task were recorded using a PST serial response button box. Production data during the repetition task were recorded on either a head-mounted condenser microphone or a stand microphone; microphones were plugged into an AudioBuddy pre-amplifier.

Subjects

Twenty-nine subjects participated in the training study. Of these, eight were excluded (five for failure to complete the entire training series; three for experimenter error), leaving 21 subjects (mean age = 22.6, s.d. = 9.2; 15 women) in the final analysis. Subjects were recruited from a volunteer subject pool maintained by the Berkeley Phonology Lab, as well as flyers posted on the UC Berkeley campus.

Subjects were pre-screened for language background prior to enrollment. Almost all potential subjects have had some exposure to a second language in secondary or post-secondary education, so it was not possible to limit the subject pool to monolingual English speakers. However, potential subjects were excluded if they had any prior experience with Hindi or another language whose phonology contained contrasts which matched the target contrasts in the current study (e.g. a four-way VOT stop contrast, or dental or retroflex stop consonants).

Subjects were paid at a rate of $10/hour. A $20 bonus was paid in addition to the normal hourly rate at the end of session H, to encourage participants to complete the entire study.

2.4.2 Results: Discrimination data

To describe possible improvement in performance as a function of training, three separate models were fit to the data from the three test sessions (sessions A, F, and H). The first, the sensitivity analysis, uses the metric d-prime to describe improved accuracy in detecting differences, scaled by possible false hits during "same" trials. The second, the accuracy analysis, uses a logistic regression model to fit correct and incorrect responses. The third, the

reaction time model, seeks to explain any potential improvement speed in correct responses as a function of training.

Prior to all analyses, outliers were removed from the data on the basis of reaction time. All trials with reaction times less than 100 milliseconds were removed. Trials were also removed if their reaction time fell outside of 2 standard deviations of each subject's mean. Using this procedure, 93.1% of the data from test sessions (sessions A, F, and H) was retained for further analysis.

Sensitivity analysis

D-prime (d') is a discrimination metric from signal detection theory that takes into account both sensitivity (the ability of a listener to detect a difference) and bias (the tendency of a listener to respond "different" regardless of whether a difference is present or not). (For details on how d' is calculated, see Macmillan & Creelman, 1991.) In this way, d' provides protection against the possibility that accuracy in "different" trials will be falsely inflated by a listener learning to prefer "different" over "same" as a response.

For the present analysis, a d' value was calculated for each unique combination of subject (N = 21), test session (3), and contrast type (4 - place contrast, voicing contrast, place + voicing contrast, or "same"). This created a data set of 252 unique values. A linear mixed-effects regression model was constructed in `R` (version 3.1.2, R Core Team, 2014) using `lmer` function in the `lme4` package (version 1.1-7, Bates & Maechler, 2014), with d' as the dependent variable. The model included fixed effects for test session (A, F, or H) and a general measure of trial contrast type (place, voicing, place + voicing, or same). A similar model was fit with the interaction of the two fixed effects, but its inclusion did not improve the model fit ($\chi^2(4) = 3.804$, $p = 0.433$). The model also included a random intercept for subject[2]. In `lme4` syntax, the model was specified as:

```
dPrime ~ session + contrast + (1 |subject)
```

After the model was fit, it was examined for potential outliers and influential cases. A Shapiro-Wilk test for normality showed that the residuals of the original model were not normally distributed ($W = 0.9651$, $p = 0.0001$). Using the `romr.fnc` function in the `LMERConvenienceFunctions` package (A. Tremblay & Ransijn, 2013), 5 data points were identified as having residuals greater than 2.5 standard deviations above 0. When these data points were removed, the model was re-fit. The updated model were identified as being normally distributed ($W = 0.9882$, $p = 0.1274$). Density plots and Q-Q plots of the original and updated model residuals are shown in figure 2.4.

[2] More complex random effects structures were ruled out, as there was a limited number of data (9 points per subject). Preliminary models which tested the inclusion of random slopes showed high correlations between random effects, suggesting that this approach was overfitting the data.

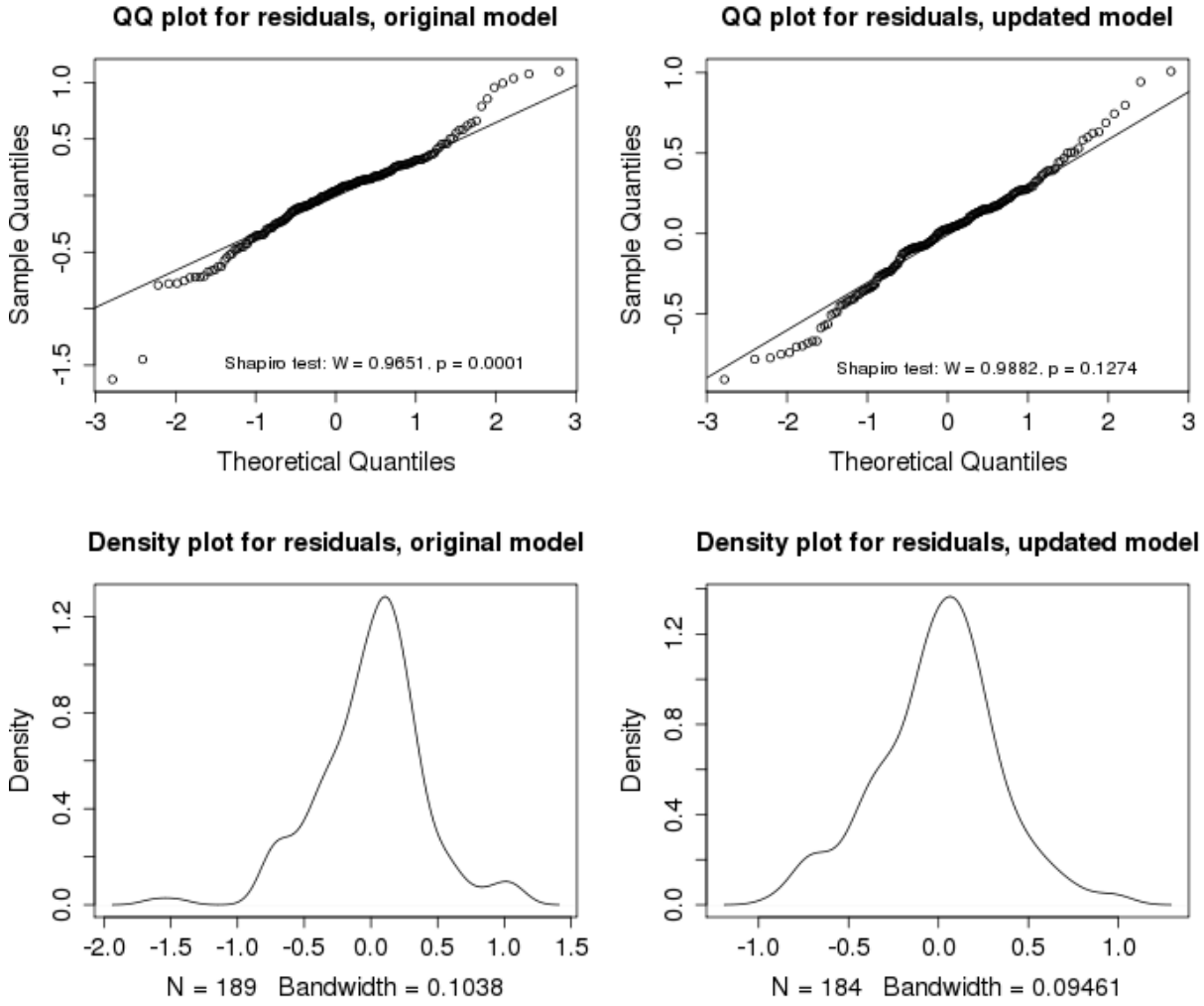


Figure 2.4: Diagnostics of sensitivity model residuals, experiment 1A.

Table 2.4 reports the fixed effects of the model. There was a significant effect of session; pairwise comparisons[3] revealed a significant increase in d' between sessions A and F (post-test - pre-test; $\beta = 0.392$, *p adj.* < 0.001) as well as between sessions H and A (re-test - pre-test; $\beta = 0.383$, *p adj.* < 0.001), but not between sessions H and F (re-test - post-test; β = -0.009, *p adj.* = n.s.). There was also a significant simple effect of contrast type. Post-hoc comparisons found significant differences between all contrast types (place + voicing - place, $\beta = 1.721$; voicing - place + voicing, β = -0.282; voicing - place, $\beta = 1.439$; all *p adj.* $<$

[3] These and all further pairwise comparisons were corrected for multiple comparisons using Tukey's Honest Significant Difference test.

predictor	β	s.e.	df	t	p
(Intercept)	1.401	0.114	39.4	12.272	< 0.001
session					
F (post-test)	0.392	0.076	164.0	5.164	< 0.001
H (re-test)	0.383	0.076	164.0	5.042	< 0.001
contrast type					
place + voicing	1.721	0.076	164.0	22.658	< 0.001
voicing	1.439	0.076	164.0	18.943	< 0.001

Table 2.4: Fixed effects of sensitivity (d-prime) model, experiment 1A. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations.

0.001). As can be seen in figure 2.5, trials with place + voicing contrasts were most readily detected, followed by voicing contrasts; place contrasts were the hardest to detect.

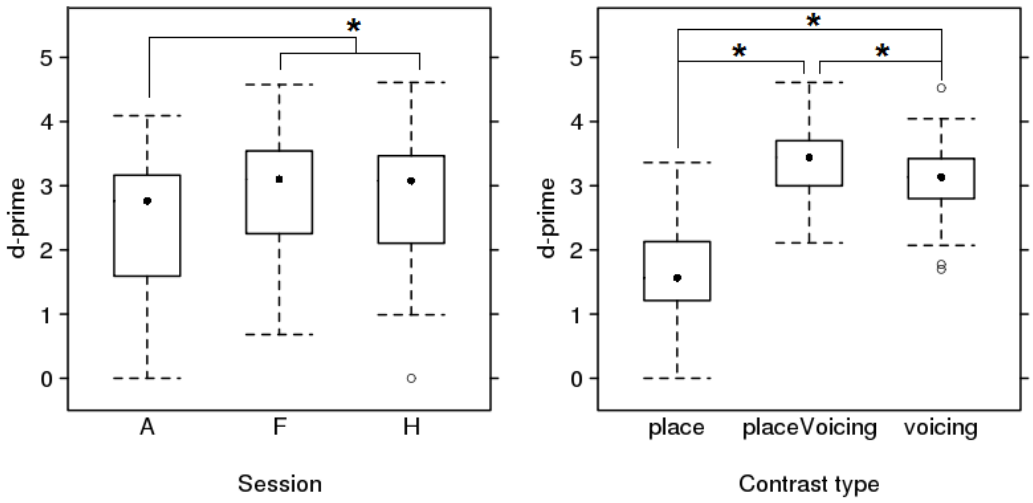


Figure 2.5: D-prime by session and contrast, experiment 1A. Significant contrasts are signaled with stars.

The sensitivity analysis reveals that learners did improve their ability to detect contrasts as a function of perceptual training (session A to session F), and retained that improvement throughout the experiment (session A to session H). However, there was no detectable improvement as a function of production training (session F to session H). The analysis also shows that the contrast types differed in their detectability. Place contrasts were the hardest to detect; voicing contrasts (when taken as a whole) were much more reliably detected. The combination of two features (place + voicing) allowed for even more reliable discrimination.

Accuracy analysis

The sensitivity analysis is a careful metric of improvement; using d' as a dependent variable ensures that enhanced discrimination is not solely a function of increased "different" responses. However, it limits the power of the analysis, because it requires averaging over large groups of data in order to calculate d'. This lack of power limits the investigation of other covariates, may have led to a failure to detect a session * contrast interaction, and certainly precludes a more complex random effects structure which could more effectively control by-subject and by-item variance.

To complement the sensitivity analysis with an analysis with more degrees of freedom, a mixed-effects logistic regression model (fit with the `glmer` function in `lme4`) was run with accuracy ("correct" or "incorrect") as the dependent variable.

Model selection proceeded in the following way: for fixed effects, all covariates which were thought to have hypothesized value as predictors or control variables were initially entered into the model, with interactions included when there was an explicit hypothesis about that interaction term. The random effects structure was initially specified with random intercepts for stimulus 1, stimulus 2, subject, and by-subject random slopes for session[4]. Once this initial model had been fit, the coefficients and estimated p-values of the fixed effects were evaluated, and any which did not significantly contribute to the model were removed. For predictors which were on the edge of significance, models were run with and without the predictor included, and compared via log-likelihood tests; if the model with the predictor did not significantly improve model fit, it was held out.

Once the fixed effects structure was determined, a sanity check model was run, which removed the random slopes term. This non-slopes model was compared to the slopes model, again with a log-likelihood test, and this confirmed that inclusion of the slopes term was justified.

The following predictors were included in the final model:

- **Session** (3-level factor, levels = A, F, H): As in the sensitivity analysis, this factor captures performance at baseline (session A, pre-test), post-perception training (session F, post-test), and post-production training (session H, re-test). Improvement as a function of either training set (perception or production) would be reflected in greater accuracy for session F or session H than the previous session.

- **Contrast** (15-level factor): This factor represents trial contrasts in a more fine-grained way than the factor used in the sensitivity analysis. In addition to place contrasts (dental vs. retroflex), specific voicing contrasts (e.g. breathy vs. voiced, aspirated vs. breathy) are each represented by a factor level. In addition, the combination of each

[4]In preliminary model exploration, by-subject random slopes for contrast type were included, both with the session random slopes and as the sole random slope. None of these models converged, suggesting that the contrast type random slopes led to model overfitting.

voicing contrast and place contrasts (e.g. "place + unaspirated vs. aspirated" for a trial with stimuli [t̪a] and [t̪ʰa]) is a factor level. Importantly, "same" trials were also included in this analysis. While not entirely equivalent to the bias measured by d', improvement in "same" trial accuracy across sessions (see session * contrast interaction) would indicate that subjects are resisting a bias towards responding "different" to all trials.

- **Session * contrast**: This interaction term did not improve the model in the sensitivity analysis; however, this may have been an issue of power. It is hypothesized that some contrasts will show more evidence of learning than others, either because (1) some contrasts are better targeted by the design of the training or because (2) some contrasts may start at a higher level of discriminability than others, and may hit a ceiling earlier in the learning process.

- **Days to completion** (numeric covariate, range = 7 - 29): Due to scheduling constraints, subjects varied in the length of time needed to complete all eight sessions. It was hypothesized that a longer time to complete the study could adversely impact accuracy, as training/exposure would be reinforced less recently than subjects who completed all sessions in a shorter span of time.

- **Trial count** (numeric covariate, range = 0.00 - 4.31)[5]: Trial count indexes the order of a trial within a session (each had 432 trials). If learning is detectable over the course of a session (and not just through consolidation of learning after a session has completed), then accuracy should increase as trial number increases.

- **Random effects**: Random intercepts were included for stimulus 1 and stimulus 2 (the two stimuli in each trial), to capture any variance attributable to non-categorical differences in the stimuli. A by-subject random slope for session was included, to capture any variance across subjects in learning across sessions[6].

In addition, **vowel** (/a/, /i/, or /u/) and **trialCount * session** were examined during model selection. These predictors did not improve the model and were ultimately excluded.

The final model specification was:

```
correct ~ session * contrast + trialCount + daysToCompletion + (1 |stim1) +
        (1 |stim2) + (1 + session |subject), family = ''binomial''
```

[5]The model failed to converge properly when this predictor was entered with its original range of 0-431; scaling it down by a factor of 100 allowed the model to reach convergence.

[6]This was specified in the model syntax as `(1 + session |subject)`. There was no separate `(1 |subject)`, as that redundantly specified the random intercept for subject and prevented convergence. A decorrelated combination of these two - `(1 |subject) + (0 + session |subject)` was also attempted, but this too failed to converge

predictor	β	**s.e.**	***t***	***p***
(Intercept)	0.622	0.301	2.065	0.039
Session (base level: A)				
F	1.366	0.193	7.066	< 0.001
H	1.281	0.203	6.316	< 0.001
Contrast (base level: Aspirated vs. breathy)				
Place	-1.640	0.136	-12.090	< 0.001
Same	2.189	0.165	13.241	< 0.001
Aspirated vs. voiceless	3.126	0.256	12.189	< 0.001
Breathy vs. voiced	1.816	0.189	9.608	< 0.001
Voiced vs. voiceless	-1.518	0.164	-9.259	< 0.001
Place + aspirated vs. breathy	0.396	0.145	2.724	0.006
Place + aspirated vs. voiced	3.483	0.291	11.951	< 0.001
Place + aspirated vs. voiceless	2.210	0.209	10.567	< 0.001
Place + breathy vs. voiced	0.377	0.156	2.161	0.031
Place + breathy vs. voiceless	0.469	0.161	2.905	0.004
Place + voiced vs. voiceless	-1.518	0.164	13.241	< 0.001
Days to completion	-0.024	0.012	-2.018	0.044
Trial count	0.069	0.018	4.775	< 0.001

Table 2.5: Selected coefficients of accuracy model for significant simple effects only, experiment 1A. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations. Interaction terms are excluded from this table.

Significant simple effects of the model are reported in table 2.5. There was a significant effect of session ($F(2) = 9.317$, $p < 0.001$) as well as a significant effect of contrast type ($F(14) = 212.454$, $p < 0.001$). Because these two factors entered into an interaction, the discussion of their contributions is limited to the breakdown of that interaction below.

The accuracy analysis was able to control for within-session and across-session variability that was not examined in the sensitivity analysis. A significant negative effect of days to completion ($\beta = -0.024$, $t = -2.018$, $p = 0.044$) revealed that subjects who took longer to complete the training were less accurate than those who completed the study in fewer days. A positive effect for trial count ($\beta = 0.069$, $t = 4.775$, $p < 0.001$) was also found, confirming that subjects became more accurate as a session went on, indicating within-session learning. Importantly, because explicit feedback was not present during the test sessions included in this analysis, this suggests that implicit learning via exposure was the driver of this effect.

The significant session * contrast interaction addresses the hypothesis that different contrasts may be more or less affected by training. Table 2.6 shows each session * contrast interaction level of interest (i.e. all within-contrast session comparisons), and their significance as assessed by post-hoc pairwise comparisons. It is apparent that most contrasts saw improvement as a result of perception training (pre-test to post-test) and that these

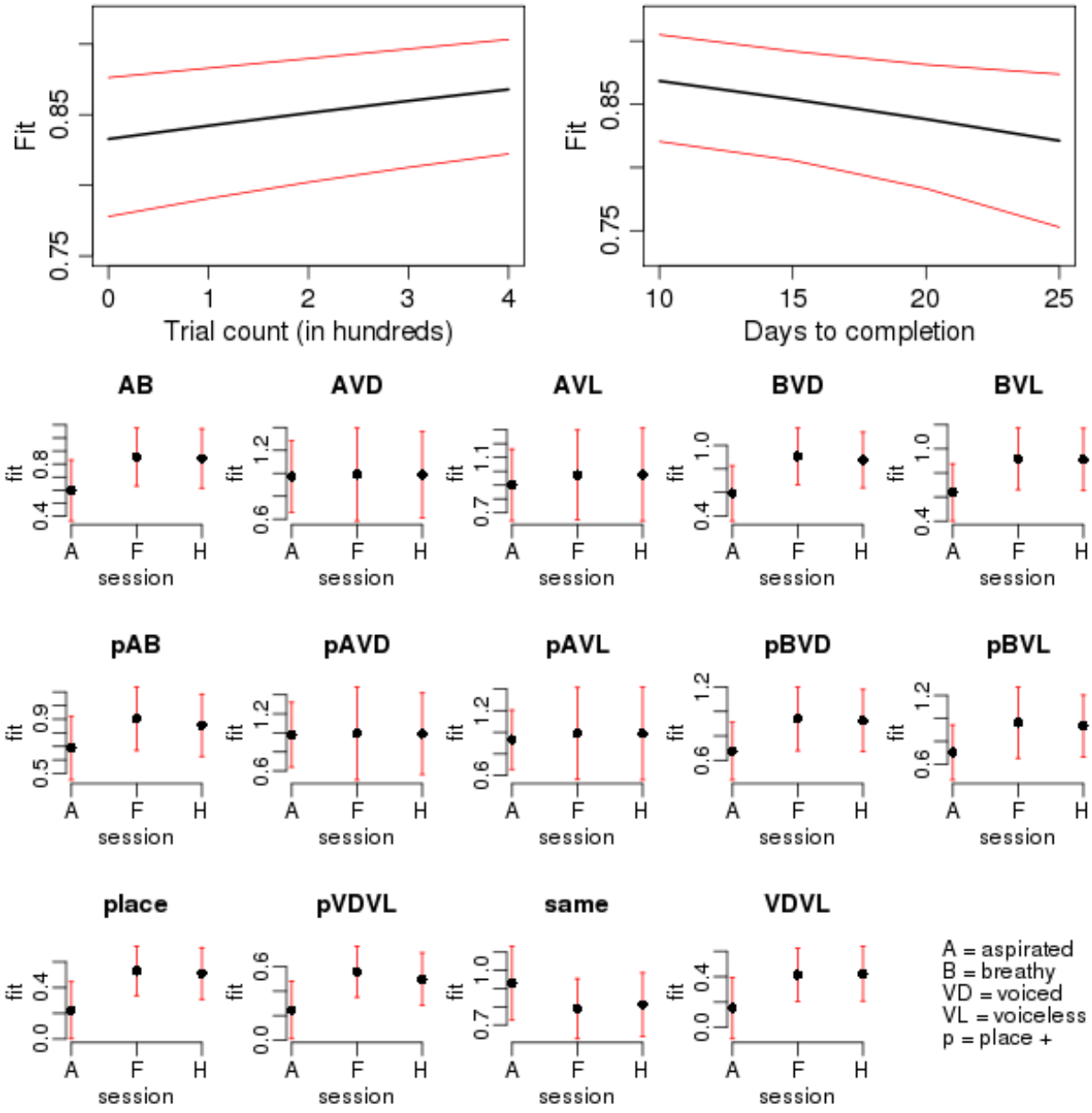


Figure 2.6: Partial effects plot for the fixed effects in the accuracy model, experiment 1A.

improvements were maintained through the duration of the study (pre-test to re-test). It is also clear that production training (post-test to re-test) had no impact, aside from a single marginal effect, on any contrast. This reinforces the conclusion from the sensitivity analysis that production training did not contribute to increased discrimination.

Contrast	**Pre-test to post-test** (Perception training)	**Pre-test to re-test** (All training)	**Post-test to re-test** (Production training)
Increased accuracy			
place	*	*	
aspirated vs. breathy	*	*	
aspirated vs. voiced	(.)		
aspirated vs. voiceless	*	*	
breathy vs. voiced	*	*	
breathy vs. voiceless	*	*	
voiced vs. voiceless	*	*	
place + aspirated vs. breathy	*	*	(.)
place + aspirated vs. voiced			
place + aspirated vs. voiceless	*	*	
place + breathy vs. voiced	*	*	
place + breathy vs. voiceless	*	*	
place + voiced vs. voiceless	*	*	
Decreased accuracy			
same	*	*	

Table 2.6: Contrast * session pairwise comparisons, accuracy model, experiment 1A. The first column shows improvement as a result of perception training (session A - F). The second column indicates improvement over the course of the whole study (session A - H). The third column shows improvement from production training (session F - H). All listed contrasts which were significant at p *adj.* < 0.05 (as assessed with the Tukey correction for multiple comparisons) are indicated with *. Marginal comparisons, with p *adj.* between 0.05 - 0.10, are marked with (.).

Two contrast trials did not show significant improvement at any stage: aspirated vs. voiced trials, and place + aspirated vs. voiced trials. This can be attributed to the fact that /t^{h}/ and /d/ map cleanly on to the English categories /t/ and /d/, and thus this contrast was already accurately perceived by learners who are native speakers of English. This is supported by the error data from trials where subjects incorrectly responded "same" to a voicing contrast (figure 2.7). In this subset of the data, there were comparatively fewer errors to voiced vs. aspirated trials than other voicing contrasts.

The comparisons also revealed that accuracy in "same" trials dropped over the course of the study - the type of bias that the d' analysis is designed to control. As a result, the interpretation of the post-hoc comparisons must be interpreted with some caution, as this finding suggests that subjects were learning to respond "different" as a result of training.

Discussion The accuracy analysis reinforces the finding from the sensitivity analysis; namely, discrimination increased from pre-test to post-test, but not from post-test to re-test

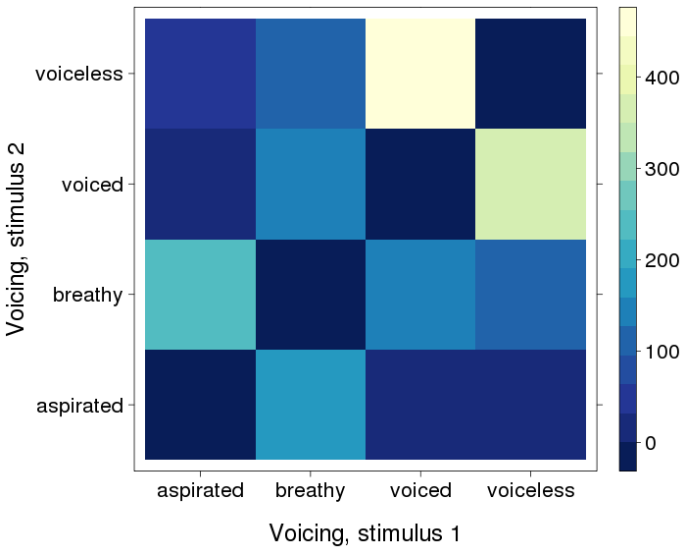


Figure 2.7: Number of voicing contrast trials identified as "same" trials, experiment 1A.

(except for a single marginal contrast). In addition, the fine-grained contrast information permitted by the increased power of the model reveals that most, but not all contrasts, saw an improvement - although those with high discriminability at pre-test did not improve over time. This finding is tempered by the finding that accuracy for "same" trials decreased over the experiment, suggesting that subjects may have been learning a "different" bias.

Finally, the analysis revealed two important points about the time course of learning. First, improved accuracy can be seen over the course of a session, not just between sessions. Because no explicit learning metrics were introduced during the sessions included in this analysis, it appears that exposure is the key factor to this effect. This may suggest that as subjects are listening to trials, they are building or refreshing a stronger categorical representation that allows them to perform more accurately on later trials. Second, a longer time complete the entire session was associated with lower accuracy, suggesting that performance is strengthened by recent exposure to the target sounds.

Reaction time analysis

Another possible metric of improvement in detecting a contrast is speed of response. Even if a listener can detect a different from early in training, he or she may become faster at making a judgment through increased experience with the target contrasts. To test this, a linear mixed-effects model was run on all trials during test session (A, F, H) where subjects gave the correct response, with reaction time (log-transformed) as the dependent variable.

Model selection proceeded in the same way as described for the accuracy analysis, with

predictor	β	**s.e.**	t	p
(Intercept)	6.784	0.058	117.527	< 0.001
Session (base level: A)				
F	-0.537	0.063	-8.490	< 0.001
H	-0.650	0.063	-10.284	< 0.001
Contrast (base level: Aspirated vs. breathy)				
Place	0.121	0.049	2.441	0.015
Same	-0.182	0.038	-4.740	< 0.001
Aspirated vs. voiced	-0.086	0.044	-1.938	0.053
Voiced vs. voiceless	0.180	0.065	2.768	0.006
Place + aspirated vs. voiced	- 0.129	0.044	-2.931	0.003
Place + aspirated vs. voiceless	-0.086	0.045	-1.940	0.052
Trial count	- 0.075	0.006	-12.269	< 0.001

Table 2.7: Selected coefficients of reaction time model for significant or marginal simple effects only, experiment 1A. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations. Interaction terms are excluded from this table.

the same initial candidate predictors considered[7]. The final model specification was the following:

```
log(RT) ~ session * contrast + trialCount * session +
   (1 |stim1) + (1 |stim2) + (1 + session |subject)
```

The reaction time model differed from the accuracy model in two key ways: there was no effect of days to completion, and there was a significant interaction between session and trial count. While the models are not directly comparable, as the reaction time model was constructed on a subset of the data (correct trials only), this suggests that accuracy and speed may operate as partially distinct indicators of learning which are impacted in different ways.

Coefficients for significant simple effects are reported in table 2.7. The effect of session was again significant ($F(2, 32.8) = 85.527$, $p < 0.001$), indicating that reaction time differed between sessions. There was also an effect of contrast type ($F(13, 2797.7) = 31.872$, $p < 0.001$), indicating that some contrasts had quicker responses than others. The contrast * session interaction will be discussed in more detail below.

[7]Model criticism for this analysis followed the same procedure as for the sensitivity analysis. However, while the procedure identified 391 data points (2.0% of the total data set) with residuals greater than 2.5 standard deviations above 0, the removal of these points did not sufficiently improve the normality of the residuals, as assessed with a two-tailed Kolmogorov-Smirnoff test (original model, $W = 0.1684$, $p < 0.0001$; updated model, $W = 0.1803$, $p < 0.0001$). As a result, the original model is reported.

Contrast	**Pre-test to post-test** (Perception training)	**Pre-test to re-test** (All training)	**Post-test to re-test** (Production training)
Place	*	*	
Aspirated vs. breathy	*	*	
Aspirated vs. voiced	*	*	
Aspirated vs. voiceless	*	*	
Breathy vs. voiced	*	*	
Breathy vs. voiceless	*	*	
Voiced vs. voiceless	*	*	
Place + aspirated vs. breathy	*	*	
Place + aspirated vs. voiced	*	*	
Place + aspirated vs. voiceless	*	*	
Place + breathy vs. voiced	*	*	
Place + breathy vs. voiceless	*	*	
Place + voiced vs. voiceless	*	*	
same		*	

Table 2.8: Contrast * session post-hoc comparisons, reaction time model, experiment 1A. The first column shows decreased speed as a result of perception training (session A - F). The second column indicates decreased speed over the course of the whole study (session A - H). The third column shows decreased speed from production training (session F - H). All listed contrasts which were significant at p *adj.* < 0.05 are indicated with *.

There was a simple effect of trial count, which revealed that reaction times decreased as trial count increased (β = -0.075, t = -12.269, $p < 0.001$), indicating a general speeding up over the course of the session. There was also a significant trial count * session interaction; pairwise comparisons revealed that the effect of trial count was different between sessions F and A ($\beta = 0.051$, $t = 6.52$, p *adj.* < 0.001) and sessions H and A ($\beta = 0.038$, $t = 4.836$, p *adj.* < 0.001), but not between sessions H and F (β = -0.013, t = -1.719, p *adj.* = n.s.). This indicates that the slope of the trial count effect was more pronounced in session F and H than in session A; in other words, the decrease in reaction time over the course of the session was more pronounced in these later sessions.

The session * contrast interaction is visualized in figure 2.8 and summarized in table 2.8, which catalogs all within-contrast pairwise comparisons of session. For nearly every contrast, reaction time decreased from pre-test to post-test (session A - F), as well as from pre-test to re-test (session A - H), indicating that subjects' responses got faster as a function of training. For "same" trials, the effect only reached significance from pre-test to re-test. No contrast increased from post-test to re-test, indicating that production training did not contribute to faster responses.

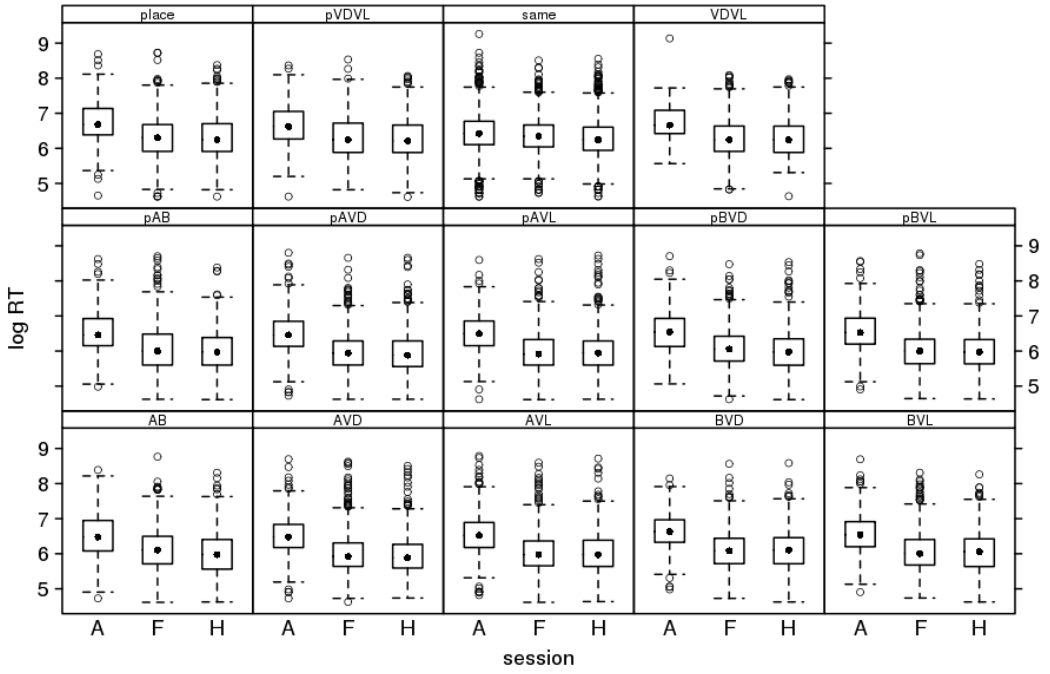


Figure 2.8: Reaction time by session and by contrast, experiment 1A.

2.4.3 Results: Production data

Production data was collected from the pre-test, post-test, production training, and re-test sessions, to assess subjects' ability to articulate the target sounds. Twenty subjects were included in the production data analysis. (One subject included in the discrimination data analysis was excluded from the production data analysis due to recording errors.) Across 20 subjects, 7263 tokens were recorded, transcribed, and analyzed. (For detailed information about transcription procedures, see section 2.4.1.) Tokens were excluded from analysis if there was too much background noise, if the attempt to produce the syllable resulted in a consonant other than a stop, or if the production of the syllable was cut off prior to the end of the production.

Acoustic analysis of the production data mirrored the acoustic analysis of the stimuli (see section 2.4.1). Results for properties assessing subjects' ability to produce the target voicing contrasts (VOT analysis) and the target place of articulation contrast (formant analysis, burst analysis) are discussed below.

Voicing: VOT analysis

Two measures of voice onset time (VOT) were taken from syllables produced during the repetition task: positive VOT, or aspiration, and negative VOT, or pre-voicing. Positive VOT was defined as the time from the onset of the burst to the onset of the vowel. Negative VOT

was defined as the time from the onset of any visible pre-voicing to the onset of the burst. In practice, almost all syllables (99.8%, 7245 of 7263) had a positive VOT period identified, as nearly all stop productions had at least an identifiable burst. Many fewer syllables were identified as having negative VOT (15.6%, 1132 of 7263). All syllables produced by subjects during the repetition task were consonant-vowel (CV) syllables; as a result, there was no closure duration measurement (defined as the duration of closure between two vowels in an intervocalic stop) comparable to the one measured for VCV stimuli. Extreme values (defined as any VOT measurement above 250 ms) were removed from the data set prior to analysis.

To analyze the VOT data, two mixed-effects models were constructed - one for the negative VOT data, and one for the positive VOT data. In each case, VOT was modeled as a function of the interaction of session (pre-test, post-test, production training, or re-test) and voicing type (voiceless, voiced, aspirated, or breathy), with a random intercept for subject to control for individual variation in speaking rate. Post-hoc comparisons using Tukey's HSD test were used to compare within-voicing differences between sessions, to examine whether positive or negative VOT changed from session to session in any voicing category. Five specific comparisons were inspected: pre-test to post-test, post-test to production training, production training to re-test, pre-test to production training, and pre-test to re-test (the latter two to test any differences between baseline performance and later performance). By-voicing VOT data is visualized in figure 2.9.

Prevoicing is not a feature of utterance-initial stop consonants in English. Therefore, if subjects learned new information about Hindi voicing targets and were able to apply it to their own productions, then it is expected that negative VOT durations in breathy and voiced syllables, which both have prevoicing, would increase as a function of training. Post-hoc comparisons revealed that for negative VOT (pre-voicing), differences were confined to breathy syllables, with an increase in negative VOT from pre-test to production training ($\beta = 31.581$, $t = 4.443$, $p\ adj. = 0.0007$) and from post-test to production training ($\beta = 24.339$, $t = 3.559$, $p\ adj. = 0.0244$). This indicates that compared to baseline and post-perceptual-training performance, breathy voiced consonants produced during production training had longer prevoicing, meaning that subjects were able to achieve this voicing target more reliably during this session. The lack of any effect of the re-test session suggests that this may not have been robustly maintained after training was complete. There was no similar increase in negative VOT for voiced syllables, which reveals that subjects did not successfully apply prevoicing to this category of sounds.

For positive VOT, the relevant category of interest is breathy consonants. The Hindi aspirated and voiceless categories map their positive VOT straightforwardly onto English /t/ and /d/, respectively. Hindi's breathy category tends to assimilate to the English /d/, but the English /d/ has a short-lag positive VOT (typically 10-30 ms), while breathy consonants in Hindi have a long positive VOT (the "breathy" or voiced aspirated portion of the consonant). If subjects learn to produce breathy consonants more accurately as a function of training, then their productions of these consonants should show longer positive VOT after training, as they transition from an English-like /d/ to a Hindi-like breathy consonant.

However, there was no significant difference in positive VOT duration for the breathy

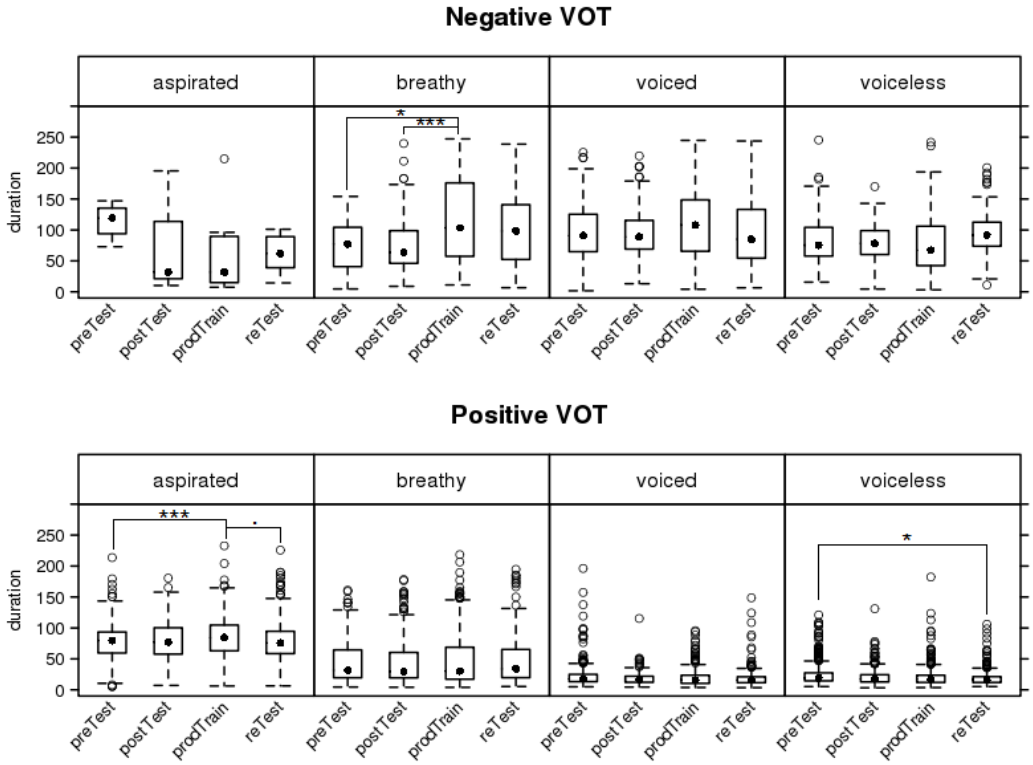


Figure 2.9: Positive and negative VOT durations for production data by voicing type, experiment 1A. Significant within-voicing differences between sessions are indicated with significance stars (. = $p < 0.1$, * = $p < 0.05$, *** = $p < 0.001$).

consonants after any training session, suggesting that subjects did not effectively learn to produce this feature of breathy consonants. The only significant or marginal differences in positive VOT came for aspirated consonants (pre-test to production training, β = 7.519, t = 4.686, p *adj.* = 0.0003; post-test to production training, β = 5.837, t = 3.594, p *adj.* = 0.028; production training to re-test, β = -5.194, t = 03.210, p *adj.* = 0.096) and voiceless consonants (pre-test to re-test, β = -5.577, t = -3.466, p *adj.* = 0.044). These categories mirror English categories in their VOT durations, so these differences reflect greater precision in articulation of familiar categories, rather than the learning of new categories.

Place of articulation

Two metrics were used to assess the acoustics of place of articulation of the Hindi stimuli: spectral properties of the burst, and formant frequencies at vowel onset and midpoint. Those analyses are reproduced below with the repetition data, with the aim of assessing how similar subjects' productions were to capturing the acoustic properties found in the stimuli spoken by a native Hindi speaker.

Session	Overall accuracy	Dental accuracy	Retroflex accuracy
Pre-test	50.96%	43.35%	58.47%
Post-test	57.14%	57.92%	56.36%
Production training	60.57%	60.35%	60.78%
Re-test	49.10%	50.68%	47.51%

Table 2.9: Classification accuracy of place of articulation for repetition data in experiment 1A. Each token was classified as dental or retroflex as a function of F2 onset, F3 onset, and F3 midpoint.

Formant analysis Following the analysis of the stimuli (see section 2.4.1), linear discriminant analysis was used to classify each unaspirated token as dental or retroflex. The same properties used to classify the stimuli were used for the repetition data: F2 onset, F3 onset, and F3 midpoint. Classification accuracy at all four sessions where repetition data was collected is reported in table 2.9.

Classification accuracy at pre-test was at chance (50.96%) across both place categories, indicating that subjects did not reliably distinguish between articulation of dental and retroflex consonants. Accuracy improved to 57.14% at post-test, suggesting that the perceptual training had contributed to some modest improvement in place of articulation production targets. Classification accuracy was highest during production training (60.57%), when subjects had access to visual cues and explicit instruction about articulation. It should be noted that the classification accuracy for the native speaker-produced stimulus was 83.3% - subjects did not reach this effective ceiling during their best session, indicating that there is still room for their articulatory targets to improve; however, their performance during this session does indicate an ability to distinguish between the two targets. Perhaps surprisingly, classification dropped back to chance at post-test (49.10%), indicating that subjects did not retain what they had learned with explicit articulatory instruction once those cues were removed.

Burst analysis The burst analysis for the stimuli considered eight spectral properties of the stop burst: centroid, standard deviation, skew, kurtosis, and their bark-transformed counterparts, as an alternative means to investigating place of articulation across all voicing categories. Following that analysis, Wilcoxon rank-sum tests were run on the dental-retroflex distinction in the repetition data for each of these metrics in each of the four sessions. Significant dental-retroflex differences ($p < 0.05$) are summarized in table 2.10, and by-session metrics are visualized in figure 2.10.

The tests summarized in table 2.10 indicate that many of the acoustic characteristics which distinguish dental and retroflex consonants in the stimuli are also present in the repetition data. Centroid and skew are reliably distinguished in both their linear and log transformations across sessions; bark (but not linear) skew is as well. The feature which evolves the most over training is standard deviation, which does not reliably distinguish dentals and retroflexes in the production data except at production training (linear and

	Pre-test	Post-test	Production training	Re-test
centroid	*	*	*	*
standard deviation			*	
skew	*	*	*	*
kurtosis				
bark centroid	*	*	*	*
bark standard deviation		*	*	
bark skew	*	*	*	*
bark kurtosis	*	*	*	*

Table 2.10: Summary of burst spectra analyses, experiment 1A. Stars indicate that a significant dental-retroflex distinction was found for that metric in that session, as assessed by Wilcoxon rank-sum tests.

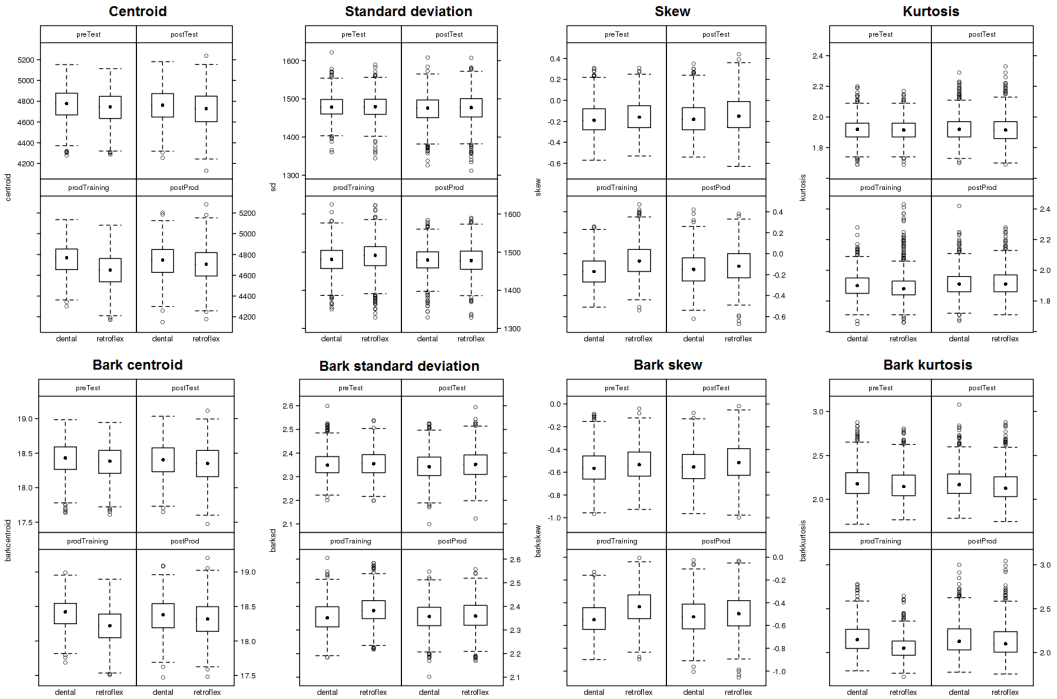


Figure 2.10: Burst spectra properties for repetition data, experiment 1A.

bark) and post-test (bark only).

To complement the individual tests of each spectral moment, an LDA classification analysis was performed for the reproduction data, with all four spectral moments predicting place of articulation at each session. The classification results are summarized in table 2.11. Results were very similar for the linear and bark metrics, so the linear metrics are reported here to be consistent with the stimuli analysis.

Session	Overall accuracy	Dental accuracy	Retroflex accuracy
Pre-test	53.92%	59.35%	48.41%
Post-test	55.05%	61.09%	49.08%
Production training	66.29%	70.42%	62.17%
Re-test	53.81%	65.37%	42.17%

Table 2.11: Classification accuracy of place of articulation for repetition data in experiment 1A. Each token was classified as dental or retroflex as a function of four spectral moments of the burst (centroid, skew, standard deviation, and kurtosis).

As with the formant analysis, the highest classification accuracy was achieved during the production training session, with 66.29% of stimuli corrected classified as dental or retroflex. Again, this is not as high as the classification of the stimuli for these metrics (75.1%), but it is a modestly successful classification. Once again, the major benefits of production training appear not to carry over to the re-test, where accuracy drops to 53.81%. Interestingly, while the formant analysis was a better predictor of retroflex segments, the burst analysis appears to favor the classification of dental segments.

2.4.4 Discussion

The analyses of experiment 1A reveal that subjects successfully improved their discrimination of nearly all target contrasts as a function of perception training (sessions B - E). Both accuracy and reaction time showed performance improvement from pre-test to post-test; contrasts which did not show improvement were restricted to those which were already phonemic in English, the speakers' native language.

Conversely, production training did not seem to affect performance as measured in any of the three analyses; there were no appreciable differences in sensitivity, accuracy, or reaction time between sessions F and H. This suggests that the articulatory information provided during production training did not contribute to the fledgling category representations that were learned during perception training.

The production data suggests that subjects were able to make some changes in their articulation of the target sounds as a function of training, but these changes were not uniform across all metrics investigated. For the voicing contrast (arguably the harder articulatory feature to learn in the present paradigm), the biggest improvement came in breathy consonants, which showed increased negative VOT as a function of training. This change was driven by production training, not perceptual training. Voiced segments showed no difference in positive or negative VOT, suggesting that subjects did not form new articulatory representations of these targets. Some enhancement of positive VOT after training in voiceless (shorter VOT) and aspirated (longer VOT) segments, which are shared with English, may indicate that subjects were refining even familiar categories, perhaps in an attempt to enhance contrasts between the segments.

The place of articulation analysis of the production data indicate that, on the basis of two sets of acoustic features (burst spectral moments and formant transitions), the dental-retroflex contrast was strongest during production training, indicating that subjects were able to approximate the difference between the two segments when they had explicit information about their articulation. The formant analysis also suggests some improvement at post-test, signaling that perceptual training may have made some contribution to subjects' developing articulatory representations of this novel contrast. In both metrics, the improvement shown during production training was not present at re-test, indicating that what was learned was relatively fleeting.

2.5 Experiment 1B: Control study

Experiment 1A was designed to integrate multiple cues (exposure, feedback, adaptive fading, and articulatory information) into a single paradigm. A limitation of this approach is that these interventions are not directly dissociable, as implemented in that study. In particular, there is an open question about whether exposure alone, absent the other cues, would provide any improvement in discrimination. There is reason to suspect that exposure may be relevant: Pegg and Werker (1997) found that exposure in the native language to two phones which are allophones - but which are not phonemically contrastive - can lead to above-chance discrimination, but does not lead to distinct identification. Passive exposure certainly plays a role in infant acquisition (e.g. Maye, Werker, & Gerken, 2002); however, it is less clear to what extent adults can learn perceptual information without explicit or implicit attention (Seitz & Watanabe, 2003). Experiment 1B was thus designed as a control study to address whether passive exposure could lead to any detectable improvement in discrimination for adults in a short-term phoneme learning paradigm.

2.5.1 Methods

Experiment 1B was run in four sessions (see table 2.12). Sessions A, C, and D were identical to the pre-test, post-test, and re-test sessions (A, F, H - see table 2.3) of experiment 1A. Session B of experiment 1B was the same as training session 4 (session E) of experiment 1A, except that there was no feedback in the session. These four sessions were selected because they all use the CV natural stimuli. Restricting the study to this stimulus set controls stimulus complexity, thereby eliminating the adaptive fading manipulation which was present in experiment 1A.

While production data was collected in sessions A, C, and D of experiment 2B in order to be comparable to experiment 1A, there was no production training, and production data was not of interest in this control study. Therefore, the following results report only on results from the discrimination task.

Session code	Session	Tasks	Stimuli	Feedback?
A	Pre-test	discrimination; repetition	CV natural	no
B	Exposure	discrimination	CV natural	no
C	Post-test	discrimination; repetition	CV natural	no
D	Re-test	discrimination; repetition	CV natural	no

Table 2.12: Structure of experiment 1B.

Subjects

Nine subjects (median age = 21, range = 18 - 30; 7 female) participated in the experiment. As in experiment 1A, all subjects reported English as their native language, and had no prior experiment with Hindi, Urdu, or languages with a dental-retroflex contrast or a four-way VOT contrast.

2.5.2 Results

As in experiment 1A, d' was used as a dependent variable to examine a bias-free metric of discrimination. Model selection preceded as follows: the data was first fit with fixed effects of session and general contrast type (place, voicing, or place + voicing), as well as the interaction of session and contrast; it also included a random effect for subject. The interaction was not significant and subsequently removed. When the model was re-fit without the interaction term, the simple effect of session was found to be marginally significant ($F(3,94) = 2.24$, $p = 0.088$). Another model was fit, eliminating the session predictor. The model retaining session as a predictor was a marginally-better fit ($\chi^2 = 6.844$, $p = 0.077$). Because the effect is on the border, and of potential theoretical interest, it was ultimately retained for the analysis below. The final model specification was:

```
dPrime ~ session + contrast + (1 | subject)
```

Coefficients for the fixed effects of the model are reported in table 2.13. The simple effect of session was, as noted above, marginally significant; post-hoc tests revealed that this was driven solely by the difference between session C and A (β = -0.233, t = -2.544, p *adj.* = 0.0534). Importantly, the sign of the effect was negative, indicating that accuracy *decreased* between session A and session C.

The effect of contrast was also significant ($F(2, 94) = 400.74$, $p < 0.001$). Post-hoc comparisons revealed differences between all three types: place + voicing vs. place ($\beta = 2.058$, $t = 25.911$, p *adj.* < 0.001), voicing vs. place ($\beta = 1.814$, $t = 22.833$, p *adj.* < 0.001), and voicing vs. place + voicing (β = -0.245, t = -3.078, p *adj.* = 0.006). This finding replicated the finding of experiment 1A, that place + voicing trials were most accurately perceived, followed by voicing trials, with place trials the least likely to be discriminated.

predictor	β	s.e.	df	t	p
(Intercept)	1.472	0.127	17.040	11.509	< 0.001
Session (base level: session A)					
session B	-0.156	-0.092	94	-1.700	0.093
session C	-0.233	-0.092	94	-2.544	0.013
session D	-0.122	-0.092	94	-1.326	0.188
Contrast (base level: place)					
place + voicing	2.058	0.079	94	25.911	< 0.001
voicing	1.181	0.079	94	22.833	< 0.001

Table 2.13: Fixed effects of the sensitivity (d-prime) model, experiment 1B. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations.

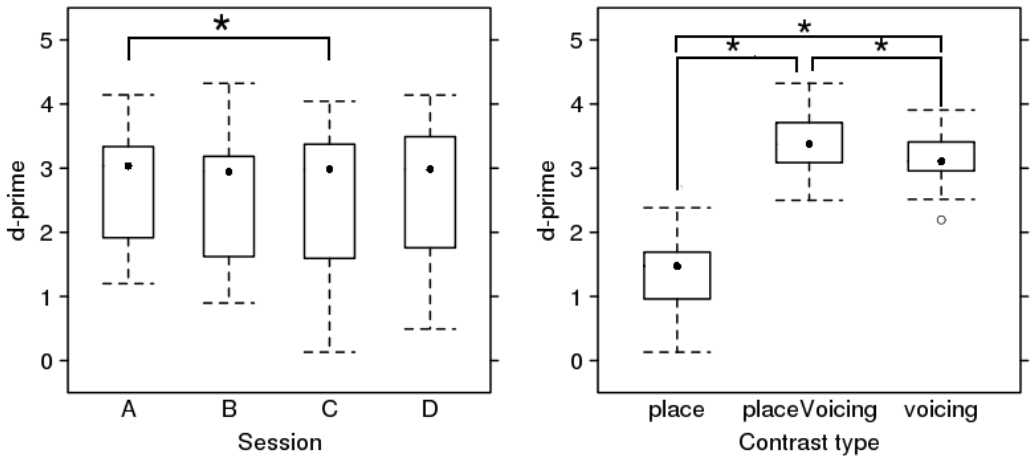


Figure 2.11: D-prime by session and by contrast, experiment 1B.

2.5.3 Discussion

The sensitivity analysis suggests that exposure alone is not sufficient to improve category learning. In fact, if anything, repeated exposure seems to have *decreased* sensitivity to the target contrasts, at least between the first and third sessions (although this result should be interpreted with caution, due to the marginal nature of this effect). In light of the failure to find improvement in discrimination ability, the simple effect of contrast can be interpreted as a reinforcement of the discriminability of each contrast type at baseline. Trials with two phonetic cues - place and voicing - were the most likely to be discriminated, with voicing trials slightly less detectable, but still generally discriminated. Place trials were most difficult, with a mean d' value of 1.34 ($s.d. = 0.52$) across all sessions.

Due to the lack of a positive effect of session, the analysis of the data from experiment 1B was restricted to the sensitivity analysis.

2.6 Experiment 1C: Retention study

Experiments 1A and 1B describe the nature of short-term category learning (as explored in the current experimental paradigm), but they cannot capture how well categories were retained after training had ended. As a result, a retention study was designed to address this question. Subjects enrolled in experiment 1A were invited to return to the lab approximately one month after the completion of session H, and take a final test of their discrimination and production performance. Comparison of performance on session A and the retention session will indicate what learning, if any, was retained.

2.6.1 Methods

The retention session (hereafter referred to as session I) was a single-session study, run in the same manner as any test session from experiment 1A. Subjects completed the AX discrimination task without feedback, and then the repetition task, with the CV natural stimuli. All subjects from experiment 1A were invited back to participate in the retention study approximately one month after their last testing session. Seven subjects enrolled and completed the retention study.

2.6.2 Results

Because only seven subjects returned for the retention study, there are not enough data points for a sensitivity (d') model following the ones conducted for experiments 1A and 1B. Therefore, the analysis of the retention data is restricted to the accuracy and reaction time analyses.

The goal of the retention analysis is to compare performance one month after testing ended to (1) baseline (pre-test, session A of experiment 1A), to see if performance remained elevated after significant time had passed, and (2) the final session (re-test, session H), to directly measure what changes had taken place in the representations in the intervening month. To that end, the retention (session I) data was compared in the following analyses to the data from session A and H, for those subjects who completed the retention session. Discussion of the models is restricted to the parameters of direct interest for this question (namely, the effect of session and any of its interactions).

Accuracy analysis

Using the procedures for model fitting described above, the best model for the accuracy data was the following:

```
correct ~ vowel + session * contrast + session * trialCount + (1 |stim1) +
        (1 |stim2) + (1 + session |subject), family = ‘‘binomial’’
```

There were simple effects of vowel, contrast, session, and trial count; the latter three are explored in more detail below, in their interactions. The effect of vowel was driven by the contrast between /u/ and /i/ ($\beta = 1.037$, $t = 2.881$, p *adj.* $= 0.011$); stimuli with /u/ were more likely to be perceived accurately than stimuli with /i/.

Post-hoc comparisons of the session * contrast interaction (see table 2.14) revealed that most contrasts which improved from pre-test to re-test (A to H) were also maintained through the month until the retention test (A to I). In one case (place + aspirated vs. voiced trials), there was heightened discrimination at the retention session which was not present at re-test. In two cases (breathy vs. voiceless and place + breathy vs. voiceless), accuracy was higher at retention than at re-test. As with experiment 1A, accuracy in "same" trials decreased across all three sessions.

Investigation of the session * trial count interaction initially showed a significant increase from session A to session H ($\beta = 0.120$, $t = 2.679$, $p = 0.007$); however, this contrast did not hold up under post-hoc correction for multiple comparison. Furthermore, there was no significant interaction of trial count with session I, so the interaction is not of much interpretive interest for the present analysis.

Reaction time analysis

The reaction time data from the retention study was best fit with the following model[8]:

```
log(RT) ~ session * contrast + session * trialCount + (1 |stim1) + (1
                  |stim2) + (1 + session |subject)
```

There was a significant negative effect of trial count (β = -0.053, t = -4.930, $p < 0.001$), indicating that reaction times sped up as a testing session went on. The significant session * trial count contrast indexes how the effect of trial count varied as a function of session. There was no difference in the trial count effect between sessions A and H ($\beta = 0.017$, t = 1.187, p *adj.* = n.s.), but there was a difference between sessions A and I ($\beta = 0.046$, t =3.281 , p *adj.* $= 0.003$), as well as a marginal effect between session H and I ($\beta = 0.030$, t = 2.209, p *adj.* $= 0.069$). The positive coefficient of these terms indicate that session I had a steeper slope for the effect of trial count than session H or session A - that is, the increasing speed as a function of trial count was more pronounced in the retention session.

The results of the session * contrast interaction are summarized in the lower half of table 2.14. The interaction appears to be entirely driven by the difference between "same" and "different" trials - all "different" trials showed reduced reaction times from pre-test to re-test, and from pre-test to retention. There was no effect in "same" trials. No effect was found

[8]As in experiment 1A, outliers were identified as data points with residual values greater than 2.5 standard deviations. Removal of these 151 data points (2.27% of the data set) did not improve normality of the residuals (original model, $W = 0.1577$, $p < 0.0001$; updated model, $W = 0.1734$, $p < 0.0001$).

ACCURACY MODEL			
Contrast	**Pre-test to re-test**	**Pre-test to retention**	**Re-test to retention**
Increased accuracy			
Place	*	*	
Aspirated vs. breathy	*	*	
Aspirated vs. voiced	(.)		
Aspirated vs. voiceless	*	*	
Breathy vs. voiced	*	*	
Breathy vs. voiceless	*	*	*
Place + aspirated vs. breathy	*	*	
Place + aspirated vs. voiced		*	
Place + aspirated vs. voiceless			
Place + breathy vs. voiced	*	*	
Place + breathy vs. voiceless	*	*	*
Place + voiced vs. voiceless	*	*	
Voiced vs. voiceless	*	*	
Decreased accuracy			
Same	*	*	*
REACTION TIME MODEL			
Contrast	**Pre-test to re-test**	**Pre-test to retention**	**Re-test to retention**
Decreased reaction time			
Place	*	*	
Aspirated vs. breathy	*	*	
Aspirated vs. voiced	*	*	
Aspirated vs. voiceless	*	*	
Breathy vs. voiced	*	*	
Breathy vs. voiceless	*	*	
Place + aspirated vs. breathy	*	*	
Place + aspirated vs. voiced	*	*	
Place + aspirated vs. voiceless	*	*	
Place + breathy vs. voiced	*	*	
Place + breathy vs. voiceless	*	*	
Place + voiced vs. voiceless	*	*	(.)
Voiced vs. voiceless	*	*	
No difference			
Same			

Table 2.14: Contrast * session pairwise comparisons, experiment 1C. The first column shows improvement/decline in performance as a result of training (session A - H). The second column indicates maintenance of improvement over baseline after one month (session A - I). The third column shows change from end of training to one month after (session H - I). All listed contrasts which were significant at p *adj.* < 0.05 (as assessed with a correction for multiple comparisons) are indicated with *. Marginal comparisons, with p *adj.* between 0.05 - 0.10, are marked with (.).

(aside from a single, marginal effect) from re-test to retention, indicating that subjects were not speeding up or slowing down after a month without training.

2.6.3 Discussion

Taken together, the retention analyses indicate that information learned in the initial training study (sessions A - H) were maintained at a testing session (session I) one month after the final mandatory session. Not only was performance heightened at the retention session above pre-test levels, but there was no detectable decline in either speed or accuracy from re-test to the retention session. This suggests that the novel categories learned by subjects were maintained very well even in the absence of reinforcement.

2.7 General discussion

Experiments 1A, 1B, and 1C probe different aspects of the early stages of novel phoneme acquisition. Experiment 1A provides characterization of the contribution of multiple cues combining to build at least a preliminary representation of a novel category in a relatively short time frame. Experiment 1B provides a control by demonstrating that exposure alone, in the absence of other cues, is unlikely to be helpful in a short-term paradigm. Experiment 1C demonstrates that fledgling categories acquired in a short-term paradigm can persist for at least a month after training has concluded.

2.7.1 Relevant cues to learning novel contrasts

The results of experiment 1A confirm that the combination of cues employed in the perception training paradigm - adaptive fading, repeated exposure, and feedback - improved discrimination of the target categories in Hindi. These contrasts are well-known to be challenging for native English-speaking learners (e.g. Bradlow, 2008), so the fact that improvement was detectable in a short-term study reinforces the argument of previous researchers (e.g. Goudbeek et al., 2008; McCandliss et al., 2002) that these components are important cues to overcoming native language phoneme biases. The effect of production training in experiment 1A, conversely, was non-significant.

Concerning exposure as a singular cue, there was no effect of learning in experiment 1B. In that study, subjects were exposed repeatedly to the target contrast but received no other reinforcing information about the target categories. This result is perhaps not surprising - subjects received no information, explicit or implicit, about the target contrasts, and so they arguably had no motivation to change their response patterns as a result of exposure alone. Even if distributional information can lead to non-explicitly-cued contrasts in some cases (Pegg & Werker, 1997; Maye et al., 2002), the current paradigm may not have provided sufficient exposure to induce such learning in adults.

2.7.2 Retention, maintenance, and recency

The analysis of experiment 1C suggests that improvement in discrimination over the course of training (as measured during session H) can be retained for at least a month (session I) after training and testing has concluded. Performance during the retention session was higher than baseline levels (session A). Just as importantly, there was no significant difference found between session H and session I, suggesting that the level of performance attained at the end of training was maintained at a comparable level in the intervening month. This suggests that the learning that took place over the course of the experiment was not ephemeral or due solely to recent exposure; at some level, the category information that subjects learned appears to persist, even without explicit reinforcement.

This finding may seem at odds with the effect of days to completion found in experiment 1A - in that study, longer times to complete the study resulted in a penalty in accuracy, suggesting a role for recency after all. A possible confound that may explain this discrepancy is the role of subject motivation. Experiment 1C was an opt-in session, and only 7 of 21 subjects who completed the main study chose to return. It is possible that those who chose to return for the retention study were unusually motivated in some way. If so, the same commitment that caused them to return to the lab may have also pushed them to take the task more seriously, leading to better initial learning of the target contrasts.

This study therefore leaves open the role of learner motivation in the acquisition of novel phonemes. While there is ample evidence that motivation is a key factor in second language acquisition generally (e.g. Gardner & Lambert, 1959; Masgoret & Gardner, 2003), none of the subjects in the current study expressed interest in acquiring Hindi. The current findings suggest that general task motivation may play a role in learning outcomes even when there is no expectation of following through to learn the target language.

2.7.3 Variation in learnability

While similar patterns in learning across most contrast types, a few differences between types were notable. In the sensitivity analysis of experiment 1A, the general differences in discriminability between broad contrast classes (place, voicing, place + voicing) were very clear - place trials were much more difficult for learners than trials with a voicing contrast, or with multiple phonetic cues. This reinforces the notion that some non-native contrasts are more challenging for learners than others.

When contrast types were broken down into finer-grained categories, similar learning patterns emerged for most types - generally, responses were more accurate and faster as a function of perception training (but not production training), even though contrasts started at different baseline levels. Thus, it seems that the training strategies employed in experiment 1A were general enough to have a positive impact on multiple types of contrasts, with different acoustic cues to those target contrasts.

However, a few contrasts (aspirated vs. voiced, place + aspirated vs. voiced) failed to show improvement - an unsurprising result, as that contrast maps onto the English /t/ vs.

/d/ contrast. (It is somewhat surprising that the aspirated vs. voiceless contrast showed improvement, as this also maps onto the same English contrast under certain allophonic conditions. However, the aspirated vs. voiced contrast is even more acoustically dispersed, which may account for the difference in performance on the two contrasts.)

It is reasonable to suspect that not all contrasts will respond to training in the same way. Herd et al. (2013) compared perceptual training with a form of production training in which English subjects tried to match visualized acoustic features of target Spanish sounds, through comparison of waveforms and spectrograms of their own speech and a native speaker productions. (This paradigm was based off the work of Hirata, 2004.) The two procedures led to different outcomes for different contrasts - the /ɾ/-/r/ was better-identified after production training, while the /d/-/ɾ/ contrast was only improved after perception training (in fact, the latter failed to improve for subjects who got both types of training). Because the independent involvement of perception and production were not tested in the present paradigm, it is not possible to directly compare these effects in the present study. However, the relative ordering of perception and production training could contribute to asymmetrical effects, with some contrasts benefiting from early perceptual training, and others suffering due to relatively late articulatory instruction.

2.7.4 The perception-production link

The current study found no consistent evidence of articulatory information influencing performance on the discrimination task. A strong interpretation of this finding would be that information about production is not useful at this stage of learning for the formation of perceptual categories, an argument that would seem to contradict the predictions of the Perceptual Assimilation Model (Best et al., 2009), as well as the findings of Catford and Pisoni (1970). A less stringent (if less theoretically exciting) conclusion is that the methodology presented here did not permit the incorporation of the information presented during the production training session. This could be either because (1) the session itself was not long or detailed enough, or because (2) the production training always occurred after the perception training, and subjects may have reached their greatest performance within the context of the paradigm before they had even gotten to the production training session.

If the failure to find an effect of production training *is* real - and not just an artifact of methodological design - then it suggests that the type of information that can be scaffolded to build a category representation is restricted in some ways. This issue will be explored in more detail in Chapter 3.

There was also not much evidence for perceptual training aiding production in experiment 1A. The biggest difference that could be attributable to perceptual training was for place of articulation, as measured by formant transitions. Overall classification accuracy of syllables produced by subjects as dental or retroflex improved from 50.96% to 57.14% from pre-test to post-test, which may indicate some improvement in articulatory representations of this place contrast as a function of discrimination training. This improvement was not present for the burst spectra properties cataloging place of articulation, as well as the VOT properties cuing

the voicing distinction. Therefore, the extent to which this any perceptual-to-articulatory transfer was present in the current paradigm, it was limited and constrained to a specific property and feature.

It is perhaps not surprising that place of articulation was the property which showed more success on this metric. Informal conversations with subjects after training indicated that the place contrast was easier for them to implement than the voicing contrast. The articulatory targets for the dental-retroflex contrast (tongue tip placement) were easier to rapidly acquire than the more complex laryngeal and timing gestures required for the four-way VOT contrast. This suggests an interesting relationship between articulatory ease and phonological assimilation during acquisition: the place contrast may be easier for non-native speakers to produce in the current study, but it was more difficult for them to perceive. There must have been some perceptual support for this contrast underlying the ability of subjects to make distinct articulations for the dental and retroflex categories.

An alternative perspective is that acoustic-phonetic detail, rather than phonological categories *per se*, is the better predictor of the perception-to-production transformation for non-native speakers (C. Wilson et al., 2014). This may speak to the dissociation between the burst cues and formant cues for the place of articulation contrast in the present data set: the latter, but not the former, seems to suggest improvement as a function of perceptual training. Perhaps English speakers are more attuned to those properties of place of articulation cued by the formant transitions, and are able to use these cues more rapidly to detect, and ultimately reflect, a difference between dental and retroflex place of articulation.

2.7.5 A note on category terminology

Throughout this chapter, the term "fledgling category" has been used. This terminology is deliberately cautious, and used to avoid the suggestion that the training paradigm presented here created robust, fully-established categories. Clearly, subjects were not performing at ceiling after training, especially for the dental-retroflex contrast (see figure 2.5), and so it would be erroneous to assume that the representations that *have* begun to form bear all the hallmarks of truly robust categories. Even if discrimination performance was at ceiling, this paradigm did not test identification, a second hallmark of classic accounts of categorical perception (Liberman, Harris, Hoffman, & Griffith, 1957; Chang et al., 2010). The working hypothesis of this chapter is that the ability to discriminate target contrasts is a necessary, but not sufficient, condition for the ultimate establishment of a robust category which shows all the hallmarks of categorical perception. The procedures presented here are one way to "jump start" the development of a category, but they are unlikely to be the sole means needed to reach robust categories.

The invocation of this terminology raises the question of what constitutes a robust, fully-fledged category. Research on adult second-language phoneme acquisition is concerned with at least two major issues: the challenge of overcoming the biases of native language categories, and the linking of perceptual and articulatory accuracy. Based on these, one working definition of a fully-fledged category could be the following: a phonetic category

which is not primarily influenced by a category in a separate phonological system, and one which is robust in both the perception and production domains.

Regarding perceptual robustness, the dual benchmarks of categorical perception - reduced discrimination within-category, and enhanced discrimination across-category (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Diehl et al., 2004; Chang et al., 2010) - are useful metrics. These would indicate that both discrimination and identification are necessary to test the robustness of a perceptual category. With respect to articulatory robustness, it is well-known that accents in a second language often persist even the speaker is fluent (for discussion of factors influencing accents in a second language, see Piske, MacKay, & Flege, 2001). Therefore, native-like pronunciation may not be a reasonable benchmark for the adult learner. One possible metric that could be used is the degree of consistency in pronunciation of articulatory targets over time. With consideration for other sources of variation in articulation (e.g. speaking rate, coarticulation, prosodic context), a consistent and fairly tight distribution of articulations of a particular target category may indicate that a speaker has settled on a relatively stable articuatory representation of that category.

2.8 Conclusion

The experiments in this chapter explored the ability of new learners to improve their perceptual and articulatory targets of several non-native phoneme contrasts in a set of a short-term training studies. Together, the studies showed that a combination of feedback, exposure, and adaptive fading contributed to improvement in perceptual discrimination, but that articulatory information, and exposure absent all other cues, did not further improve discrimination. Production accuracy was improved primarily in the place of articulation dimension, with some support coming from perceptual learning, but more improvement as a function of production training. Together, these findings suggest that certain combinations of cues, but not all cues (or all cues in isolation), can effectively build preliminary category representations in a relatively short period of time.

Improvements in discrimination were retained for a small subset of subjects one month after training had completed. This retention indicates that, while fledgling, these newly-learned categories are somewhat resistant to decay as a function of time.

The failure to detect improvement as a function of production training is somewhat surprising, given past research suggesting that this can be an effective cue for perceptual learning. The source of this discrepancy (methodological or mechanistic) remains an open question, and will be explored in detail in Chapter 3.

Chapter 3

Isolating the influence of articulatory learning

3.1 Introduction

The previous chapter discussed the effects of a multi-day training study on the acquisition of non-native phoneme categories by beginning learners. While the perceptual training procedure was successful in improving subjects' discrimination of the target novel contrasts, the articulatory training conferred no further benefit - discrimination abilities remained stable even after subjects learned how the target sounds are produced.

However, based on past research, there is good reason to believe that production training should exert some influence over the development of perceptual categories. This chapter explores the issue in more detail, with a series of experiments designed to test whether a possible effect of articulatory learning can be isolated outside of the larger training paradigm.

3.1.1 Theoretical accounts of the link between perception and production

A major debate in speech perception concerns the ultimate objects of representation of speech sounds. General auditory accounts of speech perception (Diehl et al., 2004) contend that perceptual and auditory systems support the perception of speech. In this view, the listener must make statistical and perceptual inferences based on accumulated experience with speech, leading to generalized categories that are flexible enough to accommodate variance in the acoustic signal.

However, another prevailing view in the field is pursued by the Motor Theory (MT) of speech perception, as well as the related Direct Realist (DR) account. These theories argue that due to the variance of the acoustic signal of speech sounds in different contexts, the ultimate representation of a phoneme is better described according to its articulatory routine (Liberman et al., 1967; Liberman & Mattingly, 1985; Galantucci, Fowler, & Turvey, 2006;

Fowler, 2008). The objects of perceptual representation, from a motor theory perspective, are the motor routines of production, explicitly linking the two processes at the representational level. While the popularity of articulatory accounts has varied over time, the idea has recently gained renewed support as an instantiation of embodied cognition (Wilson, 2002; Fischer & Zwaan, 2008), perhaps supported by the mirror neuron system (e.g. Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Watkins & Paus 2003; but see Lotto, Hickok, & Holt 2009). A related view, Analysis by Synthesis (Stevens & Halle, 1967; Kuhl, Ramírez, Bosseler, Lin, & Imada, 2014), argues that bottom-up acoustic information integrates with internal motor predictions to constrain hypotheses about the segments in an incoming acoustic signal.

The theoretical divide between acoustic and motor theories has influenced views on the acquisition of non-native phonemes, although these accounts do not necessarily fall into strict camps. Flege's Speech Learning Model (Flege, 1995; Flege et al., 1997; Guion et al., 2000) predicts that during second language acquisition, learners will struggle to perceive target phonemes which are acoustically similar to categories in the learner's native language. In this view, non-native phonemes assimilate to native categories when their acoustic properties make them plausible (if non-central) tokens of a native category. Best's Perceptual Assimilation Model (Best et al., 2001; Best & McRoberts, 2003; Best & Avery, 2007) predicts a similar challenge for discrimination, but in this account, there is a primary focus on articulatory similarity, with a particular focus on similarity in place of articulation. Depending on the class of sounds involved in a target contrast, then, the two theories can make different predictions about what types of sounds learners will struggle with the most - and, perhaps, which interventions will be most successful in helping them to overcome native language biases.

The latest instantiation of the Native Language Magnet Theory (Kuhl et al., 2008), proposes mechanisms that infants use to ignore universal phonetic contrasts and focus in on the contrasts relevant to their native language. The theory includes a role for a link between perception and production via the perceptual reinforcement that children receive as they practice articulating sounds - both from caregiver responses, and from listening to their own voices. While this view does not assume that articulatory gestures are the representational basis for phoneme categories, it assumes that the two systems can support one another. Recent work by Kuhl and colleagues (Kuhl et al., 2014) suggests that neural motor routines may be most active when percepts are less familiar, while acoustic pathways are sufficient to support perception for familiar, well-learned categories. In a related claim, recent work on neural pathways of speech perception suggest that motor routines may be recruited in noisy conditions, when perceptual targets are less certain (Davis & Johnsrude, 2007; Du, Buchsbaum, Grady, & Alain, 2014).

3.1.2 Experimental findings on perception-production links during second language acquisition

In the literature on non-native phoneme acquisition, a small set of studies have investigated the influence of training articulatory targets on perceptual discrimination or identification. In the study most directly resembling the current task, Catford and Pisoni (1970) compared native English speakers' discrimination of several non-native phonemes (/y/, /ɯ/, /ç/,/ʔ/, /q/, /k'/, and /ø/) before and after a two-hour training session. Half of participants received auditory training, which emphasized the difference between the target sounds and familiar sounds; the other half received articulatory training from a trained phonetician, who gave detailed instructions about the place, manner, and airstream mechanisms involved in producing the targets. After training, the discrimination performance of the group which had received articulation training was significantly higher than the group which had received auditory training. (Both groups were also tested on pronunciation performance; unsurprisingly, the articulation training group also outperformed the auditory training group on this task.) More modest effects were reported by Gómez Lacabex, García Lecumberri, and Cooke (2008), who found, for Spanish learners of an English full-schwa vowel distinction, comparable improvement with 3-month perceptual or articulatory training sessions over a control group, but no distinction between the training groups.

Hazan, Sennema, Iba, and Faulkner (2005) measured an indirect form of articulatory knowledge by providing learners with visual information about the articulation of target contrasts. They trained Japanese native speakers to perceive two sets of English contrasts: /b/-/p/-/v/ and /l/-/ɹ/. Learners received either audio-only training or audio-visual training; in the latter, videos of native speakers articulating the target sounds accompanied the training audio. Both types of training enhanced performance, but the audio-visual training group outperformed the audio-only group for the /b/-/p/-/v/ contrast - that is, when the articulatory information signaling the contrast was visually salient.

The benefits of learning about production targets are not restricted to segments; length and pitch contrasts can also benefit. Hirata (2004) worked with English native speakers who had 2-4 years of experience with Japanese, to improve their perception of Japanese pitch and length contrasts. In 10 training sessions, subjects listened to native speakers produce utterances, with visual prosody graphs which displayed pitch and length information as subjects listened to the speech. Subjects' perception and production of length and pitch contrasts both improved when these training stimuli were presented in sentential contexts (but not for words in isolation).

While the studies described above tend to show positive effects of production training, Schneiderman, Bourdages, and Champagne (1988) warn that the effect may not be universally positive. In their study, English-native speakers learning French worked through 12 one-hour training sessions devoted to improved perception and production of French prosody, intonation, the front rounded and nasal vowels, and the consonants /ɥ/ and /ʒ/. The first six were devoted to perception-only training; production training (a repetition task) was introduced in hour seven. Their pattern of results was complex, but suggested that an initial

improvement in perception of the non-native phonemes was disrupted by the articulatory knowledge introduced by the repetition task. They suggest that the problem may have arisen from tenuous connections between the subjects' fledgling perceptual and articulatory targets, or from incorrect hypotheses held by the learners. This latter concern echoes the concerns of several studies (Vlahou et al., 2011; Seitz et al., 2010; Lim & Holt, 2011; Gulian et al., 2007) in the literature on non-native phoneme acquisition which suggest that explicit learning may cause learners to develop incorrect hypotheses which interfere with acquisition of the true targets. (See section 2.2.1 for more discussion.) Baese-Berk (2010) also found that perceptual learning may be disrupted by production training, although in her study, the issue was resolved to some degree after several days of training.

3.2 Hypotheses

The findings in the literature indicate that, at least under certain conditions, production training should have a positive effect on the formation of perceptual categories. This raises the question of why production training had no detectable effect in experiment 1A. Because the methodologies recorded in the literature are diverse, and differ in many ways from experiment 1A, there are many candidate sources for the null effect. In the remainder of this chapter, I consider three possible hypotheses about the outcome of production training in experiment 1A, and examine each with an experiment designed to isolate that potential cause.

Hypothesis 1: The "local ceiling" effect. It is possible that there is a limit on the amount of improvement in discrimination ability that can be reasonably expected in a laboratory setting. The production training in experiment 1A took place only after several rounds of perceptual training. Therefore, it is possible that subjects had already hit their "local ceiling" prior to receiving production training. If this is true, it suggests that the potential benefits of articulatory knowledge would be more apparent on completely novice learners who had not received any prior training.

Hypothesis 2: The "inflexible instructor" effect. Catford and Pisoni (1970) included directed instruction from a trained phonetician in their study; this approach provides the opportunity for feedback between the subject and a knowledgeable experimenter. It is possible that the computer program in experiment 1A was comparatively too inflexible for a learner to benefit, because it could not answer questions or provide clarification on the material being taught.

Hypothesis 3: The exposure effect. The exposure to the articulatory information to be learned during the production training module may have been too short (approximately 15-25 minutes, depending on an individual subject's pace) to have a significant impact on a learner's representations of the target sounds. Either a lengthened session, or a longer repetition task (which reinforces the target articulatory postures and links perceptual information to productions) may give subjects more time to learn the articulatory routines presented in the training.

3.3 Methods

3.3.1 Experiment 2 series

Experiment 2 consisted of four related sub-experiments, with different types of training. Three of these were designed to address the hypotheses in section 3.2. In each of these, one of the factors identified as a possible reason for the failure of production training in experiment 1A was manipulated, to test whether a more optimal methodology would lead to improvement in discrimination ability of novice learners. In the fourth, a perception training session serves as a control.

Experiment 2A: Addressing the local ceiling. If improvement due to perceptual training masked any modest effects of production training in experiment 1A, this can be tested by extracting the production training and administering it to naïve subjects without any perceptual training. This study maintains the exact pre-test, production training, and post-test sessions from experiment 1A.

Experiment 2B: Guided training. If the production training program designed for experiment 1A was not sufficiently detailed or clear, it is possible that some subjects had questions about the target material which were not answered, or that some formed incorrect hypotheses about the targets. This version of the study alters the production training session by having a trained phonetician (the experimenter) present to practice the target sounds with each subject. The session proceeds as before, but with guided support and checking of articulatory form by the experimenter/native speaker as the subject progresses through training. The experimenter is also able to answer any questions that the subject has about correct pronunciation (e.g. "Did I pronounce that correctly?"). In this way, the experimenter is able to provide immediate feedback about pronunciation to the subject, and to flexibly reinforce correct targets.

Experiment 2C: Increased exposure. If length is the primary factor preventing improvement as a function of production training, experiment 2C can demonstrate this by lengthening the duration of the production training session. Increased exposure in this paradigm comes in the form of more repetition trials (four repetitions of each stimulus, as compared to one in the original training design). These repetition trials use visual cues learned during production training to reinforce the target articulation of each sound, and to remind subjects about the distinctness of each category.

Experiment 2D: Controlling for perceptual learning. Experiments 2A-2C are not directly comparable to experiment 1A, as subjects received multiple days of discrimination training prior to the post-test session. Experiment 2D controls for this by running the first day of discrimination training (session B of experiment 1A) during the training session. This manipulation allows for a direct comparison of the efficacy of perceptual training and articulatory training.

Session code		Session	Tasks	Stimuli	Performance feedback?
A		Pre-test	discrimination; repetition	CV natural	no
B		Training			
	Exp. 2A		production training	CV natural	no
	Exp. 2B		guided production training	CV natural	yes
	Exp. 2C		long production training	CV natural	no
	Exp. 2D		discrimination training	VCV careful	yes
C		Post-test	discrimination; repetition	CV natural	no

Table 3.1: Structure of all studies in the experiment 2 series.

3.3.2 Procedures and stimuli

All versions of experiment 2 consisted of three sets of tasks (see table 3.1). The pre-test (session A) and post-test (session C) sessions were drawn directly from sessions A and F of experiment 1A. The training session (session B) varied between sub-experiments, as described in section 3.3.1. In all versions of the experiment, all three sessions were run in a single two-hour period, with optional breaks of a few minutes between each session. The stimuli in all versions of experiment 2 were identical to those in experiment 1.

3.3.3 Subjects

Sixty subjects were recruited to participate in experiment 2, 15 per sub-experiment. As in the experiment 1 series, subjects were pre-screened to ensure that they were native English speakers, and that they had no prior experience with Hindi or another language containing the phonemic contrasts of interest. In addition, potential subjects were excluded if they had studied linguistics in any detail beyond the introductory level; this was done to reduce variability between subjects in how much prior knowledge about phonetics was being brought to the tasks.

3.4 Results

Analysis of the experiment 2 data followed the same approach as the one developed for experiment 1. Three analyses were run: a sensitivity analysis with d' as the dependent variable, an accuracy analysis using logistic regression, and a reaction time analysis.

As in experiment 1, outliers were removed from the data on the basis of reaction time: trials with reaction times less than 100 milliseconds were removed. Trials were also considered outliers if their reaction time fell outside of 2 standard deviations of the mean for each subject. This procedure retained 95.2% of the data.

3.4.1 Sensitivity analysis

The sensitivity metric d' was used as the dependent variable for the sensitivity analysis. A d' value was calculated for each independent combination of subject (60), test session (2), and contrast type (3), for a total of 360 unique values. A linear mixed-effects regression model was constructed in `R` (version 3.1.2, R Core Team, 2014) using `lmer` function in the `lme4` package (version 1.1-7, Bates & Maechler, 2014), with d' as the dependent variable. The model included fixed effects for contrast type (place, voicing, or place + voicing), session (pre-test or post-test), and the interaction of contrast and session. The effect of training type (2A - production, 2B - guided production, 2C - long production, or 2D - discrimination) was explored in an alternative model. While the model AIC value was lower for the model which included training type as compared to the model which did not, the model comparison did not reach significance or improve model fit as assessed by the log likelihood criterion (model with training type, AIC = 342.81; model without training type, AIC = 337.72; $\chi^2 = 0.8048$, $p = 0.8243$).

A similar criterion was used for assessing the validity of including the interaction between session and contrast type; model comparisons with and without the interaction showed a marginal improvement in model fit, as assessed by the log likelihood criterion (model with interaction, AIC = 337.72; model without interaction, AIC = 338.63; $\chi^2 = 4.913$, $p = 0.086$). Because the interaction was of theoretical interest, and because further post-hoc testing revealed significant interactions between the variables, the interaction between session and contrast type was retained.

A by-subject random slope for session was also included. (There was not a sufficient number of data points to include a more elaborated random effect structure, e.g. a by-subject session * contrastType random slope). In `lme4` syntax, the model was specified as:

```
d-prime ~ session * contrastType + (1 + session |subject)
```

The original model specified by this formula had a handful of outliers (even after the original data set had been pruned on the basis of reaction time), and its residuals were not normally distributed, as identified by a Shapiro-Wilk test for normality (W = 0.9829, p = 0.0003). To identify potential influential cases, the `romr.fnc` function in the `R` package `LMERConvenienceFunctions` (A. Tremblay & Ransijn, 2013) to flag values with standardized residuals outside of 2.5 standard deviations from 0. Eleven of 354 data points (3.1%) met this criterion and were removed. The model was subsequently re-fit, and the residuals of the new model were determined to be normally distributed (W = 0.9953, p = .3796). Q-Q plots and density plots for the model residuals before and after outlier removal are visualized in figure 3.1.

Simple effects for the sensitivity model are reported in table 3.2. The effect of contrast type indicates that, as in experiment 1, trials with a voicing contrast ($\beta = 1.584$, $t = 33.370$, $p < 0.0001$) and two phonetic features ($\beta = 1.776$, $t = 37.417$, $p < 0.0001$) were more likely to be detected than trials with a place of articulation contrast. The effect of session indicates

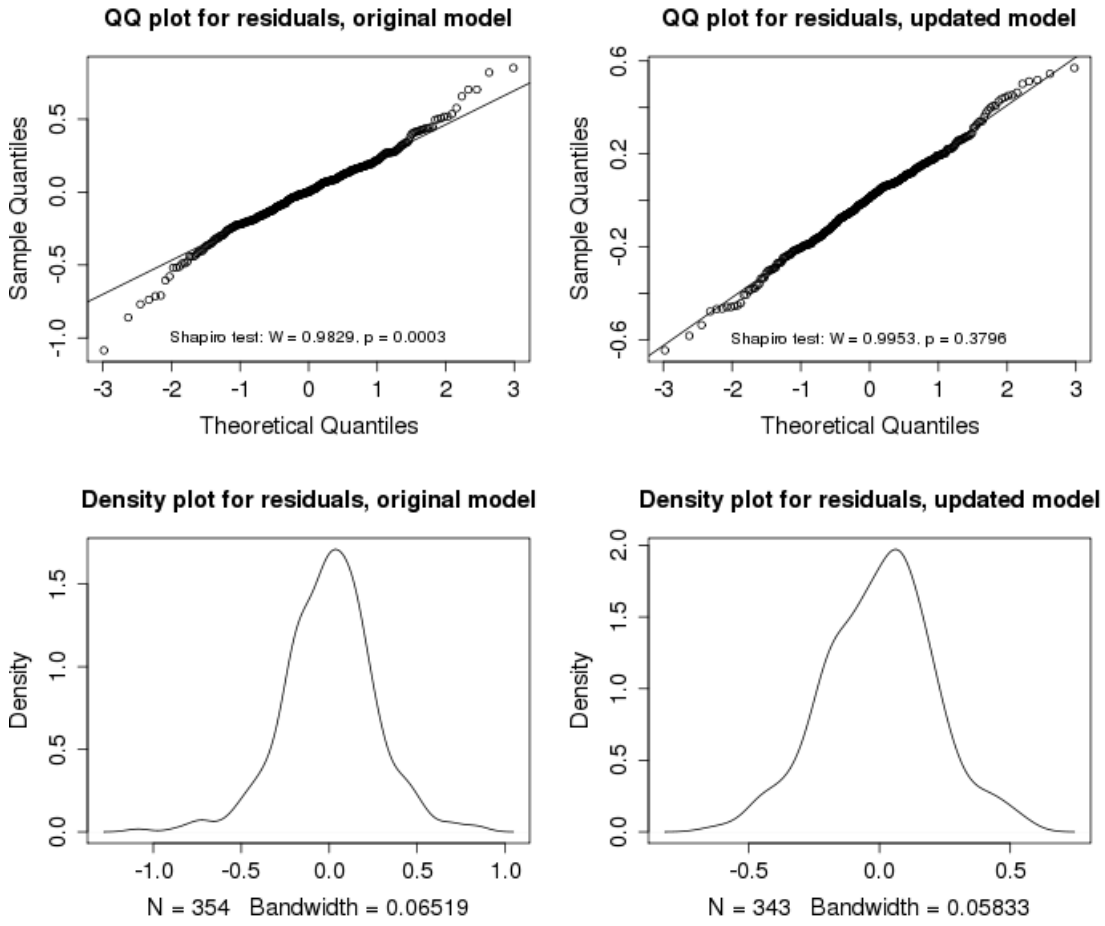


Figure 3.1: Diagnostics of sensitivity model residuals, experiment 2.

that for the reference level of contrast type (place), there was a significant improvement from pre-test to post-test. Post-hoc comparisons (Tukey's HSD) of the interaction between contrast type and session (visualized in figure 3.2) indicate that this improvement from pre-test to post-test was restricted to place trials ($\beta = 0.204$, $t = 3.343$, *adj.* $p = 0.010$); improvements in voicing trials ($\beta = 0.098$, $t = 1.652$, *adj.* $p = 0.550$) and place + voicing trials ($\beta = 0.129$, $t = 2.176$, *adj.* $p = 0.239$) did not reach significance.

This analysis suggests that training was effective in improving subjects' discrimination of the place of articulation contrast, but did not improve discrimination of contrasts with a voicing distinction, or both features simultaneously. The lack of an effect of training type suggests that the presence of training was more influential than the specific implementation (duration, quanity, or perceptual vs. articulatory information).

predictor	β	s.e.	df	t	p
(Intercept)	1.252	0.053	115.640	23.756	< 0.0001
session (*base level: pre-test*)					
Post-test	0.205	0.061	150.44	3.343	0.0010
contrast type (*base level: place*)					
place + voicing	1.776	0.047	224.620	37.417	< 0.0001
voicing	1.584	0.047	224.620	33.370	< 0.0001
session * contrast type (*base level: Pre-test/place*)					
Post-test/place + voicing	-0.076	0.066	224.030	-1.149	0.2517
Post-test/voicing	-0.107	0.664	223.230	-1.608	0.109

Table 3.2: Fixed effects of sensitivity (d-prime) model, experiment 2. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations.

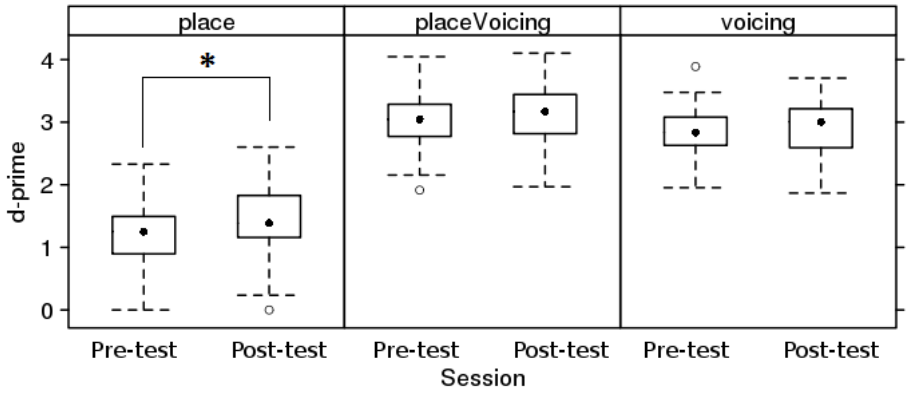


Figure 3.2: D-prime by contrast type * session, experiment 2. Significant contrasts are signaled with significance stars (. $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$) as determined by pairwise comparisons (adjustment with Tukey's Honest Significant Difference Test).

3.4.2 Accuracy analysis

The sensitivity analysis revealed that performance in discriminating the target contrast improved after training. However, it suggested that this improvement was limited to contrasts in place of articulation; in addition, there was no detectable differences between the different types of training. To confirm these findings and ensure that they were not attributable to a lack of power, a logistic regression model was constructed. While this approach cannot control for response bias in the same way as the d' analysis, it may allow for more sensitive contrasts to be detected, as it uses all data points individually (rather than collapsing groups into a single sensitivity score). As implemented, this approach also allows for a more fine-grained investigation of contrast type, by taking specific voicing contrast pairs (e.g. breathy

vs. aspirated) into account.

The model was fit with fixed effects for **vowel** (/a/, /i/, /u/) **session** (pre-test, post-test), **contrast type** (14 levels, corresponding to all pairwise sets of feature contrasts, including "same" trials), and **training type** (production training, long production training, guided production training, perception training). Interactions were included between session and contrast type, and session and training type. The model included random intercepts for subject, stimulus A, and stimulus B, and by-subject random slopes for session. In `lme4` syntax, the model was specified as:

```
correct ~ vowel + session * contrastType + session * trainingType +
(1|subject) + (1|stimulusA) + (1|stimulusB) + (0 + session|subject)
```

Coefficients of significant fixed effects are reported in table 3.3. Partial effects for the model are visualized in figure 3.3. The only fixed effect which did not enter into an interaction was vowel; post-hoc pairwise comparisons (Tukey's HSD) revealed that there was a significant difference in accuracy between syllables with /u/ and syllables with /i/ ($\beta = 0.7618$, $t = 2.600$, $p\ adj. = 0.025$).

The main variables of interest - session, contrast type, and training type - all entered into interactions, and are best inspected with post-hoc pairwise comparisons, reported in table 3.4. These comparisons test whether performance improved (a) as a function of the differing training paradigms (trainingType * session) and (b) in different featural contrasts (contrastType * session).

The effect of training type reveals the efficacy of the varying training paradigms in improving subjects' ability to detect the target contrasts. An initial post-hoc pairwise comparison (Tukey's HSD) of training type found no difference between any of the training types, across sessions (all $p\ adj. > 0.05$). However, inspecting the training type * session interaction (Table 3.4, upper panel) revealed that accuracy improved in the production training ($\beta = 0.505$, $t = 4.376$, $p\ adj. < 0.001$), guided production training ($\beta = 0.762$, $t = 6.584$, $p\ adj. < 0.001$), and perception training ($\beta = 0.862$, $t = 7.243$, $p\ adj. < 0.001$) paradigms, but not in the long production training paradigm ($\beta = 0.260$, $t = 2.268$, $p\ adj. = 0.312$). This result indicates that subjects who received both perception training and subjects in the certain production conditions improved their discrimination of the target contrast. But somewhat paradoxically, subjects in the long production training paradigm did not show improvement, even though the (basic) production training group did. This is surprising because the latter paradigm is shorter, but otherwise identical, to the former.

The interaction of contrast type and session (table 3.4, lower panel) indicates which contrasts showed a change in discrimination as a function of training. Most contrasts did show a statistically-significant improvement; the two which did not (aspirated vs. voiced and place + aspirated vs. voiced) map onto the English /t/ vs. /d/ distinction, and thus were likely well-discriminated even during the pre-test session. However, it should be noted that "same" trials showed a decrease in accuracy ($\beta = -0.691$, $t = -8.664$, $p\ adj. < 0.001$) from pre-test to post-test. This likely indicates that to some extent, subjects adopted a bias

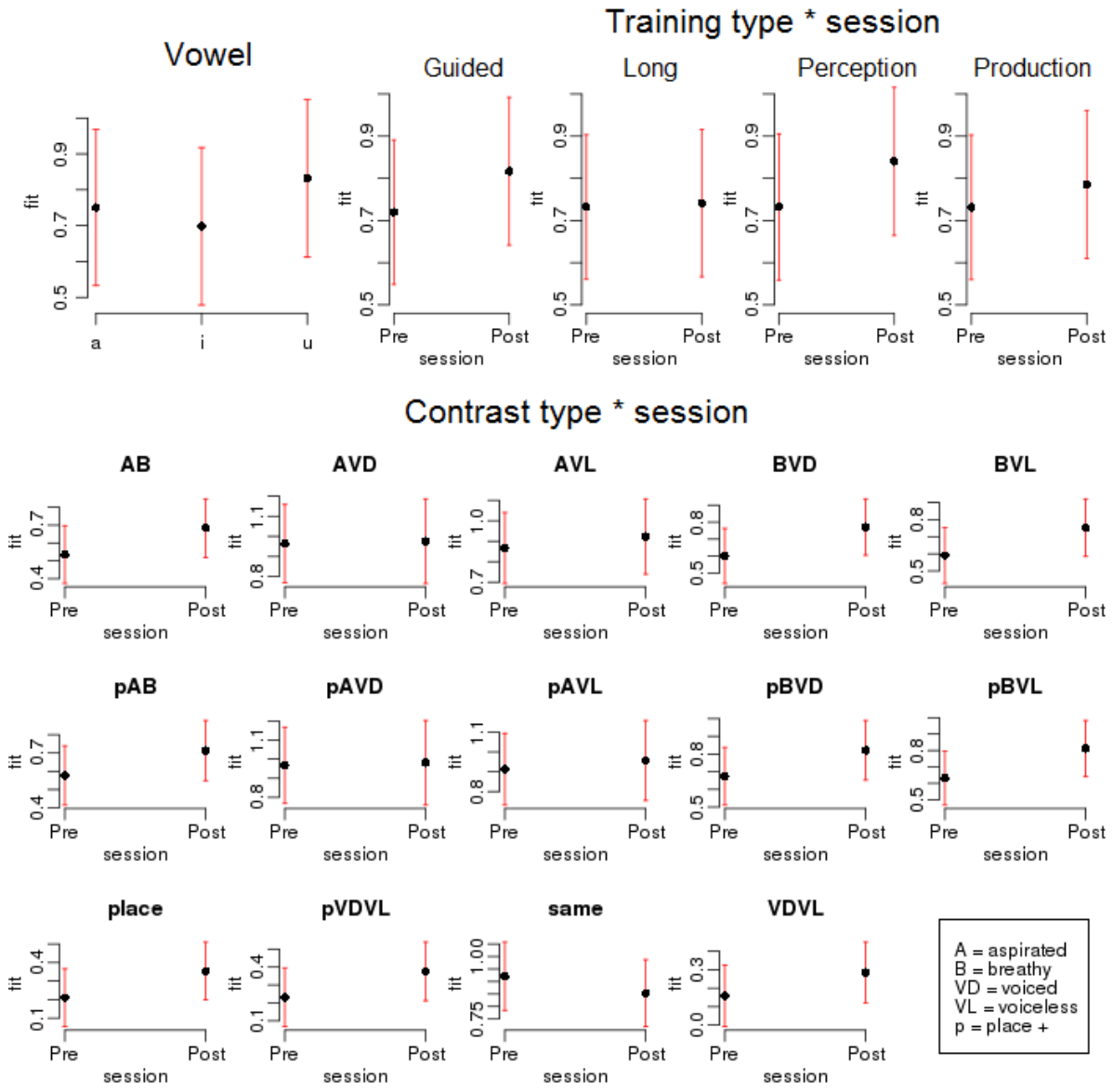


Figure 3.3: Partial effects of accuracy model, experiment 2.

towards "different" responses. Because the accuracy analysis (unlike the d' analysis) does not control for this bias, the results of this analysis should be interpreted with caution.[1]

In summary, the accuracy model confirms the finding from the d' analysis that discrimi-

[1]Another issue concerns whether this bias developed over the course of the session. However, inclusion of a trial number predictor (in either the fixed or the random effects) prevented the model from converging.

predictor	β	**s.e.**	t	p
(Intercept)	0.016	0.251	0.062	0.949
Vowel (base level: /a/)				
i	-0.253	0.292	-0.902	0.367
u	0.499	0.292	1.705	0.088
Session (base level: pre-test)				
post-test	0.797	0.139	5.40	< 0.001
Contrast (base level: aspirated vs. breathy)				
aspirated vs. voiced	3.136	0.142	22.155	< 0.001
aspirated vs. voiceless	1.735	0.108	16.008	< 0.001
breathy vs. voiced	0.268	0.091	2.960	0.003
breathy vs. voiceless	0.225	0.092	2.442	0.015
voiced vs. voiceless	-1.813	0.105	-17.247	< 0.001
place + aspirated vs. breathy	0.170	0.081	2.093	0.036
place + aspirated vs. voiced	3.261	0.148	22.027	< 0.001
place + aspirated vs. voiceless	2.208	0.121	18.236	< 0.001
place + breathy vs. voiced	0.586	0.092	6.365	< 0.001
place + breathy vs. voiceless	0.397	0.093	4.257	< 0.001
place vs. voiced vs. voiceless	-1.342	0.101	-13.301	< 0.001
place	-1.468	0.080	-18.466	< 0.001
same	0.059	0.157	0.374	0.709
Training type (base level: guided production training)				
long production training	0.062	0..157	0.387	0.699
perception training	0.063	0.159	0.394	0.693
production training	0.059	0.157	0.374	0.709

Table 3.3: Selected coefficients of accuracy model for significant or marginal simple effects, experiment 2. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations. Interaction terms are excluded from this table.

nation improved as a function of training. The accuracy model detected a difference between training types, with no improvement shown in the long production training group, but improvement in all other training groups. In addition, most contrasts showed improvement after training, with the exception of two which map onto the English phoneme contrast. However, the accuracy analysis also revealed a post-training bias for "different" responses (*i.e.* decreased accuracy for "same" trials), and these findings must be interpreted with that bias in mind.

SESSION * STUDY TYPE				
Comparsion (post-test - pre-test)	β	**s.e.**	***t***	***p***
Production training	0.505	0.115	4.376	0.0003
Guided production training	0.762	0.116	6.584	< 0.0001
Long production training	0.260	0.115	2.268	0.3117
Perception training	0.862	9.119	7.243	< 0.0001
SESSION * CONTRAST TYPE				
Comparison (post-test - pre-test)	β	**s.e.**	***t***	***p***
Aspirated vs. breathy	0.633	0.098	6.459	< 0.0001
Aspirated vs. voiced	0.433	0.297	2.204	0.9155
Aspirated vs. voiceless	0.620	0.141	4.391	0.0037
Breathy vs. voiced	0.810	0.102	7.936	< 0.0001
Breathy vs. voiceless	0.754	0.106	7.096	< 0.0001
Voiced vs. voiceless	0.757	0.105	7.234	< 0.0001
Place + aspirated vs. breathy	0.605	0.099	6.124	< 0.0001
Place + aspirated vs. voiced	0.538	0.213	2.526	0.7309
Place + aspirated vs. voiceless	0.748	0.173	4.326	< 0.0001
Place + breathy vs. voiced	0.800	0.107	7.482	< 0.0001
Place + breathy vs. voiceless	0.927	0.112	8.302	< 0.0001
Place + voiced vs. voiceless	0.697	0.099	7.028	< 0.0001
Place	0.731	0.081	8.986	< 0.0001
Same	-0.691	0.80	-8.664	< 0.0001

Table 3.4: Pairwise comparisons of interaction terms in accuracy model, experiment 2, as assessed with Tukey's Honest Significant Difference test.

3.4.3 Reaction time analysis

As in experiment 1, reaction time data was collected as an additional metric of improvement. It is possible that some contrasts will reflect a high level of discriminability, even at baseline, but that performance improvement will still be detectable as a function of faster responses after training.

The reaction time model was fit to log-transformed reaction times from correct trials, with fixed effects of session (pre-test, post-test), contrast type (all pairwise place and voicing contrasts, as well as "same" trials), and training type (production training, long production training, guided production training, perception training). Vowel (/a/, /i/, /u/) was initially entered into the model, but removed as it did not improve model fit. Interactions between session and contrast type, and session and training type, were also included. Random intercepts for stimulus A, stimulus B, and subject were included, as well as by-subject random

slopes for session[2]. The model specification was:

```
logRT ~ session * contrastType + session * trainingType + (1|subject) +
        (1|stimulusA) + (1|stimulusB) + (0 + session|subject)
```

predictor	β	**s.e.**	**df**	***t***	***p***
(Intercept)	6.534	0.008	66.000	81.198	< 0.0001
Session (base level: pre-test)					
Post-test	-0.146	0.056	87.000	-2.555	0.012
Contrast type (base level: aspirated vs. breathy)					
Aspirated vs. voiced	-0.061	0.028	16140	-2.167	0.030
Voiced vs. voiceless	0.132	0.041	18270	04.915	< 0.0001
Place + aspirated vs. voiced	-0.096	0.028	16190	-3.406	0.0007
Place	0.130	0.031	25540	4.175	< 0.0001
Same	-0.124	0.025	2882	-4.915	< 0.0001
Training type (base level: guided production training)					
Production training	-0.076	0.107	51.000	-0.714	0.479
Long production training	0.016	0.107	51.000	0.153	0.879
Perception training	-0.061	0.108	52.000	-0.566	0.574

Table 3.5: Selected coefficients of reaction time model for simple effects, experiment 2. Contrast type levels are reported only if they significantly differed from the base level. Reported degrees of freedom, t values, and p values are derived from Satterthwaite approximations. Interaction terms are excluded from this table.

Significant fixed effects are reported in table 3.5. Because all fixed effects entered into interactions, these effects are best investigated with pairwise comparisons, reported in table 3.6. The interaction of session and training type (see figure 3.4) evaluates the degree to which different training groups were faster at making a correct response after training. Post-hoc tests (Tukey's HSD) revealed that only subjects in the perception training group reduced their speed after training (β = -0.242, t = -3.868, p *adj.* = 0.002). This may indicate that only perception training was effective in reducing speed of correct judgments. However, an alternate explanation is that this finding reflects a task practice effect; subjects in the

[2]A similar outlier removal procedure to the one used for the sensitivity analysis (section 3.4.1) was explored for the reaction time model. However, while this procedure did identify 612 outlier values (1.82% of the data set), removal of them did not improve the normality of the residuals, as assessed by a two-tailed Kolmogorov-Smirnoff test, (original model $D = 0.1605$, $p < 0.0001$, updated model $D = 0.1693$, $p < 0.0001$). Because outliers > 2 SD had already been removed from the data set, and because this original pruning did not improve model fit, it was decided that the original model would be reported. However, the results of this model should be interpreted with caution as a result.

perception training group completed AX discrimination tasks for both training and testing, while subjects in the 3 production training groups had a different task during training. As a result, subjects who received perception training may simply be more practiced at responding quickly to an AX discrimination trial than the other groups.

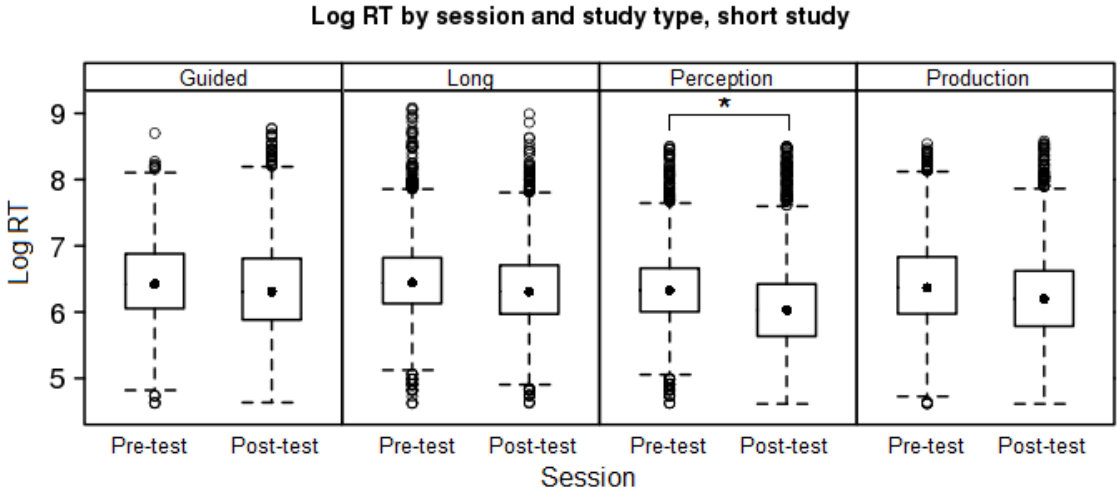


Figure 3.4: Session * training type, reaction time model, experiment 2.

The lower panel of table 3.6 reports pairwise comparisons for the interaction of session and contrast type. Most contrasts showed a significant or marginal decrease in reaction time after training. Contrasts which did not were the place + voiced vs. voiceless contrast (p *adj.* $= 0.102$) and "same" trials (p *adj.* $= 0.257$). Regarding the former, it is likely that this contrast was already relatively fast at pre-test, as it maps well onto the English /t/-/d/ distinction. The lack of an effect for "same" trials, coupled with the decrease in accuracy for these trials reported in section 3.4.2, may indicate that even when subjects correctly identified two syllables as containing the same consonant, that they had relatively more uncertainty or a "different" bias to contend with in making that judgment.

In summary, the reaction time analysis revealed that speed of response may index both improved discrimination and task practice effects. The restriction of reduced reaction times to the perception training group suggests that practice with the AX discrimination task facilitates speeded responses in the post-test session. The significant contrast type * session interaction can be interpreted as meaning that this practice effect does not mean faster responses across the board - reduced reaction times were found in those contrasts which showed enhanced discrimination, but not those which were already well-discriminated (place + voiced vs. voiceless) or those which showed reduced accuracy ("same" trials).

SESSION * STUDY TYPE				
Comparsion (post-test - pre-test)	β	**s.e.**	***t***	***p***
Production training	-0.067	0.062	-1.093	0.941
Guided production training	0.019	0.061	0.315	1.000
Long production training	-0.014	0.061	-0.227	1.000
Perception training	-0.242	0.063	-3.868	0.002
SESSION * CONTRAST TYPE				
Comparison (post-test - pre-test)	β	**s.e.**	***t***	***p***
Aspirated vs. breathy	-0.3086	0.1123	-2.747	0.0167
Aspirated vs. voiced	-0.3323	0.1110	-2.995	0.0076
Aspirated vs. voiceless	-0.3098	0.1111	-2.789	0.0149
Breathy vs. voiced	-0.3305	0.1119	-2.953	0.0094
Breathy vs. voiceless	-0.3430	0.1117	-3.071	0.0065
Voiced vs. voiceless	-0.2754	0.1162	-2.370	0.0449
Place + aspirated vs. breathy	-0.4096	0.1121	-2.761	0.0161
Place + aspirated vs. voiced	-0.3029	0.1110	-2.730	0.0176
Place + aspirated vs. voiceless	-0.3370	0.1110	-3.035	0.0071
Place + breathy vs. voiced	-0.3213	0.1117	-2.876	0.0116
Place + breathy vs. voiceless	-0.3666	0.1116	-3.286	0.0035
Place + voiced vs. voiceless	-0.2298	0.1145	-2.007	0.1018
Place	-0.2858	0.1122	-2.546	0.0288
Same	-0.1657	0.1095	-1.513	0.2573

Table 3.6: Pairwise comparisons of interaction terms in reaction time model, experiment 2, as assessed with Tukey's Honest Significant Difference test.

3.5 Discussion

The experiment reported in this chapter expands upon the findings of experiment 1, and explicitly tests the contributions of within-mode and cross-modal learning the acquisition of non-native phoneme categories. In experiment 1, learners always received perceptual training before they received articulatory training, as well as receiving more time with the former than the latter. This prevented a controlled, systematic comparison of the efficacy of each training type. In experiment 2, learners were complete novices when they received either perceptual or articulatory training, and so their effects could be more directly compared.

The results of all three analyses indicate that subjects who received perception training and subjects who received production training both improved their discrimination of the target contrasts. There was an effect of session but no effect of training type in the sensitivity (d') analysis, suggesting that subjects in all groups were more able to detect the target contrasts after training. The accuracy analysis, which has more statistical power to detect differences between groups, found that all but one production training group improved after

training (with the caveat that accuracy for "same" trials declined, suggesting a bias for responding "different" after training). The reaction time analysis only showed improvement (i.e. reduced response times) after training for the perception training group.

This last result may suggest that perception training was more effective, as all three metrics (sensitivity, accuracy, and reaction time) improved. Alternately, it may suggest a task practice effect, as the production training group both trained and tested with AX discrimination tasks. The two explanations are not mutually exclusive, and to dissociate the two, it would be necessary to devise a perceptual training task that did not use AX discrimination, in order to make it like the production training in its dissimilarity to the testing task. But even if perception training is more effective for discrimination than production training - a not-unreasonable outcome, as perception training is within-mode learning - it remains the case that production training groups also showed improvement in their ability to discriminate the target sounds after training. This indicates that cross-modal learning can contribute to fledgling representations of novel categories.

To explore the reasons why production training was ineffective in experiment 1, three types of related production training paradigms were tested. The first, (basic) production training, used the same methodology as the production training in experiment 1. Improved performance as a result of this training would suggest that the failure to detect an effect in experiment 1 could be attributed to a "local ceiling" - that is, subjects had shown as much improvement as they were likely to in the confines of the experimental paradigm by the time they had gone through perceptual training, and adding articulatory information contributed no (detectable) effect on top of this. The second type, long production training, tested the hypothesis that the training in experiment 1 was too short to show an effect, by lengthening the duration of the task. The third, guided production training, provided a higher quality of training by having subjects practice with the experimenter, ensuring that they were engaging with the task and that they had all of their questions answered.

In the sensitivity and reaction time analyses, there was no distinction between the three production training groups. The sensitivity analysis showed that subjects in all groups improved their ability to detect the target contrasts. This suggests that the presence of articulatory learning in the current paradigm was more influential than its specific implementation. The reaction time analysis showed no reduction in response times for any production training group, suggesting that this training type was equally ineffective at speeding responses across implementations. These two analyses suggest that perhaps the "local ceiling" was the most likely culprit in experiment 1 - it was not duration or quality of training that impacted the ability to detect a change in performance so much as subjects starting from a complete baseline state.

However, there was a distinction in the accuracy analysis - subjects in the basic production training and guided production training groups improved the accuracy of their responses, but subjects in the long production training group did not. This is a surprising finding - the long production training is simply an extended version of the basic production training, so if articulatory information is helpful at all, it should logically be more beneficial with increased exposure, not less. One possible explanation from this finding may have to do with subjects'

imperfect performance in the execution of the articulatory targets for each novel category. The long production training paradigm was extended during the repetition task only - that is, the portion where subjects heard a target syllable and repeated it back, with visual cues from the training present to remind them about articulatory targets. It was hypothesized that this lengthened repetition task would increase the link between subjects' representations of the perceptual category and the articulatory category, as they had the target articulation reinforced every time they heard a target token. However, it is likely that subjects' own attempts to articulate the targets were imperfect, as they were still learning to command the place and voicing features that are not present in English. Perhaps, then, the task reinforced a link between the target category and the subjects' own imperfect tokens, rather than the desired link between the categories and more abstract/ideal representations modeled by the native speaker. If that is the case, then a longer repetition task would emphasize erroneous exemplars, and potentially hinder the ability to discriminate the well-formed tokens tested in the post-test session.

The hypothesis that subjects' imperfect articulatory representations interfered with their ability to discriminate the target contrasts bears some similarity to research on implicit training during acquisition of perceptual categories. Vlahou and colleagues (2011) and Seitz and colleagues (2010) have made the argument that explicit attention directed to a target novel contrast by non-native speakers can actually have a negative impact on performance, because learners' own internal models and native language biases lead them to incorrect hypotheses about the actual target of learning. It is possible that a similar interference occurred cross-modally in the current study - if subjects' productions were sufficiently inaccurate, repeated productions would reinforce these erroneous targets and lead them to build representations that did not reflect the actual targets they were trying to discriminate. If this is true, it would suggest that cross-modal learning, like other forms of explicit training, is highly sensitive to the quality of subjects' internal representations. It is worth noting that in Catford and Pisoni (1970), much of the articulatory practice was done by practicing small articulatory movements silently; while vocalization was inevitable, the emphasis was on motor routines rather than perceptual feedback. This may have prevented the reinforcement of incorrect auditory cues that may have played a factor in the long production training group of the present study.

Even if interference were not an issue, it is not clear that additional repetitions would be as beneficial as originally hypothesized. Wright and colleagues (2015) examined effects of training duration on acquisition of a non-native contrast. While 30 trials was insufficient to show improvement in discrimination, improvement was detected after 60 trials, and no additional benefit was conferred for training with 120 or 240 trials. While it remains an open question as to whether this effect would transfer straightforwardly to articulatory training, the finding indicates that in general, more trials may not be better in a training paradigm.

It should be noted that the efficacy of the various production training methodologies - basic, long, or guided - are largely internal to the current experimental paradigm. They were designed explicitly to test hypotheses about the outcome of experiment 1, but may not be generalizable to broader contexts. For example, even if there is a ceiling effect for duration

of training in a laboratory paradigm, this certainly does not imply that more exposure to a language will not be more beneficial over the long term when an individual is learning a new language. Similarly, a more engaged and higher quality of instruction should still be more beneficial than more cursory instruction. The findings here relate primarily to the earliest levels of category acquisition, but may not generalize to the entire learning process. They are best contextualized as contributing to our understanding of what learners do when they are first becoming aware of a category distinction.

Finally, the "local ceiling" hypothesis makes an interesting prediction about the nature of category learning in its earliest stages. If it is true that subjects in experiment 1 were no longer sufficiently "novice enough" to benefit from cross-modal learning after receiving perception training, then perhaps it is the case that cross-modal learning is most effective with true novices - those who have no prior exposure to the target information at all. If a learner has no information about the number or nature of the target contrasts in a new language, perhaps any information which informs them of those targets is useful to get them over the very initial hurdle of their native language biases, including cross-modal information. Once they already have that information, more relevant domain-specific information (i.e. acoustic/perceptual information for discrimination or identification, and articulatory information for production) may be necessary to improve upon the quality of performance.

3.6 Conclusion

The experiment in this chapter was designed to follow up on the efficacy of cross-modal learning for adults acquiring novel contrasts in a second language. Completely novice learners who received a single session of training improved their discrimination of the target Hindi consonant contrasts, indicating that the very first steps in acquiring a new contrast can take place in a short time frame. Cross-modal learning (articulatory learning for discrimination) was effective in improving accuracy and sensitivity, but only within-mode learning (perceptual learning for discrimination) showed an effect on the speed of responses. Taken together with the the results of experiment 1, these findings suggest that cross-modal information may be less effective than training which targets the relevant domain, but that it can still have an impact, particularly if learners are at the very beginning stages of acquisition.

Chapter 4

Pre-attentive processing of novel phonetic categories

This chapter focuses on the question of pre-attentive responses to phonemic categories at a neural level. The experiments described in the previous chapter demonstrated that improvements in discrimination of non-native sounds are detectable after articulatory training. The experiment described in this chapter extends that work, by investigating the extent to which cross-modal (perception-production) links in phonemic learning can be detected at a cortical level. A primary goal of this experiment is to expand upon the current understanding of neural representations of nascent and still-developing phonetic categories.

4.1 Background

4.1.1 Neural representations of phonetic categories

Identifying the development of newly-forming phonemes in the brain requires a general understanding of the cortical representation of phonetic categories. Current cortical models of spoken language processing (Scott & Wise, 2004; Hickok & Poeppel, 2004, 2007) locate the primary area of phonetic and phonological processing in the posterior superior temporal lobe. The Hickok and Poeppel model specifically differentiates between spectrotemporal processing of the acoustic speech signal in the bilateral superior temporal gyri (STG) and more abstract phonological representations in the superior temporal sulci (STS). These representations then feed into dual pathways: a ventral stream that links acoustic to auditory information, and a dorsal stream that connects phonetic/phonological representations to articulatory/motor representations in the inferior frontal gyrus (IFG) and primary motor cortex.

In many studies of speech perception, it is assumed that successful spoken language comprehension is supported by the transformation of variable acoustic input to invariant phonetic categories at some point along the processing stream. For models which assume

this invariant abstract representation, this has raised the issue of locating the site of phonetic category invariance at a neural level. Recent work on native language phonetic categorization using intracranial recordings has found that the hallmarks of categorical perception - invariance for within-category tokens, and heightened sensitivity to cross-category tokens - can be detected in the posterior STG (Chang et al., 2010). This suggests that category invariance happens early in the processing stream, and may reflect tuning of the auditory cortex to the spectrotemporal features of sounds in the native language (Köver & Bao, 2010). This coincides with theoretical accounts of speech perception which argue that the basis of phonetic categories is in the acoustic properties of speech sounds (Diehl et al., 2004). In fact, tuning to the native language has even been found in the brainstem for pitch/tone processing (Krishnan, Xu, Gandour, & Cariani, 2005), suggesting that even lower-level components of the auditory pathway can become sensitive to language-specific patterns.

However, others argue that downstream regions more connected to motor processing display category invariance more clearly. Myers, Blumstein, Walsh, & Eliassen (2009) found sensitivity for both within- and across-category differences to native sounds in the left STG, while responses in the left IFG were invariant to within-category differences. Kovelman, Yip, & Beck (2011) found that category invariance in the IFG was restricted to native categories, suggesting that this region may privilege well-learned categories. Myers & Swan (2012) ran a training paradigm for non-native sounds in an fMRI study, and found pre- and post-training differences in frontal regions, particularly the bilateral middle frontal gyri, consistent with category learning. This pattern was also supported by Golestani & Zatorre (2004), who observed correlations between behavioral performance and efficiency of frontal lobe processing for non-native sounds after training. Findings of this type have been taken as evidence the idea that articulatory processing is more central to categorical perception of speech (Liberman et al., 1967; Liberman & Mattingly, 1985; Galantucci et al., 2006), as these frontal regions around Broca's area are linked to articulatory processing along the dorsal stream (Hickok & Poeppel, 2007).

In adjudicating between these two accounts, it is worth noting that at least three crucial factors could confound the comparison between the findings of these two camps. The first is the issue of temporal resolution. The findings of Chang and colleagues (2010) relied on the high spatiotemporal resolution of intracranial data, which was critical to their results - categorical patterns in their data emerged after 100 ms and then decayed again after 200 ms, a time scale that is not attainable using fMRI.

A second factor concerns analytical techniques. Neuroimaging papers focused on magnitude differences have found an advantage for category processing in the IFG (Kovelman et al., 2011; Myers & Swan, 2012; Myers et al., 2009). However, approaches which inspect finer-grained spatial patterns, such as multi-pattern voxel analysis (MVPA), have found more promising category separability in the auditory cortex (Raizada, Tsao, Liu, & Kuhl, 2010; Ley et al., 2012). One study which applied this approach to frontal regions found that MVPA could distinguish categories in Broca's area (Lee, Turkeltaub, Granger, & Raizada, 2012), suggesting that this approach may be more generally robust than subtraction-based approaches.

Finally, a task-related caveat must be considered. Responses in tasks which require an explicit judgment engage decision-related processing in addition to categorical processing, and these could become conflated when looking at responses in temporal regions for any task that requires an explicit decision. The role of explicit attention and decision will be discussed in more detail in section 4.1.3.

4.1.2 Categorization and the mismatch negativity response

While the ultimate locus of phonetic category invariance is still being debated, the on-line neural processing of categories which are new or developing provides an alternate view on the problem, by inspecting the early formation of categories that have not yet reached a stable representation. One methodology for investigating developing neural representations that has widespread use in phonetic research is electroencephalographic (EEG) recording. EEG data, while spatially diffuse, has a temporal resolution on the order of milliseconds, and shows responses which are closely time-locked to auditory events, making it an ideal candidate for investigating questions related to the rapid, on-line processing of phonetic-level information.

An experimental paradigm that is commonly used in the EEG literature on phonetic processing to demonstrate categorization of speech sounds is the auditory oddball paradigm. This paradigm leverages a component of the auditory evoked-response potential known as the mismatch negativity (MMN), which occurs when a change is detected in an otherwise repetitive auditory stream. (For recent overviews of the use of the mismatch negativity in auditory and cognition-related research, see Näätänen, 2001 and Kujala & Näätänen, 2010). The MMN response is primarily generated in the temporal cortices and dorsal pre-frontal regions (Alho, Woods, Algazi, Knight, & Näätänen, 1994; Alain, Woods, & Knight, 1998), reflecting both the contribution of auditory sensory memory from the auditory cortex and attention switch to the deviant stimulus from frontal sites (Alho, 1995).

The oddball paradigm takes advantage of this response by presenting a repeated "standard" stimulus for a significant portion of a testing session (75% to 90% of all stimuli), mixed randomly with an infrequent "deviant" stimulus (10-25% of stimuli). If the change from the frequent deviant to the infrequent standard is salient in some way, an MMN response is elicited at approximately 100-250 ms after the deviant onset (Kujala & Näätänen, 2010). In this way, the MMN response acts as a change detector, indicating whether the differences between the standard and deviant stimuli are detectable to the listener.

Cross-language differences in the mismatch negativity response

Early auditory MMN studies seemed to indicate that within-category acoustic differences could elicit an MMN (Sharma, Kraus, McGee, Carrell, & Nicol, 1993; Maiste, Wiens, Hunt, Scherg, & Picton, 1995; Kraus et al., 1995), suggesting that the response was operating at a sensory level and sensitive to any acoustic change. However, a number of recent studies investigating cross-linguistic speech perception have shown that a switch between stimuli which are not phonemic in the listener's language may not elicit an MMN. This suggests

that the MMN is not a purely sensory response, but shows some sensitivity to more abstract categories that differ between languages or dialects (Brunellière, Dufour, & Nguyen, 2011). These studies indicate that the MMN response does not always operate as a perceptual/acoustic change detector. Rather, it can be sensitive to phonemic-level processing and categorization. As a result of this property, the MMN response has proven to be a useful diagnostic for the ability of listeners to discriminate two categories. Developmental MMN data suggests that this commitment to native language categories is not strongly detectable at 7 months, but emerges by 11 months (Rivera-Gaxiola, Silva-Pereyra, & Kuhl, 2005).

In an early cross-linguistic study, Näätänen and colleagues (1997) found that the MMN response was sensitive to language-specific prototypicality. In their study, native Finnish and native Estonian speakers listened to a standard [e] (phonemic in both languages) with deviants [ö] (phonemic in both languages) and [õ] (phonemic in Estonian; not a prototypical member of any Finnish category). The response to [õ] was reduced in Finnish listeners compared to their native deviant [ö], as well as compared to the response to [õ] by Estonian speakers, suggesting that the presence or absence of a separate category [õ] mediated the MMN response. A similar study with native Hungarian speakers who either were or were not proficient in Finnish (Winkler et al., 1999) showed that second-language proficiency mediated responses to Finnish vowels. Hungarians who were fluent in Finnish showed an MMN response to the deviant [æ] as compared to the standard [e]; Hungarian speakers without Finnish experience, who do not have the [æ/e] distinction in their native language, showed no MMN at all. Peltola et al. (2003) found that the amplitude of the MMN response to a contrast was lessened for non-native speakers compared to native speakers, a finding they attribute to incomplete acquisition from classroom-based learning.

Dehaene-Lambertz (1997) designed an experiment which demonstrated both within-category invariance and cross-category categorization in a single population. In this study, native French speakers showed an MMN response for a native category change ([b] vs [d]) but not a VOT difference that fell within a single French category (long VOT [d] vs. short VOT [d]) or that used a non-native place of articulation contrast ([d̪] vs. [ɖ]). Sharma and Dorman (2000) found a similar language-specific effect in the VOT domain: Hindi (but not English) speakers showed a MMN for the distinction between prevoiced [b] and unvoiced [p] along a VOT continuum, even though both groups showed an identical N1 sensitivity for the duration of the pre-voicing (as evidenced by the latency of the N1 component).

The topography of the effect may be dependent on language experience, even for different groups which both show evidence of categorization. Zevin, Datta, Maurer, Rosania, and McCandliss (2010) compared responses to the /ɹ/-/l/ distinction by native speakers of English and speakers of Japanese (for whom the contrast is not native) who learned English as a second language. Both groups showed a reliable mismatch response, but source analyses suggested that the English native speakers had a stronger left-lateralized pole than the more central-right source in Japanese native speakers. The researchers propose that this may reflect differences in the phonetic status of the contrast in these two groups.

Indexing phonetic learning

The MMN response has been used as an index for phonetic learning in short-term training paradigms. If training can develop a fledgling representation of a category which a listener did not previously recognize as distinct from another, then a MMN response would be expected after training when it was not present before it.

Ylinen et al. (2009) used the MMN response to demonstrate a re-weighting of perceptual cues as a function of training. The English /i/-/ɪ/ distinction in English is cued by both duration and spectral cues. For Finnish speakers, who use length contrasts but not this particular spectral contrast in their native language, the duration cue tends to be more heavily weighted when distinguishing this English contrast. After training, Finnish speakers were able to more reliably distinguish a set of /i/-/ɪ/ minimal pairs (e.g. "beat" - "bit", "peach" - "pitch") that had been modified to have equal vowel durations, leaving only the spectral cues present. The post-training MMN response was similar in Finnish and English native speakers, suggesting that Finnish speakers had learned to re-weight spectral cues in a more English-like way as a function of training.

Cheng and Zhang (2013) tested the same contrast for native Chinese speakers, and included multi-talker variability, audiovisual cues, and adaptive fading in their training paradigm. The post-training MMN response indicated learning that was corroborated by behavioral discrimination and identification, in particular a shift from perception of a continuum as a continuous distribution to discrete, categorical groups. They take their results as evidence that audiovisual training can be an important contribution to the phonetic training paradigm.

Kaan, Wayland, Bao, and Barkley (2007) found native language-specific MMN responses after training to the perception of Thai tones by Thai, Mandarin, and English speakers. All listeners showed an MMN to a low falling tone deviant (compared to a level tone standard) before an after training. The MMN response to the high rising deviant showed a strongly attenuated MMN by comparison, and was increased by training only for the English group, suggesting that different languages are causing listeners to key in to different parts of the signal.

Zhang et al. (2009) used magnetoencephalography to study the mismatch field (MMF) response, the magnetic correlate of the MMN, in Japanese adults with limited exposure to English. Their paradigm was designed to mimic what they believe to be the process of first language phoneme acquisition, and employed exaggerated cues and and adaptive fading to teach the /ɹ/-/l/ distinction across 12 sessions. They found an increased MMF to the target contrast after training, and also observed a bilateral decrease in activity with learning, which they attribute to increased efficiency in processing the contrast.

It is possible that the MMN as an index of learning could even precede behavioral discrimination of a novel contrast. Tremblay, Kraus, and McGee (1998) exposed ten listeners to a novel VOT contrast over the course of nine days of training and testing. In all subjects, the first day of training (day 4) was sufficient to elicit a detectable change in the MMN response. In five subjects, improvement in behavioral identification of the contrast occurred

on the same day; four more showed behavioral improvement on following day. (One subject failed to show behavioral identification of the contrast by the end of testing, despite a reliable MMN on day 4.) In all cases, the MMN response indicating learning of the novel contrast occurred as early or earlier than the behavioral response. Tremblay and colleagues suggest that the MMN response may be driven by fast sensory-based processing which can produce an automatic response, while the decision-based behavioral response requires retrieval of this information and explicit attention, leading to a lag in performance improvement. These results further suggest that the pre-attentive MMN response is fairly consistent across individuals, and that individual variation connected with higher cognitive processes may be more apparent in behavioral responses.

4.1.3 The role of attention

The findings of Tremblay, Kraus, and McGee (1998) highlight an additional benefit for the use of the MMN response to study early category learning: the effect does not require explicit attention or overt judgment on the part of the learner. While some mismatch studies ask subjects to identify deviant trials (e.g. Sjerps, Mitterer, & McQueen, 2011), auditory oddball paradigms are often run as passive listening tasks where subjects complete a distractor visual task (e.g. Hisagi, Shafer, Strange, & Sussman, 2010) or attend to a silent film (e.g. Zevin et al., 2010). This means that subjects need not have formed overt hypotheses about the existence of categories in order for category learning to be demonstrated.

In the behavioral literature on non-native phoneme acquisition, there is some evidence that overt attention to the target contrast may not be the best way to learn a novel contrast. Several studies (Seitz et al., 2010; Lim & Holt, 2011) have found that implicit training, in which a contrast is cued but is not brought to the attention of the learner, is as successful or more successful than paradigms where explicit attention and feedback is used. Vlahou and colleagues (2011) found that implicit training outperformed explicit attention to target contrasts with performance feedback, suggesting that performance may be best when subjects are not asked to make explicit judgments - and thus form explicit (and potentially inaccurate) hypotheses - about the target contrasts. Gulian and colleagues (Gulian et al., 2007) argued that in their paradigm, explicit training *hampered* learning compared to an implicit learning strategy which used distributional cues.

However, explicit attention by early learners does not always inhibit accurate categorization. There are several studies showing positive effects of selective attention to acoustic cues that signal a target contrast (McGuire, 2008) or to the segment in a minimal pair that defines the contrast (Pederson & Guion-Anderson, 2010). Gordon and colleagues (Gordon, Eberhardt, & Rueckl, 1993) suggest that the benefit of attention to an acoustic feature may depend on whether that feature is a "strong" or "weak" cue to the relevant phonemic contrast.

One relevant aspect of behavioral studies which manipulate attention is that they require either overt judgments about either the target itself (in the case of explicit attention studies) or explicit attention to a correlated cue (in the case of implicit tasks). In both

cases, attention is being directed to some associated feature, whether or not it is the target feature. ERP data has the advantage of being able to show a response without requiring an overt behavioral response, and thus can access a more purely implicit or attention-free response. An interesting example of this was documented by Tremblay, Kraus, and McGee (K. Tremblay et al., 1998), who trained subjects to identify a VOT boundary that was previously unfamiliar to them. They measured learning in two metrics run in separate tasks: an MMN response in an auditory oddball task, and a behavioral identification response. For all subjects, the MMN indicating discrimination of the new VOT categories happened at least as early as successful behavioral identification, and in half of subjects the MMN preceded the behavioral response. This indicates that automatic and unconscious discrimination can precede an overt judgment about a categorical judgment.

Hisagi and colleagues (Hisagi et al., 2010) found a more complex pattern in ERP data. They tested Japanese and English learners on an oddball task which varied vowel length, a property that is phonemic in Japanese but not English. Modulating attention (by having subjects attend to either the auditory oddballs or to an unrelated distractor task) affected the strength of the MMN for English, but not Japanese, subjects. This finding suggests that well-learned (i.e. native) contrasts may induce automatic categorical responses, while novel or sufficiently unfamiliar contrasts may require explicit attention.

4.2 The current study

The present study is designed to build upon the existing literature in non-native phoneme acquisition using the MMN as a signal for detecting successful discrimination. It has two primary goals. The first is to expand the methodological approaches taken in the ERP literature on phoneme learning. To my knowledge, the use of articulatory/production-focused training to improve perceptual discrimination has not yet been explored with a mismatch negativity study (although the findings of Cheng & Zhang, 2013 indicate that audiovisual information can be effective, suggesting promise for the current approach). If the present study shows an increase in the MMN response as a function of production training, it will provide evidence that articulatory information can have a rapid effect on a signal that is tied to auditory/acoustic processing.

A second goal is to examine whether learning can be detected more quickly, or more sensitively, in an automatic neural response as compared to a conscious behavioral response. There is evidence that training-related changes in speech-related learning may be detectable more quickly at the neural level than in behavioral responses (K. Tremblay et al., 1998). If this holds for non-native phoneme acquisition, then performance during the MMN task may be stronger than the behavioral findings reported in experiment 2.

Session code	Session	Tasks	Stimuli	Feedback?
A	Pre-test	passive listening	CV natural (normalized)	no
B	Training	production training	CV natural	no
C	Post-test	passive listening	CV natural (normalized)	no

Table 4.1: Structure of all studies in the experiment 3 series.

4.3 Methods

4.3.1 Procedure

The structure of experiment 3 was modeled after experiment 2, in that it consisted of pre-test, training, and post-test sessions conducted in a single sitting. A summary of experiment procedures is outlined in table 4.1.

While the basic structure of the experiment was similar to experiment 2, the pre-test and post-test sessions were different; in experiment 3, these sessions were passive listening tasks. Subjects were instructed to sit quietly and watch a short silent movie (one per session; movies were selected to be comprehensible without sound and to have minimal screen text and intertitles). While they watched, the critical stimuli (see section 4.3.2) played over speakers in the sound-attenuated testing room. Subjects were instructed to ignore the syllables and focus on the movie. The stimuli were presented in a double-oddball paradigm: 80% of the stimuli presented were the standard, 10% the voicing deviant, and 10% the place deviant. Each stimulus was 424 milliseconds long, and the stimulus-onset asynchrony was 1 second. Stimuli were presented in pseudo-random order, with the restriction that each deviant had to be preceded by at least two standards. Lists of stimuli were generated independently for each subject and each block, using the pseudorandomization software Mix (van Casteren & Davis, 2006). Short break trials (20 seconds in duration) were inserted one-third and two-thirds of the way through each test block.

During the second block, a production training task was run. The training was modified from that used in experiments 1 and 2, to focus only on the contrasts critical to the critical stimuli (the dental-retroflex contrast, and the breathy-voiced contrast). As in experiments 1 and 2, subjects received explicit information about the articulation of the place of articulation and voicing contrasts, and were given visual cues to remind them of articulatory targets while completing a repetition task. In this session, the original stimuli from experiments 1 and 2 were used (see section 4.3.2 for details). No behavioral data was collected during any blocks of the experiment.

4.3.2 Stimuli

Three stimulus tokens were selected from experiment 2 as the critical stimuli for experiment 3. These stimuli were selected to provide one voicing contrast and one place contrast for a base stimulus. The contrast types (vowel identity for both types, voicing type for the

place contrast, and place feature for the voicing contrast) were chosen by looking at the pre- and post-test performance improvement on stimuli in experiment 2[1]. Contrasts were chosen when they had the maximal amount of improvement over all contrast types from pre-test to post-test.

The selected stimulus types were the following: [d̪ʲa] (standard, presented 80% of the time), [da] (voicing deviant, presented 10% of the time), and [d̻ʲa] (place deviant, presented 10% of the time). One token was then chosen for each stimulus type; selected tokens either showed strong improvement in experiment 2 from below-chance to above-chance performance, or were selected because they had the most similar vowel quality to the other tokens.

The selected tokens, as used in experiment 2, contain variations that are not critical to the target contrasts, but which could confound the result of the experiment if they are sufficient to cue subjects to the differences between two stimuli. To avoid this confound, stimuli were resynthesized to remove as much non-critical variation as possible. The pitch of each stimulus had already been flattened in the original stimulus creation (see chapter 2 for details), so the duration and amplitude were the focus of this synthesis.

First, stimuli had their durations normalized to 424 milliseconds (+/- 0.5 ms). To do this, the prevoicing period was set to 65 milliseconds in each syllable, by adding or subtracting pitch pulses from the prevoicing period. (The original [da] token did not have a very clear prevoicing period but was otherwise a better match to the standard stimulus in terms of vowel quality; for this stimulus, the prevoicing portion was spliced in from another [da] token with clearer prevoicing.) Once the prevoicing periods were equalized, the duration of the whole syllable was matched to 424 milliseconds by adding or subtracting pitch pulses from the vowel. Next, the amplitudes of the three stimuli were matched by setting the maximum RMS level in each to -20.70 dB (+/- 0.11) using the SoX (Sound eXchange) program `level`.

The training session did not use the re-synthesized stimuli described above; instead, all tokens of [d̪ʲa], [da], and [d̻ʲa] from experiment 2 (12 in total) were used. This approach exposed subjects to within-category variation, as they were in experiments 1 and 2. In addition, the use of these tokens ensured that subjects did not become overtrained on the critical stimuli. If the critical stimuli were reused as training stimuli, this may have led to MMN responses that were indicative of detection of an acoustic change, rather than a more generalized categorical response.

4.3.3 Subjects

Eighteen subjects were recruited to participate in the experiment. Of these, two were excluded from final analysis (one due to recording error, the other due to persistent low-frequency noise in ERPs which remained after data processing), leaving sixteen subjects

[1]Stimuli for experiment 3 were developed while experiment 2 was still in progress; the data contributing to stimulus selection came from 25 participants

who were included in the final analysis. All subjects were at least eighteen years of age, native speakers of English, and had no experience with Hindi or other languages with the target phonetic contrasts (dental-retroflex, breathy-voiced).

4.3.4 Data collection and processing

Recording

Recording took place in a electrically-shielded, sound-attenuated room. Subjects were seated in an armchair approximately 125 cm from a screen and speakers which presented the stimuli at 65 dB SPL. Subjects wore a 64-channel electrode cap (see figure 4.1). In addition, five external electrodes were placed outside the outer corner of each eye, below the right eye, and on each earlobe.

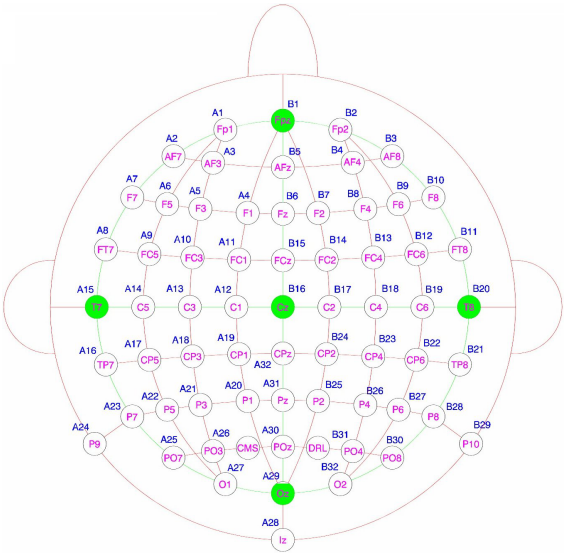


Figure 4.1: Arrangement of electrodes in 64-channel BioSemi cap, experiment 3.

The EEG data was recorded using a BioSemi ActiveTwo 64-channel EEG system and BioSemi ActiView 5.1 software. The raw EEG signal was amplified and digitized at 1024 Hz and band-pass filtered between 0.01 - 100 Hz during recording. Electrode impedance was kept below 25 kΩ; when this was not possible, electrodes with high impedance were removed during pre-processing.

Pre-processing

Recorded EEG data was pre-processed using custom scripts which implemented routines developed for the EEGLAB MATLAB Toolbox (Delorme & Makeig, 2004). For each subject and block (pre-test or post-test), the data was downsampled to 256 Hz and band-pass

filtered between 0.1 - 55 Hz. The data was referenced to the average of the two earlobe channels. Muscle artifacts were removed using the Automatic Artifact Removal (AAR) toolbox (Gomez-Herrero et al., 2006). To remove artifacts attributable to eye blinks, independent components analysis (ICA) was performed on the data set; this procedure allowed eye blinks to be isolated and removed from the signal without requiring the removal of entire epochs where a blink occurred. Following ICA, channels which showed excessive noise or which were noted as having high impedance (> 25 kΩ) were removed and replaced by the interpolation of signals from neighboring electrodes. Following this, the data set was split into trial types (standard, place deviant, or voice deviant) and segmented into epochs starting 100 ms before stimulus onset and ending 1000 ms after onset. After segmentation, trials within each trial type were removed if their average power exceeded 5.5 standard deviations above the mean, as these were considered to be likely outliers. This procedure retained 83.2% of trials across trial types and both sessions. The data was z-scored and normalized to the baseline period (-100 to 0 ms before trial onset), and then averaged across trial type (standard, place of articulation deviant, or voicing deviant), within-subject and within-session. Finally, a second band pass filter was applied (1 - 30 Hz) for ERP analysis and visualization.

4.3.5 Data visualization

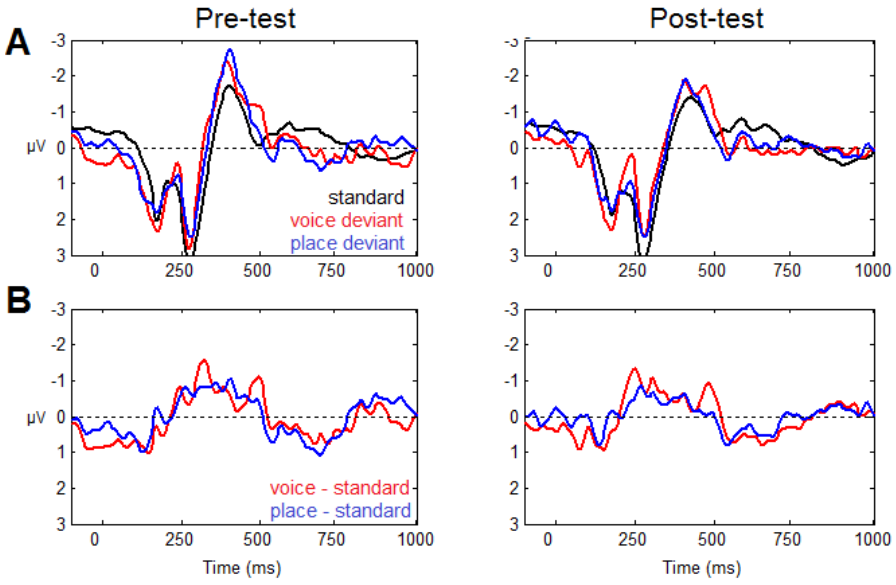


Figure 4.2: Sample (a) ERPs and (b) difference plots for channel FC1 from the pre-test and post-test sessions.

Analysis of the data consisted of comparisons of the averaged response to standard trials, voice deviant trials, and place deviant trials in the pre-test and post-test. Figure 4.2 models the plots that will be presented alongside the analysis in section 4.4, with channel FC1 as

an example electrode. In (a), average evoked responses to the standard, place deviants, and voice deviants are plotted, compared to a baseline period of 100 ms prior to stimulus onset (time 0). Of interest is whether the deviant lines significantly diverge in direction or amplitude from the standard line - or from one another - particularly in the 200-400 ms window which is relevant for the MMN response. In (b), difference plots show the standard response subtracted from each deviant response, to emphasize the difference between them. A deviation from 0 in these plots suggests a difference in response between the standard and the deviant stimulus, which would be indicative of a difference in how the stimulus types were processed. Note that in all visualizations presented here and below, negativities are plotted up and positivities down along the y-axis.

4.4 Results

4.4.1 ERP data

Grand average ERPs for the standard [d̪ʲʰa], voice deviant [d̪a], and place deviant [d̠ʲʰa] are plotted in figures 4.3 (pre-test) and 4.4 (post-test). Some visual separation of the trial types is apparent in frontal channels around 250 ms, and again shortly after 500 ms. There appears to be a slight amplitude difference between the pre-test and post-test sessions, with stronger responses in pre-test ERPs.

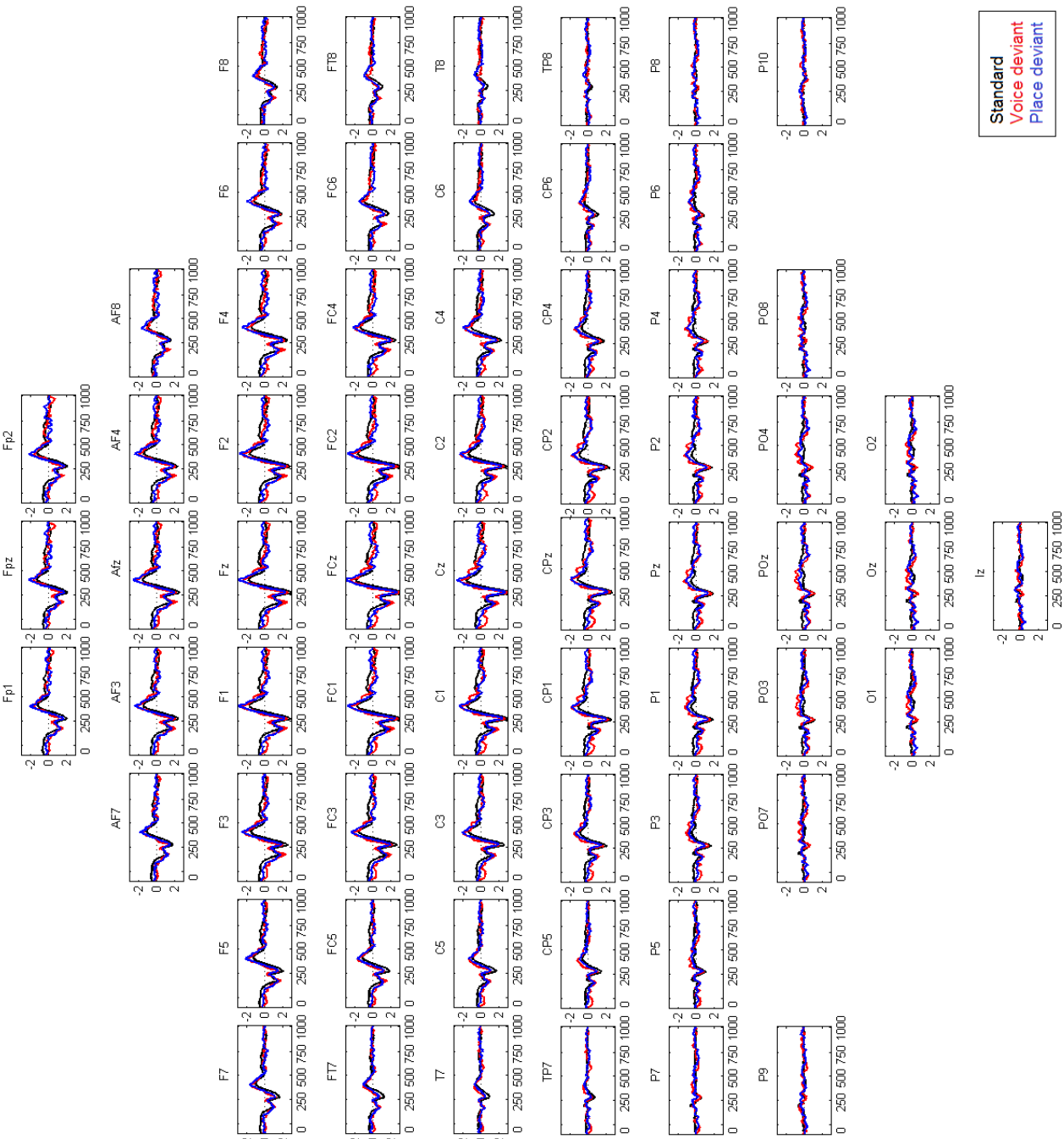


Figure 4.3: Pre-test ERPs for all channels.

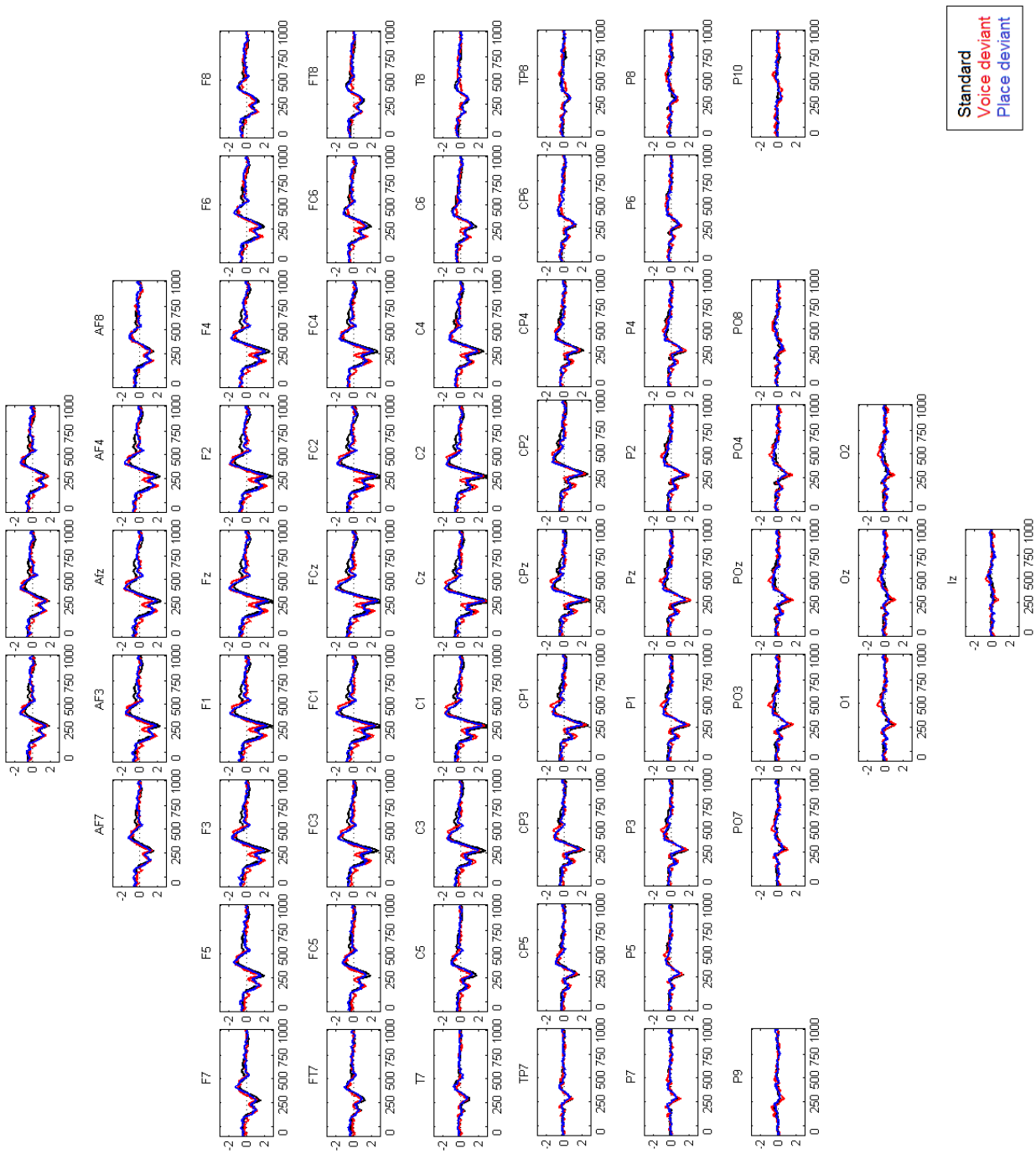


Figure 4.4: Post-test ERPs for all channels.

Inspection of the ERPs in figures 4.3 and 4.4 reveals the lack of a clear N1 component (a negativity centered around 100 ms after stimulus onset) which is typically reported for auditory stimuli (Kujala & Näätänen, 2010). One reason for its absence may be the acoustic structure of the stimuli (visualized in figure 4.5). The syllables used in the current experiment were all consonant-vowel syllables which had relatively long consonantal periods. The consonants had a long prevoicing period (65 milliseconds) prior to the release burst, and the standard [d̰ʱa] and place deviant [d̪ʱa] both had breathy periods as long as 60 milliseconds that preceded full modal voicing of the vowel, which is where the loudest portion of the syllable begins. There is evidence that a long amplitude rise time inhibits the amplitude of the N1 component compared to stimuli with a short rise time - for example, Thomson, Goswami, and Baldeweg (2009) found a marked difference for stimuli which had rise times of 15 or 85 milliseconds (see also Ruhm & Jansen, 1969; Onishi & Davis, 1968).

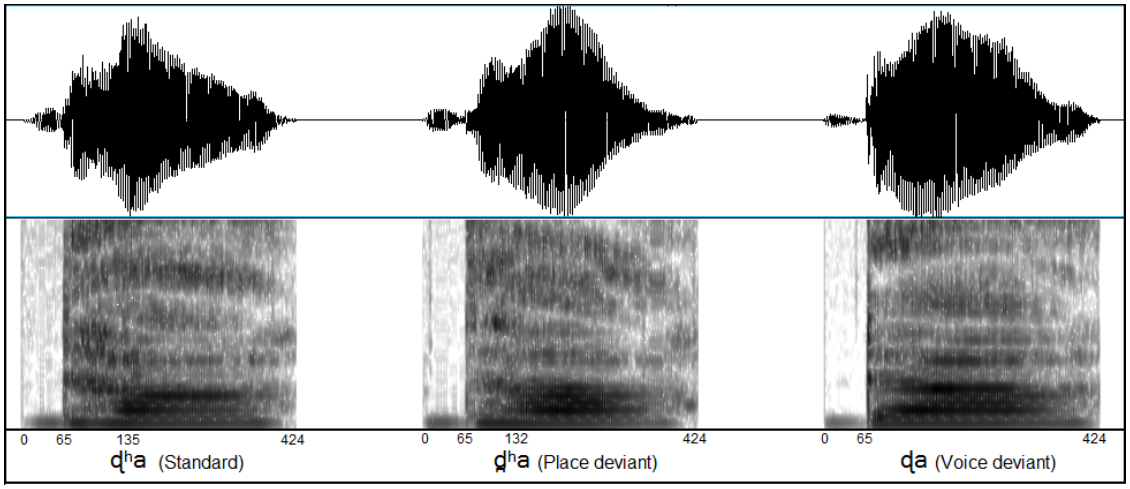


Figure 4.5: Stimuli used in experiment 3. Acoustic landmarks for stimulus onset, burst onset, breathy-voicing offset (in standard and place deviant) and stimulus offset are shown, in ms.

Difference curves, showing the subtraction of the standard response from each deviant response, are plotted in figures 4.6 and 4.7. Visually, the differences between the two deviants and the standard are most apparent in a negativity between 250-500 ms. The MMN response is typically described as occurring at 150-250 ms after change onset (Kujala & Näätänen, 2010). This onset is consistent with the stimuli of this experiment: the first 65 milliseconds of the standard and both deviants are comparable periods of prevoicing, so the first point of departure between stimuli would be at/after the stop release, leading to an expected MMN response onset no sooner than 205-305 ms. In the case of the voice deviant, the difference from the standard would be cued by the lack of breathy voicing after the stop release; for the place deviant, spectral properties of the burst and formant transitions would cue the difference.

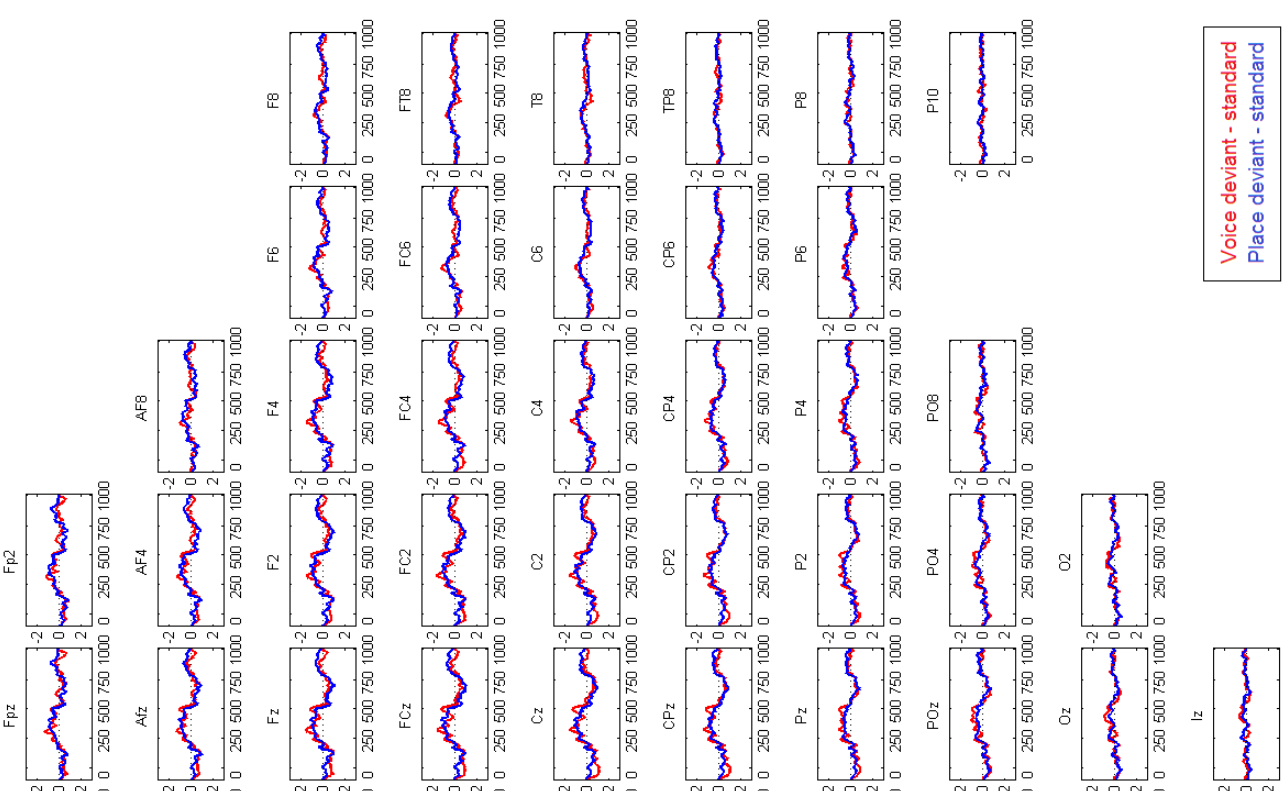


Figure 4.6: Differences curves for pre-test data. Red curves show voice deviant-minus-standard responses; blue curves show place deviant-minus-standard responses.

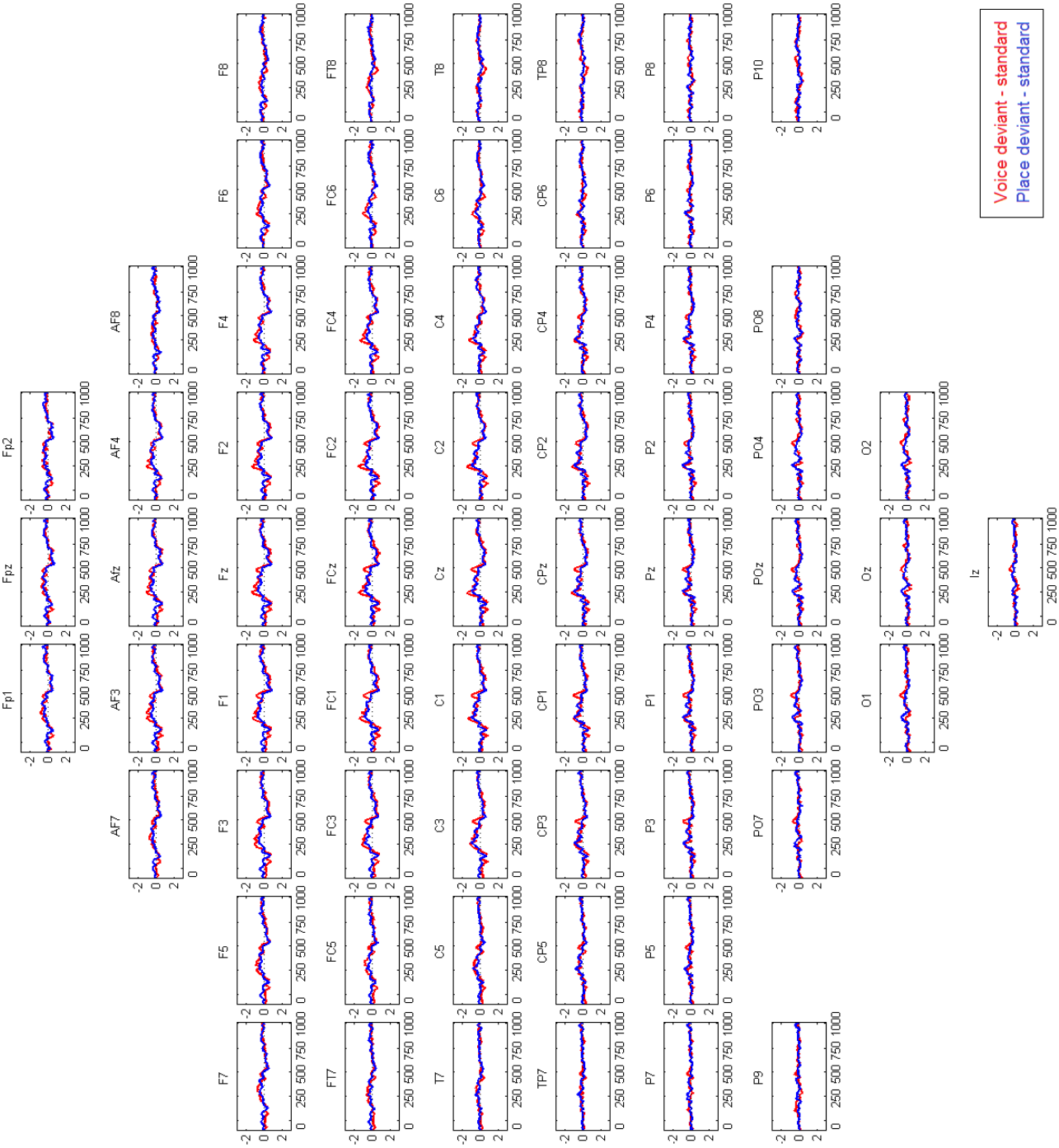


Figure 4.7: Differences curves for post-test data. Red curves show voice deviant-minus-standard responses; blue curves show place deviant-minus-standard responses.

4.4.2 Channel selection for analysis

The mismatch negativity response is typically strongest at frontocentral electrode sites (Näätänen, 1990; Dehaene-Lambertz, 1997) reflecting projections from temporal and dorsal pre-frontal regions (Alain et al., 1998; Alho et al., 1994). The current analysis uses the frontocentral electrodes selected by Brunellière et al. (2011) - FC1, FC2, FCz, C1, C2, and Cz. It also includes Fz, which is commonly investigated in MMN studies (e.g. Sharma & Dorman, 2000; Peltola et al., 2003; Ylinen et al., 2009), as well as the adjacent electrodes F1 and F2. These nine electrodes were used for the analyses reported below, which confirmed that significant by-stimulus differences in activity were found in all nine sites. Selected electrodes are highlighted in figure 4.8.

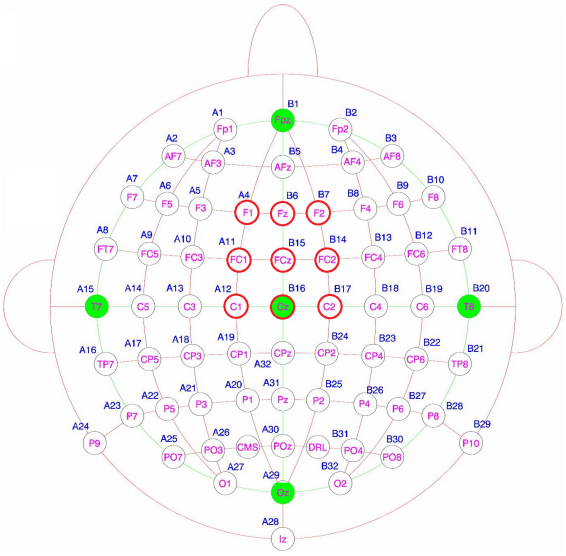


Figure 4.8: Channels selected for analysis, experiment 3. Selected channels are outlined in red.

Pre-test and post-test ERPs and difference curves for the selected channels only are reproduced in figure 4.9.

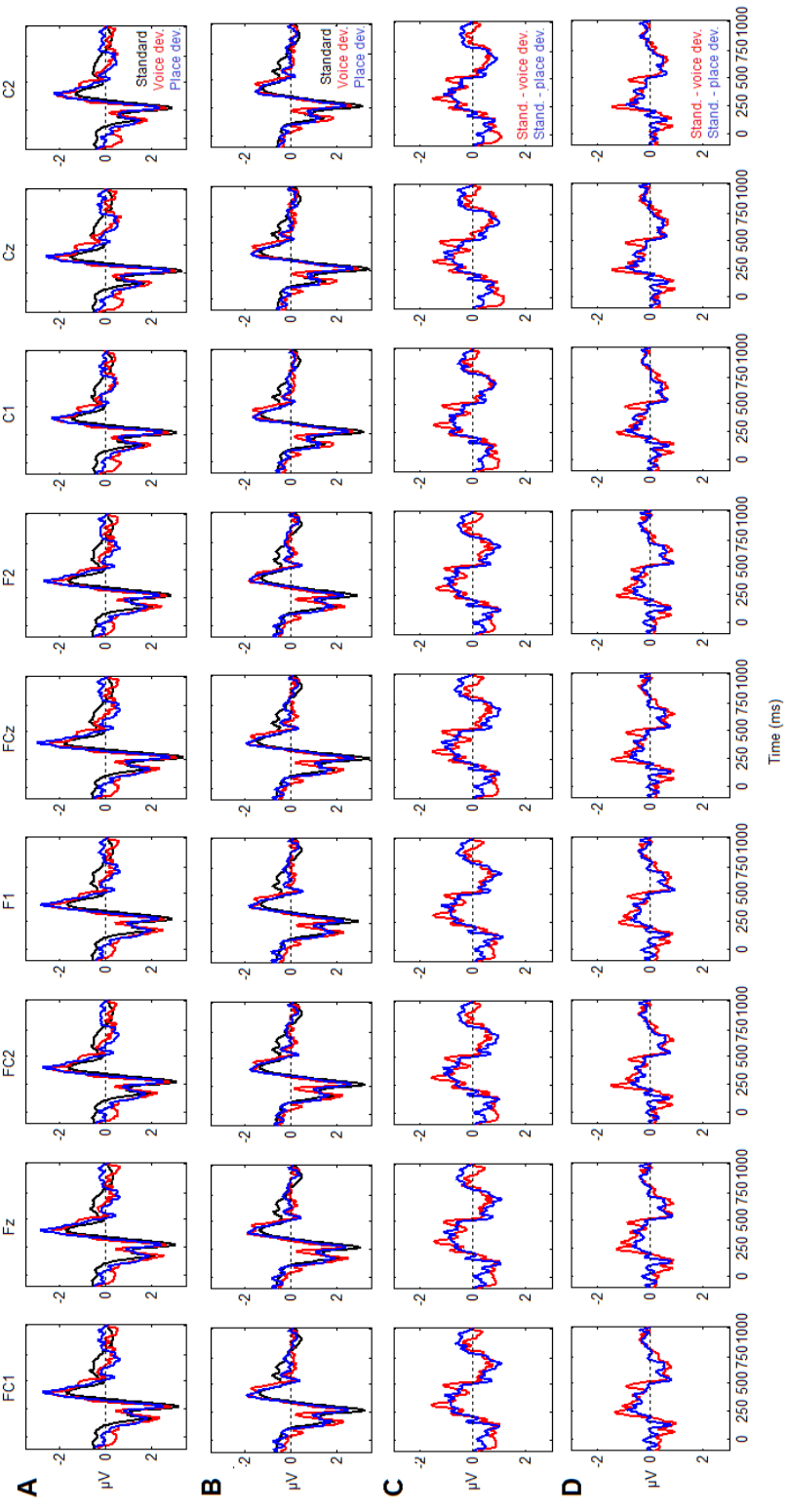


Figure 4.9: (a) Pre-test and (b) post-test ERPs, and (c) pre-test and (d) post-test difference plots, for selected channels.

4.4.3 Analysis of variance

Following the analysis procedures of several MMN studies (e.g. K. Tremblay, Krause, Carrell, & McGee, 1997; Kaan et al., 2007; Diaz et al., 2008; Brunellière et al., 2011), ANOVA was used to analyze the effects of stimulus type and test session in the channels and windows of interest. To analyze the data, a series of seven 50-millisecond windows (figure 4.10) were identified which contained the temporal region of interest. The first window ran from 150-200 ms after stimulus onset, and the last spanned 550-600 ms. An average of the activity for each trial type and both sessions was taken for each window and each critical channel. Then, for each channel/window combination, a two-way ANOVA was conducted using Matlab's `anova2` function, predicting the amplitude of activity as a function of trial type (standard, voice deviant, or place deviant) and session (pre-test or post-test). This analysis addresses three questions of interest: (1) is there a difference in overall activity that distinguishes the standard and deviants, (2) is there a difference in activity before and after training, and (3) does training affect the relationship between the standard and the deviants?

Scalp topographies for the selected analysis windows are visualized in figure 4.10. Visually, differences between the standard and the deviants is present modestly in the 200-250 ms window, and becomes more noticeable at 350 ms in the pre-test, and 300 ms in the post-test. Late differences are also detectable in the 550-600 ms window in the pre-test session. Statistical analyses of these differences are presented in the following sections.

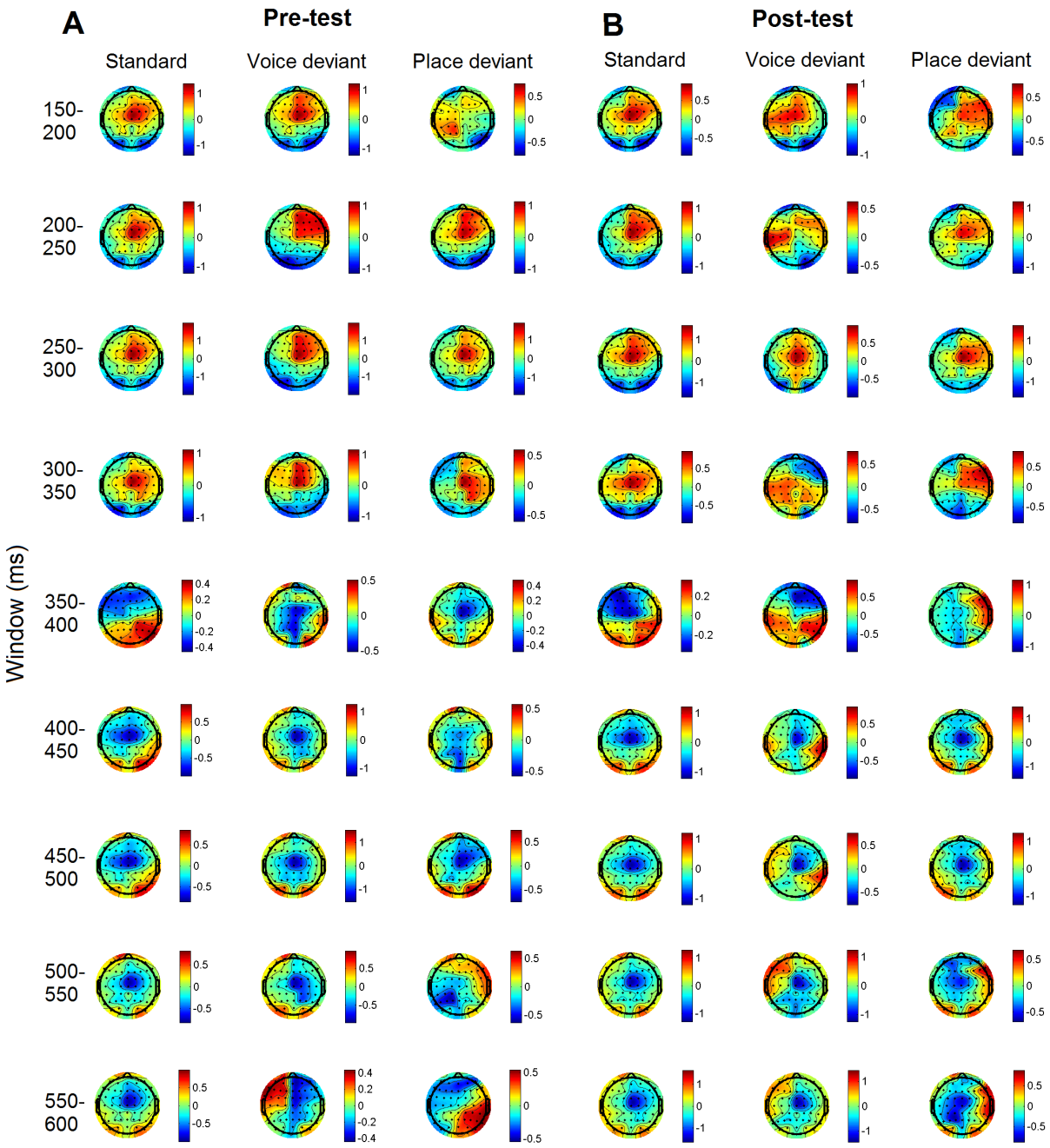


Figure 4.10: Scalp topographies for (a) pre-test and (b) post-test data. Each row represents a 50-ms window between 150 and 600 ms after stimulus onset. Activation on the scalp is shown from a top-down view, with redder regions indicating positivites and bluer regions showing negativities; the absolute scale differs for each topography. Differences between the magnitude, direction, or localization of activity between the standard (left-most column in each subplot) and the deviants (middle and right columns) suggest a processing difference between these trial types for a given window.

Main effect of trial type

The main effect of trial type describes overall differences in activity between the standard [dʲfia], voice deviant [dia], and place deviant [dfia]. Differences between trial types suggests that subjects were able to detect differences between the categories of sounds pre-attentively. The main effect of trial type for each channel/window combination is reported in table 4.2. There are some differences between trial types in a moderately early window (350-400 ms), corresponding to the differences identified in the ERPs (figure 4.9) and scalp topographies (figure 4.10). In addition, all channels show differences in a later window (450-500 ms), with some channels maintaining this difference as much as 100 ms later.

While differences between any trial types are notable, as all are predicted to assimilate to the English /d/ in untrained listeners, of particular interest is whether the deviant trials show activity that is separable from the standard trials. To assess the specific differences between each pairing of two trial types, post-hoc tests with multiple comparisons corrections were run using Matlab's `multcompare` function.

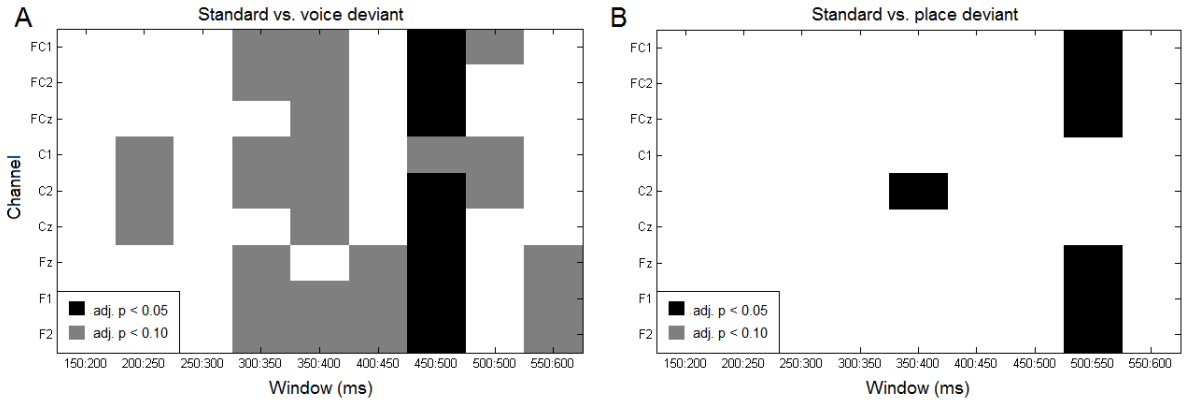


Figure 4.11: Post-hoc comparisons for the main effect of trial type, for each channel/window combination. Subplot (a) shows comparisons between the standard and voice deviant; subplot (b) shows comparisons between the standard and place deviant. Significant comparisons (after correction) are shown in black; marginal comparisons in grey. No significant comparisons were found between the place deviant and voice deviant for any time window or channel.

Results of the multiple comparison analysis are visualized in figure 4.11. Differences between the standard and voice deviant survive only in the 450-500 ms window, although many marginal comparisons in the preceding window suggest differences in a broader window from 350-500 ms. Fewer differences were found for the standard vs. place deviant comparison. With the exception of one early window (350-400 ms) in a single channel, all differences were found after stimulus offset, between 500-550 ms. There were no significant comparisons between the voice deviant and the place deviant for any channel/time window combination.

	150-200	**200-250**	**250-300**	**300-350**	**350-400**	**400-450**	**450-500**	**500-550**	**550-600**
FC1	1.26, 0.29	1.80, 0.17	1.71, 0.19	*2.45, 0.09*	*2.80, 0.07*	1.89, 0.16	**3.97, 002**	**4.11, 0.02**	*2.50, 0.09*
FC2	1.29, 0.28	1.82, 0.17	1.52, 0.23	*2.54, 0.08*	**3.25, 0.04**	2.14, 0.12	**3.12, 0.05**	**3.19, 0.05**	2.12, 0.13
FCz	0.89, 0.38	2.05, 0.14	1.40, 0.25	2.04, 0.14	2.37, 0.10	2.12, 0.13	**3.45, 0.04**	**3.20, 0.05**	2.20, 0.12
C1	1.22, 0.29	2.22, 0.11	1.59, 0.21	2.31, 0.11	*2.48, 0.09*	1.31, 0.28	**4.43, 0.01**	*2.95, 0.06*	2.09 0.13
C2	1.42, 0.25	2.34, 0.10	1.77, 0.18	*2.63, 0.08*	**3.63, 0.03**	1.31, 0.28	**3.47, 0.04**	*2.67, 0.08*	2.09, 0.13
Cz	1.54, 0.22	2.28, 0.11	1.24, 0.30	2.03, 0.14	*2.65, 0.08*	1.81, 0.17	**4.07, 0.02**	2.25, 0.11	1.62, 0.20
Fz	1.16, 0.32	1.58, 0.21	1.60, 0.21	2.40, 0.10	*2.73, 0.07*	2.20, 0.12	**3.66, 0.03**	**3.79, 0.03**	**3.08, 0.05**
F1	0.89, 0.41	1.57, 0.21	1.67, 0.19	2.55, 0.08	**3.06, 0.05**	*2.43, 0.09*	**3.78, 0.03**	**4.46, 0.01**	**3.31, 0.04**
F2	1.12, 0.33	1.45, 0.26	1.41, 0.25	2.46, 0.09	**3.18, 0.05**	*2.54, 0.08*	**3.32, 0.04**	**3.59, 0.03**	*2.83, 0.06*

Table 4.2: F-statistics (first number) and p-values (second number) for the main effect of trial type (standard, voice deviant, or place deviant), by channel (rows) and time window (columns). Bolded values indicate significant ($p <= 0.05$) window/channel pairings; italicized values indicate marginal ($p < 0.10$) pairings.

It is perhaps not surprising that there were more detectable differences between the voice deviant and standard than the place deviant and standard. The voice deviant is distinguished by its lack of breathy voicing, which is a robust acoustic difference. The place contrast, by comparison, is cued in formant transition from consonant to vowel, as well as spectral properties of the stop burst, and is somewhat more subtle. This effect is consistent with findings in experiments 1 and 2 that voicing contrasts were more apparent to listeners than place of articulation contrasts.

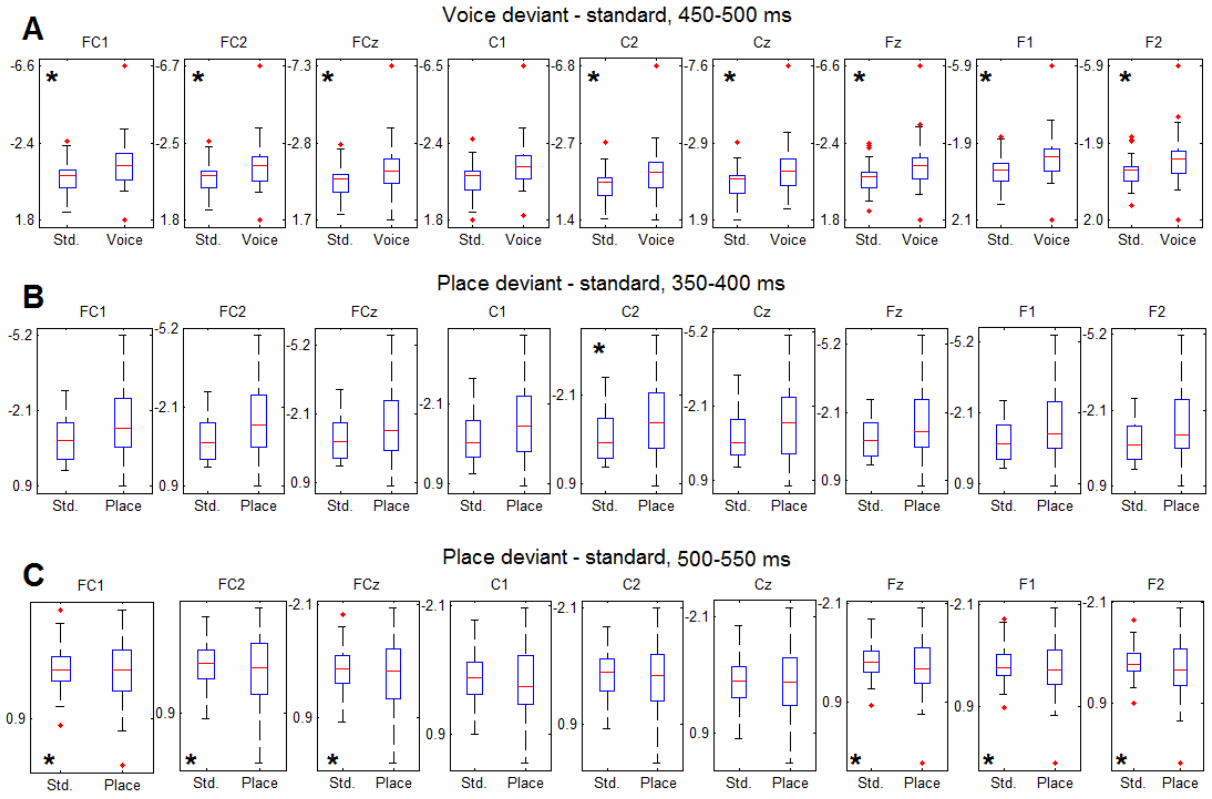


Figure 4.12: Boxplots of standard and deviants which showed significant differences in responses during particular windows, for each channel. Plot (a) compares the standard and voice deviant; plots (b) and (c) compare the standard and place deviant. Significant comparisons after multiple comparisons correction are indicated with (*).

Distributions of activity for each window which showed a significant trial type contrast after correction are shown in figure 4.12. The window that survived correction for the voice-place contrast, 450-500 ms (subplot A), shows greater negativity for the voice deviant as compared to the standard, which is consistent with an MMN response. The same is true for the place deviant at its earlier window, 350-400 ms (subplot B). For the later place-standard window, 550-600 ms, the reverse is true, with greater negativity for the standard. This reversal of positive and negative components is visible in the ERPs and particularly the difference plots (figure 4.9) for these channels.

Main effect of session

The effect of session tests whether there are overall differecnes in the magnitude of activity from pre-test to post-test. Ideally, differences from pre-test to post-test suggest an effect of training on recorded activity. A summary of the main effect of session for each combination of channel and time window is presented in figure 4.3.

As can be seen from the table, only one window (350-400 ms) showed significant differences in activity from pre-test to post-test. As shown in figure 4.13, the difference in all channels reflects an attenuation of the response from pre-test to post-test. This attenuation may reflect a habituation to the stimuli, which subjects have heard hundreds of repetitions of by the start of the post-test session.

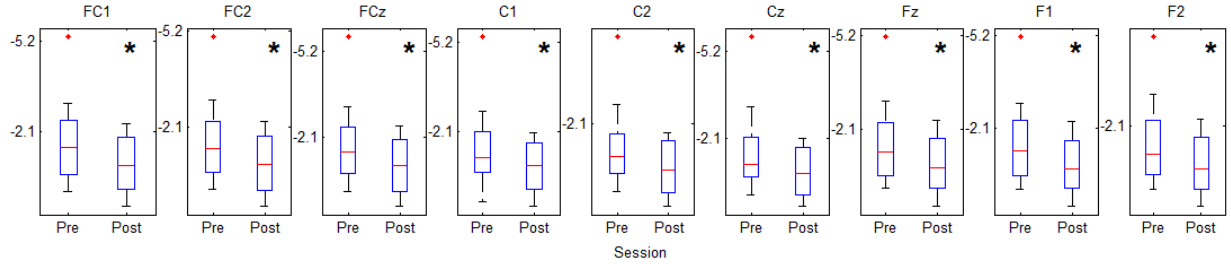


Figure 4.13: Boxplots for the window (350-400 ms) which showed a significant difference in pre-test vs. post-test activity. All 9 channels showed a significant difference (*) during this window.

Interaction of session and trial type

The primary question of this experiment concerns the interaction of trial type and session. If training has an impact on pre-attentive responses to the target stimuli, then the relationship between the three trial types (standard, voice deviant, and place deviant) should differ from before training (pre-test) to after training (post-test). In particular, it was predicted that differences between the standard and the two deviants would be more apparent after training, and not present (or minimally present) before training.

Inspection of the results of each ANOVA, however, indicated that there was no channel/time window combination for which the session * trial type interaction was significant, or even marginal ($p < 0.10$). This indicates that the relationship between the standard and the two deviants was not fundamentally affected by training. Differences reported above for trial type appear to have been present at both pre-test and post-test. In other words, any detection of the difference between the standard and the two deviants was present prior to training, and not affected by training. Similarly, the effect of session was independent of trial type - the attenuation of responses from pre-test to post-test was not affected by whether responses came after a standard, voice deviant, or place deviant.

	150-200	**200-250**	**250-300**	**300-350**	**350-400**	**400-450**	**450-500**	**500-550**	**550-600**
FC1	0.06, 0.81	0.05, 0.82	0.04, 0.85	0.86, 0.36	**6.69, 0.01**	0.93, 0.34	0.63, 0.43	0.01, 0.91	0.01, 0.94
FC2	0.02, 0.88	0.21, 0.65	0.00, 1.00	1.34, 0.25	**7.99, 0.01**	1.10, 0.30	0.48, 0.49	0.35, 0.70	0.18, 0.67
FCz	0.05, 0.83	0.25, 0.62	0.05, 0.83	0.94, 0.34	**6.96, 0.01**	0.91, 0.34	0.59, 0.44	0.05, 0.83	0.02, 0.88
C1	0.29, 0.59	0.16, 0.69	0.01, 0.93	2.20, 0.14	**6.06, 0.02**	0.07, 0.80	0.93, 0.34	0.01, 0.90	0.06, 0.81
C2	0.04, 0.85	0.32, 0.58	0.06, 0.81	*2.87, 0.09*	**7.74, 0.01**	0.16, 0.69	0.80, 0.37	0.03, 0.87	0.34, 0.56
Cz	0.14, 0.71	0.10, 0.76	0.01, 0.93	2.50, 0.12	**7.83, 0.01**	0.03, 0.86	0.89, 0.35	0.00, 0.98	0.12, 0.73
Fz	0.02, 0.89	0.09, 0.77	0.10, 0.75	0.27, 0.61	**6.04, 0.02**	1.96, 0.16	0.11, 0.74	0.10, 0.75	0.00, 0.96
F1	0.15, 0.70	0.34, 0.71	0.06, 0.80	0.36, 0.55	**6.99, 0.01**	*2.93, 0.09*	0.14, 0.71	0.18, 0.67	0.01, 0.91
F2	0.02, 0.88	0.09, 0.76	0.18, 0.67	0.31, 0.58	**6.52, 0.01**	1.82, 0.18	0.07, 0.79	0.14, 0.71	0.02, 0.90

Table 4.3: F-statistics (first number) and p-values (second number) for the main effect of session (pre-test, post-test), by channel (rows) and time window (columns). Bolded values indicate significant ($p <= 0.05$) window/channel pairings; italicized values indicate marginal ($p < 0.10$) pairings.

4.5 Discussion

Experiment 3 had two primary goals: to test whether pre-attentive responses to non-native consonants would show faster discrimination for new learners than behavioral metrics, and to measure the efficacy of production training in an auditory oddball training paradigm. In light of these, the results of the experiment are mixed. There was some sensitivity to a contrast (dental vs. retroflex place of articulation) that was very difficult to detect via behavioral responses for subjects in experiment 1 and 2. However, there was no significant change in detection of phonetic contrasts from pre-test to post-test, indicating that there was no effect of training. The implications for each goal is discussed in detail below.

4.5.1 Pre-attentive responses to novel stimuli

Results of the analysis indicate that subjects were sensitive to the difference between the standard [d̪ʰa] and the voice deviant [d̪a], as well as the standard and the place deviant [d̻ʰa]. This sensitivity was present at both pre-test and post-test, as evidenced by the lack of a significant trial type * session interaction for any time window or channel included in the analysis.

The consonants [d̪ʰ], [d̪], and [d̻ʰ] are all predicted to assimilate to the English /d/ for native English speakers, so detection of either contrast at pre-test indicates that pre-attentively, these consonants are not fully assimilated. The voice deviant is more acoustically distinct from the standard, so it is perhaps not surprising that it was more robustly distinct from the standard than the place deviant (if marginally significant time windows are taken into account). However, the place deviant was also significantly detectable, as indexed by the mismatch negativity response.

Pre-test discrimination of pairs which contrasted in place of articulation was relatively low in experiments 1 and 2. This raises a question: why was the place deviant separable from the standard in the present study? One possibility is that the contrast is dissociable at a pre-attentive auditory level, but that this is overwhelmed by English phonemic biases when an overt decision must be made. The MMN response is often taken as evidence for phonetic-level processing, as it typically shows sensitivity to the phonemic contrasts present native language of the speaker (Näätänen et al., 1997; Näätänen, 2001; Sharma & Dorman, 2000; Peltola et al., 2003; Winkler et al., 1999). However, there are two primary sources of the MMN responses - a more automatic auditory change-detector in the temporal lobe, and an attentional switch in frontal regions (Alho, 1995). Therefore, it is possible that the responses documented in the current study reflect a more acoustic or auditory level of processing, rather than higher-order phonemic categorization. Since overt behavioral decisions required of subjects in experiments 1 and 2 are undoubtedly influenced by native-language phonemic processing, then these two levels of processing may help account for the differences in pre-test performance between those studies and the current results.

Another consideration in the difference between behavioral decisions and automatic neural responses is the time lag involved in the former. A behavioral response requires a short

period of time to make a decision. The time lag inherent in that decision can be understood as a window which increases uncertainty about the percepts that were heard during the trial. Assuming time acts like noise to increase uncertainty about the percept in this way, Bayesian perceptual models predict a regression of percepts towards the prior category center - in this case, the native language category center (Feldman et al., 2009). This is consistent with the predictions that fall out of the perceptual magnet effect (Kuhl et al., 2008). Given all this, a task which involves a longer time delay will necessarily increase reliance on established categories. Therefore, aside from the requirement of an overt decision, the behavioral task is more susceptible to native language category effects than an fast, automatic neural response.

It should be noted that past studies which tested similar contrasts in speakers of languages which do not have it often failed to find evidence of a mismatch effect (Dehaene-Lambertz, 1997, for French speakers listening to the dental-retroflex contrast; Sharma & Dorman, 2000 for English speakers listening to the voiceless-voiced unaspirated contrast). Therefore, it is an open question why the effect was found for the current set of subjects.

One possibility is that the use of a non-native category as a standard caused listeners to operate in a more acoustic level of processing, as opposed to a more phonemic or categorical mode. Zevin et al. (2010) tested Japanese speakers with the English /ɹ/-/l/ distinction and found little difference between their MMN responses at a particular critical electrode and those of native English speakers (although broader topological differences did emerge), despite differences in behavioral performance. This may support a hypothesis that status of the standard affects the mode of processing, as neither token is in the Japanese phoneme inventory, and the same was true for the stimuli in the current study.

4.5.2 Production training and automatic detection

There was no effect of training on the discrimination of the three trial types; the differences between the standard and the two deviants did not differ from pre-test to post-test. This means that production training was not effective at changing subjects' responses to the target categories. What is the reason that an effect was not found in the present study?

One possibility is that there was insufficient information to teach subjects about the target contrasts in the shortened version of the training experiment used in the EEG experiment. The EEG training was modeled after the production training used in experiments 1 and 2 and identical in many ways. However, because there were fewer target contrasts in the EEG experiment, the training was necessarily shorter (subjects did not receive detailed training for the aspirated and unaspirated voiceless categories, for example). The repetition task included in the production training was also shorter as a result. All the critical targets received the same number of repetitions as in the original studies, but the smaller number of overall targets meant that there was less practice of the phonetic space in general, and less practice of the dental-retroflex contrast in particular (which would've gotten extra practice in a paradigm with more VOT contrasts, and therefore more tokens in general). While the results of the long production training group in experiment 2 suggests that overtraining may be counterproductive, it is possible that subjects in the EEG experiment were undertrained.

A second possibility is that there was not space for detectable improvement, because subjects already showed some sensitivity to the target contrasts during the pre-test. This seems unlikely for three reasons. First, it is well-known that native English speakers struggle with these Hindi contrasts (e.g. Werker & Tees, 1984; Pruitt et al., 2006; Golestani & Zatorre, 2009), and they are not always recognized even pre-attentively (Sharma & Dorman, 2000). Second, there was room for improvement in all behavioral tests of the target contrasts, so while it is possible that pre-attentive performance could be at ceiling for acoustic perception, it seems likely that there is room for improvement in a categorical listening mode. (Native Hindi speakers would be a useful control to clarify this distinction.) Third, the time windows which showed significant contrasts between the standard and each deviant were limited, and relatively late in time relative to the acoustic onset of the change. At the very least, it seems possible that the magnitude of the effect could have increased, or the timing of the mismatch could have come earlier, given effective training.

One possible limitation to detecting the efficacy of training could be the habituation of the automatic response to repeated presentation of the stimuli. The main effect of session in the analysis showed that the overall amplitude of the ERPs was reduced in the post-test session, suggesting that subjects were habituating to the overall acoustic environment. It is known that habituation to repeated stimuli can cause continuous, gradual dampening of ERPs, even well after attention is already diminished (Ritter, Vaughan, & Costa, 1968; Woods & Elmasian, 1986). This can affect the magnitude of the MMN response as well; McGee et al. (2001) note that the higher signal-to-noise ratio afforded by many repeated trials in typical MMN studies was undermined to some extent by the attenuation of the MMN response in these long experiments. Therefore, an effect of training would have to be sufficiently robust to counteract this dampening effect, and may be hard to detect otherwise.

A third possibility, which is perhaps most theoretically interesting, is that production training is an insufficient training mechanism for *pre-attentive* responses, even if it is beneficial for behavioral learning. To pursue an admittedly speculative line of thinking, perhaps it is the case that cross-modal information is most useful when leveraged in a situation which requires an explicit attentional strategy. In such a case, the novice learner may be able to recruit information from multiple domains in order to arrive at a decision that gets closest to the target category. The pre-attentive MMN response may be unable to leverage this information. Perhaps the speed of the response precludes the rapid integration of motor and acoustic information in a way that would be useful for pre-attentive categorization, at least in the early learner who has not formed tight links between these two levels of representation yet. Alternately, perhaps learners in the behavioral task are leveraging the articulatory knowledge they've learned in a very deliberate way, and explicitly trying to recall that information when making a judgment about the categorization of two stimuli. This strategy is not available to learners in the EEG task, and therefore would not benefit them in the same way.

This line of thought opens up interesting possibilities for future work. In particular, it would be worthwhile to follow up on the implementation of production training using a testing paradigm which is behavioral, but which does not employ explicit judgments about

phonetic categorization. The implicit testing paradigms used in some studies, such as the video game paradigm used in Lim and Holt (2011), may be able tease these ideas apart. In such a paradigm, subjects would have the decision time to integrate motor and acoustic representations (assuming that this is in fact a slow process), but would not be making explicit phonetic judgments, and so presumably would not be using an overtly deliberate recall strategy to do so.

4.6 Conclusion

This chapter explored pre-attentive neural responses to non-native phoneme categories, and the effect of cross-modal training on those responses. Category discrimination, as measured by the mismatch negativity response, was present for both the place of articulation contrast and the voicing contrast, and did not detectably change after training. This work suggests that pre-attentive neural responses may differ from behavioral responses in early learners of novel categories, and that those differences may have implications for models which account for the integration of acoustic and articulatory information in categorical representations of speech sounds.

Chapter 5

Conclusion

This dissertation contains three experiments, presented in chapters 2, 3, and 4, which were designed to test the efficacy of cross-modal learning in adults acquiring novel phoneme contrasts. In all cases, native English speakers were trained to discriminate (and in some cases, articulate) coronal stop contrasts in Hindi which are known to be challenging for adult English speakers. While there was generally behavioral improvement in the discrimination of these targets, the evidence for cross-modal learning is mixed.

In this chapter, I attempt to present a unified account of the findings from all three studies. First, a summary of the results from each experiment is presented. Following that, I discuss the implications of these findings for three major theories of novel phoneme acquisition (the Speech Learning Model, the Native Language Magnet Theory, and the Perceptual Assimilation Model). This is followed by an account of cross-modal learning in novice learners, which aims to unify recent accounts about the interaction between motor and acoustic representations in speech perception. Finally, methodological considerations and avenues for future research are explored.

5.1 Summary of findings

5.1.1 Experiment 1

Experiment 1A (chapter 2) implemented a multi-day training experiment to test the effects of several training techniques on the discrimination and pronunciation of Hindi coronal stops by adult native English speakers. Subjects were tested before and after four sessions of perceptual training and one session of production training. Perception training included performance feedback and adaptive training (starting with acoustically easy targets and getting progressively harder). Production training consisted of instruction on the correct articulation of targets and a repetition task with visual cues.

Perceptual training was effective at improving discrimination accuracy, replicating effects in the literature that adaptive training (Jamieson & Morosan, 1986; Pruitt, 1995; Escudero

et al., 2011) and performance feedback (Goudbeek et al., 2008; McCandliss et al., 2002; Pederson & Guion-Anderson, 2010) can assist in the acquisition of novel phoneme categories. However, the benefit of production training was not detectable, even though some previous studies (Catford & Pisoni, 1970; Hirata, 2004; Herd et al., 2013) have shown beneficial effects of production training on perception of non-native contrasts. A control experiment (experiment 1B) showed that the effect of repetition alone did not change subjects' baseline performance, suggesting that the perceptual training interventions were critical to improved discrimination. A follow-up experiment (experiment 1C) found that the benefits of training in the main experiment were retained a month after training had ended.

Changes in production accuracy were mixed. Analyses of place of articulation (burst spectra and formant transitions) suggest minor but not definitive improvements in the dental-retroflex contrast as a function of perceptual training, but stronger accuracy during production training. However, this improvement was not retained into post-training testing (the re-test session). Voice onset time analyses suggest that the main change in non-native targets was the increase in pre-voicing for breathy consonants during production training, but this was also not maintained after training ended.

Taken together, the results of experiment 1 suggest that subjects were able to improve some aspects of their perceptual and articulatory targets for the Hindi stop categories. However, these improvements were primarily driven by within-mode training (perceptual training for discrimination, articulatory training for production). There was not strong evidence for cross-modal learning in either domain.

5.1.2 Experiment 2

Experiment 2 (chapter 3) was designed to probe the lack of cross-modal learning in experiment 1. It focused specifically on perceptual learning, and used multiple training groups with different paradigms to test the effect of a single session of training on discrimination. One group (production training) received the production training procedure from experiment 1A. A second (long production training) received that same training, but with a lengthened repetition task. A third (guided production training) completed the production training task with the experimenter present to answer questions. A fourth (perception training) received the first day of perception training (easiest stimuli, with performance feedback) from experiment 1A. All groups were tested before and after training.

Overall, there was improvement in discrimination performance from pre-test to post-test. Most groups showed improvement in the accuracy data, with the exception of the long production training group, whose inaccurate productions may have reinforced category targets that did not align with the actual Hindi categories. These results suggest that production training is not necessarily less effective than perception training. Given this, the failure of cross-modal training to improve discrimination in experiment 1A may have been because subjects had already improved substantially as a function of perceptual training, and had hit a "functional ceiling" for performance within the confines of the experiment. However, the result from the long production training group suggests that the benefits of

articulatory learning may be mediated by the accuracy of subjects' fledgling articulatory representations of category targets.

5.1.3 Experiment 3

Experiment 3 (chapter 4) was designed to further test the limits of cross-modal training. If it can be beneficial for novice learners in behavior, can those benefits also be detected pre-attentively? EEG data was recorded from subjects before and after production training with a limited set of stimuli (one breathy-voiced standard, one place of articulation deviant, and one voicing deviant). An oddball paradigm was used, in order to see if mismatch negativity (MMN) responses would indicate pre-attentive detection of the standard-deviant contrasts after training.

The results did not suggest an effect of training. MMN responses were present in both pre-test and post-test sessions, with stronger MMNs for the voice deviant than the place deviant. This suggests that the place of articulation contrast may be detectable pre-attentively, even though performance in behavior on this contrast is typically low. However, it also suggests that the benefits of cross-modal training may not be detectable in pre-attentively, and may be restricted to overt behavioral metrics.

5.2 Perceptual bias patterns and theoretical accounts revisited

The summary in section 5.1 describes the overall pattern of results across stimulus types, but there was a substantial amount of variation in how well subjects performed on the discrimination of particular contrasts. One consistency across all three studies is that the place of articulation contrast (dental vs. retroflex) was most challenging. Within the voicing contrasts, there was a range of performance. Subjects performed best on the voiceless-aspirated contrast, had more trouble with the voiced-breathy and breathy-aspirated contrast, and struggled most with the voiced-unaspirated contrast. Performance was better when two cues (a place feature and a voicing feature) distinguished the stimuli.

A straightforward way to understand the overall bias patterns in subjects' behavior in studies 1 and 2 is to consider which non-native sounds would assimilate to each of the two coronal stop categories in English, /d/ and /t/. Since the English contrast does not have a place of articulation distinction, both the dental and retroflex would fall into the English alveolar category. Because there is not a category distinction along this axis within the English coronal stops, it is perhaps unsurprising that the place contrast is quite difficult for English speakers. This leaves the voicing contrast, which has more interesting bias patterns. For onset-initial stops, the Hindi /t/ and /t^h/ (both dental and retroflex) map straightforwardly onto English /d/ ([t]) and /t/ ([t^h]), respectively. The Hindi /d/ voicing category is an allophone of the English /d/ in intervocalic context, and thus it maps onto

English /d/ as well. Because Hindi /t/ and /d/ both map onto English /d/, it is easy to understand why they are regularly confused in the behavioral results.

The breathy category /d^h/ has features of both English /t/ (post-release aspiration) and the allophone of the English /d/ (voicing during the closure); in principle, it could map onto either. In fact, we do see evidence that the breathy category is confused with all the other voicing types (chapter 2, figure 2.7), although it is most often confused with /t^h/.

The bias patterns observed in this study have relevance to the three major theoretical accounts of novel phoneme perception. The findings from each experiment are discussed with respect to each theory in the following sections.

5.2.1 The Speech Learning Model

The basis of predicted assimilations in the Speech Learning Model (SLM, Flege, 1995; Guion et al., 2000; Baker et al., 2002) is phonetic distance between targets. Guion et al. (2000) used goodness-of-fit judgments by native speakers to quantify phonetic distance of Japanese phonemic categories, and found that these were good predictors for the difficulty of some, but not all non-native contrasts. The studies reported in this dissertation did not assess phonetic distance in this way. Perceptual confusion data from German, Japanese, and Hindi native speakers reveal that retroflex segments are the most likely segments to be confused with their non-retroflex counterparts (Dev, 2009), which is consistent with the difficulty that learners had with the place of articulation contrast in the current experiment. The confusion data presented in experiment 1 (chapter 2, figure 2.7) for the voicing contrasts would need to be corroborated with data from native Hindi speakers in a noisy environment (as in Miller & Nicely, 1955) to see if these patterns also fall out from phonetic distance predictions.

Although SLM is primarily focused on second-language learners and not naïve listeners, one part of the model that is consistent with the framework outlined in this dissertation is the contention that discrimination must precede robust category formation. The studies described in chapters 2, 3, and 4 were all conducted with individuals who had no exposure to Hindi prior to the start of the experiment. Listeners in baseline conditions failed to behaviorally discriminate many of the target contrasts. Improvements after training in experiments 1 and 2 indicate the beginning of category learning, but robust categorical perception has not yet been achieved. These studies are situated at the "precursor to categories" stage. It would be informative to follow participants further to track the development of robust category representations after more experience with the language.

5.2.2 The Native Language Magnet Theory

The Native Language Magnet Theory (NLM, Kuhl, 2000; Iverson et al., 2003; Kuhl et al., 2008) emphasizes the acoustic salience of contrasts as critical to acquisition of phonetic contrasts in the first language. Once established, these categories pull incoming percepts towards them, an account that is consistent with the assimilation patterns of the current experiment. The focus on acoustics as key to the establishment of these categories may help

explain the bias patterns across the different contrast types in the present experiments. The acoustic distinctions between the voicing contrasts (presence or absence of prevoicing and aspiration/breathy voicing) are longer and more prominent than the cues that signal the place of articulation contrast (burst spectral properties, formant transitions). Consistent with this, the voicing contrasts were more reliably discriminated than the place contrast in the behavioral experiments (chapters 2 and 3), and the voice deviant was more robustly distinct from the standard than the place deviant in the EEG experiment (chapter 4).

One consideration for the assessment of NLM is the relative effectiveness of perception training and production training. Subjects who had already received perceptual training (experiment 1, chapter 2) did not receive an additional benefit from articulatory instruction. At minimum, this suggests that prior acoustic knowledge of a contrast may minimize the additive effectiveness of articulatory information for building new representations. A stronger interpretation would be that acoustic information is more effective than articulatory information for adult novel phoneme acquisition; this, however, was not supported by the findings of experiment 2 (chapter 3), where there was not a detectable difference between perception training and most types of production training.

A basic tenet of NLM is that infant-directed speech amplifies the acoustic cues critical to native phoneme contrasts, and that this acoustic signposting is what helps infants develop native categories. This raises the question of whether exaggerated acoustic cues would be beneficial to adult learners as well. The adaptive training paradigm used in experiment 1 (chapter 2) could be interpreted as being consistent with this principle. There is some evidence that exaggerated cues may be helpful for lexical learning in adults (Golinkoff & Alioto, 1995, but see Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011). This idea is also consistent with Hebbian learning accounts (McCandliss et al., 2002; Vallabha & McClelland, 2007), which argue that some overt cue must enhance the target contrast to the extent that it is recognized by the listener, or else the bias towards native language categories will cause assimilation.

5.2.3 The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM, Best et al., 2001; Best & McRoberts, 2003; Best et al., 2009), in concert with the Articulatory Organ Hypothesis (AOH) predicts assimilation to native categories on the basis of shared articulatory organs for two targets in a contrast. If a non-native target matches an existing native category with the same primary articulator, it is more likely to assimilate to that category than a non-native target which matches on another feature (e.g. manner of aritculation). This neatly explains difficulty with the place of articulation contrast in the current experiment - both dental and retroflex stops in Hindi are apical coronal stops.

The predictions for voicing contrasts are somewhat more difficult to tease out in the AOH framework. Since the primary articulator for all of the target segments is the tongue tip, one might expect all of the voicing contrasts to be difficult, as they are all within-organ contrasts. Clearly, there are differences in the difficulty of these contrasts, and so

some component of laryngeal control must be considered. Best and Hallé (2010) indicate that the timing of laryngeal gestures are relevant to assimilation patterns, and that these can be predicted with reference to articulatory phonology. They found differences between English and French speakers' perception of non-native onset contrasts which varied in their integration of voicing and lateral features (Hebrew /tl/-/dl/, Zulu /ɬ/-/ɮ/, and Tlingit /tɬ/-/dɮ/), which they attribute to the relative phasing of coordinated laryngeal gestures in these complex onsets. While a similar explanation has not yet been fleshed out for segments within a single paradigm (such as the Hindi voicing paradigm), this explanation plausibly accounts for the present data as well.

However, there is one aspect of the present research that is more difficult to align with PAM. If assimilation of non-native sounds is grounded in articulatory gestures, it is reasonable to expect that instruction in the target gestures aid in repairing the bias towards native language categories. In other words, production training should be an effective strategy - more so than perceptual training, perhaps - if the assumptions of PAM are correct. Given that the efficacy of production training was only found in certain contexts in the present work, this casts some doubt on the role of articulatory gestures as the primary mediators of native language biases.

5.3 The perception-production link and the novice learner

A major goal of this dissertation was to evaluate the relationship between perceptual and articulatory representations during the early stages of category learning. This was investigated through the use of cross-modal learning: perceptual training for production tasks, and articulatory training for discrimination tasks. The data was somewhat mixed with respect to whether this type of learning has a facilitatory effect. Behavioral discrimination of non-native contrasts seemed to benefit from production training subjects were exposed to the first type of learning (experiment 2, chapter 3), but not if prior perceptual training had occurred (experiment 1, chapter 2). Pre-attentive discrimination did not reveal an effect of articulatory training (experiment 3, chapter 4), unlike (some of) the overt behavioral data. The effects of perceptual training on accurate pronunciation of non-native targets (experiment 1, chapter 2) were moderate at best.

Catford and Pisoni's (1970) classic experiment found that articulatory training was *more* robust than acoustic training for non-native contrasts, which seems to contradict the present findings that cross-modal learning is somewhat fragile. However, other studies have found variability in the contrasts which respond well to cross-modal vs. within-mode training (Herd et al., 2013), and not every experiment shows an obvious benefit (Schneiderman et al., 1988; Baese-Berk, 2010). Clearly there is still work to be done on the conditions and procedures under which cross-modal learning is most advantageous.

The interaction between perception and production has been a hot topic in speech percep-

tion for decades, particularly due to debate between motor theory/direct realist accounts of phonetic representations (Liberman et al., 1954, 1957; Liberman & Mattingly, 1985; Galantucci et al., 2006; Fowler, 2008) and accounts which posit an acoustic basis for these representations (Massaro & Chen, 2008; Diehl et al., 2004). These debates have intensified in recent years due to research into mirror neurons (Kohler et al., 2002; Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Watkins & Paus, 2004) and embodied cognition (M. Wilson, 2002; Fischer & Zwaan, 2008) more generally, which posit a neural mechanism by which perception and conceptual understanding are mediated through simulation of the corresponding actions in cortical regions adjacent to the motor cortex. (For responses to mirror neuron accounts of speech, see Lotto, Hickok, & Holt, 2009; Hickok, 2009.)

While much of the debate centers around the nature of established phonetic categories, it is equally important to consider how perceptual and articulatory systems interact in newly-forming phonetic representations. Some accounts have considered a mutually-facilitatory role for the two representational systems. Newer formulations of the Native Language Magnet theory (Kuhl et al., 2008) include a perception-production link as an explanation for how young infants' first attempts to produce speech can help reinforce the perceptual categories they are trying to learn, and vice-versa. Very recent work from this group (Kuhl et al., 2014) has relied on the Analysis by Synthesis account of speech perception (Stevens & Halle, 1967), arguing that motor representations can constrain incoming acoustic information to aid predictions about the phonetic segments in an acoustic signal.

Even researchers who are skeptical of strong versions of motor theory allow for links between perceptual and motor representations (Hickok & Poeppel, 2007), and acknowledge that these connections can be particularly useful in cases of adverse listening conditions. Hickok, Houde, and Rong (2011) posit that motor representations may serve as a mechanism for attentional modulation, and support speech perception in adverse listening conditions (e.g. noise, or certain types of speech disorders) where acoustic information alone may not suffice for accurate decoding of the signal.

Could such a principle apply during language acquisition as well? Supporting this notion, Kuhl et al. (2014) found that motor representations were stronger than auditory representations for non-native sounds in older infants and adults, while auditory representations were stronger than motor representations for native sounds. Seven-month-old infants relied on each system equally. This suggests that when categories are new or still forming, motor systems may play a substantial part in supporting representations, but that these may diminish in importance relative to perceptual/acoustic representations as categories strengthen.

This account is consistent with the behavioral findings reported in experiment 1 and experiment 2. In experiment 2, completely novice learners benefited from production training when they had received no other form of learning. In experiment 1, when subjects had received prior perceptual training, articulatory information failed to confer an additional benefit. Could it be that the shift away from motor support can occur in the short time frame (5-6 days) that it appeared to in experiment 1? More data would be needed to confirm that this is in fact the mechanism at work. However, in the broadest terms, it seems reasonable to propose that cross-modal information is most beneficial for the most novice

learners.

5.3.1 All information is good information?

If cross-modal information benefits novice learners the most, a related line of speculation is that for the listener who has no knowledge of prior categories, all sources of information about the existence of a category are good sources of information. In other words, learners will use whatever evidence they can to reinforce a fledgling category. This can be seen as analogous to the argument of Hickok et al. (2011) concerning sensorimotor integration during noisy conditions, and can even be seen in Bayesian terms as a means to overcoming uncertainty about a signal (Feldman et al., 2009).

This formula is clearly simplistic, and is already subject to a caveat from the data presented in this dissertation. Recall that subjects in experiment 2 who received extra articulation practice (the long production training group) performed worse than others; in this case, their own motor representations did not help to reinforce the target categories. It is likely that there is an element of accuracy that is needed to mediate the "all information" principle - these representations must approximate the target categories in order to help reinforce them. If subjects in the long production training group were producing consonants that reflected English categories rather than the target Hindi ones, they were reinforcing and strengthening L1 categories, which is counterproductive to the acquisition of the L2 categories.

It should be noted that young infants' productions, such as those described in Kuhl et al. (2014), are unlikely to be very accurate by the standards of adult targets. However, in the case of first-language acquisition, there are no pre-existing L1 categories, and so there is no bias towards interfering categories from another system, as there is in the case of adult second-language acquisition. Therefore, inaccurate infant productions are unlikely to be *systematically* biased in a way that interferes with the accurate acquisition of target categories. Infants also have more time than the adults in experiment 2 to resolve the conflict between their articulatory representations and the target acoustics.

A second caveat is that no cross-modal benefit was found in experiment 3, in the preattentive detection of novel categories. In that experiment, subjects had no overt behavioral task associated with perception. One possibility is that cross-modal learning may not be helpful for a novice learner in a condition where the whole system is not engaged. That is, in a case where only an automatic response is required, the cross-modal representation may not be engaged to the degree that it is when a subject is overtly recruiting all information available in order to make a decision. Perhaps some sort of active engagement is required for subjects to make an initial connection between the two levels of representation.

Taken together, these patterns suggest the following trajectory for cross-modal integration in novel category acquisition. For completely novice learners, acoustic and motor representations are mutually beneficial in supporting the development of fledgling categories, assuming that learners are acquiring with and engaging both systems. If an existing category system is in place (i.e. for adult learners with a first language), accuracy is especially im-

portant; in the case of individuals without prior category biases (i.e. infants acquiring their first language), there is more leeway for support from imperfect motor representations. As learners begin to form and stabilize the novel categories, within-mode information (acoustics for perception, articulation for production) becomes more heavily weighted, and provides the main support for these representations, with less influence from cross-modal situations except in cases with adverse conditions (e.g. noise, speech disorders).

5.4 Future directions

The work presented here opens up new questions about the most effective approaches to non-native phoneme training. Below, I discuss several directions that future research could take to follow up on topics explored in this dissertation.

5.4.1 Cross-paradigm learning

A topic that was raised by the production data in experiment 1, but was not fully explored, is the issue of featural learning across a paradigm. The production data in this study suggests that certain non-native articulatory gestures were acquired more accurately in some target categories than others. In particular, the pre-voicing component of breathy stops were more accurately produced during production training, as evidenced by an increase in negative VOT for these targets during that session. However, a comparable increase was not found for voiced stops, which also have this feature.

Why was this articulatory gesture partially acquired for one category of stops, but not another? It is beyond the scope of the current work to fully explore the articulatory constraints that make this a challenging component of articulatory acquisition. However, perhaps the complexity of the required laryngeal gestures might make it easier for subjects to produce pre-voicing in the presence of breathy voicing, as opposed to in segments which have short-lag positive VOT. Future work which more fully explores generalization of articulatory features from one category to another within a phonological paradigm would be valuable. One such study on generalization might compare (a) learners who get instruction on two categories which share a novel articulatory gesture (e.g. [d̰] and [d̰ʱ], as in the current study), (b) learners who are trained on one category in a specific context and tested on generalization to a novel context within that learned category (e.g. learning [d̰ʱa] and testing on [d̰ʱi] and [d̰ʱu]), and (c) learners who are trained on one category and tested on generalization to another category which shares that feature (e.g. learning [d̰ʱ] and testing on [d̰]). These conditions would tease apart the extent to which generalization occurs at all in early articulatory acquisition, and the extent to which feature generalization can occur *across* categories within a phonological paradigm which is new to the learner.

5.4.2 Neural correlates of fledgling categories

As discussed in section 5.3, the interaction of perceptual and articulatory learning in phoneme acquisition is part of a broader line of research about the ultimate nature of phonetic representations. Many lines of research will be needed to illuminate this issue; one fruitful direction comes from recent developments into the neural representations of speech. Traditional neuroimaging techniques are limited in the ability to resolve either the temporal scale needed to investigate speech at a segmental level (a challenge for fMRI) or the fine spatial precision necessary to localize distinct processing systems (a limitation of EEG, including the current work). But recent developments in MEG, combined MRI-EEG and MRI-MEG systems, and intracranial recordings (such as electrocorticography, or ECoG) are making it possible to study the spatiotemporal correlates of neural representations in more precise detail.

Intracranial recordings in particular present a promising direction for studying phonetic categorization. These recording situations are rare, as they are only available during surgical interventions for particular neurological conditions. However, when available, they provide a high signal-to-noise ratio that makes it possible to look at the high-gamma range of the recorded cortical signal, which is argued to reflect time-locked task-related neuronal activity informative for cognitive processing (Crone, Boatman, Gordon, & Hao, 2001; Edwards, Soltani, Deouell, Berger, & Knight, 2005; Flinker, Chang, & Barbaro, 2011). Importantly, these recordings provide high spatial and temporal resolution, which makes it possible to localize stimulus-locked responses with millisecond-level precision, and as fine as 4 mm spatial resolution.

High temporal resolution is critical for this topic, as speech is a rapid time-varying signal with rapid event timescales on the order of tens of milliseconds. Fine spatial resolution, coupled with high temporal resolution, may help to illuminate the question of phonetic representations because it would provide the opportunity to compare responses in auditory and motor regions of the cortex to individual speech categories. Recent work has demonstrated the ability to analyze ECoG responses for evidence of category invariance in native language phonetic categories at individual recording sites (Chang et al., 2010). Similar techniques applied to novel categories, at both early and later stages of learning, could reveal the extent to which acoustic and articulatory representations are being recruited to discriminate non-native contrasts. This would be illuminating in cases where discrimination fails (i.e. where there is assimilation to native targets), as it would be possible to observe the assimilation of sepctrotemporal representations for multiple non-native categories being reduced to a single category at some point along the processing pipeline. It would also be informative for cases where acquisition is underway, as it may be possible to disentangle the contribution of motor and acoustic representations to the process of discrimination.

5.4.3 Novice vs. highly-proficient learners

All studies in this dissertation focus on novice learners, and the "fledgling" categories they are developing after a few hours or a few days of training. There is considerable work in the literature on both novice learners (Best et al., 2001; Golestani & Zatorre, 2009; Pruitt et al., 2006; Lim & Holt, 2011) and more proficient adult second-language learners (Akahane-Yamada et al., 1996; Flege et al., 1997; Hattori & Iverson, 2009; Diaz et al., 2008) who struggle with second-language phoneme categories. In some cases, challenges persist even after substantial training. Nevertheless, it must be the case that substantial learning and experience with the target language changes the nature of second-language categories. As a result, there is more work to be done to describe the differences between fledgling and more robust L2 categories.

A fruitful area for further research on this topic is the effects of cross-modal learning on novice and proficient learners. In section 5.3, I suggested that the effectiveness of cross-modal learning may diminish with increased exposure to the target language. However, that hypothesis was based on data which only observed novice learners (either with a single session of training, as in experiment 2, or with only a handful of training days, as in experiment 1A). Does this linear relationship correctly predict learning outcomes across the longer trajectory of L2 phoneme acquisition? Given the relatively limited number of studies on cross-modal training for perceptual learning, there is little data to currently speak to the longer-term patterns of this type of learning.

If a linear prediction is correct, then this suggests that cross-modal learning is only an effective tool for "jump-starting" acquisition. Once category learning has begun, within-domain approaches are more efficient tools for fine-tuning the system and developing categories beyond their fledgling state and into robust, well-formed representations. This would imply a very limited role for cross-modal learning, and support the notion that it primarily exists as a support system during adverse conditions (e.g. novel learning, noisy conditions, and cases of language disorders).

If the linear prediction is incorrect, perhaps a U-shaped trajectory is more appropriate. Such a trajectory would predict an effect of cross-modal learning at the beginning of learning, a decrease in its effectiveness during the middle stages of learning, and then an increase in its importance again as high proficiency is reached. In the latter case, cross-modal learning might act as a fine-tuning mechanism for the system, strengthening categories and getting them closer to targets that would be produced by native speakers of the language. It is possible that this fine tuning could act to polish the performance of speakers who are highly proficient in their second language in most levels of linguistic representation, but who still retain a significant accent and L1 perceptual biases.

Studies on highly-proficient bilinguals will be critical to testing these two hypotheses. Data in this vein would speak to the most effective strategies for second-language learners at multiple stages of proficiency. It would also shed light on the nature of phonetic representations more generally, and the extent to which perceptual and articulatory representations support one another throughout the course of language acquisition in adulthood.

5.5 Conclusion

The experiments in this dissertation explored perceptual and articulatory learning during non-native phoneme acquisition. Concerning the interaction between acoustic and articulatory representations, these studies present a mixed picture, with some cross-over between representations in the two domains that depend on the context of learning and the nature of the task. Cross-modal learning appears to be most effective for the most novice learners, and in situations where an overt behavioral response is required.

There were three primary contributions from this set of studies. The first was to add data to the literature on cross-modal learning for phonetic categories - particularly in the production-to-perception direction, which is relatively understudied. The second was a proposal of a tentative trajectory for fledgling category formation. The third was to provide additional commentary on the debate over the acoustic and articulatory bases for phonetic representations. Our understanding of these topics will benefit from future work on cross-modal learning at different stages of the acquisition process, more fine-grained investigations of the neural correlates of speech perception, and a deeper look into the trajectory of phonetic learning across the span of second language acquisition.

References

Akahane-Yamada, R., Tohkura, Y., Bradlow, A. R., & Pisoni, D. B. (1996). Does training in speech perception modify speech production? *Proceedings of the Fourth International Conference on Spoken Language Processing*, 606–609.

Alain, C., Woods, D. L., & Knight, R. T. (1998). A distributed cortical network for auditory sensory memory in humans. *Brain Research*, *812*, 23–37.

Alho, K. (1995). Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear and Hearing*, *16*(1), 38–51.

Alho, K., Woods, D., Algazi, A., Knight, R., & Näätänen, R. (1994). Lesions of frontal cortex diminish the auditory mismatch negativity. *Electroencephalography and Clinical Neurophysiology*, *91*, 353–362.

Arteaga, D. L. (2000). Articulatory phonetics in the first-year Spanish classroom. *The Modern Language Journal*, *84*(3), 339–354.

Baese-Berk, M. M. (2010). *An examination of the relationship between speech perception and production* (Unpublished doctoral dissertation). Northwestern University.

Baker, W., Trofimovich, P., Mack, M., & Flege, J. E. (2002). The effect of perceived phonetic similarity on non-native sound learning by children and adults. In *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 36–47).

Bates, D., & Maechler, M. (2014). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. (R package version 1.1-7)

Best, C. T., & Avery, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In J. E. Flege, O.-S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins Publishing Company.

Best, C. T., Goldstein, L., Tyler, M. D., & Nam, H. (2009). Articulating the Perceptual Assimilation Model (PAM): Perceptual assimilation in relation to articulatory organs and their constriction gestures. *The Journal of the Acoustical Society of America*, *125*(4), 2758–2758.

Best, C. T., & Hallé, P. A. (2010). Perception of initial obstruent voicing is influenced by gestural organization. *Journal of Phonetics*, *38*(1), 109–126.

Best, C. T., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*, *46*(2-3), 183–216.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America*, *109*(2), 775–794.

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*(3), 305–330.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*(4), 1001–1017.

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer (version 5.1.13).* Retrieved from `http://www.praat.org`.

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training Japanese adults on the English /r/-/l/ contrast. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (Vol. 1, pp. 287–308). Amsterdam/Philadelphia: John Benjamins.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohikura, T. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310.

Brunellière, A., Dufour, S., & Nguyen, N. (2011). Regional differences in the listener's phonemic inventory affect semantic processing: A mismatch negativity (MMN) study. *Brain and Language*, *117*(1), 45–51.

Catford, J., & Pisoni, D. B. (1970). Auditory vs . Articulatory in Auditory Training Exotic Sounds. *The Modern Language Journal*, *54*(7), 477–481.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Emergence of Categorical Speech Representation in the Human Superior Temporal Gyrus. *Nature Neuroscience*, *13*(11), 1428–1432.

Cheng, B., & Zhang, Y. (2013). Neural plasticity in phonetic training of the /i-I/ contrast for adult Chinese speakers. *The Journal of the Acoustical Society of America*, *134*(5), 4245–4245.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, *112*(4), 565–582.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1), 132–147.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, *8*(4), 919–924.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, *27*(4), 769–773.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

Dev, A. (2009). Effect of retroflex sounds on the recognition of Hindi voiced and unvoiced stops. *AI & Society*, *23*(4), 603–612.

Diaz, B., Baus, C., Escera, C., Costa, A., & Sebastián-Gallés, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning. *Proceedings of the National Academy of Sciences*, *105*(42), 16083–16088.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Reviews of Psychology*, *55*, 149–179.

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences*, *111*(19), 7126–7131.

Earle, S., & Myers, E. (2013). The effect of sleep on learned sensitivity to a non-native phonetic contrast. *The Journal of the Acoustical Society of America*, *134*(5), 4107–4107.

Edwards, E., Soltani, M., Deouell, L. Y., Berger, M. S., & Knight, R. T. (2005). High gamma activity in response to deviant auditory stimuli recorded directly from human cortex. *Journal of Neurophysiology*, *94*(6), 4269–4280.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–12.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, *15*(2), 399–402.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–82.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology*, *61*(6), 825–850.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 233–277.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470.

Flinker, A., Chang, E., & Barbaro, N. (2011). Sub-centimeter language organization in the human temporal lobe. *Brain and Language*, *117*(3), 103–9.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115–123.

Fowler, C. (2008). The FLMP STMPed. *Psychonomic Bulletin & Review*, *15*(2), 458–462.

Galantucci, B., Fowler, C., & Turvey, M. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second-language acquisition. *Canadian Journal of Psychology*, *13*(4), 26–72.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–23.

Golestani, N. (2014). Brain structural correlates of individual differences at low-to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism*, *18*(1), 6–34.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506.

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2-3), 55–67.

Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. *Journal of Child Language*, *22*(3), 703–726.

Gómez Lacabex, E. (2009). Relationship between perception and production in non-native speech. In *Phonetics and Phonology in Iberia.*

Gomez-Herrero, G., De Clercq, W., Anwar, H., Kara, O., Egiazarian, K., Van Huffel, S., & Van Paesschen, W. (2006). Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In *Proceedings of the 7th Nordic Signal Processing Symposium* (p. 130-133).

Gordon, P., Eberhardt, J., & Rueckl, J. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, *25*, 1–42.

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

Guion, S., Flege, J., Akahane-Yamada, R., & Pruitt, J. (2000). An investigation of current models of second language speech perception: The case of Japanese adults perception of English consonants. *The Journal of the Acoustical Society of America*, *107*, 2711–2724.

Gulian, M., Escudero, P., & Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. In *ICPhS XVI* (pp. 1893–1896). Saarbrücken.

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469–79.

Hazan, V., & Sennema, A. (2007). The effect of visual training on the perception of non-native phonetic contrasts. *ICPhS XVI*(August), 1585–1588.

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378.

Herd, W., Jongman, A., & Sereno, J. (2013). Perceptual and production training of intervocalic /d, flap, r/ in american english learners of spanish. *The Journal of the Acoustical Society of America*, *133*(6), 4247–4255.

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, *21*(7), 1229–1243.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407–422.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67–99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Hirata, Y. (2004). Computer Assisted Language Learning Computer Assisted Pronunciation Training for Native English Speakers Learning Japanese Pitch and Durational Contrasts for Native English Speakers Learning Japanese. *Computer Assisted Language Learning*, *17*(3-4), 357–376.

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research*, *1360*, 89–105.

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*, 37-58.

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

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: acquisition of the English /delta/-/theta/ contrast by francophones. *Perception & Psychophysics*, *40*(4), 205–15.

Kaan, E., Wayland, R., Bao, M., & Barkley, C. M. (2007). Effects of native language and training on lexical tone perception: an event-related potential study. *Brain Research*, *1148*, 113–22.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant–vowel syllables. *The Journal of the Acoustical Society of America*, *72*(2), 379–389.

Kewley-Port, D., Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America*, *73*(5), 1779–1793.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, *297*(5582), 846–848.

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587.

Kovelman, I., Yip, J. C., & Beck, E. L. (2011). Cortical systems that process language, as revealed by non-native speech sound perception. *Neuroreport*, *22*(18), 947–50.

Köver, H., & Bao, S. (2010). Cortical Plasticity as a Mechanism for Storing Bayesian Priors in Sensory Perception. *PLoS ONE*, *5*(5), 1–7.

Kraus, N., McGee, T., Carrell, T. D., King, C., Tremblay, K., Nicol, T., & Nkol, T. (1995). Central Auditory System Plasticity Associated with Speech Discrimination Training. *Journal of Cognitive Neuroscience*, *7*(1), 25–32.

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–7.

Kuhl, P. K., Conboy, B., & Coffey-Corina, S. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*, 979–1000.

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J.-F. L., & Imada, T. (2014). Infants brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences*, 11238-11245.

Kujala, T., & Näätänen, R. (2010). The adaptive brain: a neurophysiological perspective. *Progress in Neurobiology*, *91*(1), 55–67.

Lacabex, E. G., García Lecumberri, M., & Cooke, M. (2008). Identification of the contrast full vowel-schwa: training effects and generalization to a new perceptual context. *Ilha do Desterro*, *55*, 173–196.

Lai, Y.-h. (2009). Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese. *Language and Cognitive Processes*, *24*(7-8), 1265–1285.

Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. S. (2012). Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. *The Journal of Neuroscience*, *32*(11), 3942–8.

Levitt, A., Best, C., Goldstein, L., & Carpenter, A. (2006). English listeners perceptual assimilations for Zulu sounds: Evidence in support of the articulatory organ hypothesis. In *Acoustical Society of America* (Vol. 120, p. 3174).

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., & Formisano, E. (2012). Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex. *Journal of Neuroscience*, *32*(38), 13273–13280.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, *68*(8), 1–13.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lim, S.-J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–405.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*(2), 874–886.

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110–114.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

Macmillan, N., & Creelman, C. (1991). Detection theory: a user's guide. *Cambridge CUP Archive*, *6*.

Maiste, A. C., Wiens, A. S., Hunt, M. J., Scherg, M., & Picton, T. W. (1995). Event-related potentials and the categorical perception of speech sounds. *Ear and Hearing*, *16*(1), 68–89.

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta–analysis of studies conducted by gardner and associates. *Language Learning*, *53*(1), 123–163.

Massaro, D., & Chen, T. (2008). The motor theory of speech perception revisited. *Psychonomic Bulletin & Review*, *15*(2), 453–457.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavioral Research Methods*, *44*(2), 314-324.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McCandliss, B. D., Fiez, J. a., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective & Behavioral Neuroscience*, *2*(2), 89–108.

McGee, T., King, C., Tremblay, K., Nicol, T., Cunningham, J., & Kraus, N. (2001). Long-term habituation of the speech-elicited mismatch negativity. *Psychophysiology*, *38*(4), 653–658.

McGuire, G. (2008). *Selective attention and English listeners' perceptual learning of the Polish post-alveolar sibilant contrast.*

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, *20*(7), 895–903.

Myers, E. B., & Swan, K. (2012). Effects of Category Learning on Neural Sensitivity to Non-native Phonetic Categories. *Journal of Cognitive Neuroscience*, *24*(8), 1695–708.

Näätänen, R. (1990). Automatic and attention-dependent processing of auditory stimulus information. *Behavioral and Brain Sciences*, *13*(02), 261–288.

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*, 1–21.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., ... Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, *385*, 432–434.

Onishi, S., & Davis, H. (1968). Effects of duration and rise time of tone bursts on evoked V potentials. *The Journal of the Acoustical Society of America*, *44*(2), 582–591.

Pederson, E., & Guion-Anderson, S. (2010). Orienting attention during phonetic training facilitates learning. *Journal of the Acoustical Society of America*, *127*(2), EL54–9.

Pegg, J. E., & Werker, J. F. (1997). Adult and infant perception of two English phones. *The Journal of the Acoustical Society of America*, *102*(6), 3742–3753.

Peltola, M. S., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., & Näätänen, R. (2003). Native and foreign vowel discrimination as indexed by the mismatch negativity (MMN) response. *Neuroscience Letters*, *352*(1), 25–28.

Perrachione, T., Lee, J., Ha, L., & Wong, P. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*(1), 461–472.

Pimsleur, P. (1963). Discrimination training in the teaching of French pronunciation. *The Modern Language Journal*, *47*(5), 199–203.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an l2: A review. *Journal of Phonetics*, *29*(2), 191–215.

Protopapas, A., & Calhoun, B. (2000). Adaptive phonetic training for second language learners. In P. Delcloque (Ed.), *Proceedings of the 2nd International Workshop on Integrating Speech Technology in Language Learning* (pp. 31–38). University of Abertay, Dundee, UK.

Pruitt, J. S. J. (1995). *The perception of Hindi dental and retroflex stop consonants by native speakers of Japanese and American English* (Unpublished doctoral dissertation). University of South Florida.

Pruitt, J. S. J., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, *119*(3), 1684–96.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Raizada, R. D. S., Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: Prediction of individual differences. *Cerebral Cortex*, *20*(1), 1–12.

Ritter, W., Vaughan, H. G., & Costa, L. D. (1968). Orienting and habituation to auditory stimuli: a study of short terms changes in average evoked responses. *Electroencephalography and Clinical Neurophysiology*, *25*(6), 550–556.

Rivera-Gaxiola, M., Silva-Pereyra, J., & Kuhl, P. K. (2005). Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. *Developmental Science*, *8*(2), 162–72.

Ruhm, H., & Jansen, J. (1969). Rate of stimulus change and evoked response. 1. Signal rise-time. *Journal of Auditory Research*, *9*(3), 211–216.

Sadakata, M., & McQueen, J. M. (2011). The role of variability in non-native perceptual learning of a Japanese geminate-singleton fricative contrast. In *Interspeech 2011* (pp. 873–876).

Schneiderman, E., Bourdages, J., & Champagne, C. (1988). Second-language accent: The relationship between discrimination and perception in acquisition. *Language Learning*, *38*(1), 1–19.

Scott, S. K., & Wise, R. J. (2004). The functional neuroanatomy of prelexical processing in speech perception. *Cognition*, *92*(1), 13–45.

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*(3), 435–43.

Seitz, A. R., & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature*, *422*(6927), 36–36.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *The Journal of the Acoustical Society of America*, *107*(5), 2697–2703.

Sharma, A., Kraus, N., McGee, T., Carrell, T., & Nicol, T. (1993). Acoustic versus phonetic representation of speech as reflected by the mismatch negativity event-related potential. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *88*(1), 64–71.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, *49*(14), 3831–3846.

Song, J. H., Skoe, E., Wong, P. C. M., & Kraus, N. (2008). Plasticity in the adult human auditory brainstem following short-term linguistic training. *Journal of Cognitive Neuroscience*, *20*(10), 1892–1902.

Stevens, K. N., & Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, *3*, 215–233.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. *Models for the Perception of Speech and Visual Form*, 88–102.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r,l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145.

Studdert-Kennedy, M., & Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. *Studies in the Evolution of Language*, *3*, 235–254.

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation. *The Journal of the Acoustical Society of America*, *94*(3), 1256–1268.

Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *The Journal of the Acoustical Society of America*, *90*(3), 1309–1325.

Tees, R. C., & Werker, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology*, *38*(4), 579-590.

Terrace, H. S. (1963). Discrimination learning with and without "errors". *Journal of the Experimental Analysis of Behavior*, *6*(I), 1–27.

Thomson, J. M., Goswami, U., & Baldeweg, T. (2009). The ERP signature of sound rise time changes. *Brain Research*, *1254*, 74–83.

Tremblay, A., & Ransijn, J. (2013). LMERConvenienceFunctions: A suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions. *R package*, *version 2*, 919–931.

Tremblay, K., Kraus, N., & McGee, T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, *9*(16), 3557–3560.

Tremblay, K., Krause, N., Carrell, T. D., & McGee, T. (1997). Central auditory system plasticity : Generalization to novel stimuli following listening training. *Journal of the Acoustical Society of America*, *102*(6), 3762–3773.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(1), 53–73.

van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, *38*(4), 584–589.

Vlahou, E., Protopapas, A., & Seitz, A. (2011). Implicit Training of Nonnative Speech Stimuli. *Journal of Experimental Psychology: General*, *141*(2), 1 – 19.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033.

Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *Journal of Cognitive Neuroscience*, *16*(6), 978–987.

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, 349–355.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Wilson, C., Davidson, L., & Martin, S. (2014). Effects of acoustic - phonetic detail on cross-language speech production. *Journal of Memory and Language*, *77*, 1–24.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., ... Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology*, *36*, 638–642.

Woods, D. L., & Elmasian, R. (1986). The habituation of event-related potentials to speech sounds and tones. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *65*(6), 447–459.

Wright, B. A., LeBlanc, E. K., Conderman, J. S., & Coburn, C. S. (2015). Contributions of practice with feedback and testing without feedback to learning of a non-native phonetic contrast. *The Journal of the Acoustical Society of America*, *137*(4), 2384.

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2009). Training the Brain to Weight Speech Cues Differently : A Study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332.

Zevin, J. D., Datta, H., Maurer, U., Rosania, K. A., & McCandliss, B. D. (2010). Native language experience influences the topography of the mismatch negativity to speech. *Frontiers in Human Neuroscience*, *4*, 212.

Zhang, Y., Koerner, T., Miller, S., Grice-Patil, Z., Svec, A., Akbari, D., ... Carney, E. (2011). Neural coding of formant-exaggerated speech in the infant brain. *Developmental Science*, *14*(3), 566–81.

Zhang, Y., Kuhl, P. K., & Imada, T. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage*, *46*, 226–240.

Appendix A

Production training script

This appendix contains the text and images used in the production training sessions of experiments 1A and 2. An abridged version which focused on the dental-retroflex contrast and the breathy-voiced contrast was developed for experiment 3. All material was presented using custom scripts constructed in OpenSesame (Mathôt et al., 2012). Line breaks indicate separate slides.

Sagittal sections were adapted from figures prepared by Robert Mannell (http://clas.mq.edu.au/phonetics/phonetics/consonants/oral_stops.html).

Thank you for participating in today's study.

Please turn off your cell phone if you have one with you.

When you are ready, click "ok" to start.

In today's task, you are going to learn how to produce several sounds in Hindi. Some of them are like sounds that you know in English; others may be new to you.

First, you'll be introduced to some information about the sounds, and then you'll practice repeating sounds made by a native Hindi speaker.

Do your best to follow along during the tutorial. Much of the information may be new to you, but if you practice and follow along, you'll learn it more easily.

Part 1: Place of articulation

Place of articulation is a term in linguistics that is used to describe the place that a consonant is made in your mouth.

Some consonants are made with the lips (try saying "bee"), others with the tip of the tongue at the roof of the mouth (say "ta"), and so on.

Hindi has two places of articulation that are similar to English "t".

They may both sound a lot like "t" to you, but they're slightly different than the "t" in English - and also slightly different from each other.

The first place of articulation is called "dental". It is a sound much like "t", but with the tip of the tongue slightly forward in the mouth, touching the back of the upper teeth.

Trying saying an English "tah" first, and notice how the tip of your tongue briefly touches the hard ridge behind your upper teeth.

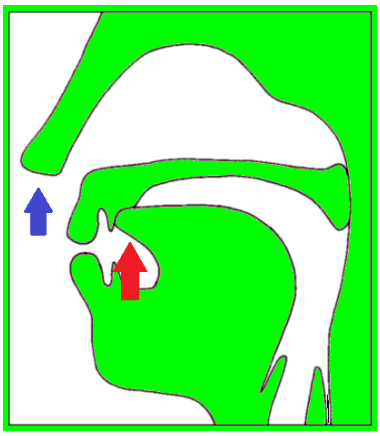
Now, say "tah" again, but instead of touching the tip of the tongue to that hard ridge, touch it to the back of the top teeth. You just said a dental sound.

Now try to say "tee" and "too" with a dental "t" sound.

Don't worry if the dental "t" doesn't sound very different to you than an English"t". Just focus on the position of your tongue, and make sure you can feel the difference.

A visual may help you remember the tongue placement of a dental sound.

Here is a picture of the inside of your mouth when you say a dental "t".



Look at this picture as if you were viewing the inside of your mouth from the side. To help orient you, a blue arrow is pointing to your nose.

Notice how the tip of the tongue touches the upper teeth (indicated with a red arrow).

Study the picture, then press any keyboard button to continue.

Throughout this experiment, we'll use the color green to refer to dental sounds.

Anytime you see a green picture of the tongue, or see green text, remember that you'll be making a dental consonant, with the tip of your tongue against the back of the upper teeth.

Practice saying "tah", "tee", and "too" with dental "t"s a few times. When you feel comfortable with that, press "ok" to continue.

Now we'll learn the second place of articulation. These are sounds called "retroflex" consonants.

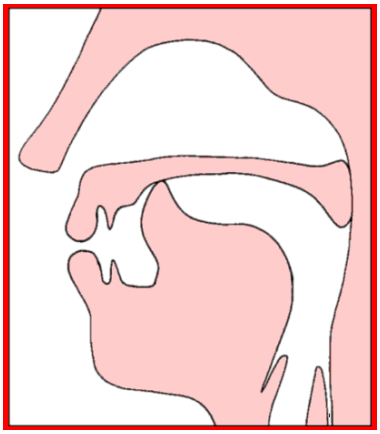
To try making a retroflex "t", curl your tongue back a bit so that the underside of the tip of the tongue touches the roof of the mouth. You should touch the roof of the mouth just behind that hard ridge where you would say an English "t".

Try saying "tah" with a retroflex "t".

Remember, the tip of your tongue curls back so that the underside touches the roof of your mouth.

Here's a picture of your tongue in the mouth during a retroflex "t".

Notice the tongue placement and how the tongue tip curls back.



Press any key to continue.

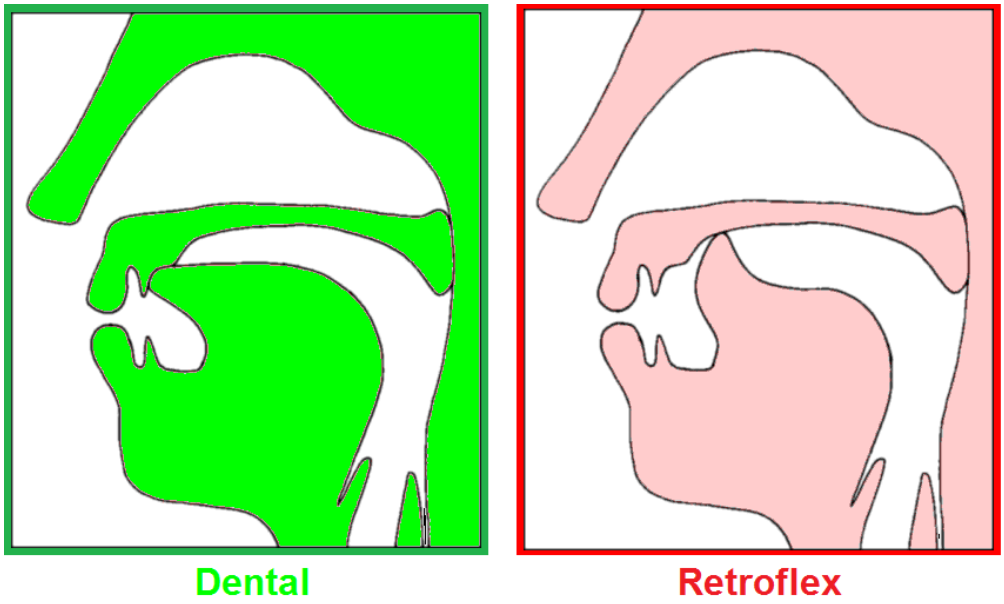
Throughout this experiment, we'll use the color "red" to refer to retroflex sounds.

Anytime you see a red picture of the tongue, remember that you'll be making a retroflex consonant, with the tip of the tongue curled back.

Practice saying "tah", "tee", and "too" with retroflex "t"s a few times.
When you feel comfortable with that, press "ok" to continue.

Here are our dental "t" and retroflex "t" side by side.

Notice the difference in tongue placement in each.



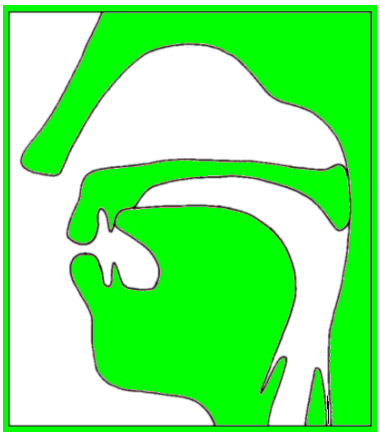
Press any key to continue.

Now you're going to practice saying some dental and retroflex "t"s.

They will be color-coded, and you'll see the picture of the tongue corresponding to each place of articulation, so that you remember how to say each.

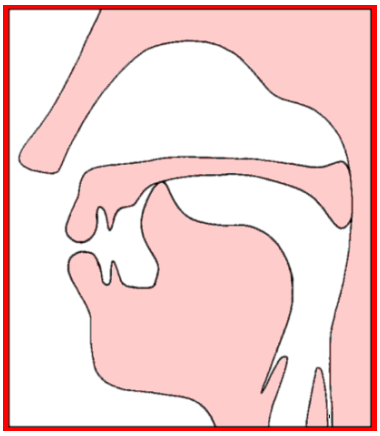
When you're ready, press the "start" button to begin.

Say “tah”



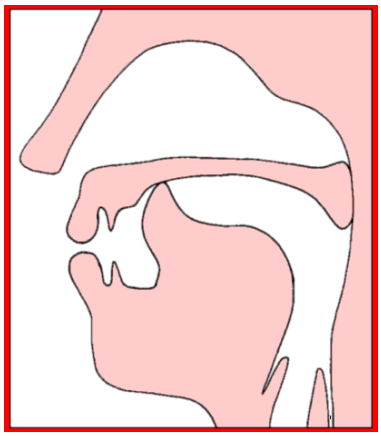
Press any key to continue.

Say “too”



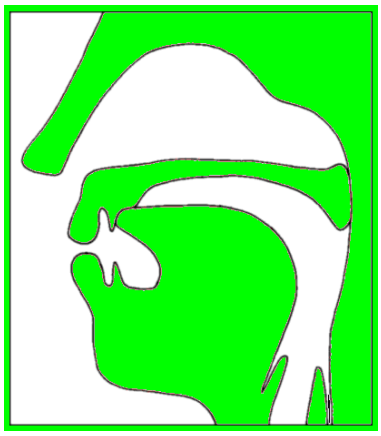
Press any key to continue.

Say “tee”



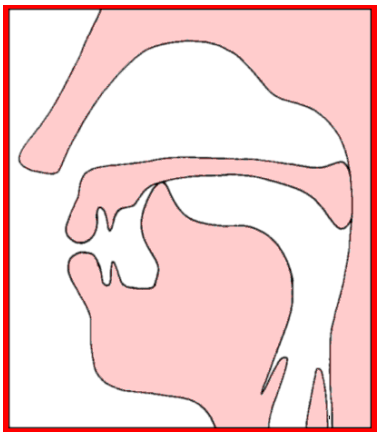
Press any key to continue.

Say “too”



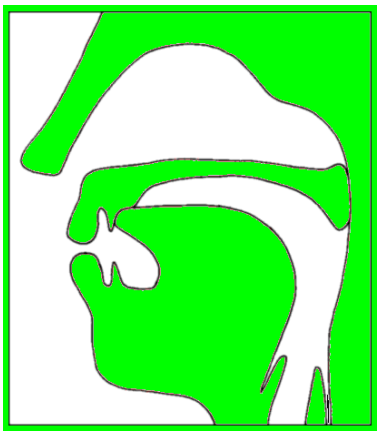
Press any key to continue.

Say “tah”



Press any key to continue.

Say “tee”



Press any key to continue.

Part 2: Voicing

Voicing is a term about consonants that roughly describes how air passes through the mouth when you say them.

Hindi has four different types of voicing.

These types of voicing can be combined with the places of articulation you just learned, for a total of 8 consonants in our training set.

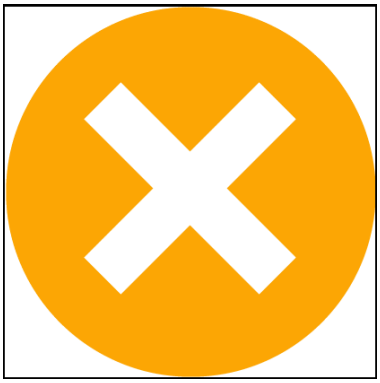
Luckily, you already know two types of voicing from English. One is just like the way that you say "d".

Say "dah", "dee", "doo". (Easy, right?)

Now try saying them with the places of articulation you learned earlier.

Dental: "dah", "dee", "doo"
Retroflex: "dah", "dee", "doo"

Whenever you're asked to say a sound like this, we'll show you this image:



(You'll see why shortly.)

So when you see the orange X, your voicing will be just like the English "d".

Press any key to continue.

The second type of voicing we'll learn is like the English "t".

Say "tah", "tee", "too".

Now dental: "tah", "tee", "too"
Now retroflex: "tah", "tee", "too"

We want you to notice something about "t". Place your hand close to the front of your mouth, and say "t- t- t-".

Do you feel a little puff of air on your hand each time you say it?

Good. Try putting a piece of paper in front of your mouth and say "t- t- t-" again.

The paper should move a little bit.

That little puff of air is a voicing feature of the sound t.

In the experiment, we'll remind you of that little puff of air moving that sheet of paper with this picture:

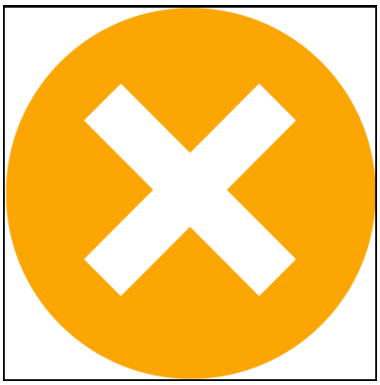
Whenever you see it, you'll make a sound with voicing like t.

Press any key to continue.

One of the ways that t is different from d is that d does NOT have that little puff of air.

So when you see this picture for t, think "puff of air".

When you see the orange X for d, think "no puff of air".



Press any key to continue.

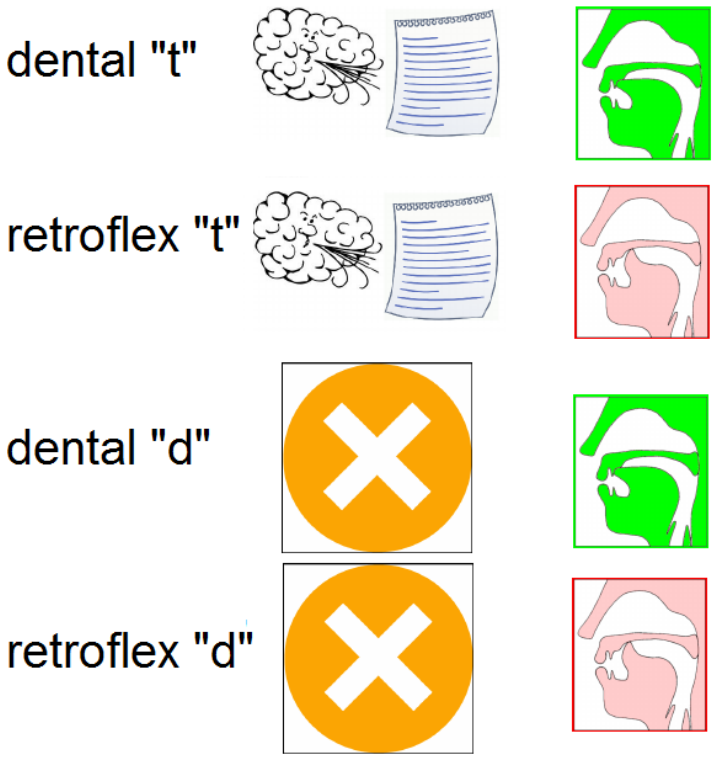
Try saying "tah dah tah dah" with the piece of paper or your hand in front of your mouth a few times.

Notice the puff of air during "t", but not "d"?

Good; you've already learned two of the four Hindi voicing types.

Let's review the places of articulation you learned earlier with these two voicing types.

Try saying each of these:



Press any key to continue.

Nice work. Now we'll learn the third voicing type.

This type sounds a lot like a "d", but with one important difference:

Your vocal cords will make sound - they'll be vibrating - while your mouth is still closed, before you make the "d" sound.

How do you know if your vocal cords are vibrating?
Try humming right now.

While you're humming, place your fingers lightly over your throat. Do you feel a buzzing? That's how you know that your vocal cords are vibrating.

That's the feeling you want to feel when you say this new sound.
We'll type it as "~d" to remind you about the vibration before the "d" sound.

Try saying: "~dah ~dee ~doo".

To remind you about this voicing, you'll see this whenever you should produce a ~d sound:

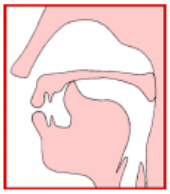
You don't have to place your hand exactly like this; however you can feel the vibration is fine.

Press any key to continue.

Now try it with the places of articulation you learned:

dental "~d"
~dah ~dee ~doo

retroflex "~d"
~dah ~dee ~doo



Press any key to continue.

The difference between "d" and "~d" can be a little tricky.
Try switching back and forth a few times here, with dental articulation:

"dah" "~dah" "dee" "~dee" "doo" "~doo"

Feel free to keep your fingers at your throat to help you tell when your vocal cords are vibrating.

Good, we're almost there!

The fourth type of voicing combines the buzzing vocal cords you just learned with the puff of air you already know from "t".
In this voicing, you should feel vibration at your throat throughout the consonant, and you should ALSO have a puff of air after release.

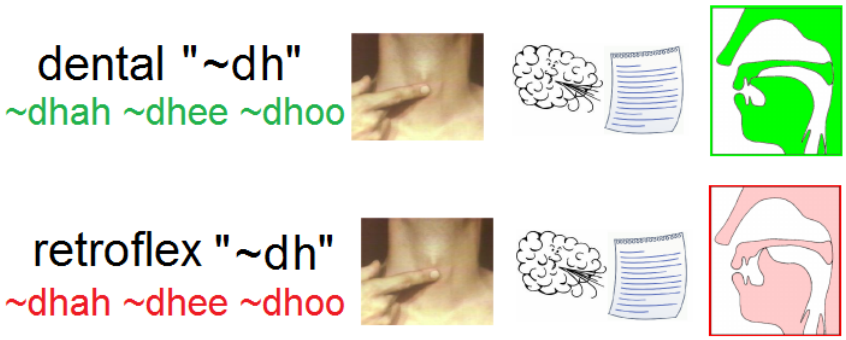
We'll call this sound "~dh", with the ~ reminding you about the vibration, and the "h" reminding you of the puff of air.

Try saying "~dhah", "~dhee", "~dhoo".

To remind you about the voicing of "~dh", we'll show you both of these pictures:

Press any key to continue.

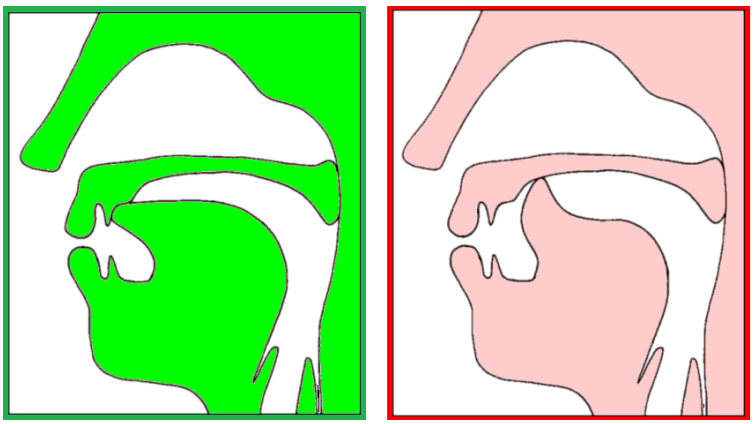
Now try it with the places of articulation you learned:



Press any key to continue.

You've learned everything you need to know!
Let's review the types of sounds you'll be producing today.

Every sound will have one of two places of articulation:



Make sure that you use the color and the tongue diagram to remind you which to use.

You shouldn't ever be using the English place of articulation for the task ahead. Every sound will be either dental or retroflex.

Press any key to continue.

Remember that you'll also be using one of 4 voicing types.

t no buzzing, puff of air 

d no buzzing, no puff of air 

~d buzzing, no puff of air 

~dh buzzing, puff of air  

Press any key to continue.

You'll now be completing a repetition task.
In this task, you'll hear a Hindi speaker say each of the sounds you've learned, and you'll repeat after her as best you can.

We'll keep using the colors and pictures throughout the experiment, to remind you of how to pronounce each sound.

If you like, you can take a short break for a minute or two.
When you're ready, click on the button to start the repetition task.