Acquisition of allophony from speech input by adult learners

by

Masaki Noguchi

B.A. Humanity, Soka University, 2000

M.A. Anthropology, Tulane University, 2007

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Linguistics)

The University of British Columbia

(Vancouver)

May 2016

Abstract

Sound systems are a basic building block of any human language. An integral part of the acquisition of sound systems is the learning of allophony. In sound systems, some segments are used as allophones, or contextually-conditioned variants of a single phoneme, and learners need to figure out whether given segments are different phonemes or allophones of a single phoneme. There is a growing interest in the question of how allophony is learned from speech input (e.g., Seidl and Cristia, 2012). This dissertation investigates the mechanisms behind the learning of allophony. Whether given segments are different phonemes or allophones of a single phoneme is partly determined by the contextual distribution of the segments. When segments occur in overlapping contexts and their occurrences are not predictable from the contexts, they are likely to be different phonemes. When segments occur in mutually exclusive contexts, and their occurrences are predictable from the contexts (i.e., they are in complementary distribution), the segments are likely to be allophones. This dissertation starts with the hypothesis that allophonic relationships between segments can be learned from the complementary distribution of the segments in input.

With data collected in a series of laboratory experiments with adult English speakers, I make the following claims. First, adults can learn allophonic relationships between two segments from the complementary distribution of the segments in input. The results of Experiment 1 showed that participants learned to treat two novel segments as something like allophones when they were exposed to input in which the segments were in complementary distribution. Second, the learning of allophony is constrained by the phonetic naturalness of the patterns of complementary distribution. The results of Experiment 2 showed that the learning

of allophony happened only when learners were exposed to input in which relevant segments were in phonetically natural complementary distribution. Third, the learning of allophony involves the learning of the context-dependent perception of relevant segments. The results of Experiment 3 showed that, through exposure to input, participants' perception of the relevant segments became more dependent on context such that they perceived the segments as being more similar to each other when they heard the segments in phonetically natural complementary contexts.

Preface

This research project was conceived and designed by Masaki Noguchi with assistance from Carla L. Hudson Kam, Gunnar Ólafur Hansson, and Molly Babel. Data collection was performed by Masaki Noguchi, and data analyses (including statistical analyses) were performed by Masaki Noguchi with assistance from Carla L. Hudson Kam and Gunnar Ólafur Hansson. This research project was funded by NSERC Discovery Grant (Individual) to Carla L. Hudson Kam "Constraints on language acquisition and how they change (or don't) with age." All experiments presented in this dissertation were approved by the University of British Columbia's Research Ethics Board [certificate #H12-02287].

The following is a list of presentations and publications in which various parts of this dissertation were first introduced.

- The results of Expriment 1 (Chapter 2) were first presented as a poster at The 14th Conference on Laboratory Phonology in Tachikawa, Tokyo (Noguchi and Hudson Kam, 2014b). The poster was created by Masaki Noguchi with assistance from Carla L. Hudson Kam.

- The results of Experiment 2 (Chapter 3) were first presented as a poster at The 39th Annual Boston University Conference on Language Development in Boston, MA (Noguchi and Hudson Kam, 2014a). The poster was created by Masaki Noguchi with assistance from Carla L. Hudson Kam.

- The results of Experiment 3 (Chapter 4) were first presented as a poster at 2015 Annual Meeting of Phonology in Vancouver, B.C. (Noguchi and Hudson Kam, 2015b). The poster was created by Masaki Noguchi with assistance from Carla L. Hudson Kam and Gunnar Ólafur Hansson.

- The results of the experiment presented in Appendix A were first presented in the proceedings of Acoustics Week in Canada 2015 (Noguchi and Hudson Kam, 2015a). The paper was written by Masaki Noguchi with assistance from Carla L. Hudson Kam.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere gratitude to my supervisors, Prof. Carla L. Hudson Kam and Prof. Gunnar Ólafur Hansson, for the continuous support of my dissertation and related research, for their patience, motivation, and immense knowledge. Besides my supervisors, I would like to thank Prof. Molly Babel for her insightful comments and encouragement.

Chapter 1

Introduction

Sound systems are a basic building block of any human language. A major component of sound systems is the inventory of *phonemes*. Phonemes are the categories of sounds that are used to make lexical contrasts (e.g., Trubetzkoy, 1969; Twaddell, 1935). A great deal of research has investigated the acquisition of phonemes. Studies on infant speech perception have demonstrated that infants start categorizing speech sounds according to the inventory of phonemes in their target language during the first year of life even though they have a very limited amount of lexical knowledge (e.g., Kuhl et al., 1992; Werker and Tees, 1984). Studies have suggested that infants can induce phonetic categories from statistical information in input: specifically, the frequency distribution of the sounds in acoustic space (e.g., Maye et al., 2002). These phonetic categories include representations for individual sounds or *segments*.[1]

The acquisition of phonemes is more complex than the learning of segments (e.g., Werker and Curtin, 2005). One of the complexities in the acquisition of phonemes is the learning of *allophony*. In sound systems, every segment belongs to a phoneme, but a phoneme may comprise multiple segments. In other words, some segments are used as variants of a single phoneme. These variants are called *allophones*. In order to acquire phonemes, infants need to know whether given segments are separate phonemes or allophones of the same phoneme. There is a

[1] Following Pierrehumbert (2003, p.118), I use the term "segments" to refer to temporally discrete units of speech that are equivalent to IPA symbols.

growing interest in the question of when and how infants learn allophony in their target language (see Seidl and Cristia, 2012, for a review). Studies have suggested that infants start learning allophony in their target language in the first year of life (e.g., Seidl et al., 2009). However, the mechanisms behind the learning of allophony are yet to be understood.

Whether two given segments are separate phonemes or allophones is partly determined by the distribution of the segments in particular environments. Specifically, when the segments occur in the same contexts, and thus their occurrences are unpredictable, they are likely to be separate phonemes. By contrast, when the segments occur in mutually exclusive contexts, and thus their occurrences are predictable, they are likely to be allophones (e.g., Hall, 2009; Jones, 1950; Trubetzkoy, 1969). Therefore, researchers have argued that allophonic relationships between segments can be learned from the relative predictabilities of the segments in particular environments (e.g., Peperkamp et al., 2006a).

In this dissertation, I investigate the mechanisms behind the learning of allophony. Studies show that human learners are sensitive to the frequency distribution of sounds in acoustic space (e.g., Maye, 2000; Maye et al., 2002). Studies also show that human learners are sensitive to statistical dependencies between linguistic items presented in a sequence (e.g., Saffran et al., 1996a,b), and they probably use this ability to learn regularities in the distribution of segments across environments (i.e., *phonotactic distributions*) (e.g., Chambers et al., 2003; Onishi et al., 2002). Here, I hypothesize that human learners can learn allophonic relationships between sounds based on these two different kinds of distributional information. They can learn how to categorize sounds based on the frequency distribution of the sounds in acoustic space. But they can also learn the contextual distribution of the categories or segments at the same time and use their knowledge about the contextual distribution to treat the categories as separate phonemes or allophones. If the segments occur in mutually exclusive contexts, and their occurrences are predictable, they are likely to be treated like allophones. But if the segments occur in overlapping contexts, and their occurrences are not predictable, they are likely to be treated like separate phonemes.

With the data collected in a series of laboratory experiments presented in the following chapters, I will make the following claims: (1) adults can learn allo-

phonic relationships between segments from the contextual distribution of the segments in input (Experiment 1), (2) the learning of allophony is constrained by the phonetic naturalness of the patterns of contextual distribution (Experiment 2), and (3) the learning of allophony involves the learning of context-dependent perception of relevant segments (Experiment 3).

1.1 Phonological relationships: Phonemic contrasts vs. allophony

In phonological analysis, sounds are classified into categories at both segmental and sub-segmental levels. At the segmental level, sounds are classified into segmental categories (or simply segments). Categorization at the segmental level assumes that continuous speech can be analyzed as a string of temporally discrete segment-sized units. However, such an assumption has been questioned by some researchers on both phonetic and phonological grounds (see Ladd, 2014, for a recent critical review on segment-based representations in phonology). Arguments have also been put forward against the idea that segments are the basic unit of speech processing (see Klatt, 1979; Port, 2010; Port and Leary, 2005, for arguments against the primacy of segment-sized units in speech processing). Despite these criticisms, the notion of segments is still widely used in phonetics and phonology, and segments are the basic building blocks of most theories of phonological relationships.

At the sub-segmental level, sounds are classified into sub-segmental categories like *features*. While some theories assume that features are tied to segmental units (e.g., *distinctive features*: Chomsky and Halle, 1968; Clements, 1985; Jakobson et al., 1951), others assume that they are not (e.g., *articulatory gestures*: Browman and Goldstein, 1989, 1992). Features are sub-segmental in a sense that a segment comprises multiple features, but categorization based on a feature may comprise multiple segments (i.e., a *natural class*).

Segments in sound systems can stand in different types of phonological relationships with each other. When differences between segments make lexical contrasts, the relationship between the segments is *contrastive* (i.e., the segments are separate phonemes). In other words, the substitution of one segment for another in

a word can change the meaning of the word. This also means that the distribution of these segments is not conditioned by particular environments; these segments can occur in the same contexts (i.e., they are in *non-complementary distribution*) and their occurrences are not predictable from the contexts. In English, for example, substituting [u] for [i] in the word *beat* [bit] results in another word *boot* [but], and this is concomitant with the fact that [u] and [i] can occur in the same contexts, after [b] and before [t]. Therefore, [u] and [i] are in a phonemically contrastive relationship; they are separate phonemes, /u/ and /i/.[2]

When differences between segments do not make lexical contrasts, the relationship between the segments is *allophonic* (i.e., the segments are allophones or variants of a single phoneme). The distribution of allophones is usually conditioned by particular environments; they occur in mutually exclusive contexts (i.e., they are in *complementary distribution*) and their occurrences are predictable from the contexts. In English, for example, so-called "light" [l] and "dark" (velarized) [ɬ] are in complementary distribution; light [l] occurs in syllable-initial position as in the word *leaf* [lif], and dark [ɬ] occurs in syllable-final position as in the word *feel* [fiɬ] (Ladefoged and Johnson, 2014, p.73). Therefore, light [l] and dark [ɬ] are allophones of a single phoneme /l/ in English. In this case, substituting [ɬ] for [l] in syllable-initial position, or presenting [ɬ] in the inappropriate context, affects the processing of [ɬ] (e.g., slows down phoneme monitoring) but does not change the identity of [ɬ] as /l/ (e.g., Lin, 2011).

Sometimes, multiple segments are used interchangeably as surface realizations of a single phoneme: *free variation* (e.g., Trubetzkoy, 1969). For example, in English, [t] and [t^h] can occur in the same context (e.g., word final position), but the former can substitute the latter in a word without changing the meaning of the word (e.g., both [k^hæt] and [k^hæt^h] are acceptable realization of a word *cat*). Free variation is not usually considered to be a part of allophony, at least of the type discussed in this dissertation (Hall, 2009, p.11).

Recent approaches view phonological relationships as a probabilistic phenomenon rather than a categorical dichotomy between contrastive and allophonic (Hall, 2009, 2012, 2013b; Peperkamp et al., 2006a). Hall, for instance, claims that

[2] In phonology, segments and allophones are transcribed within brackets ([ ]), and phonemes are transcribed within slashes (/ /).

phonological relationships are determined on the basis of the relative predictabilities of the occurrences of two segments in particular environments; the more the occurrences of two segments are predictable, the more allophonic the relationships between the segments are. This probabilistic approach has a significant advantage over the traditional approach, particularly in the analysis of so-called intermediate phonological relationships (Hall, 2009, 2012, 2013b). In Japanese, for example, the alveolar sibilants [s] and [(d)z] and the alveolopalatal sibilants [ɕ] and [(d)ʑ], respectively, are in such an intermediate relationship. They are in complementary distribution in a subset of the lexicon (old Japanese words). While the alveolar sibilants occur before the vowels [a], [e], [o], and [ɯ], the alveolopalatal sibilants occur before the vowel [i]. But the complementarity is weakened in the other subsets of the lexicon (old loanwords from Chinese, recent loanwords, and onomatopoetic words); there the alveolopalatal sibilants also occur before the vowels [a], [e], [o], and [ɯ]. Under the traditional approach, the relationship between the alveolar sibilants and alveolopalatal sibilants is neither contrastive nor allophonic. Under the probabilistic approach, as shown in the Figure 1.1, the relationship can be defined as falling anywhere between a perfect phonemic contrast and perfect allophony depending on how much the distribution of the alveolar sibilants and the distribution of the alveolopalatal sibilants overlap with each other (Hall, 2013a).

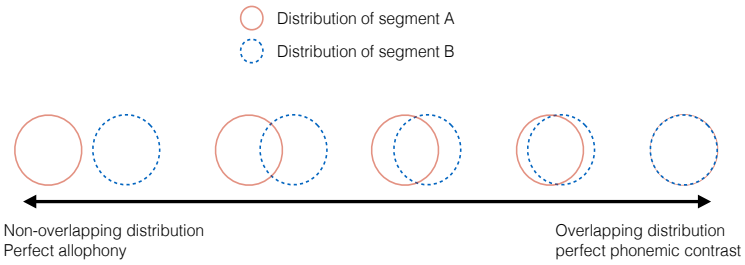


Figure 1.1: Continuum between allophony and phonemic contrast (based on Figure 1 in Hall, 2012)

1.2 The perceptual effect of phonological relationships

Whether segments are contrastive or allophonic significantly affects the way the segments are perceived. Since allophones are variants of a single phoneme category, listeners tend to be less sensitive to acoustic differences between allophones. For instance, listeners discriminate native allophones with less accuracy and longer latency than native phonemes (Beddor and Strange, 1982; Boomershine et al., 2008; Harnsberger, 2001; Peperkamp et al., 2003; Whalen et al., 1997).[3] However, comparing phonemes and allophones in terms of their discriminability within a single language has a potential drawback. Since the same set of segments cannot be both phonemes and allophones in the same language, any such comparison must be made between different sets of segments. This makes the interpretation of any differences between phonemes and allophones difficult. This problem is overcome by comparing phonemes and allophones across languages (Boomershine et al., 2008; Johnson and Babel, 2010). One such study looked at the perception of the voiced alveolar stop [d] and alveolar tap [ɾ] by Spanish and English speakers. Crucially, these two segments are phonemes in Spanish but allophones in English (i.e., the tap [ɾ] occurs intervocalically and the voiced stop [d] occurs elsewhere) (Boomershine et al., 2008). The results of the study showed that English speakers perceived these segments as being more similar to each other than Spanish speakers did. This clearly indicates that the same set of segments are perceived differently by speakers of different languages depending on whether the segments are contrastive or allophonic in their native languages.

Despite the perceptual effects of allophonic relationships that have been demonstrated, listeners are not completely insensitive to acoustic differences between native allophones. Listeners' sensitivity to allophonic variation is affected by task variables as well. For example, Pegg and Werker (1997) demonstrated that listeners show better sensitivity to allophonic variation in their native language when tested with a task that allows them to compare test stimuli at the level of auditory processing (e.g., the AX paradigm). In such a task, listeners' responses

[3] Phonological relationships are gradient (see Section 1.1). Hall (2009) argues that listeners' sensitivity is also gradient; listeners are less sensitive to acoustic differences between segments that are more allophonic. For brevity, in what follows I will just say "allophones" when what is intended is "allophones of a single phoneme".

are based more on the acoustic properties of the stimuli than on the categorization of the stimuli (Fujisaki and Kawashima, 1969, 1970; Pisoni, 1973).

Studies have also demonstrated that context plays an important role in the perception of allophones (Peperkamp et al., 2003; Whalen et al., 1997). For example, listeners' sensitivity to allophonic variation depends on whether the allophones are presented in the appropriate contexts or not. Specifically, the discrimination of native allophones is harder when they are presented in appropriate contexts than when they are presented in inappropriate contexts (Peperkamp et al., 2003). Other studies have demonstrated that listeners use their knowledge about allophonic variation in various speech processing tasks (e.g., Church, 1987). For example, listeners use allophonic variation as a cue for syllable boundaries in speech segmentation (Christie Jr., 1974; Nakatani and Dukes, 1977). Christie Jr. (1974) demonstrated that as the amount of aspiration in [t] in [asta] increases, English speakers' judgments about the location of a syllable boundary changes from [a.sta] to [as.ta]. This is because the aspirated [t^h] and the unaspirated [t] are allophones; while the aspirated [t^h] occurs in syllable-initial position, the unaspirated [t] occurs in a consonant cluster following [s], and the presence or absence of aspiration in [t] serves as a cue for the location of a syllable boundary.

In sum, phonological relationships significantly affect listeners' sensitivity to acoustic differences between segments. Specifically, listeners are sensitive to acoustic differences between segments that are contrastive in their native language, but they are less sensitive to acoustic differences between segments that are allophonic in their native language.

1.2.1 Reduced sensitivity to allophonic differences: an information theoretic account

Hall (2009) recently proposed an information theoretic account of the perceptual effect of phonological relationships. In information theory (Cover and Thomas, 2006; Shannon and Weaver, 1949; Shannon, 1951), the amount of information carried by a set of messages is measured in *entropy*, an index of how predictable the messages are. When the messages are more predictable, they are less informative and their entropy is lower. When the messages are less predictable, they are more informative and their entropy is higher. In an information theoretic view on hu-

man cognition, the amount of information in stimuli, or the predictability of the stimuli, significantly affects efficiency in the processing of the stimuli. When the amount of information is larger, or the stimuli are less predictable, the processing requires more cognitive resources. By contrast, when the amount of information is smaller, or the stimuli are more predictable, the processing requires fewer cognitive resources (e.g., Hyman, 1953; Pierce, 1980; Wickens, 1981). With speech stimuli, studies have demonstrated that higher predictability of stimuli facilitates speech processing in a high-demand task (e.g., Moray and Taylor, 1958; Treisman, 1960, 1964, 1965). For example, in Treisman (1964), English speakers heard two passages simultaneously in a binaural recording and were asked to shadow one of the passages. Participants performed better in the shadowing task when the semantic predictability of the words in the passage was higher (i.e., adjacent words showed more English-like transitional probabilities).

In speech perception, researchers have argued that listeners modulate selective attention according to the predictability of stimuli (Astheimer and Sanders, 2009, 2011). For example, listeners have a tendency to attend to the sounds that occur in word onset position (e.g., Connine et al., 1993; Marslen-Wilson and Zwitserlood, 1989). According to Astheimer and Sanders (2009, 2011), this is because the sounds that occur at the onset of a word are less predictable than the same sounds that occur within a word (e.g., the transitional probability between adjacent syllables is lower across a word boundary than within a word: Saffran et al. 1996a). Astheimer and Sanders (2009, 2011) argue that this selective attention is beneficial for speech perception. Perceiving speech sounds requires the processing of rapidly changing acoustic information, and attending to the position in which the occurrences of sounds are unpredictable helps listeners to process the acoustic information in detail and to resolve uncertainty about the sounds. However, when the occurrences of sounds are predictable, listeners may not need to process the acoustic information in detail because some parts of the information may be predictable and redundant. In this way, listeners can minimize the amount of resources needed to process predictable sounds.

In a similar vein, Hall (2009) argues that listeners become less sensitive to allophonic variation because allophones are more predictable and less informative, and listeners allocate less resources to process allophones (i.e., they become less

attentive to the acoustic properties of allophones). According to Hall,

> "for pairs of sounds (X and Y) for which there is a low degree of uncertainty (in context C), it is less crucial for mature language users to pay particular attention to acoustic and articulatory cues used to differentiate X and Y in C because these cues are redundant with the information provided by C" (Hall, 2009, p. 117).

1.3 Acquisition of phonological relationships

1.3.1 Early acquisition of phonemic contrasts

A great deal of research has investigated the early acquisition of phonemes by infants. Infants show signs of the *categorical perception* of speech sounds from an early age. Categorical perception is an phenomenon in which observers' sensitivity to physical differences between stimuli is largely determined by the way the stimuli are categorized; they are sensitive to the differences between stimuli that are classified into two different categories but are less sensitive to the same degree of difference between stimuli that are classified into a single category (see Goldstone and Hendrickson, 2010; Harnad, 2005, for recent reviews). In speech perception, listeners are sensitive to acoustic differences between sounds that are classified into two different phonemes but are less sensitive to the same degree of acoustic difference between sounds that are classified into a single phoneme (Fry et al., 1962; Liberman et al., 1957, 1961, 1967).[4]

Eimas et al. (1971) demonstrated that 1- and 4-month-old English-learning infants perceived a bilabial stop voicing continuum (i.e., a Voice Onset Time (VOT) continuum) categorically; they discriminated a pair of stimuli with 20 ms difference in VOT from a certain region of the continuum (+20 ms vs. +40 ms), but not pairs of stimuli with the same amount of difference in VOT from other regions of

[4]Note, however, that categorization is not the sole factor that determines listeners' sensitivity. As mentioned in Section 1.2 above, task variables also determine listeners' sensitivity. For example, listeners show better sensitivity to within-phoneme acoustic variation in a task that allows them to compare stimuli at the level of auditory processing (e.g., an AX discrimination task with a short interstimulus interval (ISI)). Schouten et al. (2003) have argued that categorical perception is something that emerges from the interaction of listeners' linguistic knowledge and task variables.

the continuum (-80 ms vs. -60 ms and 0 ms vs. $+20$ ms). Lasky et al. (1975) demonstrated that 4- to 6.5-month-old Spanish-learning infants also perceived a bilabial stop continuum categorically; they discriminated the stimuli with -60 and -20 ms VOT and the stimuli with $+20$ and $+60$ ms VOT, but not the stimuli with -20 and $+20$ ms VOT. Since the boundary between Spanish voicing categories falls between -20 and $+20$ ms, the results of Lasky et al. (1975) suggest that language experience has little effect on determining the 4 to 6.5-month-old Spanish-learning infants' sensitivity to the VOT differences, and yet these infants show different levels of sensitivity to the same degree of difference in VOT from different regions of the continuum. Some studies have suggested that what seems to be category effect in the perception of a VOT continuum actually arises from natural sensitivities of the auditory system, not from listeners' knowledge about categories. For example, Pastore et al. (1988) demonstrated that the auditory system has different levels of sensitivities to different timing relationships between acoustic events.

Similarly, Eimas (1974) demonstrated that 3-month-old English-learning infants perceived a stop place continuum (i.e., a formant transitions continuum) categorically; they discriminated a pair of stimuli with differences in F2 and F3 transitions that straddled the boundary between alveolar [dæ] and velar [gæ], but did not discriminate a pair of stimuli with the same amount of differences in F2 and F3 transitions that did not straddle the boundary. These studies suggest that infants are born with the ability to perceive speech sounds categorically.[5]

Infants initially show good sensitivity to acoustic differences between a wide

[5]Some studies have suggested that categorical perception is a property of the auditory system. For example, Cutting and Rosner (1974) demonstrated that adult listeners perceived non-speech audio stimuli categorically. In their study, they used a continuum of sawtooth waveforms that differed in the rise time from 0 ms to 80 ms. The stimuli with a rapid onset sounded like the plucking of a string instrument whereas the stimuli with a slow onset sounded like the bowing of the same string instrument. The results of the study showed that adult English speakers perceived the continuum categorically; their categorization of the stimuli as either "plucking" or "bowing" showed a non-linear categorization curve, and their performance in a discrimination task was determined by the categorization. Jusczyk et al. (1977) replicated Cutting and Rosner (1974)'s results with 2-month-old infants, showing that 2-month-old infants perceived the continuum categorically as well. Other studies have even suggested that categorical perception is not a unique property of the human auditory system, showing categorical perception of human speech sounds by non-human animals like rhesus monkeys (Waters and Wilson, 1976), Japanese macaques (Kuhl and Padden, 1983), and chinchillas (Kuhl and Miller, 1978).

range of speech sounds, including ones that are not different phonemes in the language of their environment (e.g., Streeter, 1976; Trehub, 1976; Werker and Tees, 1983, 1984). Trehub (1976) demonstrated that 1- and 4-month-old English-learning infants can discriminate the alveolar fricative [z] and retroflex fricative [ʐ] from Czech. Werker and Tees (1983, 1984) demonstrated that 6- to 8-month-old English-learning infants can discriminate various non-native phonemes, such as the dental stop [d̪] and retroflex stop [ɖ] from Hindi and the velar ejective stop [k'] and uvular ejective stop [q'] from Nłeʔkepmxcín (Thompson River Salish).

Infants' sensitivity, however, changes over the first year of life according to the inventory of phonemes in their target language (*perceptual reorganization*: e.g., Werker and Tees 1984). They maintain or gain good sensitivity to acoustic differences between sounds that are different phonemes in their target language but become less sensitive to differences between sounds that are not. For example, Werker and Tees (1984) demonstrated that 6- to 8-month-old English-learning infants can discriminate Hindi dental [d̪] and retroflex [ɖ], but 10- to 12-month-old infants cannot. Similarly, Tsushima et al. (1994) demonstrated that 6- to 8-month-old Japanese-learning infants can discriminate English [l] and [ɹ], but 10- to 12-month-old infants cannot. These findings suggest that infants learn the sounds that are phonemes (specifically consonants) in their target language as separate categories between 6-to-8 months and 10-to-12 months of age. Once they learn the categories, their sensitivity to acoustic differences between speech sounds becomes more dependent on their knowledge about the categories.

Compared to consonants, infants seem to learn vowels a little earlier (e.g., Polka and Werker, 1994; Polka and Bohn, 1996). A series of studies by Kuhl and colleagues demonstrated that infants start showing *perceptual magnet effects* in the perception of vowels by 6 months of age (Grieser and Kuhl, 1989; Kuhl, 1991; Kuhl et al., 1992). Perceptual magnet effects are a kind of prototype effect in speech perception and are considered to be indicative of listeners' knowledge about the internal structure of sound categories. When listeners compare two physically different sounds from the same category, the discrimination of these sounds is harder when one of the sounds is a category prototype and the other is a non-prototype than when both of them are non-prototypes. This is because prototype stimuli work as a perceptual magnet and perceptually assimilate non-prototypical

stimuli (a phenomenon referred to as *perceptual warping*: Kuhl 1991; Kuhl and Iverson 1995), and this reduces the perceived distance between prototypical stimuli and non-prototypical stimuli.

Kuhl et al. (1992) tested 6-month-old English-learning infants and 6-month-old Swedish-learning infants on their perception of English and Swedish vowels. While both English and Swedish have the high front unrounded vowel [i] as a phoneme, Swedish also has the high front rounded vowel [y] as a phoneme. The results of the study demonstrated that while English-learning infants showed a stronger magnet effect with the English [i] prototype, Swedish-learning infants showed a stronger magnet effect with the Swedish [y] prototype. In other words, while English-learning infants performed worse in discriminating the English [i] prototype from its within-category non-prototypical variants than in discriminating the Swedish [y] prototype from its within-category non-prototypical variants, Swedish-learning infants performed worse in discriminating the Swedish [y] prototype from its within-category non-prototypical variants than in discriminating the English [i] prototype from its within-category non-prototypical variants.[6] These findings suggest that infants learn the categories vocalic sounds that are used as phonemes in their target language by 6 months of age.

Perceptual reorganization may help infants to acquire their target language. Specifically, attending to the sounds that are used to make lexical contrasts (i.e., phonemes) may facilitate the learning of words. For example, Kuhl et al. (2008) demonstrated that infants' sensitivity to acoustic differences between non-native phonemes at 7.5 months of age predicts the rate of their vocabulary growth

[6]Kuhl et al. (1992) tested the prototypicality of the vowels with adult speakers. Adult English speakers and Swedish speakers were presented with the English [i] prototype and the Swedish [y] prototype and were asked to decide whether the vowel is used in their language, to decide which category the vowel belongs to, and how well the vowel represents the category using a scale from "1" (poor) to "7" (good). Adult English speakers categorized the [i] prototype as a good exemplar of the English /i/ (with an average rating of 5.4), but they responded that the [y] prototype is not used in their language. Adult Swedish speakers categorized the [y] prototype as a good exemplar of the Swedish /y/ (with an average rating of 4.7). They responded that the [i] prototype is used in their language but is ambiguous with regard to the category; they categorized the [i] prototype as the Swedish /e/ with an average rating of 2.6 or the Swedish /i/ with an average rating of 1.8 (Kuhl et al., 1992, footnote 6). The finding that Swedish speakers did not categorize the English [i] prototype as a good exemplar of the Swedish /i/ suggests that the acoustic properties of the high front unrounded vowel /i/ are quite different between these two languages.

in the next two years. Specifically, those who were more sensitive to differences between non-native phonemes at 7.5 months acquired fewer words by 24 months.

Perceptual reorganization, however, is not the acquisition of phonemes per se. Phonemes are the categories of sounds that are used to make lexical contrasts. Therefore, the acquisition of phonemes involves not only the learning of categories but also the learning of their contrastive function. Studies have suggested that, for infants at the stage of early vocabulary development, being sensitive to acoustic differences between sounds that are phonemes in their target language does not necessarily mean that they understand the contrastive function of the sounds.

For example, Stager and Werker (1997) reported that 14-month-old English-learning infants reliably discriminated two English phonemes, the bilabial stop /b/ and alveolar stop /d/, but they failed to learn a pair of novel words that differed from each other in one segment, where the differing segments were the phonemes /b/ and /d/ (a *minimal pair*), /bɪ/ and /dɪ/, in an audio-visual word learning task. Similarly, Thiessen (2007) reported that 15-month-old English learning infants reliably discriminated two English phonemes, the voiced stop /d/ and voiceless stop /t/, but they failed to learn a minimal pair, /dɔ/ and /tɔ/, in an audio-visual word learning task. These studies suggest that despite the ability to discriminate segments that are distinct phonemes in their target language, the infants of this age group have not yet acquired the contrastive function of the segments that are phonemes in their target language.[7]

[7]Interestingly, Thiessen (2007, 2011b) found that 15-month-old English-learning infants successfully learned the lexical contrast between /dɔ/ and /tɔ/ when they were exposed to non-minimal pairs, the words that differed from each other in more than one segment (e.g., /dɔgo/ and /tɔbo/), as well as minimal pairs. Feldman and colleagues have argued that for infants of this age group the presentation of a phonemic contrast in a minimal pair or in overlapping lexical contexts makes the perception of the contrast harder while the presentation of a phonemic contrast in a non-minimal pair or in non-overlapping lexical contexts makes the perception of the contrast easier because the non-overlapping lexical contexts serve as a cue for the phonemic contrast. According to this view, the infants in Thiessen (2007, 2011b) learned the lexical contrast between /dɔ/ and /tɔ/ when they were exposed to non-minimal pairs as well because non-overlapping lexical contexts in those non-minimal pairs helped them to establish a more robust phonetic contrast between /t/ and /d/ (Feldman et al., 2011, 2013a,b). Alternatively, Rost and McMurray (2009, 2010) have argued that the infants in previous studies (e.g., Stager and Werker, 1997; Thiessen, 2007) failed to learn the novel minimal pairs because they failed to differentiate the critical phonemes due to the lack of within-category variability in the input (e.g., the infants in Thiessen's study were trained with a single exemplar of each category). Rost and McMurray demonstrated that 14-month-old English-learning infants learned a novel minimal pair, /buk/ and /puk/, in an audio-visual word learning task with input that showed a large

1.3.2 Early acquisition of allophonic variation

Compared to the large number of studies on the acquisition of phonemic contrasts by infants, there are relatively few studies on the acquisition of allophonic variation by infants. The existing studies have demonstrated that infants become less sensitive to acoustic differences between native allophones at the same time as they learn native phonemes (Dietrich et al., 2007; Hohne and Jusczyk, 1994; Seidl et al., 2009). For example, Seidl et al. (2009) compared French-learning infants and English-learning infants on their perception of vowel nasality. The difference between oral and nasal vowels is contrastive in French but allophonic in English (i.e., the nasalized vowels occur when followed by a nasal consonant). In their study, Seidl et al. exposed French-learning infants (11-month-old) and English-learning infants (4-month-old and 11-month-old) to an artificial language in which the type of the coda consonant in CVC syllables was determined by the nasality of the preceding vowel (e.g., the fricatives occurred after a nasal vowel and the stops occurred after a oral vowel). Their prediction was that if infants can discriminate the nasal and oral vowels, they should be able to learn the phonotactic patterns. The results of the study showed that the 4-month-old English-learning infants and the 11-month-old French-learning infants learned the phonotactic patterns but the 11-month-old English-learning infants did not, suggesting that the older English-learning infants had become less sensitive to the acoustic differences between the nasal and oral vowels.[8]

Although infants' sensitivity to acoustic differences between native allophones starts declining towards the end of the first year of life, this does not necessarily mean that they become completely insensitive to allophonic variation. Jusczyk et al. (1999) demonstrated that 10.5-month-old English-learning infants are able to use their knowledge about allophonic variation to segment speech. In English, the voiceless aspirated stop [t^h] and voiceless unreleased stop [$t^\neg$] are allophones of /t/. The former occurs in syllable onset position as in the first /t/ of the word *nitrate*

amount of within-category variability (e.g., multiple tokens by multiple talkers: Rost and McMurray 2009 and multiple exemplars by a single talker: Rost and McMurray 2010).

[8]It is also possible that the older English-learning infants had already learned the contextually-conditioned distribution of the nasal and oral vowels in English and their L1 phonological knowledge interfered with the learning of the artificial phonotactic patterns that involved an illegal configuration (i.e., the nasal vowels occurring before an non-nasal consonant) (e.g., Finn and Hudson Kam, 2008).

[ˈnʌɪ.tʰɹeɪtˀ], and the latter occurs in syllable coda position as in the first /t/ of the phrase *night rate* [ˈnʌɪtˀ.ɹeɪtˀ]. Jusczyk et al. reported that in a speech segmentation task 10.5-month-old infants reliably placed a syllable boundary before /t/ when the /t/ was aspirated but not when the /t/ was unreleased.

In sum, infants learn to categorize speech sounds according to the phoneme inventory of their target language by the end of the first year of life. They maintain or gain sensitivity to acoustic differences between segments that are distinct phonemes in their target language. At the same time, infants become less sensitive to acoustic differences between allophones in their target language.

1.3.3 Late acquisition of non-native phonemic contrasts

As a consequence of perceptual reorganization and subsequent phonological acquisition, the perception of sounds from other languages that are not used as separate phonemes in the native language becomes significantly more difficult for adults (e.g., Best, 1995; Flege, 1995).

According to the *Perceptual Assimilation Model* (PAM), the degree of difficulty in the perception of the difference between non-native phonemes is determined by how the non-native phonemes are mapped onto native phoneme categories (Best et al., 1988, 2001; Best, 1995). Within the framework of the PAM, cross-language speech perception can be classified into five different types. In the first type, two non-native phonemes are mapped onto a single native phoneme, and the difference is hard to perceive (*single-category assimilation*, e.g., Hindi dental [d̪] and retroflex [ɖ] for English speakers: Tees and Werker 1984). In the second type, two non-native phonemes are mapped onto two different native phonemes, and the difference is easy to perceive (*two-category assimilation*, e.g., Zulu voiced lateral fricative [ɮ] and voiceless lateral fricative [ɬ] for English speakers: Best et al. 2001). In the third type, one of two non-native phonemes is perceived as a good exemplar of a native phoneme while the other is perceived as a poor exemplar of the same native phoneme, and the difference is relatively easy to perceive (*category goodness assimilation*, e.g., Zulu aspirated stop [kʰ] and ejective stop [kʼ] for English speakers: Best et al. 2001). In the fourth type, two non-native phonemes are mapped onto somewhere in between native phonemes, and the ease of percep-

tion of the difference depends on the proximity to native phonemes (*uncategorized*, e.g., English [l] and [ɹ] for Japanese speakers: Guion et al. 2000). In the fifth type, non-native phonemes are not mapped onto any native phonemes, and the difference is easy to perceive (*non-assimilation*, e.g., Zulu apical click [ǀ] and palatal click [ǃ] for English speakers: Best et al. 1988).

Despite the difficulty that adults experience in the perception of some non-native phonemes, studies have demonstrated that perceptual training can significantly improve their perception (Hirata et al., 2007; Iverson and Evans, 2007, 2009; Iverson et al., 2012; Kingston, 2003; Logan et al., 1991; Lively et al., 1993). For example, adult Japanese speakers have great difficulty in perceiving English liquids [l] and [ɹ] (Gillette, 1980; Goto, 1971; Miyawaki et al., 1975; Mochizuki, 1981). This is largely because of the absence of these English sounds in Japanese. Japanese has an alveolar tap [ɾ] that is similar to English [l] and [ɹ]. However, the relationship between Japanese [ɾ] and English [l] and [ɹ] is complicated. Takagi (1993) reported that Japanese speakers perceive Japanese [ɾ] as being more similar to English [l] than to English [ɹ]. Sekiyama and Tohkura (1993) reported that Japanese speakers perceive English [ɹ] as Japanese [ɾ], [ɯɻ], and [g]. Finally, Yamada and Tohkura (1992) reported that Japanese speakers perceive stimuli that are intermediate between English [l] and [ɹ] as [ɯɻ]. The other factor that significantly affects the perception of the difference between English [l] and [ɹ] by Japanese speakers is perceptual cue weighting. Iverson et al. (2003) reported that while English speakers primarily attend to F3 to perceive the difference, Japanese speakers attend to F2. Since F3 is the dominant acoustic cue that differentiates English [l] from [ɹ], attending to the wrong cue makes perception of the difference difficult for Japanese speakers (Lotto et al., 2004).

Studies have demonstrated that Japanese speakers' perception of English [l] and [ɹ] can be significantly improved through perceptual training (Logan et al., 1991; Lively et al., 1993, 1994). In Logan et al. (1991), Japanese speakers were trained to identify [l] and [ɹ] produced by multiple talkers in various phonetic contexts. After three weeks of training, participants showed a significant improvement in the correct identification of [l] and [ɹ] in both trained and novel words. Lively et al. (1993) examined the effect of talker variability in training stimuli and demonstrated that learners who were trained with stimuli produced by multiple talkers

generalized their learning to test stimuli produced by a novel talker, but those who were trained with stimuli produced by a single talker did not. Finally, Lively et al. (1994) demonstrated that the training had a long-lasting effect; Japanese speakers maintained the effect of the training over three to six months after the last training session. Studies have also demonstrated that Japanese speakers can shift their attention to F3 in perceiving [l] and [ɹ] through training (e.g., Ingvalson et al., 2012; Lim and Holt, 2011).

1.3.4 Late acquisition of non-native allophonic variation

The learning of non-native allophonic variation by adults has been much less studied, but there is some evidence showing that adults can learn allophonic variation in a second language (L2). Darcy et al. (2007, 2009) tested the recognition of French words by L2 French learners (native speakers of English). In French, stop consonants assimilate to the voicing of a following obstruent across word boundaries. For example, the word *bott* is realized as [bot] in isolation but as [bod] before a voiced obstruent. In other words, stop consonants have two variants, voiceless and voiced, depending on whether or not a voiced obstruent follows. As a consequence of this contextual variation, words like *bott* have two different pronunciations, and this complicates the recognition of the words. Native French speakers, however, know the contextual variation and have no problem in recognizing such words. For example, when they hear [bod] before a voiced obstruent, they can infer that the voicing of the final stop is the result of assimilation to the following voiced obstruent and the underlying form is voiceless (a process referred to as *compensation for assimilation*). In this way, they can recognize the word *bott* from [bod] occurring before a voiced obstruent. In Darcy et al. (2007, 2009), both beginning and advanced French learners correctly recognized assimilated words, but advanced learners more consistently than beginning learners. These findings suggest that adults can learn L2 allophonic variation, but that this depends on the amount of experience with the target L2.

1.4 Learning mechanisms

1.4.1 Distributional learning of sound categories

As discussed in Section 1.3.1, infants learn how to categorize speech sounds according to the inventory of phonemes in their target language by the end of the first year of life. Although it is unlikely that they have acquired phonemes with their functional values at this point, they seem to have acquired phonetic categories of sounds, such as segments. Researchers have argued that infants can learn phonetic categories from statistical information in input, specifically, the frequency distribution of sounds in acoustic space (*distributional learning of sound categories*: Kuhl, 1994; Lacerda, 1995, 1998; Maye, 2000; Pierrehumbert, 2003).

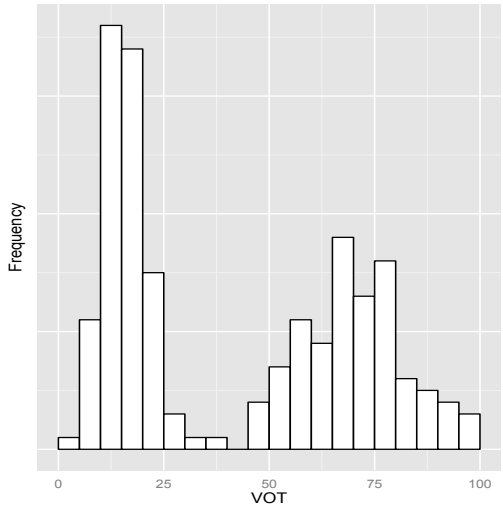
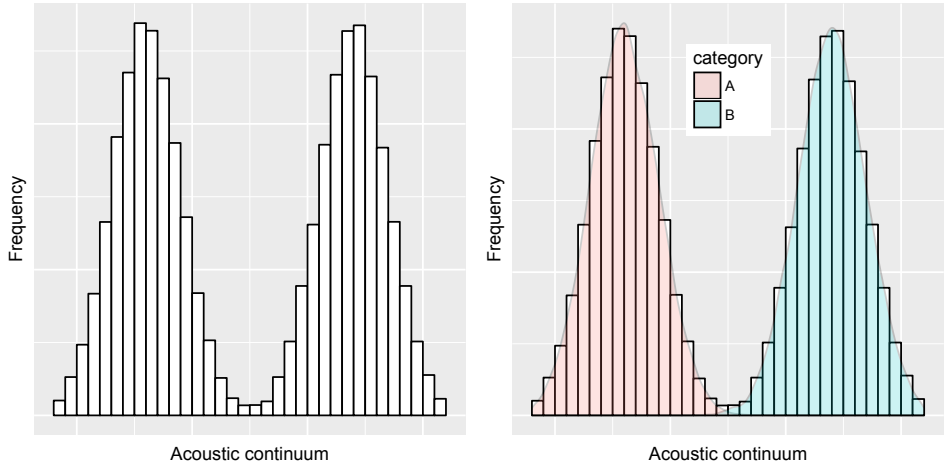


Figure 1.2: Frequency distribution of VOT values in English (200 samples generated on the basis of the data in Allen and Miller, 1999)

Naturally produced speech comes with an infinite amount of variability. However, the variability is not completely random. The sounds produced in a language are systematically distributed in acoustic space according to the kinds of sounds used in the language (Abramson and Lisker, 1964; Hillenbrand et al., 1995; Peterson and Barney, 1952). Figure 1.2, for instance, shows the frequency distribution of VOT values from word-initial stop consonants in English. In this figure, two sep-



(a) Distribution of sounds **(b)** Learned categories

Figure 1.3: Inference of categories from distributional shape

arate frequency peaks are clearly visible—one at around 10-15 ms and the other at around 65-75 ms—and these two frequency peaks represent two stop voicing categories in English, the voiceless unaspirated stops (or *short-lag*) and the voiceless aspirated stops (or *long-lag*).

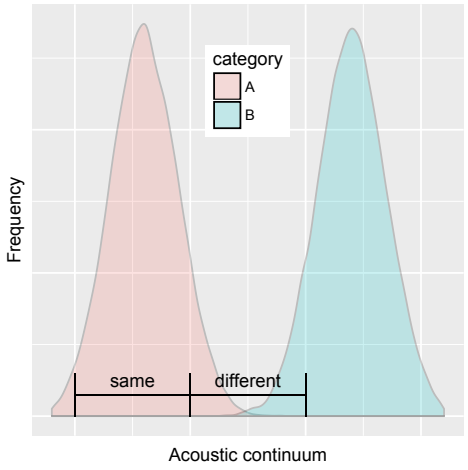


Figure 1.4: Category effect

The distributional learning hypothesis assumes that human learners are sensitive to this kind of distributional information in input. They can keep track of the frequencies of sounds that occur in the input and learn the frequency distributions of the sounds in acoustic space. Once they learn the distributions, they can use the frequency peaks as categories or form categories as abstract representations based on the frequency peaks, and start using their knowledge about the categories to classify the sounds. For instance, Figure 1.3a shows the frequency distribution of sounds along a hypothetical acoustic continuum. The distribution has two frequency peaks (bimodal distribution). The distributional learning hypothesis predicts that when learners are exposed to this input, they should be able to learn two categories based on the number of frequency peaks in the input. Figure 1.3b shows the two categories that should be learned from this input.[9]

In the distributional learning hypothesis, it is assumed that the learning of categories has a signifiant impact on speech perception. Specifically, it affects learners' sensitivity to acoustic differences between sounds. Learners become more sensitive to acoustic differences between sounds that are classified into separate categories and/or less sensitive to the same degree of acoustic difference between sounds that are classified into the same category. Figure 1.4 shows that once learners acquire knowledge about the categories, they start treating two sounds that are classified into two categories as different sounds and two sounds that are classified into the same category as the same sound. As a result, they become more sensitive to the acoustic differences that straddle the boundary between two categories and/or less sensitive to the same degree of acoustic difference that happen within the same category.

Experimental studies have demonstrated that infants are sensitive to this kind of distributional information (Cristia et al. 2011a, Maye et al. 2002, 2008, Yoshida et al. 2010, cf. Pons et al. 2006). For example, Maye et al. (2002) exposed two groups of English-learning infants (6-month-olds and 8-month-olds) to syllables taken from an 8 step-continuum between the prevoiced [da] and voiceless unaspirated [ta] in which VOT was systematically manipulated. For one group,

[9]This inductive learning process has been implemented in various computational models (De Boer and Kuhl, 2003; Guenther and Gjaja, 1996; Lin, 2005; McMurray et al., 2009; Vallabha et al., 2007).

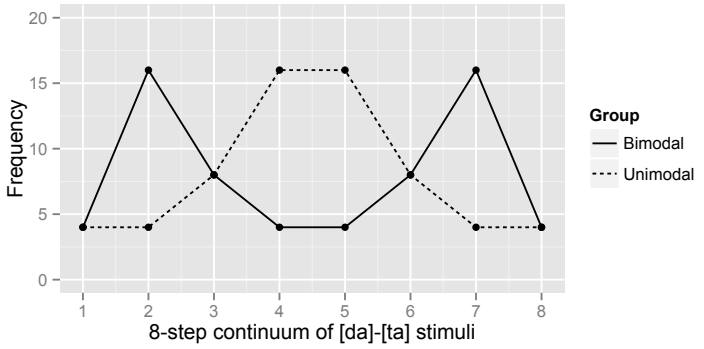


Figure 1.5: Bimodal vs. Unimodal distribution of [da]-[ta] stimuli (Based on Figure 1 in Maye et al., 2002)

the frequency distribution of the eight syllables had two separate peaks (*Bimodal group*). For the other group, the frequency distribution had only one peak (*Unimodal group*) (Figure 1.5). After exposure, infants in the bimodal group discriminated the test stimuli taken from the end points of the continuum (i.e., the canonical tokens of [da] and [ta]), but infants in the unimodal group did not. These results suggest that while infants in the bimodal group learned to classify the syllables into two separate categories, infants in the unimodal group learned to classify the syllables into a single category.

Experimental studies have demonstrated that adults are also sensitive to this kind of distributional information. These studies have tested the distributional learning of stop voicing categories (Maye, 2000; Maye and Gerken, 2000, 2001; Hayes-Harb, 2007), vowel categories (Gulian et al., 2007; Goudbeek et al., 2008), fricative place categories (Pajak, 2012; Pajak and Levy, 2012), consonantal duration categories (Pajak, 2012; Pajak and Levy, 2012), and tonal categories (Ong et al., 2015). Maye (2000), for example, exposed two groups of adult English speakers to syllables taken from 8-step stop voicing continua (e.g., between the prevoiced [da] and voiceless unaspirated [ta]) in which VOT, F1 transition, and F2 transition were systematically manipulated.[10] For one group, the frequency distri-

[10]Note that acoustic space is multi-dimensional. A phonetic contrast between two categories can be made by multiple acoustic cues, and the distributional learning of sound categories involves the integration of these acoustic cues (Toscano and McMurray, 2010). Most of the previous experimental

bution of the stop consonants showed a bimodal shape (Bimodal group). For the other group, the frequency distribution of the stop consonants showed a unimodal shape (Unimodal group). After exposure, participants in the bimodal group showed significantly better sensitivity to acoustic differences between the prevoiced and voiceless unaspirated stops, compared to participants in the unimodal group. These results suggest that while participants in the bimodal group learned to classify the stop consonants into two separate categories, participants in the unimodal group learned to classify the stop consonants into a single category.

These studies suggest that distributional learning is a learning mechanism used by both infants and adults to learn sound categories. However, this does not mean that distributional learning is equally effective for learners from all age groups. As learners acquire more knowledge about their L1 phonetics and phonology, it significantly affects the effectiveness of the learning. For example, Yoshida et al. (2010) tested the distributional learning of the prevoiced [da] and voiceless unaspirated [ta] by 10-month-old English-learning infants. After exposure, 10-month-old infants showed different levels of sensitivity to acoustic differences between [da] and [ta] depending on the distributional information in their input, but compared to 6- and 8-month-old infants, 10-month-old infants required a longer exposure to show the same amount of learning. This is probably because 10-month-old infants have already acquired the English voicing categories and their L1 knowledge interferes with the learning of novel voicing categories.

Studies have also demonstrated the impact of adults' L1 phonological knowledge on the distributional learning of novel sound categories (Pajak, 2012; Pajak and Levy, 2012). Pajak and Levy (2012) tested the distributional learning of consonantal duration categories (short consonants vs. long consonants) by adult speakers of Korean and Mandarin. While the difference in segmental duration is phonemic for vowels in Korean, but not for consonants, the difference in segmental duration is not phonemic at all in Mandarin. Therefore, Pajak and Levy expected that Korean speakers would be more sensitive to the distribution of consonants along a segmental duration continuum because they already know that the difference in

studies on the distributional learning of sound categories have focused on the role of a single acoustic cue, but studies have suggested that different acoustic cues have different weights in terms of their contribution to the learning of sound categories (e.g., Cristia et al., 2011a).

segmental duration is phonemic for vowels in their native language (i.e., their L1 phonological knowledge biases their perception such that it predisposed them to attend to segmental duration). The results of the study showed that Korean speakers were indeed more sensitive than Mandarin speakers to the difference between the bimodal and unimodal distributions of consonants along a segmental duration continuum. These results suggest that the effectiveness of distributional learning by adults is affected by their L1 phonological knowledge.

Another difference between infants and adults is found in the robustness of the learning, specifically the ability to generalize what they learn from input to novel stimuli. Maye and colleagues exposed infants to input that supported the learning of novel voicing categories for stop consonants from one place of articulation (e.g., the alveolar prevoiced [d] and alveolar voiceless unaspirated [t]). After exposure, the infants not only learned the voicing categories they were exposed to, they also generalized the categories to an unfamilar place of articulation (e.g., the velar prevoiced [g] and velar voiceless unaspirated [k]) (Maye and Weiss, 2003; Maye et al., 2008). Maye and colleagues also tested adults' ability to generalize the same voicing categories from one place of articulation to another place of articulation, but the adults did not generalize (Maye, 2000; Maye and Gerken, 2001). Later, Pajak (2012) demonstrated that adults generalized what they learned from input to novel stimuli. In her study, adults learned novel segmental duration categories for consonants from one manner class (e.g., short sonorant [l] and long sonorant [l:]) from the input and generalized the newly learned categories to consonants from a different manner class (e.g., short fricative [s] and long fricative [s:]).

From these limited data, it seems that adults have less robust abilities to generalize what they learn from input to novel stimuli. One possible explanation is a difference between infants and adults in their cognitive capacity; infants have a smaller cognitive capacity, and this restriction forces them to learn new categories in a more feature-based manner (e.g., prevoiced vs. voiceless unaspirated), while adults have a larger cognitive capacity, and this enables them to learn new categories in a more item-based manner (e.g., alveolar prevoiced stop vs. alveolar voiceless unaspirated stop) (e.g., Newport, 1988, 1990). Whatever the source of this difference in generalization is, it remains true that both infants and adults are able to learn the categories of sounds from the frequency distribution of the sounds

in input.

1.4.2 Distributional learning of allophony

The limitation of the distributional learning hypothesis is that it explains the learning of phonetic categories such as segments but does not explain the acquisition of sound systems. In sound systems, some segments are used as allophones or context-dependent variants of a single phoneme. Studies on the development of infants' speech perception suggest that infants are learning the segments used in their target language and the allophonic relationships between some of these segments at the same time (Dietrich et al., 2007; Hohne and Jusczyk, 1994; Seidl et al., 2009). Researchers have argued that allophony can be learned from statistical information in input (Peperkamp et al., 2006a). The kind of information infants would have to track to learn allophony is quite different from that required to learn segments. Because some allophones occur in mutually exclusive contexts, their occurrences are predictable from the contexts. Therefore, by learning the relative predictabilities of segments in particular environments, learners should be able to figure out whether the segments are allophones.

Note, however, that in natural languages, segments that are in complementary distribution are not always allophones. For example, in French, the bilabial approximant [ɥ] always occurs as the last consonant of an initial consonant cluster (e.g., *pluie* [plɥi]), while the mid-low front rounded vowel [œ] always occurs in a closed syllable (e.g., *peur* [pœʁ]). This means that [ɥ] and [œ] are in complementary distribution; the former occurs before a vowel and the latter occurs before a consonant. However, these two segments are not allophones in French (Peperkamp et al., 2006a). Therefore, the learning of allophony from the statistical information in input should be constrained in some way such that not just any segments that are in complementary distribution are learned as allophones.

Peperkamp et al. (2006a) proposed two constraints on the learning of allophones. The first one requires that potential allophones are phonetically similar to each other. The second one requires that the contextual distribution of potential allophones is phonetically natural in a sense that the allophones and their respective contexts are phonetically similar to each other (i.e., the distribution assumes assim-

ilatory patterns). In the above example, [ɥ] and [œ] are not allophones in French because they are phonetically too different to be allophones and/or the occurrences of [ɥ] before a vowel and [œ] before a consonant are not natural in the above sense.

Studies have demonstrated that infants learn phonotactic regularities in their target language at the same time as they learn the categories of speech sounds (Archer and Curtin, 2011; Jusczyk et al., 1993; Jusczyk and Luce, 1994; Nazzi et al., 2009). For example, Jusczyk et al. (1993) demonstrated that 6- and 9-month-old English-learning infants show a sensitivity to the difference between phonotactically legal and illegal forms in English. Jusczyk and Luce (1994) further demonstrated that 6- and 9-month-old English-learning infants show a similar sensitivity to phonotactic probabilities in English; they prefer to listen to nonce words that are phonotactically more probable (e.g., [kæz]) over nonce words that are phonotactically less probable (e.g., [guʃ]). Therefore, it is possible that infants can integrate their knowledge about segments and their phonotactic distribution and figure out whether the occurrences of the segments are predictable in particular environments.

Artificial language learning experiments have demonstrated that infants have a robust ability to learn phonotactic regularities in input with a relatively small amount of exposure (Chambers et al., 2003, 2011; Cristia and Seidl, 2008; Cristia et al., 2011b; Saffran and Thiessen, 2003; Seidl and Buckley, 2005). For example, Saffran and Thiessen (2003) demonstrated that 9-month-old English-learning infants learned novel first-order phonotactic patterns—positional restrictions on the occurrence of certain classes of segments (e.g., voiceless stops in syllable onset position and voiced stops in syllable coda position)—with a fairly small amount of exposure (two minutes).

If infants have such a robust ability to learn phonotactic regularities in input, the question arises as to whether they can use their knowledge about the phonotactic distribution of segments across environments to infer allophonic relationships between the segments. As far as I know, there is only one study that has tested the learning of allophony by infants from the phonotactic distribution of segments in artificial input. White et al. (2008) exposed 8.5- and 12-month-old English-learning infants to input in which a pair of segments (e.g., [b] and [p]) occurred in overlapping contexts ([b]-initial words and [p]-initial words occurring after [na] and [rot]: e.g., *na bevi, na pevi, rot bevi, rot pevi*), but the other pair (e.g., [z] and

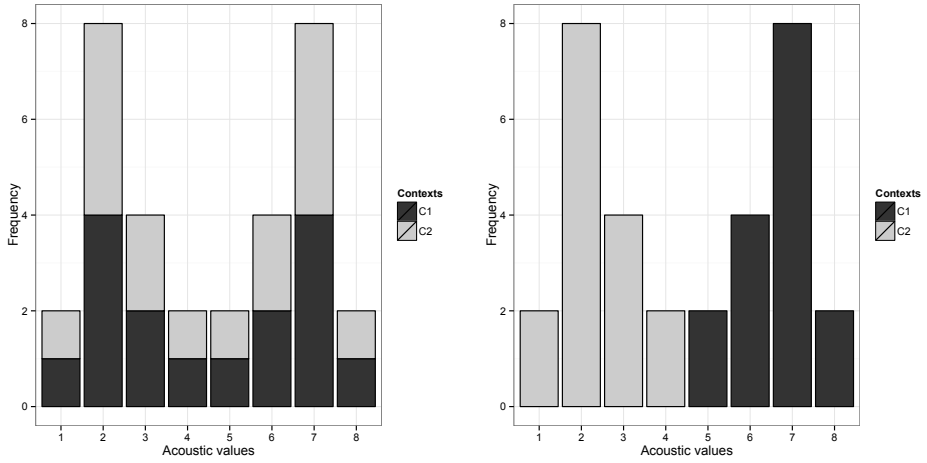
[s]) occurred in mutually exclusive contexts ([z]-initial words occurring after [na] and [s]-initial words occurring after [rot]: e.g., *na zuma, rot suma*). In other words, phonotactic regularities in the input implied that [b] and [p] are contrastive and [z] and [s] are allophonic. After exposure, infants were presented with the alternation of a [b]-initial word and its [p] initial counterpart (e.g., *rot poli, na boli, rot poli, na boli, ...*) and the alternation of a [z]-initial word and its [s]-initial counterpart (e.g., *rot sadu, rot sadu, na zadu, rot sadu,...*). White et al. predicted that if infants had learned [b] and [p] as distinct phonemes and [z] and [s] as allophones, they should hear the [b]-[p] alternation differently from the [z]-[s] alternation. Specifically, they should not hear the [z]-[s] alternation as alternation because they treat [z] and [s] as allophones. After exposure, both 8.5- and 12-month-old infants listened longer to the [z]-[s] alternation. White et al. interpreted the results as a novelty effect; since infants had been exposed to a wide variety of stimuli presented in a random order during exposure, the presentation of alternating allophones (i.e., the repetition of the "same" sound) triggered a novelty effect. The results of the study suggest that infants can learn to treat pairs of sounds differently depending on their phonotactic distribution in input.

Adults also have a robust ability to learn phonotactic regularities (Dell et al., 2000; Onishi et al., 2002; Chambers et al., 2010). For example, Onishi et al. (2002) demonstrated that adult English speakers learned first-order phonotactic patterns—positional restrictions on the occurrence of certain classes of segments (e.g., [b, k, m, t] in syllable onset position and [p, g, n, tʃ] in syllable coda position)—with a fairly small amount of exposure (120-130 items). Onishi et al. (2002) further demonstrated that adult English speakers learned more complex second-order phonotactic patterns, where the occurrence of certain segments in certain positions was conditioned by the types of neighbouring segments (e.g., [b] in syllable onset position and [p] in syllable coda position if the vowel nucleus is [æ], but [p] in onset position and [b] in coda position if the vowel nucleus is [ɪ]).

Peperkamp et al. (2003) tested the learning of allophony by adults from the phonotactic distribution of segments in artificial input. The input used in their study contained two different kinds of distributional information: (1) a bimodal frequency distribution of sounds that implied the categorization of the sounds into two segments, and (2) a phonotactic distribution of the segments that implied ei-

ther a phonemic contrast or allophony between the segments. They exposed two groups of adult French speakers to input in which the frequency distribution implied the categorization of fricative sounds into the voiced uvular fricative [ʁ] and voiceless uvular fricative [χ]. For one group, the phonotactic distribution was not conditioned by context; both [ʁ] and [χ] occurred before voiced and voiceless consonants. This implied a phonemic contrast between the segments. For the other group, the phonotactic distribution was conditioned by context; [ʁ] occurred before voiced consonants and [χ] occurred before voiceless consonants. This implied an allophonic relationship between the segments. All participants were tested on the discrimination of [ʁ] and [χ] before and after exposure.

Peperkamp et al. (2003) predicted that participants in the second group would learn the target segments as allophones and thus become less sensitive to acoustic differences between the segments. The results of the study, however, were not so clear. First, participants in the second group showed significantly better sensitivity to acoustic differences between [ʁ] and [χ] compared to participants in the first group already in the pre-test. This makes the interpretation of any possible learning effects difficult. Second, participants in both groups showed significant improvement in sensitivity after exposure. Those who were in the first group showed a numerically larger improvement, but the difference between the groups was not statistically significant. The interpretation of this possible learning effect, however, is complicated by the fact that [ʁ] and [χ] are allophones in French; the voiced [ʁ] occurs before voiced consonants and voiceless [χ] occurs before voiceless consonants. Therefore, it is not entirely clear whether the difference between the groups was due to the learning of different phonological relationships, or simply the interference of participants' L1 phonological knowledge with the processing of the exposure stimuli. Participants in the second group could have learned [ʁ] and [χ] as allophones in the artificial language, but it is also possible that they processed [ʁ] and [χ] in the exposure stimuli as allophones from the beginning since they occurred in contexts that conformed to the complementary distribution of these segments in French, and no learning happened. To date, there is no convincing evidence that adults can learn allophony from the phonotactic distribution of segments in input.



(a) Overlapping distribution **(b)** Complementary distribution

Figure 1.6: Non-complementary distribution vs. complementary distribution

1.5 Outline of dissertation

In this dissertation, I investigate the mechanisms behind the learning of allophony. Specifically, I test the learning of an allophonic relationship between two novel segments by adults. Adults have the ability to learn phonetic categories or segments from the frequency distribution of sounds in acoustic space. They also have the ability to learn the phonotactic regularities in the input. However, there does not yet exist any convincing evidence demonstrating that adults can learn phonological relationships between segments from the phonotactic distribution of the segments.

Figure 1.6a shows the frequency distribution of eight sounds taken from a hypothetical acoustic continuum. The bimodal shape of the distribution implies the categorization of the eight sounds into two separate segments (i.e., sounds 1 - 4 belong to one segment, and sounds 5 - 8 belong to the other segment). In this figure, all of the eight sounds occur in two different contexts (C1 and C2) with equal frequency. Therefore, the implied segments are also occurring in two different contexts with equal frequency. This means that the segments are occurring in overlapping contexts (i.e, they are in non-complementary distribution), and their occurrences are not predictable from the contexts. When learners are exposed to this kind of input, they should learn that the sounds are categorized into two seg-

ments and that the segments are potentially contrastive.[11]

Figure 1.6b shows the same bimodal distribution of the same eight sounds. In this figure, the first half of the eight sounds consistently occur in one context (i.e., sounds 1 - 4 occurring in context C2), and the second half consistently occur in the other context (i.e., sounds 5 - 8 occurring in context C1). This means that the implied segments are occurring in two mutually exclusive contexts (i.e., they are in complementary distribution), and their occurrences are predictable from the contexts. When learners are exposed to this kind of input, they should learn that the sounds are categorized into two segments but the segments are allophonic, representing a single phoneme category.

The difference between the learning of phonemic contrast and the learning of allophonic variation should be reflected in many aspects of learners' speech perception. In this dissertation, I focus on learners' sensitivity to acoustic differences between target segments because this is the measure that has been used to assess the learning of novel sound categories in most of the previous studies on the distributional learning of sound categories. If learners learn the the target segments as allophones, they should show reduced sensitivity to acoustic differences between the target segments.

In Experiment 1, adult English speakers were exposed to input in which the frequency distribution of novel fricative sounds implied the categorization of the sounds into two segments, the retroflex fricative [ʂ] and alveolopalatal fricative [ɕ], and the phonotactic distribution of the segments implied either a phonemic contrast or allophony between the segments. In one condition, participants were exposed to input in which the occurrences of the segments were not predictable from the contexts (*non-complementary condition*; cf. Figure 1.6a). In another condition, participants were exposed to input in which the occurrences of the segments were predictable from the contexts (*complementary condition*; cf. Figure 1.6b). The results of Experiment 1 suggest that participants in the complementary condition learned to treat the novel fricatives as something like allophones. This supports the hypothesis that adults can learn allophonic relationships between segments from the phonotactic distribution of the segments in input.

[11] Since no semantic information is taken into consideration, it remains unknown whether the segments are learned as phonemes.

Experiment 2 examined whether the learning of allophony is constrained by the phonetic naturalness of the patterns of complementary distribution. Studies on artificial language learning have demonstrated that the learning of phonological patterns can be constrained or biased by the phonetic naturalness of the patterns; phonetically motivated patterns are more learnable than phonetically unmotivated ones (Carpenter, 2010; Schane et al., 1975; Wilson, 2003, 2006). Experiment 2 tested whether the phonetic naturalness of the complementary distribution of the retroflex [ʂ] and alveolopalatal [ɕ] in input affects the learning of these fricatives as allophones. The results of Experiment 2, alongside the results of Experiment 1, indicate that adults can learn allophonic relationships between two segments only when the patterns of complementary distribution are phonetically natural.

In order to explain how phonetic naturalness affects the learning of allophony, I explore the role of perceptual biases and propose a hypothesis about the mechanisms behind the acquisition of reduced sensitivity to acoustic differences between allophones (the *context effects hypothesis*). The hypothesis is that the learning of allophony involves the context-dependent perception of sounds in the input. When listeners hear sounds presented in different contexts, they perceive the sounds differently. Specifically, listeners perceive the instances of two different segments as being more similar to each other when they hear the sounds in phonetically natural complementary contexts than in phonetically unnatural complementary contexts. Therefore, when learners are exposed to input in which the instances of two target segments are occurring in phonetically natural complementary contexts, the context-dependent perception of the sounds affects the aggregate distribution of the sounds in auditory space such that frequency peaks are closer to each other than they actually are and the boundary between the categories is less clear. The learning of such an aggregate distribution leads to the learning of less distinct categories, or it may even lead to the learning of a single category. Experiment 3 tested whether adult learners' perception of the retroflex [ʂ] and alveolopalatal [ɕ] is affected by context in the way I assume. The results showed that learners' perception became significantly more dependent on context after exposure; learners perceived the instances of these two segments as being more similar to each other when the sounds were presented in phonetically natural complementary contexts than in phonetically unnatural complementary contexts.

In Chapter 2, I present the details of Experiment 1. In Chapter 3, I present the details of Experiment 2. The presentation of the results of Experiment 2 is followed by a discussion where I propose the the context effects hypothesis about the learning of allophonic relationships. In Chapter 4, I present the details of Experiment 3. Finally, in Chapter 5, I will discuss the results of the experiments in a larger context.

Chapter 2

Experiment 1: Distributional learning of allophony

2.1 Introduction

Experiment 1 tested whether adults can learn an allophonic relationship between two novel segments. The experiment was designed as a modified version of previous experiments on the distributional learning of sound categories by adults (e.g., Maye, 2000; Maye and Gerken, 2000; Peperkamp et al., 2003). In this experiment, adult English speakers were exposed to input in which the frequency distribution of novel sounds showed a bimodal shape, implying that the sounds are classified into two segments. While the shape of the frequency distribution was kept the same in the input for all participants, the phonotactic distribution of the segments differed between two experimental conditions. In the first condition, the segments occurred in overlapping contexts and the occurrences of the segments were not predictable from the contexts (*non-complementary condition*). In the second condition, the segments occurred in mutually exclusive contexts and the occurrences of the segments were predictable from the contexts (*complementary condition*). I predicted that participants in the non-complementary condition would learn to treat the target segments as something like distinct phonemes and maintain or gain sensitivity to acoustic differences between the segments. By contrast, participants in the complementary condition would learn to treat the target segments as some-

thing like allophones and become less sensitive to acoustic differences between the segments.

2.2 Target segments

The target segments used in this experiment were two post-alveolar voiceless fricatives from Mandarin, the retroflex [ʂ] and alveolopalatal [ɕ]. While English has only one class of post-alveolar fricatives, the palato-alveolar fricatives ([ʃ], [ʒ]), Mandarin has two, and the phonetic contrast between these two classes of fricatives in Mandarin is known to be difficult for English speakers to acquire (e.g., Chao, 1948). In this section, some background information about these Mandarin sounds will be provided.

2.2.1 Post-alveolar fricatives in Mandarin

Mandarin has two classes of post-alveolar sibilants. The first class is often called *retroflex* in the literature ([ʂ, ʈʂ, ʈʂʰ]) (Chao, 1948; Duanmu, 2007). However, articulatory studies have demonstrated that these sounds are not really retroflex (Hu, 2008; Ladefoged and Wu, 1984; Lee, 2008; Lee-Kim, 2014; Noguchi et al., 2015a; Proctor et al., 2012; Toda and Honda, 2003). For example, using X-ray images and palatograms, Ladefoged and Wu (1984) demonstrated that the articulation of so-called retroflex sibilants in Mandarin does not involve the retroflexion of the tongue tip as is usually observed in the articulation of truly retroflex consonants in other languages. Rather, their articulation involves the formation of a short and slack constriction channel in the post-alveolar region using the upper part of the tongue front and the formation of a large front cavity. Similar observations have been made in later studies using other imaging techniques such as MRI (Proctor et al., 2012; Toda and Honda, 2003) and ultrasound (Lee-Kim, 2014; Noguchi et al., 2015a). Due to the absence of retroflexion, researchers have used different labels for this class of Mandarin post-alveolar sibilants, such as *flat post-alveolar* (Ladefoged and Maddieson, 1996; Toda and Honda, 2003) and *apical post-alveolar* (Lee, 2008; Proctor et al., 2012). Here, I will use the traditional label *retroflex* for convenience.

The second class of post-alveolar sibilants is often called *palatal* ([ɕ, tɕ, tɕʰ])

(Chao, 1948; Duanmu, 2007). The articulation of these sibilants in Mandarin involves the formation of a long and narrow constriction channel over the post-alveolar and palatal regions using the tongue blade and the tongue body. The formation of the palatal constriction is achieved by the forward and upward movements of the tongue body and the advancement of the tongue root (Hu, 2008; Ladefoged and Wu, 1984; Lee, 2008; Lee-Kim, 2014; Proctor et al., 2012; Toda and Honda, 2003). Researchers have used different labels for this class of Mandarin post-alveolar sibilants, such as *alveolopalatal* (Ladefoged and Maddieson, 1996) and *anterodorsal post-alveolar* (Lee, 2008). Here, I will use the label *alveolopalatal* for convenience.[1]

The phonetic contrasts between fricatives from different places of articulation are characterized by differences in a number of acoustic properties. Principal among them are the spectral shape of frication noise and formant transitions into the following vowel (Gordon et al., 2002; Hughes and Halle, 1956; Jongman et al., 2000; McMurray and Jongman, 2011; Wilde, 1993). For sibilant fricatives, the spectral shape of frication noise is largely determined by the size of the front cavity; the smaller the front cavity is, the higher the centroid frequency is (Heinz and Stevens, 1961). In English, for example, while the alveolar fricatives ([s] and [z]) have a spectral peak around 7000 Hz, the palato-alveolar fricatives ([ʃ] and [ʒ]) have a spectral peak around 4000 Hz (Jongman et al., 2000).

There are a number of studies on the acoustics of sibilant fricatives in Man-

[1]The phonological relationships between alveolopalatal consonants and consonants from other places of articulation have been a major issue in the literature on Mandarin phonology. Synchronically speaking, alveolopalatal consonants and some other consonants are in complementary distribution; alveolopalatal consonants occur before high front vowels ([i, y]) or glides ([j, ɥ]) while dental, retroflex, and velar consonants do not. Phonologists have proposed various hypotheses for the phonological status of alveolopalatal consonants: (1) allophones of dental consonants (Duanmu, 2007; Hartman, 1944), (2) allophones of velar consonants (Chao, 1948), or (3) underlying phonemes (Cheng, 1973). Recently, Lu (2011) examined how much the hypothesized phonological status of alveolopalatal consonants affects Mandarin speakers' perception. In her study, Lu compared Mandarin and Korean speakers' ratings of the similarity between the dental [s] and alveolopalatal [ɕ]. Crucially, these two sounds are in complementary distribution in both languages, but they alternate with each other in phonological processes only in Korean (i.e., these two sounds are more allophonic in Korean than in Mandarin in the sense that the allophony is supported by distribution and alternation in Korean but only by distribution in Mandarin). The results of the study showed that Korean speakers rated the sounds as being more similar to each other than Mandarin speakers did, suggesting that the allophonic relationship between [s] and [ɕ] is not as well established in Mandarin as it is in Korean.

darin (Chang, 2013; Chiu, 2009; Lee, 2011; Li, 2008; Li et al., 2007; Svantesson, 1986). For example, Lee (2011) reported that the dental/alveolar [s] has a centroid frequency between 7000 and 11000 Hz, the retroflex [ʂ] has a centroid frequency between 3000 and 6000 Hz, and the alveolopalatal [ɕ] has a centroid frequency between 5000 and 10000 Hz. These values reflect how these fricatives are produced. Dental/alveolar fricatives are produced with an anterior constriction and a small front cavity. This is reflected in the concentration of spectral energy in the higher frequency region. Retroflex fricatives, by contrast, are produced with a posterior constriction and a large front cavity. This is reflected in the concentration of spectral energy in the lower frequency region. Moreover, a short and slack constriction channel for the production of retroflex fricatives allows a higher degree of acoustic coupling between the front and back cavities and further increases spectral prominence in the lower frequency region (Stevens et al., 2004). Alveolopalatal fricatives are produced with a constriction channel that stretches from the post-alveolar to palatal regions, but the long and narrow constriction channel reduces the amount of the acoustic coupling and prevents the increase of spectral prominence in the lower frequency region. This is reflected in the concentration of spectral energy in the intermediate frequency region.

In a CV string, formant transitions between the consonant and the vowel systematically change according to the place of the articulation of the consonant (e.g., Sussman et al. 1991; Wilde 1993; cf., Fowler 1994). For sibilant fricatives, for example, studies on Polish fricatives, which are similar to the ones in Mandarin, have reported that F2 transitions systematically change according to the class of the sibilant fricatives (Nowak, 2006; Zygis and Padgett, 2010). Nowak (2006) measured F2 transitions at 25 ms and 75 ms of vowels following three sibilant fricatives in Polish, the dental [s̪], retroflex [ʂ], and alveolopalatal [ɕ]. Nowak reported that the vowel [a] showed a slight F2 falling after the dental [s̪] and retroflex [ʂ] but a steep F2 falling for the alveolopalatal [ɕ].

For Mandarin fricatives, Chiu (2009) measured F2 transitions from the onset to the midpoint of the vowel [a] after three sibilant fricatives in Taiwan Mandarin, the dental/alveolar [s], retroflex [ʂ], and alveolopalatal [ɕ]. Chiu reported that F2 transitions after these three fricatives showed a falling contour, and the transition after the alveolopalatal [ɕ] showed the steepest fall (see Table 2.1).

Class	IPA	Onset F2 − Mid F2 (Hz)
Dental/alveolar	[s]	64.18
Retroflex	[ʂ]	170.05
Alveolopalatal	[ɕ]	337

Table 2.1: F2 transitions of Mandarin sibilants in [Ca] syllables (based on Table 2 in Chiu, 2009)

The differences in the spectral shape of frication noise and formant transitions are used as perceptual cues in the perception of fricatives (Delattre et al., 1962; Harris, 1954, 1958; Wagner et al., 2006; Whalen, 1981a,b, 1991), but the degree to which listeners rely on these cues varies depending on the class of fricatives and the inventory of fricatives in a given language. For example, Harris (1954) reported that English speakers rely more on frication noise in the perception of two sibilant fricatives ([s] and [ʃ]) but formant transitions in the perception of two non-sibilant fricatives ([f] and [θ]). Wagner et al. (2006) argued that perceptual cue weighting is largely determined by the inventory of fricatives. For example, English speakers rely on formant transitions in the perception of [f] because English has another non-sibilant fricative, [θ], which is spectrally similar to [f]. Dutch speakers, by contrast, rely on frication noise in the perception of [f] because Dutch does not have another non-sibilant fricative that is spectrally similar to [f]. Studies have suggested that the perceptual cue weighting is learned over the course of language acquisition. In English, for example, children rely more on formant transitions and less on frication noise than adults do (Nittrouer, 1992; Nittrouer and Studdert-Kennedy, 1987; Nittrouer and Miller, 1997a,b; Nittrouer, 2002).

For Mandarin sibilant fricatives, Chiu (2010) demonstrated that native speakers rely on different cues in perceiving different sibilant fricatives. In his study, Mandarin speakers were asked to identify three sibilant fricatives, the dental/alveolar [s], retroflex [ʂ], and alveolopalatal [ɕ], presented before congruent and incongruent formant transitions in CV syllables. Stimuli were created by cross-splicing the tokens of frication noise before the tokens of the vowel [a] with different patterns of formant transitions. The stimuli with congruent formant transitions were created by cross-splicing a token of frication noise before a token of [a] originally produced

after the same fricative sound (e.g., [s] + [(s)a]). The stimuli with incongruent formant transitions were created by cross-splicing a token of frication noise before a token of [a] originally produced after different fricative sounds (e.g., [s] + [(ʂ)a]). The results of the study showed that incongruent formant transitions affected the identification of the fricatives in some cases. Specifically, the identification of the dental/alveolar [s] and retroflex [ʂ] was fairly accurate except when these fricatives were presented before [(ɕ)a]; then they were misidentified as the alveolopalatal [ɕ]. Moreover, the identification of the alveolopalatal [ɕ] was less accurate when they were presented before [(s)a] and [(ʂ)a]. From these results, Chiu concluded that Mandarin speakers primarily attend to frication noise in perceiving the dental/alveolar [s] and retroflex [ʂ] but to formant transitions in perceiving the alveolopalatal [ɕ].

2.2.2 Post-alveolar fricatives in Mandarin and English

Since this experiment tests the learning of Mandarin post-alveolar fricatives by native speakers of English, it is crucial to understand, first, how these Mandarin sounds are different from English post-alveolar fricatives, and second, how English speakers perceive these Mandarin sounds. Toda and Honda (2003) made an articulatory comparison between English and Mandarin sibilant fricatives. In their MRI data, English palato-alveolar [ʃ] overlaps with Mandarin retroflex [ʂ] and alveolopalatal [ɕ] in terms of the size of the front cavity and the average width of the palatal constriction channel. Li et al. (2007) made an acoustic comparison between English and Mandarin sibilant fricatives and reported that English palato-alveolar [ʃ] falls in between Mandarin retroflex [ʂ] and alveolopalatal [ɕ] in terms of amplitude ratio (the difference in dB between the amplitude of the most prominent spectral peak and the amplitude of the second formant), an acoustic measure that correlates with the degree of palatalization. In other words, English [ʃ] is slightly more palatalized than Mandarin [ʂ] but less palatalized than Mandarin [ɕ]. These similarities between English [ʃ] and Mandarin [ʂ] and [ɕ] explain why the phonetic contrast between these two Mandarin sounds is particularly difficult for English speakers to acquire (e.g., Chao, 1948).

Table 2.2: Identification of Mandarin sibilant fricatives and the rating of their similarity to English fricatives by English speakers (based on Table 8 in Hao, 2012)

Stimuli	Group	% identification (similarity score)				
		/s/	/z/	/ʃ/	/ʒ/	/tʃ/
[ʂɨ]	Advanced			82 (5.84)		10 (6.29)
	Beginning			76 (5.98)		13 (5.78)
	No-exposure			72 (6.33)		14 (4.3)
[ʂu]	Advanced			85 (6.22)		
	Beginning			78 (5.98)		14 (5.9)
	No-exposure			61 (5.82)	10 (5.14)	18 (5.31)
[ɕi]	Advanced	13 (2.89)		69 (3.38)	16 (5)	
	Beginning	13 (3.89)		64 (4.76)		14 (4.6)
	No-exposure	33 (6.17)	14 (4.4)	33 (5.83)	11 (4)	
[ɕy]	Advanced			69 (3.98)	13 (5.22)	12 (3.25)
	Beginning			65 (5.68)		25 (4.83)
	No-exposure			69 (6.1)	13 (5.22)	10 (6.29)

Hao (2012) tested the perception of Mandarin sibilant fricatives by three groups of English speakers: (1) advanced learners of Mandarin, (2) beginning learners of Mandarin, (3) naive English speakers who had no previous exposure to Mandarin. Participants heard three sibilant fricatives in CV syllables, the dental/alveolar fricative (in the syllables [sɨ] and [su]), the retroflex fricative (in the syllables [ʂɨ] and [ʂu]), and the alveolopalatal fricative (in the syllables [ɕi] and [ɕy]), and were asked to label the fricative sounds using one of five phonetic symbols used to transcribe English phonemes, exemplified for them in English words (e.g., /s/ for *son*). Participants were also asked to rate the similarity between the fricative sounds in the stimuli and English fricatives that they chose on a scale of "1" (less similar) to "7" (more similar).

Table 2.2 shows the results of the identification and similarity rating. Important findings from the results are that naive English speakers in the no-exposure group perceived Mandarin retroflex [ʂ] as English palato-alveolar /ʃ/ most of the time, but their perception of Mandarin alveolopalatal [ɕ] was dependent on vowel context. They perceived the alveolopalatal fricative in [ɕy] as the palato-alveolar /ʃ/ most of the time, but they perceived the alveolopalatal fricative in [ɕi] either as

the alveolar /s/ or palato-alveolar /ʃ/. These findings suggest that the perception of these Mandarin fricatives by naive English speakers can be characterized either as single category assimilation or category goodness assimilation in the framework of the Perceptual Assimilation Model (Best et al., 1988; Hao, 2012). Both the retroflex [ʂ] and the alveolopalatal [ɕ] are mapped onto the palato-alveolar /ʃ/ (single category assimilation), but depending on the vowel context, the alveolopalatal [ɕ] is mapped onto either the alveolar /s/ or the palato-alveolar /ʃ/, which means that overall the retroflex [ʂ] fits the category of the palato-alveolar /ʃ/ better than the alveolopaltal [ɕ] does (category goodness assimilation).

Hao (2012) also tested her participants' discrimination using an AXB task. Participants compared the retroflex [ʂ] and alveolopalatal [ɕ] in the vowel contexts in which these fricatives were identified as the palato-alveolar /ʃ/ (comparing [ʂɨ] vs. [ɕy] and [ʂu] vs. [ɕy]). The results showed that naive English speakers performed well in the discrimination task (mean accuracies were higher than 75%). This goes against the hypothesis that the perception of these Mandarin fricatives by naive English speakers is characterized as single category assimilation. However, Hao speculates that the good performance in the discrimination taks is partly due to the fact that the stimuli differed not only in the fricative portion but also in the vocalic portion. Even through the vowel contrasts used in the test stimuli ([ɨ] vs. [y] and [u] vs. [y]) were non-English, participants were still able to perceive acoustic differences between these vowels and use the differences to discriminate the test stimuli (Hao, 2012, p.103).

Taken together, the results of Hao's study suggest that, although English speakers can distinguish [ʂ] and [ɕ], they may classify the sounds into one or two categories. These conclusions are complicated, however, by the fact that the discrimination stimuli differed in vowel as well as consonant, and the categorization task asked English speakers to forcefully sort the Mandarin sounds into English categories, instead of testing how many categories English listeners would naturally put the Mandarin sounds into.

Lee et al. (2012) tested the identification of Mandarin fricatives by native Mandarin speakers and English-speaking Mandarin learners using the hearing-in-noise test. In their study, participants were asked to identify Mandarin fricatives in CV syllables, the dental/alveolar [sa], retroflex [ʂa], and alveolopalatal [ɕa],

presented with different levels of speech-shaped noise. The results of the study showed both commonalities and differences between Mandarin speakers and English speakers. Overall, Mandarin speakers performed better than English speakers. With the dental/alveolar [sa], while Mandarin speakers frequently misidentified it as the retroflex [ʂa], English speakers were more likely to misidentify it as the alveolopalatal [ɕa]. With the retroflex [ʂa], both Mandarin and English speakers frequently misidentified it as the alveolopalatal [ɕa]. With the alveolopalatal [ɕa], both Mandarin and English speakers accurately identified it as the alveolopalatal [ɕa] (mean accuracies were higher than 80% in all noise conditions). The results of the Mandarin speakers conform to the claim made in Chiu (2010). Mandarin speakers primarily attend to frication noise in the perception of the dental/alveolar [s] and retroflex [ʂ]. Therefore, the masking noise obscured the spectral shape of the frication noise and made the identification of these fricatives harder. Mandarin speakers primarily attend to formant transitions in the perception of the alveolopaltal [ɕ]. Since the masking noise did not affect formant transitions as much as the spectral shape of the frication noise, the identification of the alveolopaaltal [ɕ] was less affected. The results of the English speakers are less clear. At least the finding that English speakers performed as well as Mandarin speakers in the identification of the alveolopalatal [ɕa] in all noise conditions suggests that they are learning to attend to formant transitions in the perception of the alveolopalatal [ɕ].

Some studies have suggested that English speakers' learning of non-English post-alveolar sibilant fricatives is more sensitive to information from formant transitions than information from frication noise. For example, McGuire (2007a, 2008) tested the perceptual learning of the phonetic contrast between Polish retroflex [ʂ] and alveolopalatal [ɕ], which are similar to the ones in Mandarin, by English speakers. After a period of laboratory training, participants showed a significant improvement in sensitivity to the phonetic contrast when they were trained with input in which the contrast was cued by formant transitions but not when they were trained with input in which the contrast was cued by the spectral shape of the frication noise.

In sum, Mandarin retroflex [ʂ] and alveolopalatal [ɕ] overlap with English palato-alveolar [ʃ] in terms of both articulation and acoustics. Naive English speakers perceive the retroflex [ʂ] and alveolopalatal [ɕ] as English palato-alveolar [ʃ]

most of the time, but they also perceive the alveolopalatal [ɕ] as English alveolar [s] in some vowel contexts.

2.3 Method

Experiment 1 consisted of two sessions over two consecutive days. In each session, participants were first exposed to input, and then were tested on the discrimination of the retroflex [ʂ] and alveolopalatal [ɕ]. The experiment was split into two sessions over two consecutive days because previous research has demonstrated that memory consolidation resulting from sleep facilitates the learning of speech sound categories (e.g., Fenn et al., 2003). Participants were randomly assigned to one of three conditions, one control and two experimental. Participants in different conditions were exposed to different sets of exposure stimuli, but they did the same discrimination test.

2.3.1 Participants

Sixty-two adult native speakers of English with no known speech or hearing problems participated in the experiment. All participants completed two sessions over two consecutive days and were paid $20 for their participation. All reported English to be their first and dominant language. Many of the participants were multilingual, but none of them were familiar with any language containing two or more post-alveolar fricatives as phonemes. Two participants were excluded from the analyses on the basis of their poor performance in a monitoring task during exposure (details described below). This left 20 participants in each condition.

2.3.2 Exposure stimuli

Exposure stimuli consisted of 256 bisyllabic strings. Each string comprised a context syllable followed by a critical syllable or a filler syllable. There were eight context syllables, [li], [lu], [mi], [mu], [pi], [pu], [gi], and [gu]. Context syllables were classified into two groups according to the vowel quality, one with the high front unrounded vowel [i] ([i] context) and the other with the high back rounded vowel [u] ([u] context). There were eight critical syllables drawn from a 10-step continuum between the retroflex [ʂa] and the alveolopalatal [ɕa]. There were also

four filler syllables, [t^{h}a], [ta], [fa], and [ha].

Recording and the acoustic analyses of target syllables

A male native speaker of Mandarin from Taiwan, who is also a trained phonetician, produced four repetitions of each one of eight context syllables ([li], [lu], [mi], [mu], [pi], [pu], [gi], and [gu]), two target syllables ([ʂa] and [ɕa]), and four filler syllables ([t^{h}a], [ta], [fa], and [ha]). Recording was done in a soundproof booth with an external dynamic omnidirectional microphone (SHURE SM63LB) connected to an iMac computer through a preamp (M-Audio Fast Track Pro). Recording sampling rate was 44,100Hz. The speaker produced the context syllables in a single syllable form, but the target syllables and filler syllables in a bisyllabic string with the preceding vowel [a] (e.g., [aʂa]). All of the syllables were produced with tone 1 (high level tone).

I analyzed the acoustics of target syllables looking at the spectral shape of frication noise and formant transitions into the following vowel. First, two spectral measurements, centroid frequency (i.e., center of gravity or CoG) and peak frequency, were measured from the midpoint of the frication noise. Second, the first three formants were measured at the 10 ms and 100 ms points of the following vowel. All of the acoustic measurements were made with Praat (Boersma and Weenink, 2001).

From the frication noise of the each token of the target syllables, a slice of 24 ms around the midpoint was extracted using the Hamming window function. An FFT spectrum was generated from the slice. The FFT spectrum was smoothed using Cepstral smoothing with a bandwidth of 500 Hz and CoG was measured from the smoothed spectrum. The FFT spectrum was also converted into a Long Time Averaged Spectrum (LTAS) with a bandwidth of 125 Hz and peak frequency was measured from the LTAS. Figure 2.1 shows the CoG and peak frequency of the retroflex [ʂ] and alveolopalatal [ɕ]. Both CoG and peak frequency are higher for the alveolopalatal [ɕ]. The values of CoG are within the ranges of centroid frequency that have been reported for Mandarin retroflex and alveolopalatal fricatives in previous studies (e.g. Lee, 2011).

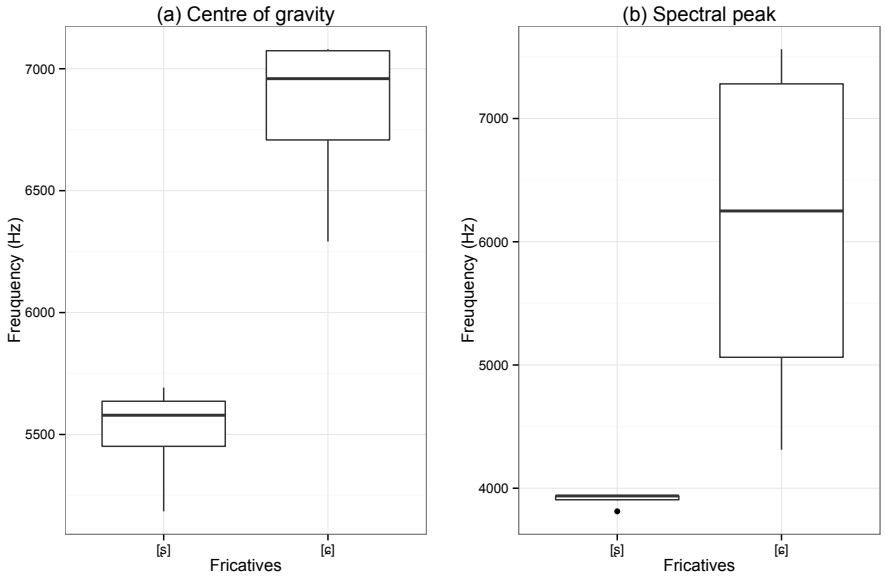


Figure 2.1: Spectral measurements of [ʂ] and [ɕ]

From the vowel of the each token of the target syllables, the first three formants were measured in a psychoacoustic scale (ERB-rate: Moore and Glasberg 1983) at 10 ms and 100 ms after the onset of voicing. Figure 2.2 shows the trajectories of the first three formants from 10 ms to 100 ms of the vowel [a] after the retroflex [ʂ] and alveolopalatal [ɕ]. The vowel after the alveolopalatal [ɕ] has lower onset F1 and higher onset F2.

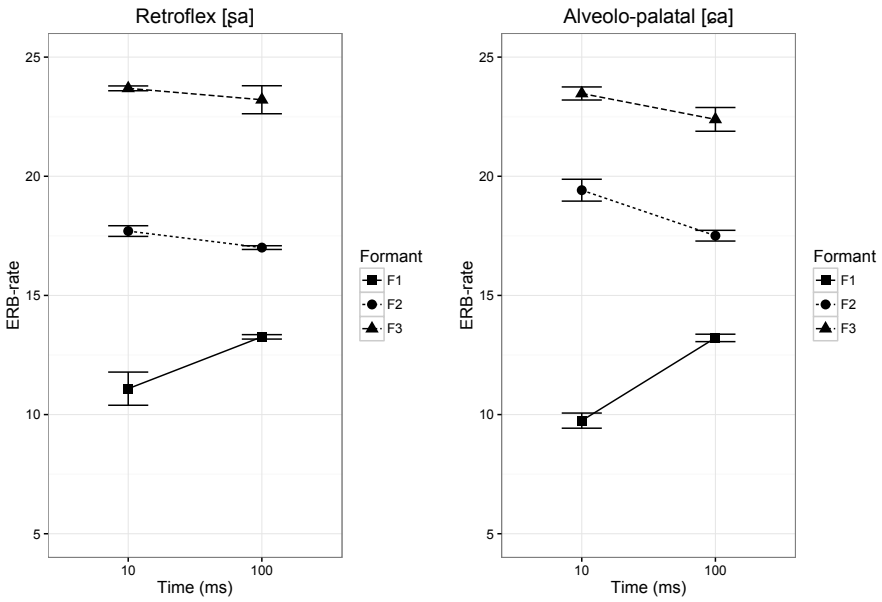


Figure 2.2: Formant transitions of [ʂa] and [ɕa] (with 95% CI)

Resynthesis of target syllables

A 10-step continuum between the retroflex [ʂa] and alveolopalatal [ɕa] was constructed by resynthesizing the original recordings of the target syllables. The resynthesis was done using a similar method as that employed by McGuire (2007a,b). First, the recordings of the target syllables were segmented into the frication noise and the vowel. Separate continua were constructed for the frication noise and the vowel. Then, the two continua were combined together to create a continuum of syllables. For the frication noise, tokens of [ʂ] and [ɕ] were selected on the basis of the quality of the recordings. From each token, 200 ms around the midpoint of the frication noise were extracted using a parabolic windowing function to smooth the onset and the offset of the extracted frication noise. After the extraction, the mean intensity of the frication noise was adjusted to be 50 dB. After the duration and the intensity were equalized, the Pitch Synchronous Overlap Add (PSOLA) method in Praat was used to interpolate a stepwise transition (10 steps) from the token of [ʂ] to the token of [ɕ] (Boersma and Weenink, 2001; Moulines and Charpentier, 1990). Figure 2.3 shows the LPC spectra of the frication noise at steps 1, 4, 7, and 10. As

the steps move from the retroflex end (step 1) to the alveolopalatal end (step 10), the prominent energy peak at around 4000 Hz is reduced and energy in the higher frequency regions increases.

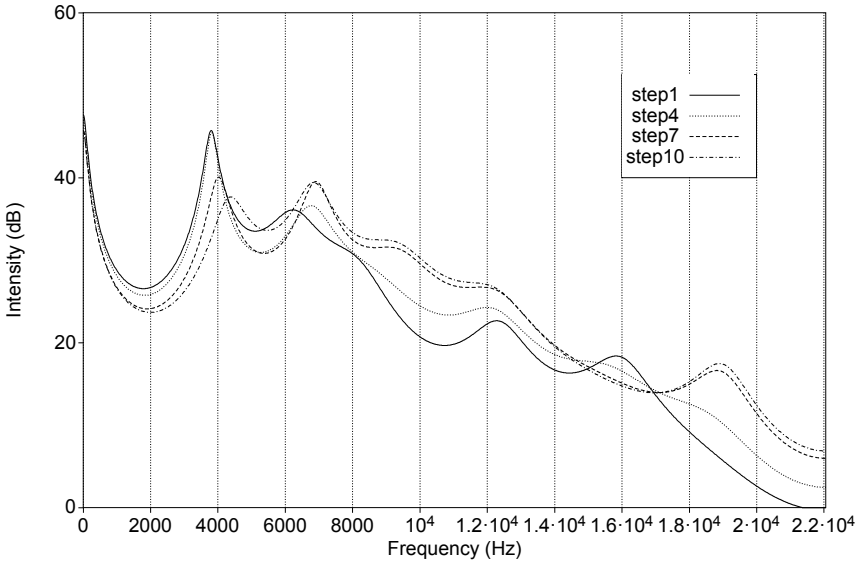


Figure 2.3: Spectra of resynthesized frication noise tokens

For the vowel, tokens of [a] after [ʂ] and [ɕ] were selected on the basis of the quality of the recordings. From each token, the initial 200 ms were extracted using a similar method used for the frication noise. First, 500 ms around the onset of voicing were extracted using a parabolic windowing function. Then, 200 ms after the onset of voicing were extracted. In this way, the intensity rise after the onset was retained to be the same as the original, but the intensity fall towards the offset was smoothed such that the two vowel tokens had similar envelope shapes towards the offset. After the extraction, the mean intensity of the vowel tokens was adjusted to be 63 dB.[2] After the duration and the intensity were equalized, the Speech Transformation and Representation by Adaptive Interpolation of Weighted Spectrogram method (STRAIGHT) was used to interpolate a stepwise transition

[2]The vowel intensity was determined based on the mean of the ratios of the vowel intensity to frication noise intensity across two fricatives in the original recordings

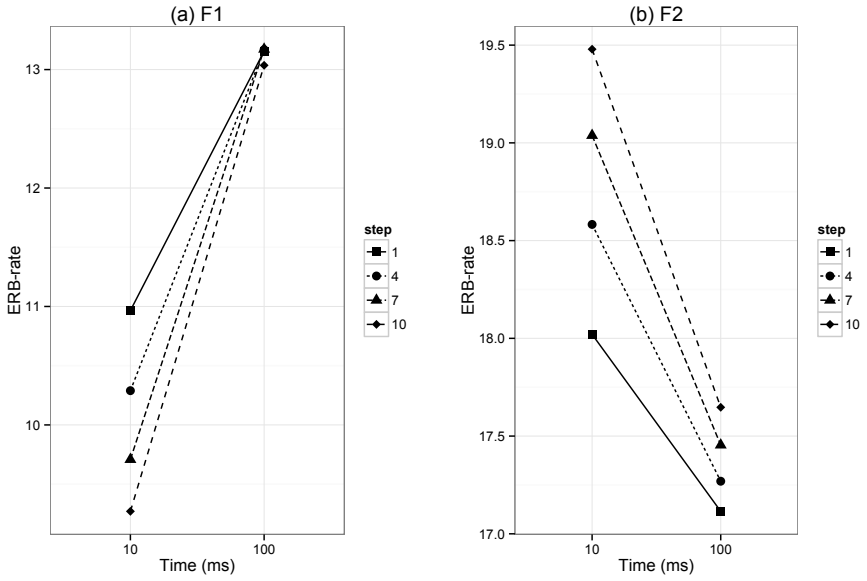


Figure 2.4: Formant transitions of resynthesized vowel tokens

(10 steps) from the token of [a] after [ʂ] to the token of [a] after [ɕ] (Kawahara et al., 1999). Figure 2.4 shows the first two formants measured at 10 ms and 100 ms of steps 1, 4, 7, and 10. As the steps move from the retroflex end (step 1) to the alveolopalatal end (step 10), the onset F1 becomes lower and the onset F2 becomes higher.

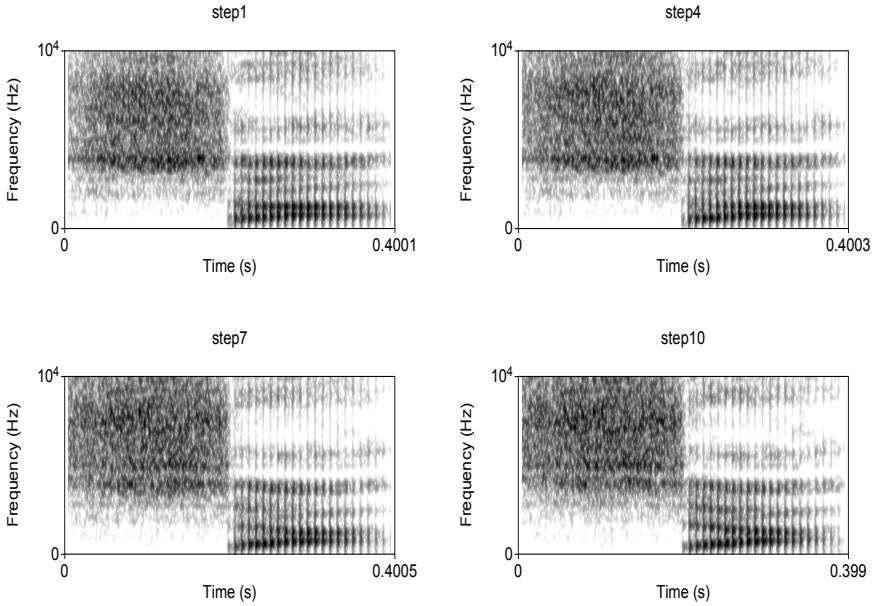


Figure 2.5: Spectrograms of steps 1, 4, 7, and 10

Finally, the frication noise continuum and the vowel continuum were combined to create a 10-step continuum from [ʂa] to [ɕa]. Figure 2.5 shows the spectrograms of steps 1, 4, 7, and 10 from the continuum. In the methods described above, the interpolations of steps between the retroflex [ʂa] and alveolopalatal [ɕa] were based purely on the acoustic properties of the end-point tokens. In order to determine the location of the perceptual boundary between [ʂa] and [ɕa], I tested the perception of the 10 syllables from the continuum by native speakers of Mandarin using an ABX discrimination task and an identification task. The results of the tests showed that the perceptual boundary is located at step 6 (Noguchi and Hudson Kam 2015a; also see Appendix A). For this experiment, eight critical syllables were selected from the continuum, four from the retroflex side of the perceptual boundary (steps 2, 3, 4, 5) and four from the alveolopalatal side of the perceptual boundary (steps 7, 8, 9, 10).

Construction of exposure stimuli

Exposure stimuli were constructed by concatenating the context syllables and critical syllables or filler syllables. In all of the stimuli, a context syllable came first, followed by a critical syllable or a filler syllable, such that the context vowel immediately preceding the target consonant. Before the concatenation, the duration of context syllables and critical syllables was manipulated so that all exposure stimuli had the same prosodic structure. The duration of context syllables and filler syllables was changed to 400 ms using a method similar to the one described above; 400 ms around the onset of the vowel was extracted using the parabolic window function. The first half (200 ms) of the extracted recording contained the acoustic signal that corresponded to the consonants, and the second half (200 ms) contained the acoustic signal that corresponded to the vowels. By doing this, context syllables, critical syllables, and filler syllables had the same vowel duration (200 ms). Moreover, all exposure stimuli had the same intervocalic interval duration (200 ms). Furthermore, mean syllable intensity was adjusted to be 55 dB for context syllables and 60 dB for critical syllables and filler syllables.

The frequencies of critical syllables were manipulated so that their aggregate distribution showes a bimodal shape (see Figure 2.6). This implies that the first half of the critical syllables (steps 2 to 5) forms the retroflex category, and the second half (steps 7 to 10) forms the alveolopalatal category. These 16 tokens of critical syllables were combined with eight context syllables ([li], [lu], [mi], [mu], [pi], [pu], [gi], and [gu]) to generate 128 bisyllabic strings (critical stimuli). Similarly, 32 tokens of filler syllables (eight tokens of [ta], [t^ha], [fa], and [ha]) were combined with eight context syllables to generate 256 bisyllabic strings (filler stimuli). Participants in the experimental conditions heard the 128 critical stimuli and one half of the 256 filler stimuli (the ones with [ta] and [t^ha]). Participants in the control condition heard the 256 filler stimuli. In all conditions, the 256 stimuli were divided into two subsets according to the consonant of the context syllables, a subset with [l] and [p] and the other subset with [m] and [g]. The first subset was used in Session 1 and the second subset was used in Session 2.

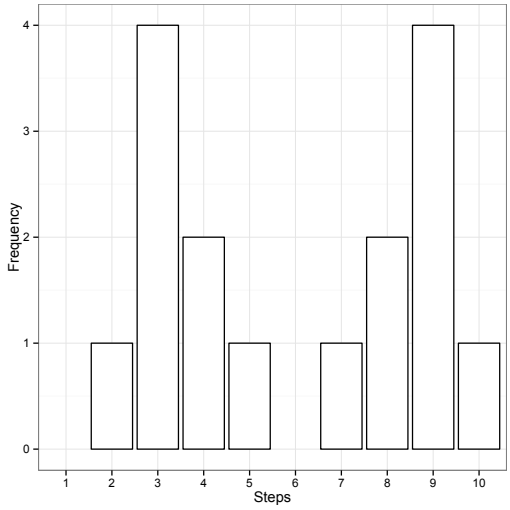


Figure 2.6: Aggregate distribution of critical syllables (Experiment 1)

2.3.3 Conditions

The experiment had three conditions, two experimental and one control. In the two experimental conditions, the phonotactic distribution of critical syllables was manipulated. In the first experimental condition (non-complementary condition), all of the critical syllables occurred in overlapping contexts: after the high front unrounded vowel [i] and the high back rounded vowel [u]. Figure 2.7 shows the distribution of the 32 tokens of critical syllables where the same number of tokens of each step occur in the [i] context and [u] context. For example, one token of step 2 occurred in the [u] context and the other token of step 2 occurred in the [i] context, for a total of two tokens at this step. In this condition, the retroflex [ʂ] and alveolopalatal [ɕ] were in a non-complementary distribution, and their occurrences were not predictable from the preceding contexts. In the second experimental condition (complementary condition), the critical syllables from the retroflex category occurred after the high back rounded vowel [u], and the critical syllables from the alveolopalatal category occurred after the high front unrounded vowel [i]. Figure 2.8 shows the distribution of the same 32 tokens of critical syllables where the tokens of step 2, 3, 4, and 5 occur in the [u] context and the tokens of step 7, 8, 9, and

10 occur in the [i] context. In this condition, the retroflex [ʂ] and alveolopalatal [ɕ] were in complementary distribution, and their occurrences were fully predictable from the preceding contexts.

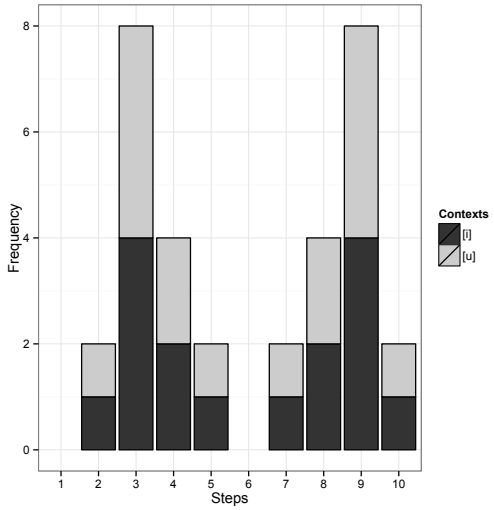


Figure 2.7: Distribution of 32 critical syllables in the non-complementary condition (Experiment 1)

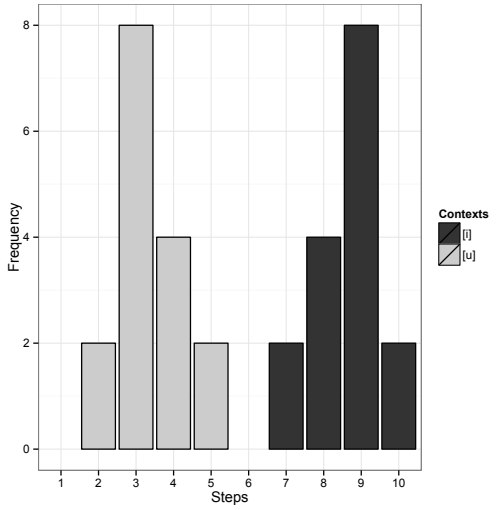


Figure 2.8: Distribution of 32 critical syllables in the complementary condition (Experiment 1)

The pattern of complementary distribution implemented in the input used in the complementary condition was defined on the basis of the relative similarities between the target segments and their respective contexts. Generally speaking, retroflex consonants share certain phonetic features with back rounded vowels (Flemming, 2003; Hamann, 2003, 2002) and palatal/palatalized consonants share certain phonetic features with high front vowels (Guion, 1996, 1998; Wilson, 2006). For example, the articulation of truly retroflex consonants involves the formation of a post-alveolar apical constriction: the lowering of the tongue front and the rising of the tongue tip towards the post-alveolar region. These gestures are facilitated by the retraction of the tongue body because it makes enough space for the lowing of the tongue front and the rising of the tongue tip. The retraction of the tongue body is also seen in the articulation of back vowels. The articulation of retroflex consonants also involves the formation of a large front cavity, which is reflected in the acoustic signal as the lowering of the onset F3 of the following vowel. This F3 lowering parallels an acoustic effect of lip rounding and/or lip protrusion that happens in the articulation of rounded vowels (Bhat, 1974; Dixit and Flege, 1991; Ladefoged and Bhaskararao, 1983). The connection between

retroflex consonants and back vowels is phonologized in some languages. For example, in Kodagu (Dravidian), retroflex consonants occur after back vowels but not after front vowels (e.g., [uɖ] is attested but [iɖ] is not: Bhat 1973; Flemming 2003; Gnanadesikan 1994).

The articulation of palatal/palatalized consonants involves the upward and forward movements of the tongue body, which are also seen in the articulation of high front vowels. The connection between palatal/palatalized consonants and front vowels is phonologized in many languages as palatalization. Palatalization usually involves three processes, tongue raising, tongue fronting, and spirantization (Bhat, 1978). For example, in Slavic, the velar stop [k] is realized as the palato-alveolar affricate [tʃ] before front vowels ([i], [œ]), and in Japanese, the alveolar fricative [s] is realized as the alveolopalatal fricative [ɕ] before the high front vowel [i] (Bateman, 2007; Bhat, 1978; Guion, 1998).

For post-alveolar fricatives in Mandarin, the connections between the retroflex [ʂ] and the high back rounded vowel [u] are not so obvious. Since the articulation of Mandarin retroflex fricatives does not involve the formation of a post-alveolar apical constriction, their articulation does not necessarily involve the retraction of the tongue body. The articulation of Mandarin retroflex fricatives involves the formation of a large front cavity. However, for the stimuli used in this experiment, the tokens of the retroflex [ʂa] and alveolopalatal [ɕa] did not show any significant difference in terms of F3 transition (see Figure 2.2).[3] By contrast, the connections between the alveolopalatal [ɕ] and the high front unrounded vowel [i] are very strong. The articulation of the alveolopalatal fricatives involves the formation of a long and narrow constriction channel over the post-alveolar and palatal regions. These articulatory gestures are achieved by moving the tongue body forward and upward, which are also seen to the articulation of the high front unrounded vowel [i]. For the stimuli used in this experiment, the presence of the palatal constriction is reflected in the low onset F1 and high onset F2 (see Figure 2.2). Overall, the occurrence of the retroflex [ʂ] next to the high back rounded vowel [u] and the

[3]One study has reported that the articulation of retroflex fricatives by some native speakers of Mandarin from Taiwan involves significant lip rounding (Chang, 2010). The speaker who produced the target syllables for this experiment is from Taiwan. However, it is not clear whether such a feature was present in his production of retroflex fricatives.

alveolopalatal [ɕ] next to the high front unrounded vowel [i] is phonetically more natural than the reverse.

Here, it should be made clear that despite the relatively natural connections described above, the particular pattern of complementary distribution implemented in the complementary condition of the current experiment is, as far as I know, not attested in any natural languages. In natural languages, there is no case in which the occurrences of retroflex and alveolopalatal fricatives are conditioned by the preceding vowels. Cases like Kodagu where the occurrences of retroflex consonants are conditioned by the preceding vowels are usually limited to stops (Bhat, 1973; Flemming, 2003; Gnanadesikan, 1994). This is probably because in the VC transition stops are strongly coarticulated with the preceding vowel, but fricatives are not (Stevens and Blumstein, 1975). Therefore, the pattern of complementary distribution implemented here does not assume any sort of typological universality.

The third condition was the control condition. In this condition, participants heard the same number of exposure stimuli as in the two experimental conditions, but the stimuli did not contain critical syllables at all; instead, they contained four filler syllables ([ta], [t^{h}a], [fa], [ha]). The control condition was included to assess native English speakers baseline sensitivity to acoustic differences between the retroflex [ʂ] and alveolopalatal [ɕ] as well as how sensitivity might be affected by repeated testing and familiarity with the test stimuli. Table 3.1 summarizes the exposure stimuli used in all three conditions.

Condition	**Context syllables**	**Critical syllables**	**Filler syllables**
Non-complementary	li-, mi-, pi-, gi-	steps 2, 3, 4, 5, 7, 8, 9, 10	-ta, -t^{h}a
	lu-, mu-, pu-, gu-	steps 2, 3, 4, 5, 7, 8, 9, 10	-ta, -t^{h}a
Complementary	li-, mi-, pi-, gi-	steps 7, 8, 9, 10	-ta, -t^{h}a
	lu-, mu-, pu-, gu-	steps 2, 3, 4, 5	-ta, -t^{h}a
Control	li-, mi-, pi-, gi-	N/A	-ta, -t^{h}a, -fa, -ha
	lu-, mu-, pu-, gu-	N/A	-ta, -t^{h}a, -fa, -ha

Table 2.3: Exposure stimuli (Experiment 1)

2.3.4 AX discrimination test

Participants' sensitivity to acoustic differences between the retroflex [ʂ] and alveolopalatal [ɕ] was tested using an AX discrimination paradigm. In each test trial, participants compared a token of [ʂa] and a token of [ɕa] (different trial) or two non-identical tokens of [ʂa] or [ɕa] (same trial). There were two types of different trials depending on the acoustic distance between the test stimuli, distant pairs and close pairs. In the distant pair trials, participants compared two stimuli that are acoustically quite different from each other (step 2 and step 10 from the continuum). In the close pair trials, they compared two stimuli that are acoustically more similar to each other (step 4 and step 8 from the continuum). There were also two types of same trials. In the retroflex category trials, participants compared two non-identical tokens of the retroflex category (step 2 and step 4). In the alveolopalatal category trials, participants compared two non-identical tokens of the alveolopalatal category (step 8 and step 10). In order to make the focus of the study less obvious, filler trials were included, where participants compared two different filler syllables ([ta] vs. [t^{h}a] and [fa] vs. [ha]) or two non-identical tokens of a single filler syllable ([ta] vs. [ta], [t^{h}a] vs. [t^{h}a], [fa] vs. [fa], and [ha] vs. [ha]). Table 2.4 summarizes the stimuli used in the AX discrimination test.

Test trial	Different trial	Distant pair	step 2 vs. step 10
		Close pair	step 4 vs. step 8
	Same trial	Retroflex	step 2 vs. step 4
		Alveolopalatal	step 8 vs. step 10
Filler trial	Different trial		ta vs. t^{h}a, fa vs. ha
	Same trial		ta vs. ta, t^{h}a vs. t^{h}a, fa vs. fa, ha vs. ha

Table 2.4: Test stimuli (Experiment 1)

Participants in all three conditions took the same AX discrimination test with the same test stimuli. Since participants in the control condition were not exposed to critical syllables at all, all of the items used in the test trials were novel stimuli for these participants. Similarly, since participants in the non-complementary and complementary conditions were exposed to only half of the filler syllables ([ta] and [t^{h}a]), the other half ([fa] and [ha]), which were used in the filler trials, were novel stimuli for these participants.

2.3.5 Design

E-Prime Professional (ver. 2.0) was used to control the presentation of stimuli and the recording of responses (Schneider et al., 2002). A session consisted of three phases: practice, exposure, and test. In the practice phase, participants did a practice AX discrimination test, in which they compared 18 pairs of English monosyllabic words. Half of the pairs contained two non-identical tokens of a single word (e.g., *cap* and *cap*), and the other half contained two different words (e.g., *cap* and *gap*). The purpose of including the practice trials was to familiarize participants with the AX discrimination task. Therefore, the structure of practice trials was identical to that of test trials (see below).

In the exposure phase, participants heard a block of 128 stimuli presented in a random order with one second interstimulus interval (ISI). They heard the block four times. Exposure stimuli were presented as "short sentences" so that each syllable in a stimulus could be treated as a word. In order to help participants stay attentive to the stimuli, a monitoring task was given to them. In each block of stimuli presentation, a monitoring stimulus (a filler stimulus with long vowels: e.g., [li:ta:]) was randomly inserted in every subblock of 16 presentations. Participants were asked to press the spacebar when they heard the instances of "slow speech".

In the test phase, participants did an AX discrimination test. There were 64 test trials, 32 different trials and 32 same trials. Half of the 32 different trials were distant pair trials, and the other half were close pair trials. Half of the 32 same trials were retroflex trials, and the other half were alveolopalatal trials. Similarly, there were 64 filler trials (32 different trials and 32 same trials). In each trial, the paired stimuli were separated by an ISI of 750 ms, which is long enough to let participants process the stimuli at a higher, non-auditory level (Pisoni, 1973; Werker and Logan, 1985). The same ISI has been used in previous studies on the distributional learning of sound categories by adults (e.g., Maye and Gerken, 2000; Pajak, 2012). Participants were given a maximum of five seconds to respond, and the trial was terminated whenever they recorded a response. Intertrial interval (ITI) was two seconds.

2.3.6 Procedure

The experiment consisted of two sessions over two consecutive days. It was conducted at the Language and Learning Lab at The University of British Columbia. On Day 1, participants came into the lab and signed the consent form. Then, Session 1 began. Participants were first given the following information about the experiment.

> *In this experiment, you will listen to someone speaking in a language that you never heard before. After listening to the speech for a while, you will be tested on what you learned from listening. During the test, you will hear the person saying two things in the language and you will decide whether they are two repetitions of the same word or two different words.*

After the introduction, participants proceeded to the practice phase. Participants were given the following instruction at the beginning of the practice phase.

> *In each trial, you will hear someone saying two things in English. They are either two repetitions of the same word or two different words. If you think that they are two repetitions of the same word, press the "SAME" key. If you think that they are two different words, press the "DIFFERENT" key.*

No feedback was provided during the practice phase. But after completing the practice trials, participants were allowed to ask for further instruction on the task if they needed. Otherwise, participants proceeded to the exposure phase. Participants were given the following instructions at the beginning of the exposure phase.

> *You will listen to someone speaking in a language that you never heard before. You will be listening to a person saying short sentences in the language. In the recordings, each sentence consists of two words. Unlike English, words in this language are very short and consist of only one syllable. Therefore a sentence could be something like "wa ko".*

Sometimes words will be pronounced very slowly, like "waaa kooo". When you hear the slow speech, press the "SPACE" bar.

After completing the exposure phase, participants proceeded to the test phase. Participants were given the following instructions at the beginning of the test phase.

We will see what you learned about the new language. The test is just like the practice that you took at the beginning. But this time, you will be tested with the new language.

In each trial, you will hear the person saying two things in the language you just heard. If you think that they are two repetitions of the same word in this language, press the "SAME" key. If you think that they are two different words in this language, press the "DIFFERENT" key.

Some words in this language are very similar to each other. Listen carefully to the way they sound. Even if you are not sure about your answer, make a guess based on what you heard in the previous listening section, and make sure that you answer all of the trials.

On Day 2, participants came back to the lab and did Session 2. Session 2 followed the same procedure as Session 1, except that participants filled a language background questionnaire and received $20 after finishing the test phase.

2.4 Results

First, participants' performance on the monitoring task was checked to see whether they were attentive to the stimuli during exposure. Participants who detected fewer than 75% of the monitoring stimuli were excluded from further analyses. This affected two participants, one in the control condition and the other in the complementary condition (Note that they were excluded prior to the end of the study and were replaced with two new participants so that the condition Ns were balanced). This left 60 participants (20 per condition) for analyses.

Responses to test trials in the AX discrimination test were converted into sensitivity or d-prime (d') scores (Macmillan and Creelman, 2004). d' is based on the

difference between the likelihood of the correct detection of a signal, that is, the likelihood of answering "yes" in the presence of a signal (*hit rate*), and the likelihood of the incorrect detection of a signal, that is, the likelihood of answering "yes" in the absence of a signal (*false alarm rate*). A larger difference between these two likelihoods means better sensitivity (i.e., the likelihood of correct detection is much higher than the likelihood of incorrect detection). The actual computation of d' takes the difference between z-transformed hit rate and z-transformed false alarm rate.

$$d' = z(\text{Hit rate}) - z(\text{False alarm rate})$$

d' scores were computed for each participant for each pair (distant or close) in each session (1 or 2). "Hit" was defined as the correct detection of a change in signal when participants heard two stimuli from two different categories (i.e., answering different in different trials), and "false alarm" was the incorrect detection of a change in signal when participants heard two stimuli from a single category (i.e., answering different in same trials). In order to compute hit rates and false alarm rates, different trials were coupled with same trials that shared the same first syllable (the A of an AX pair). For example, the different trial in which stimuli were presented in the order step 2, step 10 was coupled with the same trial in which stimuli were presented in the order step 2, step 4. In other words, hit was the detection of across-category change, and false alarm was the detection of within-category change. Following standard practice, when hit rate or false alarm rate was 0, it was replaced by $\frac{1}{2N}$, and when hit rate was 1, it was replaced by $1 - \frac{1}{2N}$ (N=number of trials for a particular trial type) (Macmillan and Creelman, 2004, p.8). All of the following statistical analyses were done in R 3.0.3 (R Core Team, 2014).

Figure 2.9 shows the mean d' scores for distant pair trials by session and condition. Figure 2.10 shows the mean d' scores for close pair trials by session and condition. A repeated-measures ANOVA was conducted on d' scores with condition as a between-participant variable and pair and session as within-participant variables. The significance level was set at $p < 0.05$. There were significant main effects of condition [$F(2,57) = 4.2,\ p = 0.019$], session [$F(1,57) = 14.153,\ p < 0.001$], and pair [$F(1,57) = 235.576,\ p < 0.001$]. There was no significant

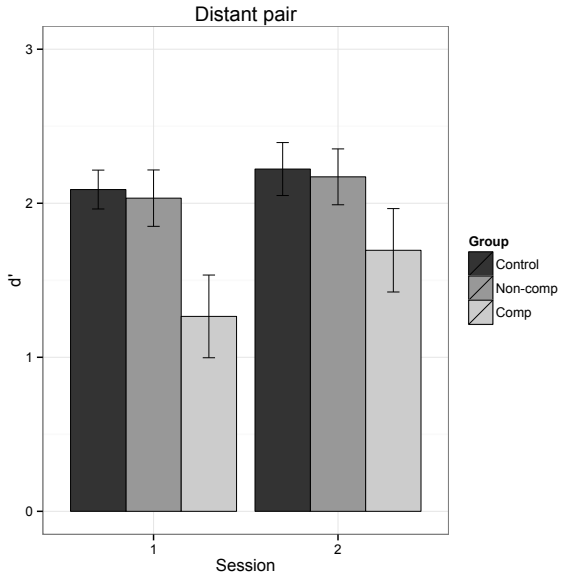


Figure 2.9: Mean d' scores for distant pair trials with 2 SE (Experiment 1)

two-way interaction between condition and pair [$F(2,57) = 1.386$, $p = 0.258$], between condition and session [$F(1,57) = 1.177$, $p = 0.838$], nor between pair and session [$F(1,57) = 3.325$, $p = 0.074$]. There was no three-way interaction [$F(1,57) = 1.642$, $p = 0.203$].

With condition, post-hoc pairwise comparison with the Holm adjustment method indicated that d' scores of the complementary condition ($M = 1.07$, $SD = 1.06$) were significantly lower than those of the control condition ($M = 1.62$, $SD = 0.86$) ($p < 0.001$) and the non-complementary condition ($M = 1.64$, $SD = 0.95$) ($p < 0.001$) but there was no significant difference between the control and non-complementary conditions ($p = 0.905$). With pair, post-hoc pairwise comparison indicated that d' scores for the distant pair trials ($M = 1.91$, $SD = 0.97$) were significantly higher than those for the close pair trials ($M = 0.97$, $SD = 0.78$) ($p < 0.001$). With session, post-hoc comparison indicated that d' scores for Session 2 ($M = 1.61$, $SD = 0.99$) were significantly higher than those for Session 1 ($M = 1.28$, $SD = 0.98$) ($p < 0.001$).

Participants in the control and non-complementary conditions showed about

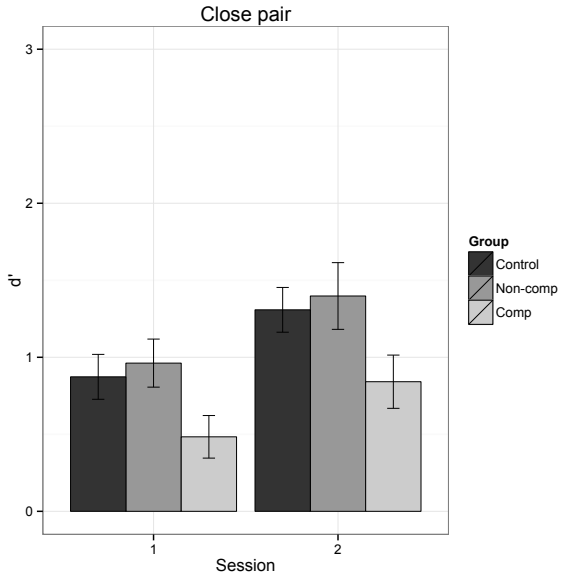


Figure 2.10: Mean d' scores for close pair trials with 2 SE (Experiment 1)

the same level of sensitivity to acoustic differences between [ʂa] and [ɕa] after exposure. If the performance of participants in the control condition reflects the baseline sensitivity, or English speakers' pre-existing sensitivity, to the acoustic differences, this finding suggests that no significant learning happened in the non-complementary condition. Exposure to the input in which [ʂ] and [ɕ] occurred in overlapping contexts did not significantly affect learners' pre-existing sensitivity to acoustic differences between [ʂa] and [ɕa]. By contrast, participants in the complementary condition showed significantly lower sensitivity compared to those in the control and non-complementary conditions after exposure. This suggests that exposure to the input in which [ʂ] and [ɕ] occurred in mutually exclusive contexts significantly reduced the learners' pre-existing sensitivity.

Participants across all three conditions showed a significant improvement in sensitivity from Session 1 to Session 2. This is probably due to familiarization with test stimuli. Since the same set of test stimuli were used in both sessions, it is possible that participants in all three conditions became more sensitive to acoustic differences between the test stimuli. If that is the case, the data from Session 1

should provide a clearer picture of the effects of the input. A follow-up analysis was conducted with just the data from Session 1. A repeated-measures ANOVA was conducted on d' scores with condition as a between-participants variable and pair as a within-participant variable. There were significant main effects of condition [$F(2,57) = 4.793$, $p = 0.012$] and pair [$F(1,57) = 165.82, p < 0.001$]. There was no significant interaction between condition and pair [$F(2,57) = 2.58$, $p = 0.085$]. Post-hoc pairwise comparison indicated that d' scores of the complementary condition ($M = 0.87$, $SD = 1.02$) were significantly lower than those of the control condition ($M = 1.48$, $SD = 0.86$) ($p = 0.011$) and the non-complementary condition ($M = 1.5$, $SD = 0.92$) ($p = 0.011$), but there was no significant difference between the control and the non-complementary conditions ($p = 0.937$). Post-hoc pairwise comparison indicated that d' scores for the distant pair trials ($M = 1.8$, $SD = 0.96$) were significantly higher than those for the close pair trials ($M = 0.77$, $SD = 0.68$) ($p < 0.001$).

In sum, participants in the control and non-complementary conditions showed the same level of sensitivity to acoustic differences between [ʂa] and [ɕa] after exposure, but participants in the complementary condition showed a significantly lower level of sensitivity after exposure.

2.5 Discussion

The results of Experiment 1 showed that a difference in the phonotactic distribution of target segments in input significantly affected learners' sensitivity to acoustic differences between the segments. Participants in the two experimental conditions were exposed to input in which the frequency distribution of novel sounds showed a bimodal shape. This implied the categorization of the sounds into two segments. In the non-complementary condition, these two segments occurred in overlapping contexts, and thus their occurrences were not predictable from the contexts. In the complementary condition, these two segments occurred in mutually exclusive contexts (i.e., they were in complementary distribution), and thus their occurrences were predictable from the contexts. The finding that participants in the complementary condition showed reduced sensitivity after exposure suggests that these participants learned to treat the target segments as something like allophones.

These results are consistent with the hypothesis that adults can learn allophonic relationships between two segments from the complementary distribution of the segments in input. However, the finding that participants in the control and non-complementary conditions did not show any significant difference may be surprising. Participants in the control condition did not get any exposure to critical syllables, while participants in the non-complementary condition got exposure to the bimodal distribution of critical syllables, which was supposed to be a robust cue for the learning of [ʂ] and [ɕ] as separate categories. Therefore, one may wonder why exposure to the bimodal distribution did not help them to improve their sensitivity.

To this point I have been implicitly assuming a model in which all sensitivities are learned. However, we know that this is not the case. Following Aslin and Pisoni (1980), Maye (2000) discusses three possible hypotheses in which exposure to input affects learners' sensitivity in a distributional learning paradigm. The first is the *maintenance hypothesis*. In this hypothesis, learners have good pre-existing sensitivity to acoustic differences between two target segments. Therefore, exposure to a bimodal distribution does not affect their sensitivity, but exposure to a unimodal distribution will decrease their sensitivity. The second is the *facilitation hypothesis*. In this hypothesis, learners have poor pre-existing sensitivity to acoustic differences between two target segments. Therefore, exposure to a bimodal distribution will improve their sensitivity, but exposure to a unimodal distribution does not affect their sensitivity. The third is the *underspecification hypothesis*. In this hypothesis, learners' sensitivity to the acoustic differences between two target segments is initially underspecified. Therefore, exposure to a bimodal distribution will improve their sensitivity, and exposure to a unimodal distribution will decrease their sensitivity.

In her own study on the distributional learning of stop voicing categories by adult English speakers, Maye (2000) compared three conditions, control, unimodal, and bimodal. In the control condition, participants did not get any exposure at all. In the unimodal condition, participants were exposed to a unimodal distribution that implied the classification of stop consonants into a single voicing category (between prevoiced and voiceless unaspirated). In the bimodal condition, participants were exposed to a bimodal distribution that implied the classification of stop consonants into two voicing categories (prevoiced and voiceless unaspirated). Af-

ter exposure, all participants were tested on the discrimination of the prevoiced and voiceless unaspirated stops. The results showed that participants in the bimodal condition performed significantly better than participants in the unimodal condition. The results also showed that participants in the control condition performed better than the participants in the unimodal condition but worse than the participants in the bimodal condition, even though the differences were not statistically significant. From these results, Maye concluded that the distributional learning happened in both directions supporting the underspecification hypothesis; while exposure to the unimodal distribution decreased learners' sensitivity, exposure to the bimodal distribution improved their sensitivity (Maye, 2000, pp.103-105).

Following studies, however, have shown more varied results. Hayes-Harb (2007) tested the distributional learning of stop voicing categories by adult English speakers, using the same stimuli used in Maye (2000). She compared several exposure conditions including ones that were equivalent to the control, unimodal, and bimodal conditions of Maye (2000). Interestingly, the results of Hayes-Harb's study showed that after exposure, while participants in the control and bimodal conditions performed at the same level in the discrimination task, participants in the unimodal condition performed significantly worse than those in the control and bimodal conditions, suggesting that learning happened only for participants in the unimodal condition.

A similar finding was reported in Pajak (2012). Pajak tested the distributional learning of consonantal duration categories (short consonants vs. long consonants) by adult English speakers. In her study, participants were exposed to the bimodal distribution of consonants along a segmental duration continuum (bimodal condition) or the unimodal distribution of consonants along the same segmental duration continuum (unimodal condition). After exposure, participants in the bimodal condition performed at chance level in the discrimination task, but participants in the unimodal condition showed a significant bias towards the "same" response. According to Pajak, while participants in the bimodal condition did not learn anything, participants in the unimodal condition did learn to classify test stimuli into a single category.

Compared to the results of these studies, the results of Experiment 1 are not unusual. While exposure to the input in the non-complementary condition did not af-

fect participants' sensitivity, exposure to the input in the complementary condition led to a reduction in participants' sensitivity. This is analogous to the maintenance hypothesis in Maye (2000), and suggests that participants in the complementary condition learned to treat the retroflex [ʂ] and alveolopalatal [ɕ] as variants of a single category.

Part of the maintenance hypothesis is good pre-existing sensitivity, which was something we see in the data from the present experiment. Participants in the control condition performed fairly well in the discrimination task. They performed particularly well in distant pair trials (step 2 vs. step 10). The mean accuracy of their responses in these trials was 75.4% ($SD = 27$) across two sessions. A possible factor that contributed to the good sensitivity of participants in the control condition is the fact that the retroflex [ʂ] and alveolopalatal [ɕ] were completely new to them. In other words, not having experienced these sounds resulted in good sensitivity. This may sound quite counterintuitive, but studies on L2 phonological acquisition have demonstrated that in some cases having less knowledge of L2 phonology allows learners to be more sensitive to acoustic differences between some L2 phonemes. Diehm (1998), for example, tested the perception of the different degrees of palatalization in Russian by Russian speakers and English-speaking Russian learners. Russian has four different degrees of palatalization, non-palatalized (CV), simple palatalization (C^{j}V), palatalized yod (C^{j}jV), and palatalized i-yod (C^{j}ijV). Surprisingly, Diehm found that English speakers did better than Russian speakers with the identification of the palatalized i-yod (C^{j}ijV); Russian speakers correctly identified C^{j}ijV only 26% of the time and misidentified it as C^{j}jC 73% of the time, while English speakers correctly identified C^{j}ijV 78% of the time and misidentified it as C^{j}jV only 15% of the time. According to Diehm, C^{j}ijV has extremely low functional load in Russian and C^{j}ijV and C^{j}jV are in a situation of near-merger. Therefore, for Russian speakers, having the functional knowledge about C^{j}ijV and C^{j}jV made them less sensitive to the acoustic information that differentiates C^{j}ijV from C^{j}jV. By contrast, English speakers do not have the functional knowledge Russian speakers do, so their perception is not biased by the linguistic knowledge; their perception is based more on the acoustic properties of C^{j}ijV and C^{j}jV. In this way, having less experience with an L2 sometimes enables learners to attend to the acoustic properties of the sounds in the L2 better. The

learning examined in Diehm (1998) is very different from the learning examined in this experiment, but it is still possible that, for participants in the control condition, not having experienced the retroflex [ʂ] and alveolopalatal [ɕ] at all could enable them to attend to the acoustic properties of these sounds better in the test trials.

2.6 Conclusion

There were three important empirical findings from the results of Experiment 1. First, adult English speakers have fairly good pre-existing sensitivity to acoustic differences between Mandarin retroflex [ʂ] and alveolopalatal [ɕ]. Second, when English speakers are exposed to input in which these two segments occur in overlapping contexts, and thus the occurrences of the segments are not predictable from the contexts, they maintain the pre-existing sensitivity. Third, when English speakers are exposed to input in which these two segments occur in mutually exclusive contexts, and thus the occurrences of the segments are predictable from the contexts, they become less sensitive to the acoustic differences. The third finding is particularly important since it suggests that the segments in complementary distribution are learned as something like allophones. Experiment 1 provided the first experimental support for the hypothesis that adults can learn allophonic relationships between segments from the complementary distribution of the segments in input.

Chapter 3

Experiment 2: Phonetic naturalness and the learning of allophony

3.1 Introduction

Experiment 1 tested whether adults can learn allophonic relationships between segments from the complementary distribution of the segments in input. Learners in two experimental conditions were exposed to the same bimodal distribution of novel sounds that implied the classification of the sounds into two segmental categories, the retroflex [ʂ] and alveolopalatal [ɕ]. The crucial difference between these two conditions was in the phonotactic distribution of the target segments. In one condition, the segments occurred in overlapping contexts and their occurrences were not predictable from the contexts (non-complementary condition). In the other condition, the segments occurred in mutually exclusive contexts such that their occurrences were predictable from the contexts (complementary condition). The results showed that learners in the non-complementary condition maintained their pre-existing sensitivity to acoustic differences between [ʂa] and [ɕa], but learners in the complementary condition showed reduced sensitivity after exposure. These results suggest that learners in the complementary condition learned

to treat[ʂ] and [ɕ] as something like allophones.

In this chapter, I address the question of how robust this learning is. I specifically ask whether the learning of allophony is constrained by the naturalness of the patterns of complementary distribution. This question relates to an issue of major importance in current research on language learning: the way in which the inductive learning of linguistic patterns is subject to constraints. Human learners have the ability to inductively learn regularities in input (e.g., statistical learning: Aslin et al. 1998, Saffran et al. 1996a, Saffran et al. 1996b). However, it has been demonstrated that inductive learning is constrained or biased such that some patterns are more learnable than others (e.g., Moreton and Pater, 2012a,b; Newport and Aslin, 2004; Saffran, 2002; Thiessen, 2011a). If allophony is inductively learned from the complementary distribution of segments in input, it is important to understand whether the learning is constrained and how.

3.2 Constraints on statistical learning

Studies have demonstrated that both infants and adults are able to segment continuous speech into sub-strings based on statistical dependencies between adjacent syllables; they extract sub-strings that have higher between-syllable transitional probabilities (e.g., Aslin et al., 1998; Saffran et al., 1996a,b). Studies have also demonstrated that statistical learning is domain-general. Human learners can segment continuous non-speech tone sequences into sub-strings based on statistical dependencies between adjacent tones (Saffran et al., 1999), and they can also learn visual patterns based on statistical dependencies between visual objects in a scene (Fiser and Aslin, 2002; Kirkham et al., 2002).

Despite the prevalence of studies showing the robust effects of statistical learning, it has been recognized that statistical learning is constrained by perceptual factors. For example, Newport and Aslin (2004) reported that adults failed to segment continuous speech into sub-strings based on transitional probabilities between non-adjacent syllables but successfully did segmentation based on transitional probabilities between non-adjacent segments of the same class (i.e., between non-adjacent consonants or non-adjacent vowels). The failure to do segmentation based on transitional probabilities between non-adjacent syllables suggests that distance be-

tween the objects over which statistical dependencies are learned affects the ease of learning (cf. Gómez, 2002). The success of doing segmentation based on transitional probabilities between non-adjacent segments of the same class suggests that similarity between the objects over which statistical dependencies are learned also affects the ease of learning (through Gestalt principles of similarity according to Newport and Aslin 2004). The similarity constraint has been reported in a study on the segmentation of non-speech tone sequences as well (Creel et al., 2004). Creel et al. reported that adults performed better in the segmentation of non-speech tone sequences based on transitional probabilities between non-adjacent tones when the tones were similar to each other (e.g., from the same pitch range).

3.3 Constraints on the learning of phonology

There is a growing number of studies identifying constraints on the learning of phonology. In artificial language learning experiments, it has been demonstrated that the inductive learning of phonological patterns is constrained such that some patterns are more learnable than others. In this section, I will provide a brief review of factors that have been considered to affect phonological pattern learning. Previous studies have investigated constraints on phonological pattern learning by both infants and adults, and these studies suggest that the learning is constrained in similar manners with both infants and adults.[1]

Some studies have suggested that the phonetic naturalness of patterns constrains the learning (Carpenter, 2010; Gerken and Bollt, 2008; Schane et al., 1975). Phonetically natural patterns are more learnable than unnatural ones. Here, natural patterns are those that can be explained with reference to aspects of speech production and/or speech perception (e.g., Blevins, 2008). For example, Schane et al. (1975) exposed adult English speakers to input in which word-final consonants were deleted in certain environments. In one condition, word-final consonants were deleted before consonant-initial words (e.g., *amuf* + *paʃi* → *amupaʃi*). In another condition, word-final consonants were deleted before vowel-initial words (e.g., *amuf* + *oga* → *amuoga*). The first pattern is more natural than the second

[1]Cristia et al. (2011c) reported some developmental changes in the effects of constraints on the learning of phonotactic patterns by infants. Specifically, the learning by 4-month-old infants is less constrained compared to the learning by 7.5-month-old infants.

one because it involves consonant cluster reduction, which is grounded in phonetic effects such as the gestural overlap between consonants in a cluster (e.g., Byrd and Tan, 1996). After exposure, learners in the first condition showed better learning performance than learners in the second condition. Carpenter (2010) exposed adult English speakers to input in which the distribution of stress was conditioned by the height of vowels. In one condition, low vowels were stressed, and high vowels were unstressed. In another condition, high vowels were stressed, and low vowels were unstressed. In natural languages, stress patterns can be sensitive to the sonority of vowels; in such cases, stress preferentially targets vowels with higher sonority (e.g., De Lacy, 2004). Since low vowels have higher sonority, the first stress pattern is more natural than the second one. After exposure, learners in the first condition showed better learning performance than learners in the second condition.

The claim that phonetic naturalness affects the learning of phonological patterns does not imply that unnatural patterns are impossible to learn. For example, Pycha et al. (2003) compared the learning of vowel harmony and vowel disharmony by adult English speakers. According to Pycha et al., vowel harmony is more natural than vowel disharmony because the former has phonetic grounding in phenomena such as vowel-to-vowel coarticulation. However, the results of the study showed that adults could learn vowel harmony and vowel disharmony equally well. Seidl and Buckley (2005) tested the learning of phonotactic patterns by 9-month-old English-learning infants. In their study, a group of infants were exposed to input in which sibilants occurred in intervocalic position and stops occurred elsewhere. Another group of infants were exposed to input in which stops occurred in intervocalic position and sibilants occurred elsewhere. The first pattern is more natural than the second one because it assumes the results of spirantization in intervocalic position, which can be explained in terms of phonetic factors such as articulatory effort (e.g., Kirchner, 1998). After exposure, infants in both groups showed learning equally well.

Studies have also demonstrated that the formal complexity of patterns constrains the learning. Here, complexity of patterns is determined by the number of phonological features and the number of operations that are required to describe the patterns (e.g., Moreton and Pater, 2012a). Specifically, patterns that apply to sets of sounds that are defined by a small number of phonological features

(i.e., natural classes) are simple (or systematic) compared to the ones that apply to sets of random sounds. Studies have demonstrated that simple patterns are more learnable than complex ones (Endress and Mehler, 2010; Kuo, 2009; Peperkamp et al., 2006b; Pycha et al., 2003; Saffran and Thiessen, 2003; Skoruppa et al., 2011; Wilson, 2003). For example, Saffran and Thiessen (2003) exposed 9-month-old English-learning infants to input in which the distribution of consonants was conditioned by syllable position. A group of infants were exposed to input in which voiceless stops ([p, t, k]) occurred in syllabe onset position and voiced stops ([b, d, g]) occurred in syllable coda position. Another group of infants were exposed to input in which a set of arbitrary consonants ([p, d, k]) occurred in syllable onset position and another set of arbitrary consonants ([b, t, g]) occurred in syllable coda position. After exposure, infants in the first group learned the phonotactic patterns, but infants in the second group did not. Peperkamp et al. (2006b) compared the learning of systematic and arbitrary phonological alternations by adult French speakers. A group of learners were exposed to input in which consonants from the same place and manner class (e.g., homorganic stops) alternated between voiced and voiceless; the consonants were voiced in intervocalic position (e.g., [nɛl pemuʃ]∼[ʁa bemuʃ]). Another group of learners were exposed to input in which the alternation involved place and manner of articulation in addition to voicing (e.g., [nɛl pemuʃ]∼[ʁa ʒemuʃ]). After exposure, learners in the first group learned the alternation, but learners in the second group did not.

Despite the robust effects of complexity, complex patterns are not impossible to learn. Some studies have demonstrated that both infants and adults can learn phonotactic patterns that apply to sets of arbitrary sounds (Chambers et al., 2003; Kuo, 2009; Onishi et al., 2002). For example, Chambers et al. (2003) demonstrated that 16.5-month-old English-learning infants learned phonotactic patterns in which [b, k, m, t] occurred in syllable onset position and [p, g, n, t, ʃ] occurred in syllable coda position.

In some of these studies, however, it is not clear whether the factor that constrains the learning is phonetic naturalness or complexity. This is because phonetically natural patterns are usually simple in their formal representations. Therefore, a comparison between systematic and arbitrary patterns can often be interpreted as a comparison between phonetically natural and unnatural patterns. For exam-

ple, Wilson (2003) exposed adult English speakers to input in which the distribution of two allomorphs ([-la] and [-na]) of a single suffix was conditioned by the consonants of the preceding words. In one condition, while [-la] occurred after CVCV words whose second consonant was either [t] or [k] (e.g., [suto-la] and [tuko-la]), [-na] occurred after CVCV words whose second consonant was a nasal (e.g., [dume-na]). In another condition, [-la] occurred after words with a nasal or [t] (e.g., [dume-la] and [suto-la]), and [-na] occurred after words with [k] (e.g., [tuko-na]). After exposure, learners in the first condition learned the distribution, but learners in the second condition did not. Here, the distribution of [-la] and [-na] in the first condition assumed a pattern of nasal harmony; the sonorant [l] was nasalized after a nasal consonant. This is phonetically natural in the sense that nasal harmony is phonetically grounded (in coarticulatory nasalization) and formally simple (or systematic) in the sense that the environment that conditioned the alternation can be defined by a single phonological feature ([+nasal]) and the alternation can be described as the change of of the sonorant from [-nasal] to [+nasal]. Indeed, such nasal consonant harmony patterns are attested in natural languages (e.g., Hansson, 2010).

Researchers have argued that what makes natural and simple patterns more learnable are inductive biases that learners are subject to (Moreton, 2008; Moreton and Pater, 2012a,b; Wilson, 2006). Wilson (2006) proposed a category of inductive biases he called *substantive biases*. According to Wilson, substance in phonology refers to "any aspect of grammar that has its basis in the physical properties of speech. These properties include articulatory inertias, aerodynamic pressures, and degrees of auditory salience and distinctiveness" (Wilson, 2006, p. 946). Substantive biases are cognitive biases that predispose learners toward those patterns that are phonetically grounded.

Wilson (2006) demonstrated that substantive biases affect the way learners make generalizations in the learning of phonological patterns. In his artificial language learning experiment, adult English speakers were exposed to input in which velar consonants are palatalized in certain environments. In one condition, velar palatalization happened only before a high front vowel (e.g., /ki/ $\rightarrow$ [tʃi]). In another condition, velar palatalization happened only before a mid front vowel (e.g., /ke/ $\rightarrow$ [tʃe]). After exposure, learners were tested on the generalization of

the rule to new vowel contexts: generalization to the mid vowel context for learners in the first condition and generalization to the high vowel context for learners in the second condition. The results showed that learners in the first condition did not generalize the rule to the mid vowel context, but learners in the second condition did generalize the rule to the high vowel context. According to Wilson, learners knew that palatalization in the high vowel context is phonetically more natural than palatalization in the mid vowel context and that the occurrence of the latter implies the occurrence of the former. Therefore, the learning of palatalization in an unnatural (or less expected) context allowed learners to infer that palatalization should happen in more natural (expected) context as well. White and Sundara (2014) reported a similar bias in the learning of phonological alternations by 12-month-old English-learning infants. In their study, a group of infants were exposed to input in which there was an alternation between a pair of relatively similar sounds (b]∼[v]). Another group of infants were exposed to input in which there was an alternation between a pair of relatively less similar sounds ([p]∼[v]). Afte exposure, learners in the first group learned the alternation they were familiarized with, but did not generalize it to a pair of less similar sounds ([p]∼[v]). By contrast, learners in the second group learned the alternation they were familiarized with, and generalized it to a pair of more similar sounds ([b]∼[v]). According to White and Sundara, infants knew that phonological alternations between phonetically similar sounds are more natural than alternations between phonetically dissimilar sounds. Therefore, learning an unnatural alternation allowed learners to infer its more natural counterpart should happen as well.

Moreton and Pater (2012a) introduced another category of learning biases, *complexity biases*. These are cognitive biases that predispose learners towards patterns that are formally simple and systematic. Moreton and Pater (2012b) further argued that substantive biases are a part of complexity biases. This is because, as discussed above, the patterns that are considered to be phonetically natural are usually simple and systematic in their phonological representations. Moreton (2012) also claimed that, unlike substantive biases, which rely on learners' knowledge about phonological substance, complexity biases are domain-general.

3.4 Constraints on the learning of allophony

Some of the studies mentioned above have investigated constraints on the learning of phonological alternations which are basically allophonic rules (Peperkamp et al., 2006b; Skoruppa et al., 2011; Wilson, 2003). However, no study has investigated constraints on the learning of allophony as a question of category learning. Experiment 2 is the first attempt to explore constraints on the learning of allophony. Specifically, it tests the effect of the phonetic naturalness of the patterns of complementary distribution in input.

In the input used in the complementary condition of Experiment 1, the retroflex [ʂ] occurred after the high back rounded [u], and the alveolopalatal [ɕ] occurred after the high front unrounded [i]. This particular pattern of complementary distribution was implemented on the basis of relative similarities between the target segments and their respective contexts. While the connections between [ʂ] and [u] are not so obvious, the connections between [ɕ] and [i] are strong; both the articulation of [ɕ] and the articulation of [i] involve the formation of a palatal constriction, and the presence of a palatal constriction is reflected in the acoustic signal as low F1 and high F2. These relative similarities between the target segments and the contexts make the complementary distribution natural in the sense that the connections between the target segments and the conditioning contexts, specifically the connections between [ɕ] and [i], may imply some sort of coarticulation (e.g., the palatality of [ɕ] has arisen as a result of coarticulation with the preceding [i]).

Here, the relative similarities between the target segments and the contexts may indirectly affect the learning of the allophonic relationship between [ʂ] and [ɕ]. As discussed earlier, studies show that the similarity between linguistic objects over which statistical dependencies are learned may affect the ease of learning (e.g., Newport and Aslin, 2004). Therefore, it is possible that the acoustic similarities between the target segments and the contexts may facilitate the learning of the complementary distribution. On the assumption that the learning of allophony crucially relies on the learning of complementary distribution, this may eventually facilitate the learning of the allophonic relationship. Alternatively, as discussed in the previous section, studies show that learners have learning biases that predispose them towards phonetically natural patterns (e.g., Wilson, 2006). Therefore, it

is possible that these learning biases may facilitate the learning of the phonetically natural complementary distribution and the allophonic relationship.

If the learning of allophony is constrained by phonetic naturalness, learners should be able to learn segments as allophones only when the segments are in phonetically natural complementary distribution. If phonetic naturalness has no effect on the learning, by contrast, learners should be able to learn segments as allophones when the segments are in complementary distribution no matter whether the pattern of the distribution is phonetically natural or not. Since the learning of an allophonic relationship between two segments from input with a phonetically-natural complementary distribution has been already tested in Experiment 1, I test whether adults can learn the same allophonic relationship from input with a phonetically unnatural complementary distribution in Experiment 2.

3.5 Methods

In Experiment 2, I tested the learning of the allophonic relationship between the retroflex [ʂ] and alveolopalatal [ɕ] from input in which these two target segments are in a phonetically unnatural complementary distribution; the tokens of the retroflex [ʂ] occurred after the high front unrounded [i] and the tokens of the alveolopalatal [ɕ] occurred after the high back rounded [u] (*complementary-unnatural condition*). If the learning is constrained by the phonetic naturalness of the patterns of complementary distribution, learners in the complementary-unnatural condition should not learn to treat the target segments as something like allophones, and thus should not show reduction in their sensitivity to acoustic differences between the segments.

The method was the same as in Experiment 1. There were two sessions over two consecutive days. In each session, I first exposed participants to input, and then tested their sensitivity to acoustic differences between the target segments using an AX discrimination paradigm.

3.5.1 Participants

Twenty adult native English speakers with no known language or hearing disorder participated in Experiment 2. All participants completed two sessions over

two consecutive days and were paid $20 for their participation. All participants reported that English was their first and dominant language. Many of them were multilingual but none of them was familiar with any language that has two or more post-alveolar fricatives as phonemes.

3.5.2 Exposure stimuli

Exposure stimuli consisted of 256 bisyllabic strings. Each string comprised a context syllable followed by either a critical syllable or a filler syllable. Syllables in the exposure stimuli were the same as the ones used in Experiment 1. There were eight context syllables grouped into two classes according to the vowel quality, [i] context ([li], [mi], [pi], and [gi]) and [u] context ([lu], [mu], [pu], and [gu]). There were eight critical syllables drawn from a 10-step continuum between [ʂa] and [ɕa] (steps 2, 3, 4, 5, 7, 8, 9, and 10). There were four filler syllables, [ta], [t^{h}a], [fa], and [ha], of which only [ta] and [t^{h}a] were used in exposure stimuli (i.e. [fa] and [ha] were used only in test stimuli).

The frequency of critical syllables was manipulated so that the aggregate distribution showed exactly the same bimodal shape with two frequency peaks as in Experiment 1 (Figure 3.1). These 16 tokens of critical syllables were combined with the eight context syllables to generate 128 critical stimuli. Similarly, 16 tokens of filler syllables (8 tokens of each of [ta] and [t^{h}a]) were combined with eight context syllables to generate 128 filler stimuli. These 256 exposure stimuli were divided into two subsets according to the consonants of context syllables, a subset with [l] and [p] and the other subset with [m] and [g]. The first subset was used in Session 1 and the second subset was used in Session 2.

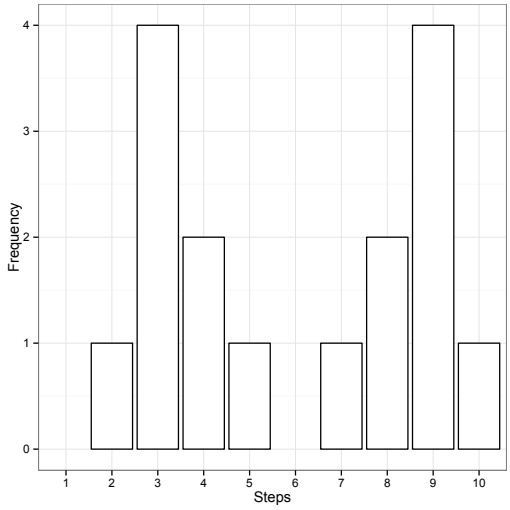


Figure 3.1: Aggregate distribution of critical syllables (Experiment 2)

The phonotactic distribution of the critical syllables was manipulated so that the tokens of the retroflex category (steps 2, 3, 4, and 5) occurred after [i], and the tokens of the alveolopalatal category (step 7, 8, 9, and 10) occurred after [u] (Figure 3.2). Table 3.1 summarizes exposure stimuli used in Experiment 2 alongside the ones used in Experiment 1.

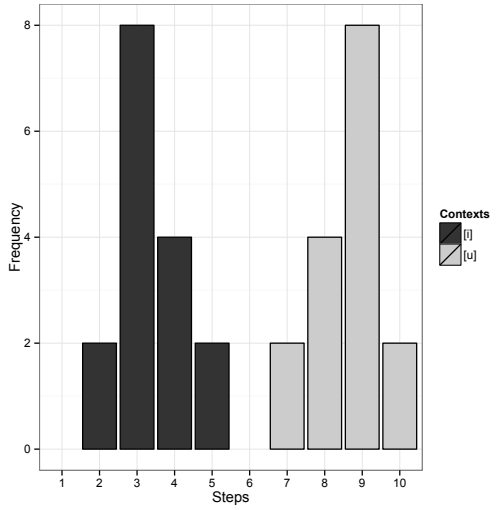


Figure 3.2: Distribution of 32 critical syllables in the complementary-unnatural condition (Experiment 2)

Condition	Context syllables	Critical syllables	Filler syllables
Non-complementary (Exp. 1)	li-, mi-, pi-, gi-	steps 2, 3, 4, 5, 7, 8, 9, 10	-ta, -t^{h}a
	lu-, mu-, pu-, gu-	steps 2, 3, 4, 5, 7, 8, 9, 10	-ta, -t^{h}a
Complementary (Exp. 1)	li-, mi-, pi-, gi-	steps 7, 8, 9, 10	-ta, -t^{h}a
	lu-, mu-, pu-, gu-	steps 2, 3, 4, 5	-ta, -t^{h}a
Complementary-unnatural (Exp. 2)	li-, mi-, pi-, gi-	steps 2, 3, 4, 5	-ta, -t^{h}a
	lu-, mu-, pu-, gu-	steps 7, 8, 9, 10	-ta, -t^{h}a
Control (Exp. 1)	li-, mi-, pi-, gi-	NA	-ta, -t^{h}a, -fa, -ha
	lu-, mu-, pu-, gu-	NA	-ta, -t^{h}a, -fa, -ha

Table 3.1: Exposure stimuli in four conditions (Experiments 1 and 2)

3.5.3 AX discrimination task

Participants' sensitivity to acoustic differences between the retroflex [ʂ] and alveolopalatal [ɕ] was tested using an AX discrimination paradigm. The test was identical to the one used in Experiment 1. In each one of the test trials, participants

compared a token of [ʂa] and a token of [ɕa] (different trials) or two non-identical tokens of [ʂa] or [ɕa] (same trials). There were two types of different trials depending on the acoustic distance between the test stimuli, distant pair and close pair. In the distant pair trials, participants compared two stimuli that are acoustically quite different from each other (step 2 and step 10). In the close pair trials, participants compared two stimuli that are acoustically more similar to each other (step 4 and step 8). There were also two types of same trials. In the retroflex category trials, participants compared two non-identical tokens of [ʂa] (step 2 vs. step 4). In the alveolopalatal category trials, participants compared two non-identical tokens of [ɕa] (step 8 vs. step 10). Table 3.2 summarizes the test stimuli used in the AX discrimination test.

Test trials	Different trials	Distant pair	step 2 vs. step 10
		Close pair	step 4 vs. step 8
	Same trials	Retroflex	step 2 vs. step 4
		Alveolo-palatal	step 8 vs. step 10
Filler trials	Different trials		ta vs. t^ha, fa vs. ha
	Same trials		ta vs. ta, t^ha vs. t^ha, fa vs. fa, ha vs. ha

Table 3.2: Test stimuli (Experiments 2: same as the ones used in Experiment 1)

3.5.4 Design and procedure

Experiment 2 followed the same design and procedure used in Experiment 1. It consisted of two sessions over two consecutive days. On Day 1, participants came into the lab and signed the consent form. At the beginning of the session, participants were told that they would hear someone speaking in a language that they had never heard before, and they would be asked about what they learned about the language after hearing the speech.[2] A session comprised three phases: practice, exposure, and test. In the practice phase, participants got some practice on the AX discrimination task. In the exposure phase, participants heard a block of 128 exposure stimuli presented in a random order four times. ISI was one second. The exposure stimuli were presented as short "sentences" so that each syllable in the

[2]See Section 2.3.6 for the actual instructions given to participants.

stimuli could be treated as a word. In order to help participants to stay attentive to exposure stimuli, they were given a monitoring task to perform while listening (see Section 2.3.5). In the test phase, there were 64 test trials. Half of the test trials were different trials, and the other half were same trials. The 32 different trials included 16 distant pair trials and 16 close pair trials. The 32 same trials included 16 retroflex category trials and 16 alveolopalatal category trials. There were 32 filler trials (16 different trials and 16 same trials). In each trial, ISI was 750 ms. Participants were given a maximum of five seconds to respond, but the trial was terminated whenever participants recorded a response. ITI was two seconds. On Day 2, participants came back to the lab and did Session 2. Session 2 followed the same procedure used in Session 1, except that participants filled out a language background questionnaire and received $20 after finishing the test phase. E-prime Professional (ver. 2.0) was used to control the presentation of stimuli and the recording of responses (Schneider et al., 2002).

3.6 Results

Before the analyses of the data, participants' performance on the monitoring task was checked to see whether they were attentive to the stimuli during exposure. All participants performed better than 75% on the monitoring task, and therefore they were all included in the following data analyses. Responses to test trials in the AX discrimination test were converted into sensitivity or d' scores (Macmillan and Creelman, 2004). d' scores were computed for each participant with each pair type (distant or close) in each session (1 or 2). In order to compute hit rates and false alarm rates, different trials were coupled with same trials that shared the same first syllable (the A of an AX pair) (see Section 2.4). Following standard practice, when hit rate or false alarm rate was 0, it was replaced by $\frac{1}{2N}$, and when hit rate was 1, it was replaced by $1 - \frac{1}{2N}$ (N=number of trials for a particular trial type) (Macmillan and Creelman, 2004, p.8). All of the following statistical analyses were done in R 3.0.3 (R Core Team, 2014).

Since the design of Experiment 2 was fully comparable to that of Experiment 1, the results of Experiment 2 were analyzed alongside the results of Experiment 1. Figure 3.3 shows the mean d' scores for distant pair trials by session (1 and

2) and condition (control (Exp. 1), non-complementary (Exp. 1), complementary-natural (Exp. 1), complementary-unnatural (Exp. 2)). Note that in order to make the distinction between the complementary condition in Experiment 1 and the complementary-unnatural condition in Experiment 2 clear, the former condition will be referred to as the complementary-natural conditon from now on. Figure 3.4 shows the mean d' scores for close pair trials by session and condition.

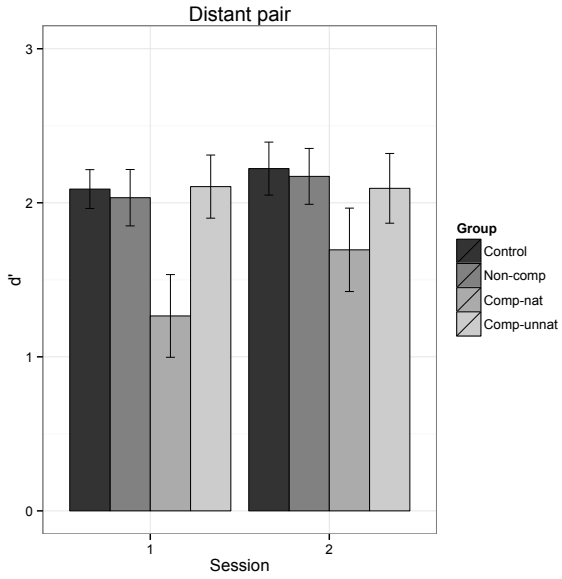


Figure 3.3: Mean d' scores with 2 SEs for distant pair (Experiments 1 and 2)

A repeated-measures ANOVA was conducted on the d' scores with condition as a between-participant variable and pair and session as within-participant variables. The significance level was set at $p < 0.05$. There were significant main effects of condition [$F(3,76) = 2.985,\ p = 0.036$], pair [$F(1,76) = 273.906,\ p < 0.001$], and session [$F(1,76) = 14.433,\ p < 0.001$]. There was a significant two-way interaction between pair and session [$F(1,76) = 5.624,\ p = 0.02$], but no significant two-way interactions between condition and pair [$F(3,76) = 1.154,\ p = 0.333$] nor between condition and session [$F(3,76) = 0.618,\ p = 0.605$]. There was no significant three-way interaction [$F(3,76) = 1.166,\ p = 0.328$].

With condition, post-hoc pairwise comparisons with the Holm adjustment

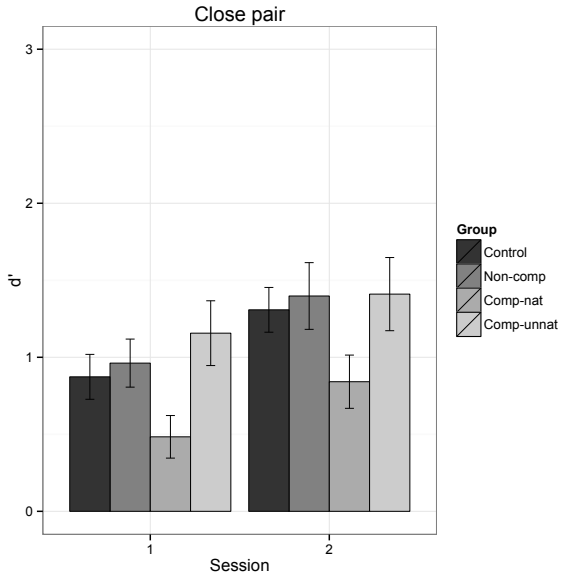


Figure 3.4: Mean d' scores with 2 SEs for close pair (Experiments 1 and 2)

method indicated that the d' scores of the complementary-natural condition ($M = 1.07$, $SD = 1.06$) were significantly lower than those of the control condition ($M = 1.62$, $SD = 0.86$) ($p = 0.001$), the non-complementary condition ($M = 1.64$, $SD = 0.95$) ($p = 0.001$), and the complementary-unnatural condition ($M = 1.69$, $SD = 1.05$) ($p < 0.001$), but there was no significant difference between the last three conditions.

Due to the interaction between pair and session, the effects of pair and session were explored separately in follow-up analyses. First, I explored the effect of pair for Session 1 and Session 2 separately. For each session, I conducted a separate repeated-measures ANOVAs on the d' scores with condition as a between-participant variable and pair as a within-participant variable. Since there was no significant three-way interaction, the effect of condition and its interactions with pair were not explored. For Session 1, the analysis yielded a significant main effect of pair [$F(1,76) = 190.519$, $p < 0.001$]. Post-hoc pairwise comparison indicated that the d' scores for distant pair trials ($M = 1.87$, $SD = 0.95$) were significantly higher than those for close pair trials ($M = 0.86$, $SD = 0.76$) ($p < 0.001$). For Ses-

sion 2, the analysis yielded a significant main effect of pair [$F(1,76) = 155.467$, $p < 0.001$]. Post-hoc pairwise comparison indicated that the d' scores for the distant pair trials ($M = 2.05$, 0.97) were significantly higher than those for the close pair trials ($M = 1.24$, $SD = 0.89$) (< 0.001).

Second, I explored the effect of session for distant pair trials and close pair trials separately. I conducted separate repeated-measures ANOVAs on the d' scores with condition as a between-participant variable and session as a within-participant variable. Since there was no significant three-way interaction, the effects of condition and its interactions with session were not explored. For distant pair trials, the analysis yielded no significant main effect of session [$F(1,76) = 3.606$, $p = 0.061$]. For close pair trials, the analysis yielded a significant main effect of session [$F(1,76) = 25.13$, $p < 0.001$]. Post-hoc comparison indicated that the d' scores for Session 2 ($M = 1.24$, $SD = 0.9$) were significantly higher than those for Session 1 ($M = 0.87$, $SD = 0.76$) ($p = 0.005$).

In sum, participants in the complementary-unnatural condition showed the same level of sensitivity as participants in the control and non-complementary conditions, and they showed significantly higher sensitivity than participants in the complementary-natural condition. Participants across conditions showed a significant improvement in sensitivity to the acoustic differences between the test stimuli used in close pair trials from Session 1 to Session 2.

3.7 Discussion

The crucial finding of Experiments 1 and 2 is that participants in the complementary-unnatural condition showed significantly better sensitivity to acoustic differences between the retroflex [ʂa] and alveolopalatal [ɕa] than participants in the complementary-natural condition. In these two conditions, participants were exposed to input in which [ʂ] and [ɕ] were in complementary distribution. The crucial difference between these two conditions was in the phonetic naturalness of the patterns of complementary distribution. In the complementary-natural condition, the target segments occurred in phonetically natural contexts. In the complementary unnatural condition, by contrast, the target segments occurred in phonetically unnatural contexts. The results of Experiments 1 and 2 together sug-

gest that allophonic relationships between two segments can be learned from input in which the segments occur in phonetically natural complementary contexts but not from input in which the segments occur in phonetically unnatural complementary contexts. This suggests that the learning of allophony, at least as indicated by a reduction in sensitivity to acoustic differences, is constrained by phonetic naturalness.

These findings add to the growing evidence for the effects of constraints on the learning of phonology (see Section 3.3). Previous studies about constraints on the learning of phonology have focused on the learning of phonological patterns, such as phonotactics and phonological alternations (including allophonic rules), but very little attention has been paid to constraints on the learning of sound categories (cf. Gilkerson, 2005). The findings of Experiments 1 and 2 provide the first evidence showing that the learning of allophony is constrained by the phonetic naturalness of the patterns of complementary distribution.

Now the question is how exactly phonetic naturalness affects the learning of allophony. In Section 3.4, I mentioned two hypotheses about the way phonetic naturalness may affect the learning of allophony. First, the acoustic similarities between the target segments and the natural contexts may facilitate the learning of the statistical dependencies between the target segments and the contexts, and thus the learning of the complementary distribution of the target segments (e.g., Newport and Aslin, 2004). On the assumption that the learning of allophony crucially relies on the learning of complementary distribution, this may indirectly facilitate the learning of the allophonic relationship between the target segments. Second, learners may have some learning biases that predispose them to favour those phonotactic patterns that are phonetically grounded (e.g., Wilson, 2006), and these biases may facilitate the learning of the complementary distribution, and thus the learning of the allophonic relationships.

The findings of Experiments 1 and 2 do not favour one or the other one of these interpretations. Therefore, the question of how exactly phonetic naturalness affects the learning of allophony remains open and further studies are needed. In the following subsections, I will propose yet another hypothesis about the way phonetic naturalness can affect the learning of allophony: *context effects hypothesis*. In this hypothesis, phonetic naturalness serves as the basis for perceptual biases

that affect the learning of sound categories in a more direct way. The hypothesis assumes that when learners hear sounds in different contexts, as was the case in the complementary-natural and complementary-unnatural conditions, they actually perceive the sounds differently (i.e., speech perception is subject to *context effects*: see below for details). More specifically, learners perceive two different sounds as being more similar to each other when they hear the sounds in phonetically natural complementary contexts than in phonetically unnatural complementary contexts.

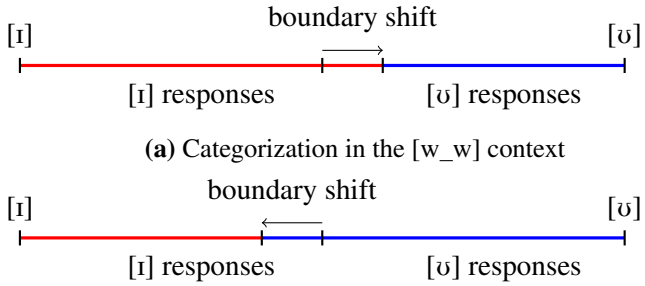
According to the distributional learning hypothesis (e.g., Maye 2000; see Section 1.4.1), learners form categories for speech sounds based on the frequency distributions of speech sounds in acoustic space. However, it is more precise to say that what they actually learn is the aggregate distribution of the sounds in auditory space. This is because their knowledge about the input is built upon their experience of the acoustic signal rather than the acoustic signal itself. This means that learners' knowledge about the aggregate distribution can be affected by the ways they perceive the sounds in the input. In the following subsections, I first explain context effects in speech perception and then how they can affect the learning of segmental categories by comparing the input used in the complementary-natural and complementary-unnatural conditions.

3.7.1 Context effects in speech perception

In the speech signal, acoustic cues for the perception of sounds are temporally distributed across segmental boundaries. For example, the perception of a stop consonant as either voiced or voiceless is cued not only by the acoustic properties of the stop consonant itself (e.g., VOT) but also by the acoustic properties of the neighbouring sounds (e.g., formant transitions into the following vowel) (e.g., Lisker, 1986). A major source of temporally distributed acoustic cues is *coarticulation*. When we produce speech, articulatory gestures required for the production of successive sounds overlap in time, and this gestural overlap systematically affects the acoustic realization of these sounds (Farnetani and Recasens, 1997). As a result, the portion of the acoustic signal that corresponds to a sound carries information about the identity of the sound as well as information about the coarticulatory effects of the neighbouring sounds. This means that acoustic cues for the perception

of a target sound are available not only from the portion of the acoustic signal that corresponds to the target sound itself but also from the portions of the acoustic signal that correspond to the neighbouring sounds because the coarticulatory effects in the neighbouring sounds provide information about the identity of the target sound that triggered the coarticulation.

Studies have demonstrated that the perception of sounds is significantly affected by manipulating the surrounding sounds (*context effects*) (Lindblom and Studdert-Kennedy, 1967; Mann, 1980; Mann and Repp, 1980; Repp and Mann, 1981; Repp, 1981; Summerfield, 1975). For example, Lindblom and Studdert-Kennedy (1967) demonstrated that the categorization of vowels is significantly influenced by consonantal contexts. In their study, adult English speakers categorized vowel tokens taken from a continuum between [ɪ] and [ʊ] occurring in two consonantal contexts: [j_j] and [w_w]. The results showed that participants were more likely to label intermediate tokens as [ɪ] in the [w_w] context and were more likely to label the same intermediate tokens as [ʊ] in the [j_j] context; the category boundary shifted towards the [ʊ] end in the [w_w] context and towards the [ɪ] end in the [j_j] context (see Figure 3.5).



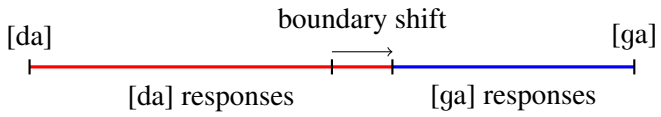
(a) Categorization in the [w_w] context

(b) Categorization in the [j_j] context

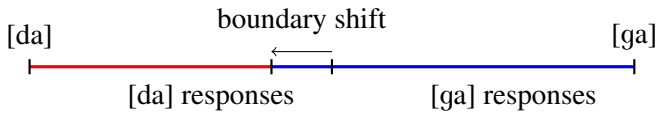
Figure 3.5: Context effects in the categorization of an [ɪ] - [ʊ] continuum

Similarly, Mann (1980) demonstrated that the categorization of consonants is significantly affected by consonantal contexts. In her study, adult English speakers categorized syllable tokens taken from a continuum between [da] and [ga] occurring in two consonantal contexts: [al_] and [aɹ_]. The results showed that participants were more likely to label intermediate tokens as [ga] in the [al_] context and

were more likely to label the same intermediate tokens as [da] in the [aɪ_] context; the category boundary shifted towards the [ga] end in the [aɪ_] context and towards the [da] end in the [al_] context (see Figure 3.6).



(a) Categorization in the [aɪ_] context



(b) Categorization in the [al_] context

Figure 3.6: Context effects in the categorization of a [da] - [ga] continuum

In the literature, there are at least two major approaches to the question of how exactly context effects arise: the articulatory approach and the auditory approach. The articulatory approach assumes that listeners know how the articulatory and acoustic realization of the same sound may vary in different contexts due to coarticulation and take the context-dependent variation into consideration in mapping the acoustic signal onto sound categories (*compensation for coarticulation*) (Lindblom and Studdert-Kennedy, 1967; Mann, 1980). There are two major theories in the articulatory approach: the *motor theory of speech perception* (Liberman and Mattingly, 1985; Liberman and Whalen, 2000) and the *direct realist theory* (Fowler, 1986, 1996, 2006). These theories both assume that the target of speech perception is the source of speech production; listeners perceive speech sounds by translating the acoustic signal into articulatory events. When listeners hear coarticulated sounds, they perceive the sounds as coarticulatory events or the interaction between the articulatory events for the target sounds and the neighbouring sounds. Therefore, depending on the articulatory events of the neighbouring sounds, the same acoustic signal may be mapped onto the articulatory events of different target sounds.[3]

[3]The major difference between these two theories is that the former assumes that the mapping of acoustic signal onto articulatory events is mediated by some sort of representation (e.g., features

According to the articulatory approach, for example, listeners in Mann (1980) know that when [d] is produced after [ɹ] the articulatory gestures for [d] and [ɹ] overlap in time. Since the place of articulation of [ɹ] is further back than the place of articulation of [d], the gestural overlap moves the place of articulation of [d] backwards. Listeners also know how this articulatory variation is reflected in the acoustic signal. Since [ɹ] has low F3, the gestural overlap results in the lowering of the onset F3 of [da]. Therefore, when they hear stimuli that are intermediate between [da] and [ga] (i.e., stimuli with too-low-to-be-[da] onset F3) after [ɹ], they are more likely to perceive the stimuli as [da] because they infer that the too-low-to-be-[da] onset F3 is a result of coarticulation with the preceding [ɹ], not an inherent property of the stimuli. Similarly, listeners know that when [g] is produced after [l], the gestures for [g] and [l] overlap in time. Since the place of articulation of [l] is further forward than the place of articulation of [g], the gestural overlap moves the place of articulation of [g] forward. They also know that this articulatory variation is reflected in the raising of the onset F3 of [ga]. Therefore, when they hear stimuli that are intermediate between [da] and [ga] (i.e., stimuli with too-high-to-be-[ga] onset F3) after [l], they are more likely to perceive the stimuli as [ga] because they infer that the too-high-to-be-[ga] onset F3 is a result of coarticulation with the preceding [l], not an inherent property of the stimuli.

The auditory approach assumes that context effects arise as a part of general contrast effects in perception. A dominant theory in this approach is the *spectral contrast theory* (Kluender et al., 2003; Lotto et al., 1997; Lotto and Kluender, 1998). It explains that context effects result from how the auditory system responds to changes in the signal. Rapid changes in the amplitude and the spectrum of the acoustic signal trigger an abrupt increase in the discharge rate of auditory nerve fibres (ANFs), but the peak in discharge rate is always followed by a gradual decay or adaptation. This happens because of the depletion of the neurotransmitter at the synapses between hair cells and ANFs in the cochlea (e.g., Smith, 1979). Adaptation happens in different frequency regions. For example, with vowel-like sounds,

and gestures) and that perception is achieved through *analysis-by-synthesis*: comparison between the incoming signal and candidates that are internally synthesized using these features and gestures. By contrast, the latter does not assume the necessity of these representations; perception is achieved by the direct mapping of acoustic signal onto articulatory events.

it happens at least in five different regions, low (below F1), F1, mid (between F1 and F2), F2, and high (above F2) (Delgutte and Kiang, 1984). One role of adaptation in speech perception is the enhancement of the spectral contrasts between successive sounds. Once ANFs are adapted, they become less responsive to the subsequent stimulation. However, the adaptation of some ANFs leads to relative exaggeration of the responses of the other unadapted ANFs to the subsequent stimulation. In this way, adaptation enhances the spectral contrasts between successive sounds (Delgutte, 1980, 1997; Delgutte and Kiang, 1984).

According to the spectral contrast theory, context effects happen because neural adaptation to the spectrum of the precursor sound modulates the perception of the spectrum of the following sound. For example, when listeners in Mann (1980) heard an intermediate syllable after [ɹ], adaptation happened with the ANFs that are responsive to the frequency range around the F3 of [ɹ], and this adaptation modulated the perception of the spectrum of the following syllable such that the weight of the spectrum shifted toward the higher end since [ɹ] has low F3. As a result, the syllables were more likely to be perceived as [da]. Similarly, when listeners heard the same intermediate syllables after [l], adaptation happened with the ANFs that are responsive to the frequency range around the F3 of [l]. This adaptation modulated the perception of the spectrum of the following syllables such that the weight of the spectrum shifted towards the lower end since [l] has high F3. As a result, the syllables were more likely to be perceived as [ga].

In sum, listeners integrate temporally distributed acoustic cues in speech perception. Therefore, the perception of speech sounds is influenced by context in systematic ways. Specifically, listeners may perceive the same stimuli as different sounds when the stimuli are presented in different contexts. In the next subsection, I will discuss how the context-dependent perception of sounds may impact the distributional learning of sound categories.

3.7.2 Context effects and the distributional learning of sound categories

According to the distributional learning hypothesis, learners form categories for sounds based on their knowledge about the frequency distributions of the sounds in acoustic space. However, it is more precise to say that what they actually learn

through exposure to the input is the aggregate distribution of the sounds in auditory space, not acoustic space. This is because learners' knowledge about the input is built upon their experience or perception of the acoustic signal in the input.

In previous experimental studies of distributional learning, participants in experimental conditions were exposed to input in which sounds were presented in the same, that is overlapping, environments. It was assumed that participants in experimental conditions perceived the sounds as being the same in all environments, and thus, the only difference between the conditions was in the frequencies of the sounds (e.g., Maye and Gerken, 2000). This is a perfectly safe assumption in previous studies. However, the mapping of sounds from acoustic space to auditory space is not trivial when it comes to the input used in Experiments 1 and 2. In the input used in the experimental conditions (overlapping, complementary-natural, complementary-unnatural), the frequency distribution of novel sounds had the same bimodal shape in acoustic space. However, the input differed in these conditions in terms of the phonotactic distribution of the sounds. This means that participants in these conditions heard the same sounds but in different sets of contexts. As discussed above, perception of speech sounds is affected by context such that the same stimulus may be perceived as different sounds when it is presented in different contexts. Therefore, when learners in different conditions hear the same sounds in different contexts, these sounds may be mapped onto different locations in auditory space. This perceptual shift in the locations of the sounds may eventually result in different shapes of the aggregate distribution of the sounds in auditory space in different conditions.

In what follows, I will explain how context effects in speech perception can affect distributional information that learners extract from input, and how that can, in turn, affect the learning of categories. To do so, I will compare the input used in the complementary-natural and complementary-unnatural conditions, so that I can highlight the difference between the phonetically natural complementary distribution and the phonetically unnatural complementary distribution.

Before going into the details of possible context effects in the perception of the novel sounds in phonetically-natural and phonetically-unnatural contexts, let us recapitulate the phonetic properties of the target segments (the retroflex [ʂ] and alveolopalatal [ɕ]) and the contexts ([i] and [u]). The articulation of Mandarin

[ʂ] involves the formation of a short and slack constriction channel in the post-alveolar region (e.g., Ladefoged and Wu, 1984; Lee-Kim, 2014; Noguchi et al., 2015a; Proctor et al., 2012; Toda and Honda, 2003). By contrast, the articulation of Mandarin [ɕ] involves the formation of a long and narrow constriction channel over the post-alveolar and palatal regions. The formation of a palatal constriction is achieved by moving the tongue body forward and upward (Hu, 2008; Ladefoged and Wu, 1984; Lee, 2008; Lee-Kim, 2014; Proctor et al., 2012; Toda and Honda, 2003). The presence of a palatal constriction in the articulation of [ɕ] is reflected in the acoustic signal as low F1 and high F2 at the onset of the following vowel. The articulation of the high front unrounded [i] involves the forward and upward movements of the tongue body, and the acoustics of [i] are characterized by low F1 and high F2. The articulation of the high back rounded [u] involves the backward and upward movements of the tongue body, and the acoustics of [u] are characterized by low F1 and low F2.

To highlight the relationship between the target segments and the contexts, Figure 3.7 shows F2 frequencies of the vowels in the context syllables and critical syllables used in the input of Experiments 1 and 2. Figure 3.7 (a) shows mean F2 frequencies at the 100 ms (50%) and 190 ms (95%) points of the vowels in the context syllables. The mean F2 frequencies of [i] are higher than the mean F2 frequencies of [u]. Figure 3.7 (b) shows F2 frequencies at the 10 ms (5%) and 100 ms (50%) points of the vowels in two critical syllables: step 2 (the canonical token of the retroflex [ʂa]) and step 10 (the canonical token of the alveolopalatal [ɕa]). The onset F2 frequency of the alveolopalatal token is higher than the onset F2 frequency of the retroflex token. Note that mean F2 frequency at the offset of [i] is higher than onset F2 frequencies of both retroflex and alveolopalatal tokens, and mean F2 frequency at the offset of [u] is lower than onset F2 frequencies of both retroflex and alveolopalatal tokens.

The contrast between the retroflex [ʂ] and alveolopalatal [ɕ] is also made by differences in the spectral shape of the frication noise. Here, I focus on the formant transitions for two reasons. First, from the learners' perspective, adult English speakers seem to rely more on formant transitions in learning non-native post-alveolar fricatives (McGuire, 2007a,b, 2008). Second, studies on context effects have demonstrated that the categorization of CV syllables from a continuum be-

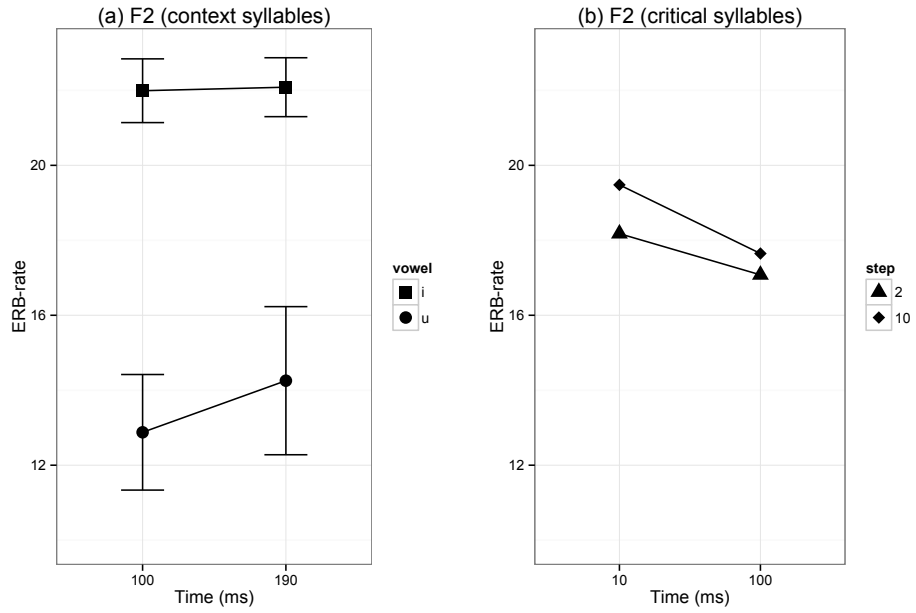


Figure 3.7: F2 transitions from context syllables to critical syllables

tween [ba] and [da], in which the onset F2 was systematically manipulated, was significantly affected by the preceding vowel (Coady et al., 2003; Holt, 1999). These results suggest that the interaction of non-adjacent formant cues is robust.

In the input used in the complementary-natural condition, the tokens of the retroflex [ʂa] occur after [u] and the tokens of the alveolopalatal [ɕa] occur after [i]. According to the articulatory approach, when participants hear a token of the retroflex [ʂa] after [u], they may infer that there is coarticulation between the unfamiliar syllable and the preceding vowel [u]. More specifically, they may infer that the onset F2 of the unfamiliar syllable has been lowered as a result of the coarticulation, and thus that the inherent onset F2 of the unfamiliar syllable is higher than it is in the acoustic signal. This makes the unfamiliar syllable seem more like [ɕa] on the [ʂa] - [ɕa] continuum. Similarly, when participants hear a token of the alveolopalatal [ɕa] after [i], they may infer that the onset F2 of the unfamiliar syllable has been raised as a result of coarticulation with the preceding vowel [i], and thus, the inherent onset F2 of the unfamiliar syllable is lower than it is in the acoustic

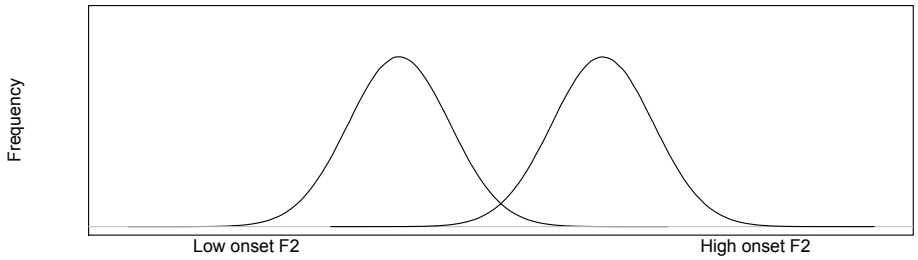
signal. This makes the unfamiliar syllable seem more like [ʂa] on the [ʂa] - [ɕa] continuum.

According to the auditory approach, when participants hear a token of the retroflex [ʂa] after [u], their perceptual system first adapts to the frequency region around the F2 of [u], and this adaptation exaggerates the relative height of the onset F2 of the following unfamiliar syllable. This makes the unfamiliar syllable seem more like [ɕa]. Similarly, when participants hear a token of the alveolopalatal [ɕa] after [i], their perceptual system first adapts to the frequency region around the F2 of [i], and this adaptation exaggerates the relative lowness of the onset F2 of the following unfamiliar syllable. This makes the unfamiliar syllable seem more like [ʂa].

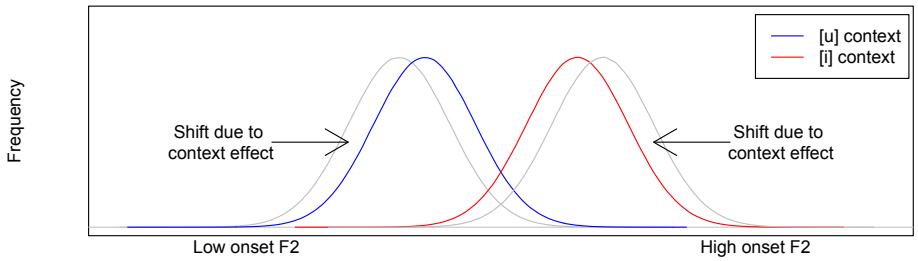
What is crucial here is that both the articulatory and auditory approaches predict that the presentation of these unfamiliar syllables in phonetically natural contexts modulates the perceived onset F2 of the syllables. The onset F2 of the retroflex token is perceived to be higher than it is in the acoustic signal, and the onset F2 of the alveolopalatal token is perceived to be lower than it is in the acoustic signal.

Figure 3.8 demonstrates how these context effects may affect the aggregate distribution in auditory space. Figure 3.8a shows a schematic representation of the frequency distribution of the unfamiliar syllables in acoustic space in the input used for the complementary-natural condition. We can see that the distribution has two peaks, implying that the syllables are classified into two categories, the retroflex [ʂa] to the lower end and the alveolopalatal [ɕa] to the higher end of the onset F2 continuum. Figure 3.8b shows how context effects may affect the perception of these unfamiliar syllables when they are presented in phonetically natural contexts, the tokens of [ʂa] after [u] and the tokens of [ɕa] after [i]. As discussed above, the context effects may shift the location of the tokens of [ʂa] to the higher end of the onset F2 continuum and the location of the tokens of [ɕa] to the lower end of onset the F2 continuum in auditory space. Figure 3.8c shows the hypothetical representation of the aggregate distribution of these unfamiliar syllables in auditory space. As a result of the context effects, two frequency peaks shows a large overlap, and the aggregate distribution has a shallower trough between the peaks than the bimodal distribution in acoustic space.

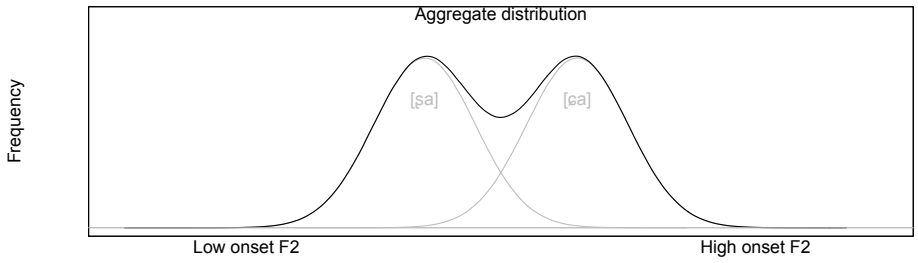
In the input used in the complementary-unnatural condition, by contrast, the



(a) Bimodal distribution in acoustic space



(b) Context effects in complementary natural contexts



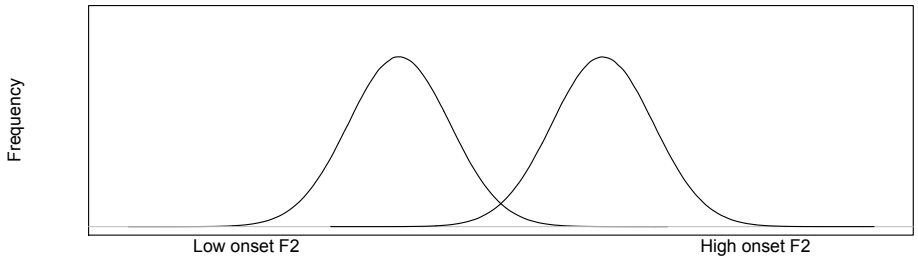
(c) Aggregate distribution in auditory space

Figure 3.8: Complementary-natural condition

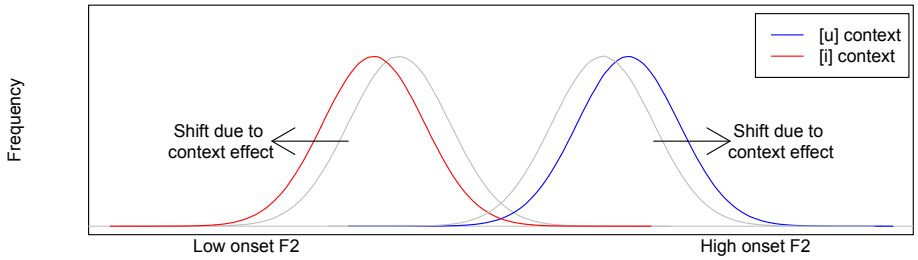
tokens of the retroflex [ʂa] occur after [u], and the tokens of the alveolopalatal [ɕa] occur after [i]. According to the articulatory approach, when participants hear a token of [ʂa] after [i], they may infer that the onset F2 of the unfamiliar syllable has been raised due to coarticulation with the preceding [i], and thus, they may infer that the inherent onset F2 of the unfamiliar syllable is lower than it is in the acoustic signal. Similarly, when participants hear a token of the alveolopalatal [ɕa] after [u], they may infer that the onset F2 of the unfamiliar syllable has been lowered due to coarticulation with the preceding [u], and thus, they infer that the onset F2 of the unfamiliar syllable is higher than it is in the acoustic signal.

According to the auditory approach, when participants hear a token of retroflex [ʂa] after [i], their perceptual system first adapts to the frequency region around the F2 of [i], and this adaptation exaggerates the relative lowness of the onset F2 of the following unfamiliar syllable; the onset F2 is perceived to be lower than it is in the acoustic signal. Similarly, when participants hear a token of the alveolopalatal [ɕa] after [u], their perceptual system first adapts to the frequency region around the F2 of [u], and this adaptation exaggerates the relative height of the onset F2 of the following unfamiliar syllable; the onset F2 is perceived to be higher than it is in the acoustic signal.

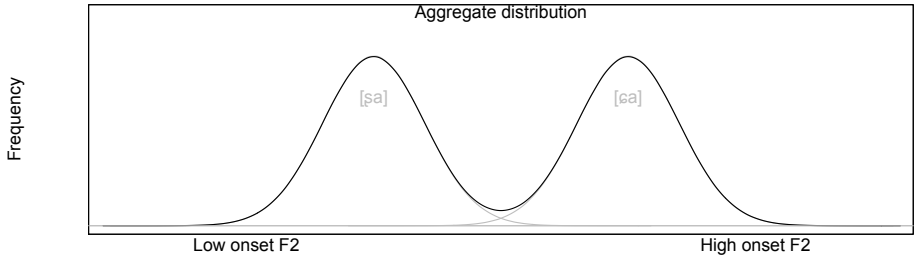
Figure 3.9 demonstrates how these context effects may affect th aggregate distribution in auditory space. Figure 3.9a shows a schematic representation of the frequency distribution of the unfamiliar syllables in acoustic space in the input used for the complementary-unnatural condition. The input implies that the syllables are classified into the retroflex and alveolopalatal categories. Figure 3.9b shows how context effects may affect the perception of these unfamiliar syllables when they are presented in phonetically unnatural contexts, the tokens of [ʂa] after [u] and the tokens of [ɕa] after [i]. As discussed above, the context effects may shift the location of the tokens of [ʂa] to the lower end of the onset F2 continuum and the location of the tokens of [ɕa] to the higher end of the onset F2 continuum in auditory space. Figure 3.9c shows the hypothetical representation of the aggregate distribution of these unfamiliar syllables in auditory space. As a result of the context effects, the overlap between the frequency peaks is very small, and the aggregate distribution has two peaks that are clearly distinguishable from each other.



(a) Bimodal distribution in acoustic space



(b) Context effects in complementary unnatural contexts



(c) Aggregate distribution in auditory space

Figure 3.9: Complementary-unnatural condition

If the context effects happen as they were described above, the input differs in terms of the shape of the aggregate distribution in auditory space between the complementary-natural and complementary-unnatural conditions. The aggregate distribution may be bimodal in both conditions, but the trough between the peaks is shallower in the input used in the complementary-natural condition (Figure 3.8c) than in the input used in the complementary-unnatural condition (Figure 3.9c). If learners form categories for speech sounds based on the shape of the aggregate distribution of the sounds in auditory space, the categories that the learners in the complementary-natural and complementary-unnatural conditions should be different from each other. Specifically, the learners in the complementary-natural condition learn an aggregate distribution in which two frequency peaks are more closely spaced and thus less distinct. As a result, the learners are more likely to interpret the distribution as unimodal, representing a single category. By contrast, the learners in the complementary-unnatural condition learn an aggregate distribution in which two peaks are distantly spaced and more distinct. As a result, the learners are more likely to interpret the distribution as bimodal, representing two categories.

These differences may explain the outcomes of the complementary-natural and complementary-unnatural conditions in Experiments 1 and 2. The participants in the complementary-natural condition were more likely to learn a single category for the novel fricative sounds and therefore to become less sensitive to acoustic differences between [ʂ] and [ɕ]. By contrast, the participants in the complementary-unnatural condition were more likely to learn two separate categories for the novel fricative sounds and therefore to maintain their sensitivity. In other words, the input used in the complementary-natural condition provided less robust distributional cues for the learning of the two categories than the input used in the complementary-unnatural condition.

This hypothesis may explain the difference between the complementary-natural and complementary-unnatural conditions. However, it does not explain the absence of difference between the complementary-unnatural and non-complementary conditions. The results of Experiments 1 and 2 demonstrated that participants in these two conditions showed the same level of sensitivity to acoustic differences between [ʂ] and [ɕ] after exposure. However, the input used in the non-complementary condition contained the novel sounds occurring in both natu-

ral and unnatural contexts. In other words, half of the exposure stimuli provided less robust cues for the learning of two categories, and the other half provided more robust cues for the learning of two categories. This means that the non-complementary condition should be intermediate between the complementary-natural and complementary-unnatural conditions in terms of the learnability of the phonetic contrast between [ʂ] and [ɕ]. However, that was not the case.

A possible explanation for the absence of difference between the complementary-unnatural and non-complementary conditions is the learning of the predictable occurrences of the target segments. Despite the fact that the input used in the complementary-unnatural condition provided more robust distributional cues for the learning of two categories, it did not help learners to improve their sensitivity to acoustic differences between the target segments because they were also detecting the predictable occurrences of the target segments at the same time. In other words, the aggregate distribution of the novel fricative sounds in auditory space could have helped learners to improve their sensitivity to acoustic differences between the target segments, but the learning of the predictable occurrences of the target segments wiped out the effect of the aggregate distribution.

Another explanation is that the context-dependent perception of critical syllables was the primary factor that determined the learners' sensitivity to acoustic differences between the target segments, but learners in the non-complementary and complementary-unnatural conditions were just showing a ceiling effect. Since participants in the control condition showed fairly good sensitivity to acoustic differences between the retroflex [ʂ] and alveolopalatal [ɕ] (especially when they compared the canonical tokens of these segments), it is possible that the amount of exposure was not enough for learners in the non-complementary and complementary-natural conditions to improve their sensitivity.

3.8 Conclusion

In this chapter, I presented the results of Experiments 2 and compared these to the results of Experiment 1. The results suggest that adults can learn the allophonic relationship between two novel segments, the retroflex [ʂ] and the alveolopalatal [ɕ], only when they are exposed to input in which the tokens of these segments oc-

cur in phonetically natural complementary contexts, but not when they are exposed to input in which the tokens of these segments occur in complementary contexts that are phonetically natural. These findings indicate that the learning of allophony is constrained by the phonetic naturalness of the patterns of complementary distribution. In order to account for the role of phonetic naturalness, I proposed a hypothesis about the mechanisms behind the learning of allophony (the context effects hypothesis). According to this hypothesis, learners in the complementary-natural condition became less sensitive to acoustic differences between [s̪] and [ɕ] because they actually learned less distinct categories or even a single category, and this happened through the the learning of the aggregate distribution of the novel fricative sounds with less distinct peaks in auditory space as a result of the context-dependent perception of the novel fricative sounds.

Chapter 4

Experiment 3: Learning of context-dependent perception of novel sounds

4.1 Introduction

In Chapter 3, I proposed a hypothesis about the mechanisms behind the learning of allophony. The hypothesis is that the context-dependent perception of sounds in input affects the aggregate distribution of the sounds in auditory space. Specifically, when learners hear tokens of two target segments in phonetically natural complementary contexts, they perceive the tokens as being more similar to each other than they actually are. This affects the shape of the aggregate distribution of the sounds in auditory space such that the distance between two distributional peaks becomes smaller. This eventually leads to the learning of less distinct categories or even a single category.

A crucial assumption that underlies the hypothesis is that the perception of sounds, no matter whether they are familiar or novel, is affected by context. There is a large body of evidence for context effects in the perception of sounds in listeners' native languages (e.g., Lindblom and Studdert-Kennedy, 1967; Mann, 1980; Mann and Repp, 1980; Repp and Mann, 1981; Repp, 1981; Summerfield, 1975).

However, our knowledge about context effects in the perception of novel (or non-native) sounds is still limited. Experiment 3 tests whether the perception of the retroflex [ʂ] and alveolopalatal [ɕ] by adult English speakers is affected by context in the ways that were discussed in Chapter 3. Specifically, it tests whether adult English speakers perceive the retroflex [ʂ] and alveolopalatal [ɕ] as being more similar to each other when they hear the sounds in phonetically natural complementary contexts than in phonetically unnatural complementary contexts.

Hao (2012) reported that naive English speakers categorized the alveolopalatal fricative in [ɕy] as English palato-alveolar /ʃ/ most of the time, but they categorized the alveolopalatal fricative in [ɕi] either as English alveolar /s/ or palato-alveolar /ʃ/ (see Table 2.2 in Chapter 2). What is interesting in these results is that naive English speakers perceived the alveolopalatal [ɕ] as being less palatalized (i.e., as /s/) before a high front unrounded vowel [i]. This suggests that they might be inferring that the palatality in the alveolopalatal [ɕ] is a result of coarticulation with the following [i] and are compensating for the coarticulation.

The question of how the perception of novel (or non-native) sounds is affected by context relates to the question of how much context effects depend on listeners' experience with specific languages. Some researchers have argued that context effects are based on listeners' general knowledge about acoustic and articulatory events (Mann, 1986; Viswanathan et al., 2010). For example, Mann (1986) demonstrated that Japanese speakers showed context effects in the categorization of a [da]-[ga] continuum when the stimuli were presented after English liquids [l] and [ɹ]; they were more likely to label intermediate tokens as [ga] after [l] and were more likely to label the same intermediate tokens as [da] after [ɹ]. Given the fact that Japanese speakers have great difficulty in categorizing these English liquid sounds (e.g., Miyawaki et al., 1975), the finding that these English sounds significantly affected Japanese speakers' categorization of the [da]-[ga] continuum was striking. According to Mann (1986), even though Japanese speakers do not have categories for these English liquid sounds, they still have access to phonetic information (e.g., formants and underlying articulatory events) from the liquid sounds and integrate the information into the perception of the stimuli from the [da]-[ga] continuum. Therefore, context effects do not necessarily rely on listeners' knowledge about the categories of sounds in specific languages.

Some researchers have further argued that context effects are based on general contrast effects in perception (Kluender et al., 2003; Lotto et al., 1997; Lotto and Kluender, 1998). In this view, context effects in speech perception arise because of the way the auditory system works. For example, Lotto and Kluender (1998) demonstrated that English speakers showed context effects in the categorization of a [da]-[ga] continuum when the stimuli were presented after the frequency modulated (FM) glides that mimicked the trajectories of the F3 of English [l] and [ɹ]. This suggests that context effects do not necessarily rely on the linguistic processing of speech sounds.

Other researchers have argued that there is some language-specificity in context effects. For example, Beddor et al. (2002) compared native speakers of English and Shona in the categorization of vowels. According to Beddor et al., both languages show non-local vowel-to-vowel coarticulation. In English, the directionality of coarticulation is symmetric; both anticipatory and carryover coarticulations happen with the same magnitude. In Shona, by contrast, the directionality is asymmetric; anticipatory coarticulation is stronger than carryover coarticulation. In perception, while English speakers showed the same amount of context effects with anticipatory and carryover coarticulations, Shona speakers showed stronger context effects with anticipatory coarticulation. These results suggest that language-specific phonetic knowledge plays a role in context effects.

If context effects are based on listeners' general knowledge about articulatory and acoustic events or general contrast effects in speech perception, the perception of non-native sounds should be affected by context in ways that are predicted by the kinds of information that are available in the acoustic signal. If context effects are based on listeners' knowledge about specific languages, the perception of non-native sounds should be affected by context only after exposure to input.

As discussed in Chapter 3, both the articulatory and auditory theories of context effects predict that English speakers would perceive the retroflex [ʂ] and the alveolopalatal [ɕ] as being more similar to each other when they hear these sounds in phonetically natural contexts than in phonetically unnatural contexts. If the context-dependent perception of these novel sounds is something that requires learning, the asymmetry between the phonetically natural contexts and phonetically unnatural contexts would emerge only after exposure to input that supports

such learning.

4.2 Methods

The experiment consisted of two sessions over two consecutive days. In Session 1, participants first did a similarity rating task and then heard input stimuli. In Session 2, the order was reversed. They first heard the exposure stimuli and then performed the similarity rating task. The fact that testing was carried out both pre- and post-exposure allows us to assess the effects of learning through exposure.

4.2.1 Participants

Twenty native speakers of English with no identified language or hearing disorder took part in the experiment. All participants were undergraduate students enrolled in linguistics courses at The University of British Columbia, and they received course credit for participation. All participants self-reported that English was their first and dominant language. Most of them were multilingual, but none of them were familiar with any language that contained two or more post-alveolar sibilants as phonemes.

4.2.2 Exposure stimuli

The exposure stimuli used in Experiment 3 were identical to the ones used in the non-complementary condition in Experiment 1 (see Section 2.3.2). They consisted of 256 bisyllabic strings. Each string comprised a context syllable followed by a critical syllable or a filler syllable. There were eight context syllables, [li], [lu], [mi], [mu], [pi], [pu], [gi], and [gu]. Context syllables were grouped into two classes according to the vowel quality; syllables with the high front unrounded vowel [i] ([i] context) and syllables with the high back rounded vowel [u] ([u] context). There were eight critical syllables taken from a 10-step continuum between the natural productions of the Mandarin retroflex [ʂa] and alveolopalatal [ɕa]. Based on the categorization of the continuum by native speakers of Mandarin, four steps from each side of the category boundary (step 6) were used as critical syllables: steps 2-5 for the retroflex [ʂ] and steps 7-10 for the alveolopalatal [ɕ] (Noguchi and Hudson Kam 2015a; also see Appendix A). The frequencies of

critical syllables were manipulated so that their aggregate distribution showed a bimodal shape with two peaks (Figure 4.1). There were two filler syllables, [t^ha] and [ta].

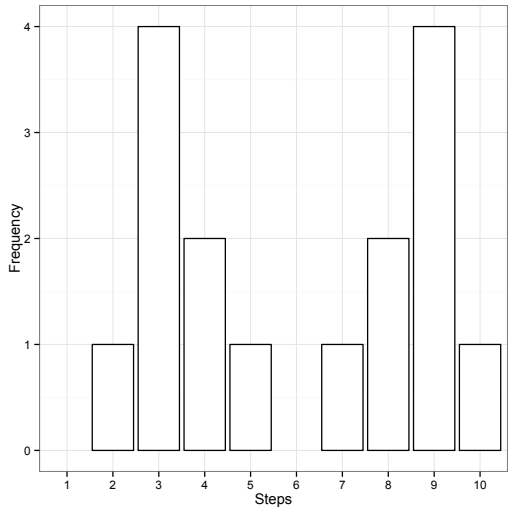


Figure 4.1: Aggregate distribution of critical syllables (Experiment 3)

Exposure stimuli were constructed by concatenating context syllables with either critical syllables or filler syllables. The sixteen tokens of critical syllables shown in Figure 4.1 were combined with 8 context syllables to generate 128 critical stimuli. Similarly, 16 tokens of filler syllables (8 tokens of [t^ha] and [ta]) were combined with 8 context syllables to generate 128 filler stimuli. These 256 exposure stimuli were divided into two subsets according to the consonant of context syllables, a subset with [l] and [p] and the other subset with [m] and [g]. The first subset was used in Session 1 and the second subset was used in Session 2. In this input, all of the critical syllables occurred in overlapping contexts. This means that tokens of the retroflex [ʂ] occurred in both the natural context ([u] context) and the unnatural context ([i] context), and the tokens of the alveolopalatal [ɕ] occurred in both the natural context ([i] context) and the unnatural context ([u] context). Figure 4.2 shows the 32 tokens of critical syllables where the same number of tokens of each step occur in the [i] context and the [u] context.

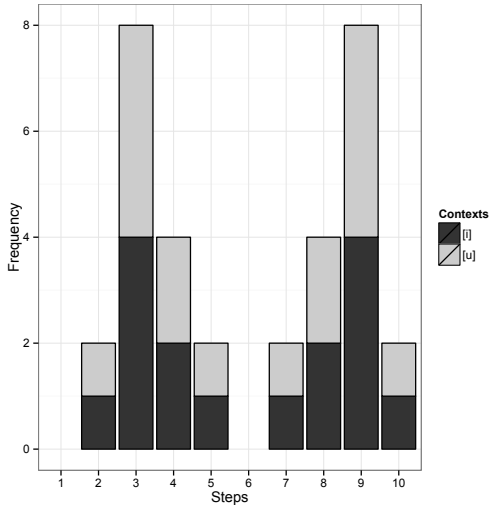


Figure 4.2: Distribution of 32 critical syllables (Experiment 3)

4.2.3 Test stimuli

Stimuli for the similarity rating task were bisyllabic strings. The first syllable was either [i] or [u] with no onset consonant. The second syllable was either a critical syllable or a filler syllable. The critical syllables were taken from the end points of the continuum used in the exposure stimuli, step 2 for the retroflex [ʂa] and step 10 for the alveolopalatal [ɕa]. The filler syllables were [t^ha] and [ta].

Stimuli were paired for the similarity rating task. For each pair, the second syllables were either the same syllable ([V.CV_1]–[V.CV_1]) or two different syllables ([V.CV_1]–[V.CV_2]). Pairs with different critical syllables were classified into three groups according to the type of context: overlapping, complementary-natural, and complementary-unnatural. In pairs with overlapping contexts, the critical syllables were presented in the same context, either after [i] or [u] (e.g., i-step 2 and i-step 10). In pairs with complementary-natural contexts, the step 2 (retroflex) syllable was presented after [u] and the step 10 (alveolopalatal) syllable was presented after [i] (u-step 2 and i-step 10). In pairs with complementary-unnatural contexts, the step 2 (retroflex) syllable was presented after [i] and the step 10 (alveolopalatal) syllable was presented after [u] (i-step 2 and u-step 10). Note that in the presentation of different critical syllables in overlapping contexts one of the critical sylla-

bles was in the natural context and the other was in the unnatural context (e.g., [u] is natural for the step 2 syllable but unnatural for the step 10 syllable).

Pairs with the same critical syllables were classified into two groups according to the types of context as well: overlapping and complementary. In pairs with overlapping contexts, the same syllable was presented twice in the same context, either after [i] or [u] (e.g., i-step 2 and i-step 2). In these pairs, a single critical syllable was presented twice, both times either in the natural (e.g., u-step 2 and u-step 2) or the unnatural context (e.g., i-step 2 and i-step 2). In pairs with complementary contexts, by contrast, one was presented after [i] and the other was presented after [u] (e.g., i-step 2 and u-step 2). In these pairs, a single critical syllable was presented twice, once in the natural context and the other time in the unnatural context. Finally, pairs with filler syllables were classified into two groups according to the type of context, overlapping and complementary. Table 4.1 summarizes the paired stimuli used in the test.

Table 4.1: Stimuli for similarity rating task

2nd syllable	Trial type	Context	Pairs	Naturalness
Critical	Different	Overlapping	i-step 2 and i-step 10	Natural for step 10
Critical	Different	Overlapping	u-step 2 and u-step 10	Natural for step 2
Critical	Different	Complementary-natural	u-step 2 and i-step 10	Natural for both
Critical	Different	Complementary-unnatural	i-step 2 and u-step 10	Unnatural for both
Critical	Same	Overlapping	u-step 2 and u-step 2 i-step 10 and i-step 10	Natural for both
Critical	Same	Overlapping	i-step 2 and i-step 2 u-step 10 and u-step 10	Unnatural for both
Critical	Same	Complementary	u-step 2 and i-step 2 u-step 10 and i-step 10	Natural for one stimulus
Filler	Different	Same	i/u-t^{h}a and i/u-ta	N/A
Filler	Different	Different	u-t^{h}a and i-ta i-t^{h}a and u-ta	N/A
Filler	Same	Same	i/u-t^{h}a and i/u-t^{h}a i/u-ta and i/u-ta	N/A
Filler	Same	Complementary	i-t^{h}a and u-t^{h}a i-ta and u-ta	N/A

The context effects hypothesis makes different predictions for the different trial types. For pairs with different critical syllables, the perceived similarity between the critical syllables should be high in complementary-natural contexts. This is because context effects would shift the perceived locations of the critical syllables on the [ʂa]-[ɕa] continuum inwards when they were heard in complementary-natural contexts. By contrast, the perceived similarity between the critical syllables should be low in complementary-unnatural contexts. This is because context effects would shift the perceived locations of the critical syllables outwards when they were heard in complementary-unnatural contexts (see Section 3.7.2). Finally, the perceived similarity between critical syllables should be intermediate in overlapping contexts. This is because context effects would shift the perceived loca-

tions of the critical syllables in the same direction; there would be no change in the perceived distance between the critical syllables. However, the presentation of different critical syllables in overlapping contexts might actually highlight the acoustic discrepancy between the syllables instead. If that is the case, the perceived similarity between the critical syllables should be low in overlapping contexts.

For pairs with the same critical syllables, a single critical syllable should be perceived differently in different contexts. This is because one of the repetitions was presented in the natural context and the other was presented in the unnatural context. In such a case, context effects would shift the location of a single critical syllable in opposite directions along the continuum, analogous to the demonstration of context effects in categorization in original studies (e.g., Lindblom and Studdert-Kennedy, 1967; Mann, 1980).

4.2.4 Design

E-prime Professional (ver. 2.0) was used to control the presentation of stimuli and the recording of responses (Schneider et al., 2002). A session consisted of two phases, exposure and test. In Session 1, the test phase was followed by the exposure phase. In Session 2, the exposure phase was followed by the test phase. In the exposure phase, participants heard a block of 128 stimuli presented in a random order with one second ISI. They heard the block four times. Exposure stimuli were presented as "words" in a foreign language. In order to help participants stay attentive to the stimuli, a monitoring task was given to them. In each block of stimuli presentation, a monitoring stimulus (a filler stimulus with long vowels:e.g., [li:ta:]) was randomly inserted in every subblock of 16 presentations. Participants were asked to press the spacebar when they heard the instances of "slow speech".

In the test phase, in each trial, participants heard a pair of bisyllabic strings and were asked to rate the similarity between the second syllables of the strings on a scale of 1 to 7, where 1 means "very similar" and 7 means "very different". Emphasis was placed on the point that they should compare the second syllables and not the whole (two-syllable) strings. They were also encouraged to use all points on the scale when rating the similarity. ISI was 750 ms. Participants were given a maximum of five seconds to respond, but the trial was terminated whenever

participants recorded a response. ITI was two seconds.

There were 48 test trials (trials with critical syllables): 24 with different critical syllables and 24 with the same critical syllables. There were eight trials with different critical syllables in overlapping contexts (two orders of presentation, two contexts, and two repetitions), eight trials with different critical syllables in complementary-natural contexts (two orders of presentation and 4 repetitions), and eight trials with different critical syllables in the complementary-unnatural contexts (two orders of repetition and 4 repetitions). Similarly, there were eight trials with the same critical syllables in overlapping contexts (two critical syllables, two vowels, and two repetitions) and 16 trials with the same critical syllables in complementary contexts (two critical syllables two orders of presentation, and four repetitions). Note that the number of trials with the same critical syllables in complementary contexts was two times the number of the trials with the same critical syllables in overlapping contexts. This is because the number of times participants heard the combinations of different context vowels was made to be the same in the trials with different critical syllables (with the contrast between complementary-natural and complementary-unnatural contexts) and the trials with the same critical syllables (without the contrast between complementary-natural and complementary-unnatural context).

There were 32 filler trials: 16 trials with different filler syllables and 16 trials with the same filler syllables. There were eight trials with different filler syllables in overlapping contexts (two orders of presentation, two vowels, and two repetitions), eight trials with different filler syllables in complementary contexts (two orders of presentation for the filler syllables, two orders of presentation for the vowels, and two repetitions). There were eight trials with the same filler syllables in overlapping contexts (two filler syllables, two vowels, and two repetitions), and eight trials with the same filler syllables in complementary contexts (two filler syllables, two orders of presentation for the vowels, and two repetitions).

4.2.5 Procedure

On Day 1, participants came into the lab and signed the consent form. In Session 1, participants first did the similarity rating task. Participants were given the following

instructions at the beginning of the test phase.

> *In each trial, you will hear a pair of words in a foreign language. Each word has two syllables. Therefore a pair will be something like "gewo baro".*
>
> *Your task is to compare the second syllables of the words ("wo" of "gewo" and "ro" of "baro" in the example above) and rate the similarity between the syllables on the scale of 1 to 7.*
>
> *In this scale, 1 means "very similar", 7 means "very different", and intermediate numbers mean intermediate similarities between "very similar" and "very different". When you rate the similarity, try to use all of the 7 levels.*

After completing the test phase, participants proceeded to the exposure phase. Participants were given the following instruction at the beginning of the exposure phase.

> *Now, you will listen to more words in this language. Your task is to listen to the recordings.*
>
> *Sometimes words will be pronounced very slowly, like "waaakooo". When you hear the slow speech, press the "SPACE" bar.*

On Day 2 participants came back to the lab and did Session 2. In Session 2, the order of the phases was reversed: the exposure phrase was first and the test phase was second. After completing the test phase, participants filled out a language background questionnaire.

4.3 Results

First, each participant's performance on the monitoring task was checked to see whether they were attentive to the stimuli during exposure. All participants did better than 75% on the monitoring task and thus were included in the analyses. Since responses on the similarity rating trials were categorical and ordered, ordinal logistic regression or cumulative link models were used to analyze the responses (Agresti, 2002; Christensen, 2015a,c).

4.3.1 Cumulative link model

When response categories are ordered on a scale, the probability that the response Y will fall at or below a response category j is called the cumulative probability of the response j.

$$P(Y \leq j) = \pi_1 + ... + \pi_j, \quad j = 1, ..., J \tag{4.1}$$

Logits of cumulative probabilities are computed for $J-1$ categories.

$$\text{logit}[P(Y \leq j)] = \log\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \log\left[\frac{\pi_1 + ...\pi_j}{\pi_{j+1} + ... + \pi_J}\right], \quad j = 1, ..., J-1 \tag{4.2}$$

A cumulative link model is a regression model for cumulative logits. It is like a binary logistic regression model for binary responses with a pair of outcomes $Y \leq j$ and $Y > j$.

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta x, \quad j = 1, ..., J-1 \tag{4.3}$$

The model consists of parameters α and β. Importantly, α is dependent on j but β is not. This means that the parameter β which specifies the effect of an explanatory variable x on the log odds of responses being at or below j is constant for all $J-1$ cumulative logits. Agresti (2002, 2010) writes the cumulative link model with a plus sign on the right hand side as shown in (4.3). With this formulation, a positive value for β means that the probability that the response Y will fall at or below j is higher for higher values of x. In other words, a positive β indicates that x has the effect of lowering the scores on the ordinal 1-to-J scale. In contrast, Christensen (2015a,c) writes the model with a minus sign on the right hand side as shown in (4.4). With this formulation, a positive value for β means that the probability that the response Y falls at or above j is higher for higher values of x.

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta x, \quad j = 1, ..., J-1 \tag{4.4}$$

Christensen (2015c) argues that the latter formulation is more intuitive because the effect of an explanatory variable x is interpreted in the same way as it is in ordinary linear regression or ANOVA models. The analyses presented below were done in R 3.0.3 (R Core Team, 2014) with the mixed effects cumulative link model function `clmm` from the `ordinal` package (Christensen, 2015b). Since the function adopts

the formulation in (4.4), I will interpret the parameters of explanatory variables accordingly.

Responses on trials with different critical syllables and trials with the same critical syllables were analyzed separately. This was because these two different types of trials had different sets of contexts; different trials had three contexts (overlapping, complementary-natural, and complementary-unnatural contexts), and same trials had two contexts (overlapping and complementary contexts). The random effects structure contained both random by-participant intercepts and random by-participant slopes for all predictor variables. However, when a model failed to converge, the structure was simplified in the following way. First, uncorrelated random intercepts and random slopes were used instead of correlated random intercepts and random slopes. If the model still failed to converge, an interaction between random slopes was excluded from the structure if the model contained an interaction between predictors; otherwise a random slope was excluded from the structure. This sort of simplification inflates Type I error rates (Barr et al., 2013). However, the failure of convergence may arise from trying to fit a model that is too complex for the data; the amount of information in the data is too small for the reliable estimation of all the parameters specified in the maximal random effect structure, and the reduction of the complexity is necessary (Bates et al., 2015).

4.3.2 Trials with different critical syllables

Figure 4.3 shows the distribution of responses on the trials with different critical syllables by context and session. The responses were analyzed with three predictors, session (Session 1 and Session 2), context (overlapping, complementary-natural, and complementary-unnatural), and the interaction between session and context. Since the levels of context are hierarchically organized (i.e. complementary-natural and complementary-unnatural can form a level "complementary" that makes a contrast against overlapping), the effect of context was examined in two steps. First, a contrast was made between overlapping and complementary. Then, another contrast was made between complementary-natural and complementary-unnatural. To do so, the predictor levels were coded using a contrast coding scheme (Helmert) (Table 4.2).

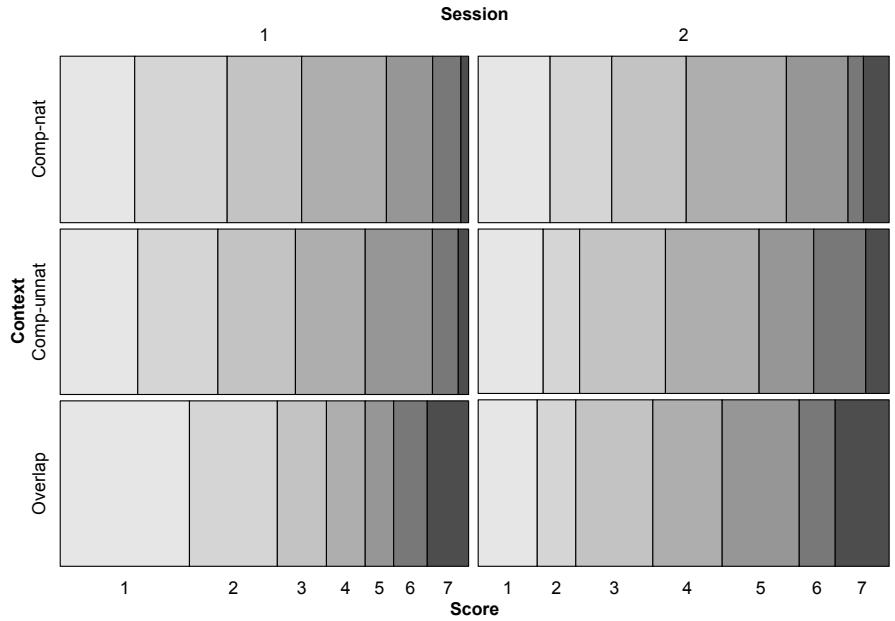


Figure 4.3: Distribution of responses to trials with different critical syllables (1="very similar" and 7="very different")

Table 4.2: Helmert contrast coding for Context

Trial type	1st comparison	2nd comparison
Overlap	0.66	0
Comp. nat.	-0.33	0.5
Comp. unnat.	-0.33	-0.5

Likelihood ratio tests were used to evaluate the statistical significance of the predictors by comparing a predictor model (a model with a predictor) and a reduced model (a model without the predictor).[1] The significance level was set at $p < 0.05$. The test results yielded no significant main effects of session [$\chi^2(1) = 2.922$, $p = 0.087$] and context [$\chi^2(1) = 2.314$, $p = 0.314$], but a significant interaction

[1]The likelihood ratio test statistic is $-2(\ell_0 - \ell_1)$, where ℓ_1 and ℓ_0 is the log-likelihood of the observed data under the predictor model and the reduced model, respectively. It asymptotically follows a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters of predictor models and their reduced counterparts (Christensen, 2015c).

between session and context [$\chi^2(2) = 21.951$, $p < 0.001$].[2]

In order to understand the nature of the interaction, follow-up analyses were conducted with the data from Session 1 and Session 2 separately. Responses were analyzed with one predictor, context (overlapping, complementary-natural, and complementary unnatural contexts, coded as shown in Table 4.2). First, the analysis of responses from Session 1 revealed that there was no significant main effect of context [$\chi^2(2) = 0.932$, $p = 0.628$]. The analysis of responses from Session 2 revealed that there was a significant main effect of context [$\chi^2(2) = 20.646$, $p < 0.001$].[3] According to the context model, the first Helmert comparison revealed that the estimated odds of giving a score at j or higher (for any $j > 1$) were 2.12 times higher in the overlapping context than in the complementary contexts ($p < 0.001$). This means that participants perceived the different critical syllables as being less similar to each other when they (the syllables) were presented in the overlapping contexts than when they were presented in the complementary contexts. The second Helmert comparison revealed that the estimated odds of giving a score j or higher (for any $j > 1$) were 1.58 times higher in the complementary-unnatural context than in the complementary-natural context ($p = 0.033$). This means that participants perceived the different critical syllables as being less similar to each other when they were presented in the complementary-unnatural contexts than when they were presented in the complementary-natural contexts, but only after exposure.

4.3.3 Trials with same critical syllables

Figure 4.4 shows the distribution of responses on the trials with same critical syllables by session and context.[4] The responses were analyzed with three predictors, session (Session 1 and Session 2), context (overlapping and complementary), and

[2]The interaction model failed to converge with the maximal random effect structure and even with the uncorrelated random intercepts and slopes. Therefore, the interaction model and the reduced model were fitted without the interaction between random slopes.

[3]The reduced model failed to converge with the maximal random effect structure and even with the uncorrelated random intercepts and the random slopes. Therefore, the context model and the reduced model were fitted without the random slopes.

[4]Remember that the number of trials with the same critical syllables in complementary contexts was two times the number of the trials with the same critical syllables in overlapping contexts (see 4.2.4).

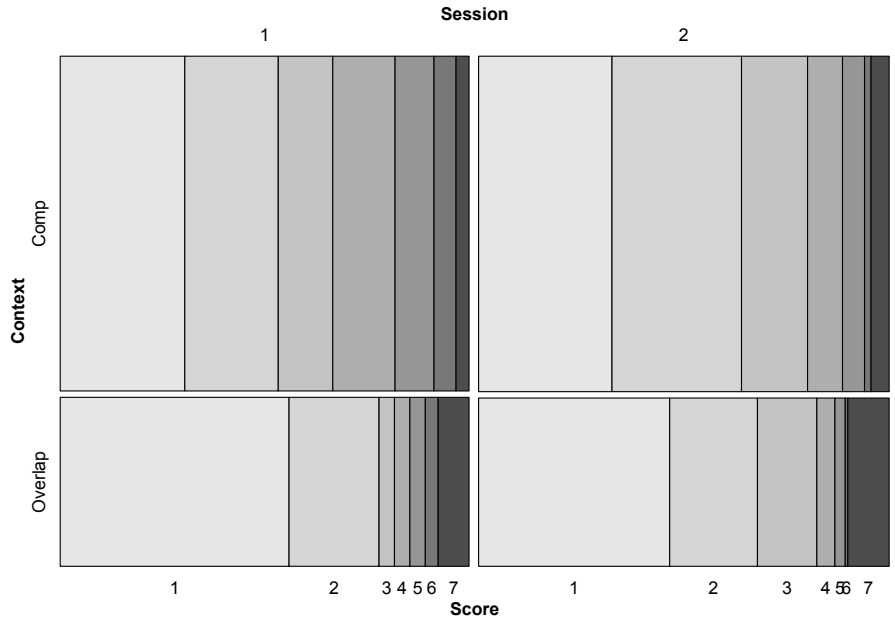


Figure 4.4: Distribution of responses to trials with the same critical syllables (1="very similar" and 7="very different")

interaction between session and context. The analyses revealed that there was no significant main effect of session [$\chi^2(1) = 0.86$, $p = 0.353$], but there was a significant main effect of context [$\chi^2(1) = 32.371$, $p < 0.001$].[5] According to the context model, the estimated odds of giving a score at j or higher (for any $j > 1$) were 2.23 times higher in the complementary contexts than in the overlapping contexts ($p < 0.001$). This means that participants perceived two repetitions of a single critical syllable as being less similar to each other when they were presented in the complementary contexts than when they were presented in the overlapping contexts. The analyses also revealed that there was a significant interaction between session and context [$\chi^2(1) = 13.254$, $p < 0.001$].[6]

[5]Both the session model and the context model failed to converge with the maximal random effects structure and even with the uncorrelated random intercepts and random slopes. Therefore, these models and their reduced counterparts were fitted without random slopes.

[6]The interaction model failed to converge with the maximal random effect structure and even with the uncorrelated random intercepts and random slopes. Therefore, the interaction model and the reduced model were fitted without random slopes.

In order to understand the nature of the interaction, follow-up analyses were conducted with the data from Session 1 and Session 2 separately. The responses were analyzed with one predictor, context (overlapping and complementary). The analysis of responses from Session 1 revealed a significant main effect of context [$\chi^2(1) = 36.96$, $p < 0.001$].[7] According to the context model, the estimated odds of giving a score at j or higher (for any $j > 1$) were 3.64 times higher in the complementary contexts than in the overlapping contexts ($p < 0.001$). The analysis of responses from Session 2 revealed that there was no significant main effect of context [$\chi^2(1) = 2.77$, $p = 0.096$].[8] Participants perceived two repetitions of a single critical syllable as being less similar to each other when they were presented in the complementary contexts than when they were presented in the overlapping contexts in Session 1, but not in Session 2.

4.4 Discussion

The results of Experiment 3 revealed the following points. First, the rating of the similarity between different critical syllables was not affected by context in Session 1, whereas it was affected by context in Session 2. In Session 2, participants perceived [ʂa] and [ɕa] as being less similar to each other when they were presented in the overlapping contexts than when they were presented in the complementary contexts regardless of whether the complementary contexts were natural or unnatural. This is different from the prediction made by the context effects hypothesis. According to the hypothesis, the perceived similarity between [ʂa] and [ɕa] should be higher in the overlapping context than in the complementary-natural context but lower in the overlapping context than in the complementary-unnatural context. However, the results showed that the perceived similarity between [ʂa] and [ɕa] was significantly lower in the overlapping context than in the complementary-natural and the complementary-unnatural contexts. This is probably because the presentation of the test stimuli in the overlapping contexts (i.e., after physically identical

[7] The context model failed to converge with the maximal random effects structure and even with the uncorrelated random intercepts and random slopes. Therefore, the context model and the reduced model were fitted without the random slope.

[8] The context model failed to converge with the maximal random effect structure and even with the uncorrelated random intercepts and random slopes. Therefore, the context model and the reduced model were fitted without the random slope.

vowels) highlighted the acoustic discrepancy between the stimuli and participants became more sensitive to the difference between the stimuli.

In Session 2, participants perceived [ʂa] and [ɕa] as being less similar to each other when they were presented in the complementary-unnatural contexts than in the complementary-natural contexts. These results followed the prediction made by the context effects hypothesis; the perceived similarity between [ʂa] and [ɕa] is higher in the complementary-natural context than in the complementary-unnatural context. The finding that participants showed a significant effect of context only after exposure is of particular importance. This suggests that they learned to perceive [ʂa] and [ɕa] in a context-dependent manner through exposure to input stimuli. Specifically, they started perceiving these syllables as being more similar to each other when they were presented in the complementary-natural contexts than in the complementary-unnatural contexts.

Second, the perception of two repetitions of a single critical syllable was significantly affected by context in Session 1. Participants rated two repetitions of [ʂa] or [ɕa] as being less similar to each other in the complementary contexts than in the overlapping contexts. However, the effect of context became non-significant in Session 2. The difference between the complementary and the overlapping contexts was expected under the context effects hypothesis. When two identical tokens of [ʂa] or [ɕa] were presented in overlapping contexts, context effects should shift the location of the tokens in the same direction along the continuum, which should have no impact on the perceived distance between the tokens. When two identical tokens of [ʂa] or [ɕa] were presented in complementary contexts, context effects should perceptually shift the location of the tokens in different directions along the continuum, which will increase the perceived distance between the tokens. Therefore, two repetitions of a [ʂa] or [ɕa], one in the natural context and the other in the unnatural context, sould not be perceived as identical (as in the original demonstration of context effects in categorization in Mann 1980). The results of Experiment 3 suggest that participants showed the context-dependent perception of the test stimuli in the trials with the same critical syllables in Session 1.

It is interesting to note that the effect of the context became non-significant in Session 2. Does this mean that participants' perception of [ʂa] and [ɕa] became less dependent on context after exposure? Assuming that participants initially did

not have categories for the novel fricative sounds, it could be the case that their perception of the novel sounds was more based on the information in the acoustic signal. Since the acoustic signal is continuous, the processing of the information from the portions of the acoustic signal that corresponded to the novel sounds could be influenced by the information from the portions of the acoustic signal that corresponded to the neighbouring sounds. Once participants learned the categories, their perception could become more dependent on their knowledge about the categories. In other words, before learning the categories, participants were more sensitive to the interactions of acoustic information across segmental boundaries. However, this does not explain the pattern of responses on the trials with different critical syllables.

Participants did not show context effects in their responses on the trials with different critical syllables in Session 1. A possible explanation for the absence of context effects in the comparison of different critical syllables in Session 1 is that the the acoustic differences between the syllables were large enough to eliminate any potential context effects. The context effects should reduce the perceived distance between [ʂa] and [ɕa] when the syllables are presented in complementary-natural contexts. However, as we saw in Experiment 1, English speakers seem to have fairly good pre-existing sensitivity to acoustic differences between [ʂa] and [ɕa]. Therefore, it is possible that participants in Experiment 3 could initially perceive the differences between [ʂa] and [ɕa] well enough that their perception was not affected by context. After being exposed to input that supported the learning of the context-dependent perception of [ʂ] and [ɕ], significant context effects emerged in Session 2.

Another possible explanation for the pattern of responses on the trials with the same critical syllables is that participants took the (non-)identity of context syllables into consideration in the similarity rating. In other words, participants rated two identical tokens of a single critical syllable presented in the complementary contexts as being less similar to each other because they consciously or unconsciously compared disyllabic strings instead of the second syllables of the disyllabic strings. If this is the case, it would be a serious confound for any comparisons between the overlapping context and the complementary context.

In order to test whether this confound was actually present, responses on trials

with the same filler syllables were analyzed. The filler syllables used were the aspirated [t^{h}a] and the unaspirated [ta]. As far as I am aware, there is no straightforward phonetic connection between the laryngeal features in stop consonants and vowel frontness, backness, and rounding. Therefore, there should not be any context effects in the perception of the filler syllables in complementary contexts, and there should not be any difference between the overlapping context and the complementary context in terms of the perceived similarity between the filler syllables. The responses were analyzed using mixed effects ordinal logistic regression models with three predictors, session (Session 1 and Session 2), context (overlapping and complementary) and interaction between session and context. The analyses revealed that there was no significant main effect of session [$\chi^2(1)t = 1.73$, $p = 0.188$], but there was a significant main effect of context [$\chi^2(1) = 7.326$, $p = 0.007$]. According to the context model, the estimated odds of giving a response at j or higher (for any $j > 1$) were 2.88 times higher in the complementary contexts than in the overlapping contexts ($p = 0.005$). This means that participants perceived two repetitions of a single filler syllable as being less similar to each other when the syllables were presented in the complementary contexts than when the syllables were presented in the overlapping contexts. The analysis yielded no significant interaction between session and context [$\chi^2(1) = 2.329$, $p = 0.127$].[9]

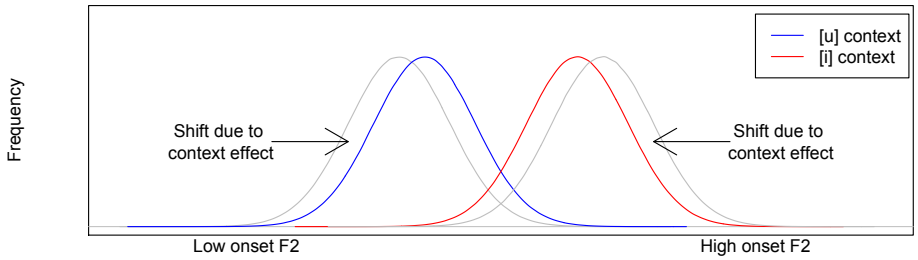
The significant effect of context is striking. Participants showed a strong bias to rate two identical tokens of a single filler syllable as being less similar to each other when the syllables were presented in different contexts even through there should not be any context effects that could explain the difference. These results suggest that at least some participants took the (non-)identity of context syllables into consideration in the similarity rating. This unfortunate confound makes the comparison between the overlapping and complementary contexts difficult. Interestingly, responses on the trials with different critical syllables did not show the same bias; there was no significant effect of context in Session 1. Moreover, participants perceived different critical syllables as being less similar to each other in the overlapping contexts than in the complementary contexts in Session 2.

[9]The interaction model failed to converge with the maximal random effect structure as well as with the uncorrelated random intercepts and random slopes. Therefore, the interaction model and the reduced model were fitted without interaction between the random slopes.

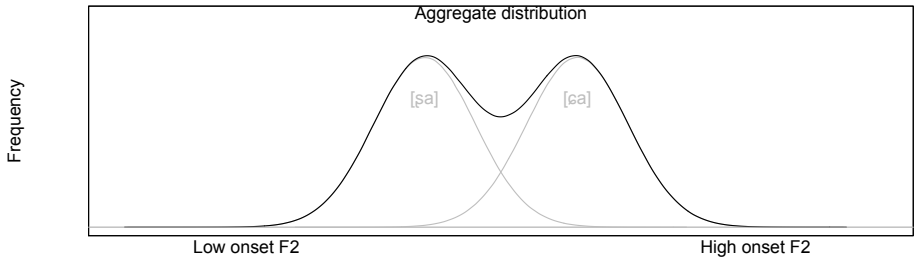
Despite the confound, the comparison between the complementary-natural and the complementary-unnatural contexts in the trials with different critical syllables remains interpretable because they differed only in the naturalness of the context. Crucially, the difference in the naturalness did not affect the similarity rating in Session 1, but it did so in Session 2. Specifically, participants perceived [ʂa] and [ɕa] as being more similar to each other when the syllables were presented in the complementary-natural contexts than when the syllables were presented in the complementary-unnatural contexts. These results suggest that, through exposure, participants learned to perceive [ʂa] and [ɕa] in a way that was predicted by the context effects hypothesis.

The finding that the context-dependent perception of novel sounds requires some learning raises a question about the generality of context effects. As discussed above, while some researchers have claimed that context effects are language-general, other researchers have claimed that there is some language specificity in context effects. The current finding supports the language specificity of context effects. However, further research is needed to understand the mechanisms supporting this learning.

The results of Experiment 3 provide support for one of the crucial assumptions on which the context effects hypothesis for allophony learning stands: Learners perceive novel sounds as more similar to each other when the segments are presented in phonetically natural complementary contexts than in phonetically unnatural complementary contexts. The results of Experiment 3 also demonstrated that the context-dependent perception of novel sounds requires learning. The results of Experiments 1 and 2 showed that learners who were exposed to input in which two novel segments were in a phonetically natural complementary distribution seemed to have learned the segments as allophones (i.e., they showed reduced sensitivity to acoustic differences between the segments after exposure), while learners who were exposed to input in which the novel segments were in a phonetically unnatural complementary distribution did not (i.e., they maintained their pre-existing sensitivity to acoustic differences between the segments). The context effects hypothesis says that this is due to the difference in the shape of the aggregate distribution of the novel sounds in auditory space. In the complementary-natural condition, participants heard tokens of two novel segments occurring only in phonetically natural



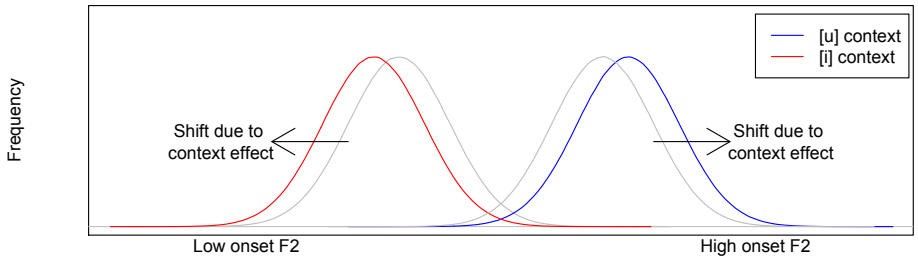
(a) Context effects in the perception of input stimuli



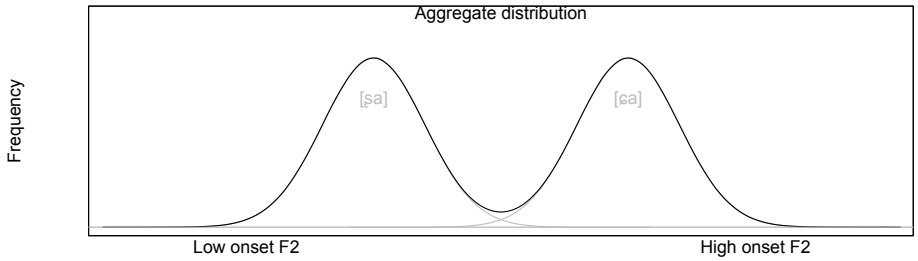
(b) Aggregate distribution in auditory space

Figure 4.5: Complementary-natural condition

contexts, and the context-dependent perception of these sounds led participants build the aggregate distribution of the sounds in auditory space with large overlap between two peaks, and thus two categories that are less distinct from each other or even a single category. Once participants built less distinct categories or a single category for the novel sounds, their perception of the novel sounds became more dependent on their knowledge about the categories (or the category) even when the sounds were presented in isolation. Therefore, participants became less sensitive to acoustic differences between the sounds (see Figure 4.5). In the complementary-



(a) Context effects in the perception of input stimuli



(b) Aggregate distribution in auditory space

Figure 4.6: Complementary-unnatural condition

unnatural condition, participants heard tokens of two segments occurring only in phonetically unnatural contexts, and thus the shape of the aggregate distribution that participants built in their auditory space had distantly separated peaks, leading to the learning of two clearly distinct categories. This is why participants in the complementary-unnatural condition maintained good sensitivity to acoustic differences between the sounds (see Figure 4.6).

In order to verify whether the learning of context-dependent perception is really a part of the mechanisms behind the learning of allophony, there are some gaps that

need to be filled. First, in Experiments 1 and 2, participants were tested on their sensitivity to acoustic differences between [ʂa] and [ɕa] presented in isolation in a discrimination task. The current experiment, by contrast, tested participants' rating of similarity between [ʂa] and [ɕa] presented in contexts. How these two different tasks are related to each other is a question that needs to be answered. Second, the current experiment was designed as a within-subject comparison. Participants were exposed to the tokens [ʂa] and [ɕa] in both the natural and unnatural contexts. In Experiments 1 and 2, by contrast, participants in the complementary-natural condition were exposed to the tokens of [ʂa] and [ɕa] only in the natural contexts, while participants in the complementary-unnatural contexts were exposed to the tokens [ʂa] and [ɕa] only in the unnatural contexts. If participants in these two conditions of Experiment 1 and 2 were actually learning the context-dependent perception of the novel sounds, they should show context effects in the similarity rating task; participants in the complementary-natural condition would perceive [ʂa] and [ɕa] as being more similar to each other than participants in the complementary-unnatural condition do. Moreover, if the learning of context-dependent perception significantly affected the learning of [ʂa] and [ɕa] as separate categories, participants in the complementary-natural and the complementary-unnatural conditions should show a significant difference in the rating of the similarity between the [ʂa] and [ɕa] even in isolation. A follow-up experiment using a between-subjects design and similarity rating in isolation would be desirable to make stronger connections between context effects and the learning of allophonic relationships. This will be a future direction of this work.

4.5 Conclusion

In this chapter, I presented the results of Experiment 3. In this experiment, I tested whether the perception of two novel sounds is affected by the contexts in which the sounds are presented and whether the context-dependent perception of the novel sounds requires learning. The results showed that context-dependent perception of the novel sounds was learned through exposure to input; participants perceived two novel sounds, the retroflex [ʂa] and the alveolopalatal [ɕa], as more similar to each other in the complementary-natural context than in the complementary-

unnatural context. The results provide support for one of the basic assumptions on which the context effects hypothesis for the learning of allophony stands: learners perceive novel sounds differently in different contexts. But, the context effects in the perception of the novel sounds still requires learning.

Chapter 5

General discussion

In this chapter, I first summarize the findings of the three experiments presented in the previous chapters. Then, I discuss the implications of these findings for the theories of sound category learning. I also discuss some remaining questions about the context effects hypothesis. Finally, I discuss some possible future directions of the research initiated in this dissertation.

5.1 Summary of findings

The goal of this dissertation was to investigate the mechanisms behind the learning of allophony. The results of Experiment 1 suggest that allophony can be learned from the complementary distribution of segments in input. There were three important findings from the results of Experiment 1. First, the results of the control condition showed that adult English speakers have fairly good pre-existing sensitivity to acoustic differences between Mandarin retroflex [ʂ] and alveolopalatal [ɕ]. Second, the results of the non-complementary condition showed that adult English speakers maintained the pre-existing sensitivity when they were exposed to input in which these segments were in non-complementary distribution (i.e., the occurrences of these segments are unpredictable from relevant contexts). Third, the results of the complementary-natural condition showed that adult English speakers became less sensitive to acoustic differences between [ʂ] and [ɕ] when they were exposed to input in which these segments were in complementary distribu-

tion (i.e. the occurrences of these segments are predictable from relevant contexts). The finding that exposure to input in which [ʂ] and [ɕ] were in complementary distribution resulted in the reduction in learners' sensitivity to the difference between these segments suggests that the segments were learned as something like allophones.

The results of Experiment 2, together with the results of Experiment 1, suggest that the learning of allophony is constrained by the phonetic naturalness of the patterns of complementary distribution. In the input used in the complementary-natural condition (Experiment 1), target segments occurred in phonetically natural complementary contexts; the retroflex [ʂ] occurred after the high back rounded vowel [u], and the alveolopalatal [ɕ] occurred after the high front unrounded vowel [i]. In Experiment 2, adult English speakers were exposed to input in which the target segments occurred in phonetically unnatural complementary contexts; the retroflex [ʂ] occurred after the high front unrounded vowel [i] and the alveolopalatal [ɕ] occurred after the high back rounded vowel [u]. The important finding from the results of Experiment 2 was that adult English speakers did not show any significant reduction in their sensitivity to acoustic differences between [ʂ] and [ɕ] when they were exposed to input in which these segments were in complementary distribution but the pattern of the distribution was phonetically unnatural.

In order to account for the role of phonetic naturalness, I proposed a hypothesis about the mechanisms behind the learning of allophony (context effects hypothesis). I hypothesized that the learning of allophony, as it is seen in the reduction in learners' sensitivity, is partly attributed to the way learners perceive the instances of the target segments during exposure. Specifically, learners perceive the instances of the target segments as being more similar to each other when they hear the sounds in phonetically natural complementary contexts. This has a significant impact on the shape of the aggregate distribution of the sounds in auditory space, and thus categories that learners build for the sounds as well. Since learners in the complementary-natural condition heard the instances of the target segments occurring only in phonetically natural contexts, the shape of the aggregate distribution in auditory space had two peaks that were closely separated from each other, and thus they formed less distinct categories or even a single category for the sounds in the input. This led to the reduction in their sensitivity to acoustic differences between

the target sounds.

Experiment 3 tested whether adult English speakers' perception of the retroflex [ʂ] and alveolopalatal [ɕ] was affected by context, and whether the context-dependent perception of these novel sounds was due to perceptual biases that adult English speakers already had or that they learned through exposure to input. In Experiment 3, adult English speakers rated the similarity between [ʂa] and [ɕa] before and after exposure to input. In the similarity rating task, learners compared [ʂa] and [ɕa] in three different types of context: overlapping context (i.e., both [ʂa] and [ɕa] were presented after the same vowel, either [i] or [u]), complementary-natural context (i.e., [ʂa] was presented after [u], and [ɕa] was presented after [i]), and complementary-unnatural context (i.e., [ʂa] was presented after [i] and, [ɕa] was presented after [u]). I predicted that learners would perceive these two novel syllables as being more similar to each other in the complementary-natural context than in the complementary-unnatural context. There were two important findings from the results of Experiment 3. First, there was no significant effect of context type on the rating of similarity in the pre-exposure test. Second, there was a significant effect of context type on the rating of similarity in the post-exposure test; participants rated [ʂa] and [ɕa] as being more similar to each other in the complementary-natural context than in the complementary-unnatural context. These results suggest that the context-dependent perception of these novel sounds requires some learning.

In sum, the experiments in this dissertation suggest that allophony can be learned from the complementary distribution of target segments in input, but that the learning is constrained by the phonetic naturalness of the patterns of the complementary distribution. I argued that the mechanisms that underlie the learning of the allophony between two segments in phonetically natural complementary contexts involve the learning of the context-dependent perception of the instances of the target segments.

5.2 The role of context in the learning of sound categories

As reviewed in Chapter 1, numerous studies have demonstrated that both infant and adult learners can learn sound categories from the frequency distribution of sounds in acoustic space. However, recent studies have pointed out that frequency

distribution is not the only cue that learners can use to learn sound categories. For example, Yeung and Werker (2009) demonstrated that 9-month-old infants are already sensitive to the functional value of novel segments. In their study, 9-month-old English-learning infants failed to discriminate Hindi dental [d̪a] and retroflex [ɖa]. But, they successfully discriminated these Hindi sounds after being presented with two novel objects paired with the dental [d̪a] and the retroflex [ɖa] respectively. These results suggest that infants as young as 9 months of age can already use semantic cues for the learning of novel segmental categories.

Another source of information for the learning of sound categories is the lexical context (Feldman et al., 2009, 2011, 2013a,b; Martin et al., 2013; Swingley, 2009; Thiessen, 2011b). Research on the role of the lexical context in the learning of sound categories has evolved from studies on the learning of minimal pairs by infants. Stager and Werker (1997) first reported that 14-month-old English-learning infants reliably discriminated the English phonemes /b/ and /d/, but they failed to learn a minimal pair that relied on the phonemic contrast in an audio-visual word learning task ([bɪ] vs. [dɪ]). Similarly, Thiessen (2007) reported that 15-month-old English-learning infants reliably discriminated the English phonemes /t/ and /d/, but they failed to learn a minimal pair that relied on the phonemic contrast in an audio-visual word learning task ([dɔ] vs. [tɔ]). These results suggest that infants' ability to discriminate native phonemes at this age does not necessarily mean that they are ready to use their ability to learn new words. They also suggest that the word learning task makes the processing of two discriminable but still similar sounds challenging for infants of this age, possibly due to cognitive capacity constraints (e.g., Werker and Fennell, 2004). Interestingly, Thiessen also reported that 15-month-old infants successfully learned the minimal pair when they were trained with a non-minimal pair along with the minimal pair ([dɔ] vs. [tɔ] and [dɔbo] vs. [tɔgu]). Thiessen (2011b) argued that the presentation of the similar sounds in very distinct lexical contexts made the sounds more differentiable and facilitated the learning of minimal pairs. This happened through the process of *acquired distinctiveness*.[1]

[1]Rost and McMurray (2009, 2010) argue that the failure of learning the minimal pairs in earlier studies is due to the lack of within-category variability in the stimuli used in the training (e.g., only one token of each category was used in Thiessen (2007)).

> "if an organism has difficulty differentiating between two similar stimuli, A and B (for example, two similar sounds), they can be repeatedly paired with two easily differentiable outcomes, X and Y (X might be punishment, and Y a reward), such that the organism consistently experiences AX and BY pairings. Over time, these pairs reinforce the orignal subtle distinction between A and B and make it easier to detect" (Thiessen, 2011b, p.1449)

Based on these findings, Feldman and colleagues developed a model of sound category learning which takes both acoustic information and lexical information into consideration (the *lexical distributional model*: Feldman et al. 2009). In this model, learners make inferences about sound categories from the distribution of acoustic values, but they make these inferences for sounds used in specific lexical items. Therefore, when there is a potential ambiguity in the acoustic information (i.e. there is a significant overlap between two distributional peaks), learners can rely on unambiguous lexical contexts, or unambiguously differentiable sounds in the lexical contexts, to overcome the ambiguity. According to this model, the contrast between [dɔ] and [tɔ] was potentially ambiguous for the infant learners in Thiessen (2007), but the contrast between the words in the non-minimal pair ([dɔbo] and [tɔgu]) was not. Therefore, the infant learners used their knowledge about the words to sort out the potential ambiguity between [dɔ] and [tɔ], and this significantly facilitated the learning of the phonetic contrast between [dɔ] and [tɔ] in the cognitively-demanding word learning task.

Feldman and colleagues conducted a series of sound category learning experiments to test whether lexical context can alone help the learning of sound categories (Feldman et al., 2011, 2013b). First, they exposed adults to syllables taken from an 8-step continuum between [tɑ] and [tɔ]. The syllables were classified into two categories (i.e., the first four steps being [tɑ] and the second four steps being [tɔ]), but the categorization was not cued by the frequency distribution of the syllables; all of the syllables occurred with the same frequency. In the non-minimal pair condition, participants heard the [tɑ] tokens and [tɔ] tokens in different lexical contexts ([gutɑ] and [litɔ] or [litɑ] and [gutɔ]). In the minimal pair condition, participants heard both [tɑ] tokens and [tɔ] tokens in the same lexical contexts ([gutɑ]

and [gutɔ] or [litɑ] and [litɔ]). Participants in both conditions were tested on the discrimination of [tɑ] and [tɔ] after exposure. The results showed that participants in the non-minimal pair condition performed significantly better than participants in the minimal pair condition. These results suggest that the non-overlapping lexical contexts facilitated the learning of the phonetic contrast between [tɑ] and [tɔ].

Feldman and colleagues also tested the effect of the lexical context in the learning of sound categories by 8-month-old English-learning infants (Feldman et al., 2013b). In this experiment, infants were familiarized with syllables taken from the same 8-step continuum between [tɑ] and [tɔ]. In the non-minimal pair condition, infants heard these syllables in non-overlapping lexical contexts ([gutɑ] and [litɔ] or [litɑ] and [gutɔ]). In the minimal pair condition, infants heard the same syllables in overlapping lexical contexts ([gutɑ] and [gutɔ] or [litɑ] and [litɔ]) (minimal pair condition). After exposure, infants were tested on the discrimination of [tɑ] and [tɔ] using a stimulus-alternation preference procedure. The results demonstrate that the infants in the non-minimal pair condition showed a significant preference for non-alternating test stimuli (e.g., [tɑ, tɑ, tɑ, ...]) as compared to alternating test stimuli (e.g., [tɑ, tɔ, tɑ,...]), but the infants in the minimal pair condition did not show any preference, indicating that the infants in the non-minimal pair condition discriminated [tɑ] and [tɔ], but the infants in the minimal pair condition did not. These results suggest that the non-overlapping lexical context facilitated the learning of the phonetic contrast between [tɑ] and [tɔ] by infants.

A somewhat similar but different line of research has developed the idea that the implicit learning of sound categories is facilitated by the explicit processing of events that are systematically correlated with the occurrences of the target sounds (*incidental learning of sound categories*: Gabay et al. 2015, Lim and Holt 2011, Seitz et al. 2010, Vlahou et al. 2012, Wade and Holt 2005). The focus of these studies is a phenomenon called *task-irrelevant perceptual learning* (TIPL). When learners perform a task that involves the processing of task-relevant stimuli, but their performance on the task or the presentation of the task-relevant stimuli is systematically correlated with the presentation of task-irrelevant stimuli, they learn the features of the task-irrelevant stimuli (Seitz and Watanabe, 2009, and references there in). TIPL is not a purely passive learning process. It is an interactive learning process in the sense that it relies on the reinforcement triggered by the learner's

active engagement in the task or the active processing of the task-relevant stimuli. Because of this interactive nature, it has been argued that TIPL is an ecologically realistic model for the learning of sound categories. In the real world, for example, sound categories are learned through word learning, and word learning involves the active processing of non-phonetic information that is systematically correlated with the occurrences of phonetic information (e.g., the properties of the referents).

In order to demonstrate the effectiveness of incidental learning in sound category learning, researchers have used various experimental paradigms. For example, Lim and Holt (2011) used a video game paradigm to test the incidental learning of the English liquids [l] and [ɹ] by Japanese speakers. In this paradigm, participants play a video game in which they shoot or capture four aliens, and each one of the four aliens is always accompanied by one of the four syllables including [la] and [ɹa]. After playing the game for about 20 minutes, participants showed a significant improvement in the identification of the English liquids. Interestingly, participants also showed an improvement in cue weighting. It is known that while native English speakers rely largely on F3 in the categorization of [l] and [ɹ], naive Japanese speakers rely more on F2. Participants in Lim and Holt (2011) became more attentive to F3 after playing the video game.

Vlahou et al. (2012) used a different paradigm to test the incidental learning of the Hindi dental [d̪a] and retroflex [ɖa] by Greek speakers. In their experiments, one group of participants was trained on the identification of [d̪a] and [ɖa] with explicit feedback, another group was trained in a TIPL paradigm. In the TIPL paradigm, participants heard the pairs of learning stimuli: two tokens of the dental [d̪a] or two tokens of the retroflex [ɖa]. Crucially, while the stimuli in the dental pair were played with the same intensity level, the stimuli in the retroflex pair were played with different intensity levels, and participants were asked to decide whether the stimuli in each pair had the same intensity level or different intensity levels. This means that participants' responses are always correlated with the type of the consonant: "same" to the dental pair and "different" to the retroflex pair. After exposure, participants who were trained in the TIPL paradigm performed as well as partipants who were trained with explicit feedback in the identification and the discrimination of [d̪a] and [ɖa]. These results suggest that incidental learning is as effective as learning with explicit feedback.

Lexical distributional learning and incidental learning are quite different learning processes, but both emphasize the role of context. In lexical distributional learning, unambiguous phonetic contrasts in the lexical context help learners to sort out potential ambiguity in input and facilitate the learning of sound categories. In incidental learning, the active processing of unambiguous task-relevant stimuli helps the learning of the properties of potentially ambiguous task-irrelevant stimuli. Here, the task itself can be characterized as the context for the learning of sound categories. Compared to the role of context in lexical distributional learning and incidental learning, the role of context in this dissertation is very different. In this dissertation, I demonstrated that the context contributes to the learning of allophony; when segments occur in phonetically-natural complementary contexts, they are likely to be learned as context-conditioned allophones.

The contrast between Feldman et al. (2011, 2013b) and this dissertation is particularly striking because these two studies implemented very similar phonotactic distributions of the target categories in input. In the input used in the non-minimal pair condition of Feldman et al.'s experiments, the target categories were in fact in a complementary distribution; the tokens of [tɑ] occurred after the vowel [i] and the tokens of [tɔ] occurred after the vowel [u], or the other way around. In the input used in the minimal pair condition, the target categories were in a non-complementary distribution; the tokens of [tɑ] and [tɔ] both occurred after the vowel [i] or [u]. If learners in Feldman et al.'s experiments were sensitive to the dependencies between the target categories and their contexts, those who were in the non-minimal pair condition could have learned [tɑ] and [tɔ] as contextually-conditioned variants of a single syllable. However, the results of Feldman et al.'s experiments showed the opposite pattern. The "complementary distribution" in the input used in the non-minimal pair condition in fact helped the learning of [tɑ] and [tɔ] as separate categories rather than the variants of a single category. How can the difference between the results of this dissertation and Feldman et al.'s studies be explained?

Feldman et al. (2011) were aware of the possibility that learners in the non-minimal pair condition of their experiment could have interpreted the lexical contexts as phonological contexts for allophonic variation between [tɑ] and [tɔ]. However, they rejected the possibility for the following reasons. In their experiment,

learners in the non-minimal pair condition were in fact grouped into two sub-conditions. In one sub-condition, learners heard [tɑ] occurring after [li] and [tɔ] occurring after [gu]. In the other sub-condition, learners heard [tɑ] occurring after [gu] and [tɔ] occurring after [li]. According to Feldman et al., the phonotactic distribution of [tɑ] and [tɔ] was more natural in the first sub-condition than in the second sub-condition. Since [ɑ] has higher F2 than [ɔ], and [i] has higher F2 than [u], the occurrence of [tɑ] after [li] and the occurrence of [tɔ] after [gu] could have been interpreted as a result of vowel-to-vowel coarticulation. With the assumption that the learning of phonology is biased towards the patterns that are phonetically natural (e.g, Wilson, 2006), Feldman et al. (2011) predicted that if learners in the non-minimal pair condition were learning the complementary distribution of [tɑ] and [tɔ], and allophonic variation between these syllables, those who were in the first sub-condition would have learned the allophony better and become less sensitive to acoustic differences between [tɑ] and [tɔ] than those who were in the second sub-condition. Feldman et al., however, did not find any significant difference between these two sub-conditions in terms of the learners' sensitivity to the acoustic differences. From these results, Feldman et al. argued that learners in their experiment were not interpreting the lexical contexts as phonological contexts and were not learning allophonic variation between [tɑ] and [tɔ].

In this dissertation, I assumed that context was used as phonological context for the learning of phonological relationships. The occurrence of different sounds in mutually exclusive contexts was used as a cue for the learning of an allophonic relationship between the sounds. Feldman et al. (2011), by contrast, claimed that context was used as lexical context for the learning of phonetic contrasts. The occurrence of different sounds in mutually exclusive contexts was used as a cue for establishing separate categories for the sounds. Therefore, the crucial difference between my claim and Feldman et al.'s claim is in the way learners interpreted the context either as lexical context or phonological context. How did the difference emerge first of all?

One possible explanation is the amount of variability in exposure stimuli. In the non-minimal pair condition of Feldman et al. (2011, 2013b)'s experiments, the tokens of target syllables (i.e., 8 syllables from a continuum between [tɑ] and [tɔ]) were presented after two syllables (e.g., [li] or [gu]). This means that there were

two types of bisyllabic strings in the input (e.g., [litɑ] and [gutɔ]). By contrast, in the complementary-natural condition of Experiment 1 of in this dissertation, the tokens of target syllables (i.e., 8 syllables from a continuum between [ʂa] and [ɕa]) were presented after four syllables (e.g., [li] and [pi] or [lu] and [pu]) in a session. This means that there were four types of bisyllabic strings in the input (e.g., [liɕa], [piɕa], [luʂa], and [puʂa]). This slight difference in the amount of variability in exposure stimuli could have directed learners in these two studies in different directions in terms of the interpretation of the context. Studies have demonstrated that the learning of regularities in input is facilitated by the presence of variability that seems to highlight the regularities under certain conditions (e.g., Gómez, 2002). A recent study on the acquisition of phonotactics by infants has suggested that the acquisition of native phonotactic patterns is determined by the type frequency of the patterns or the number of different items in which the patterns are instantiated in input rather than the token frequency of the patterns or the frequency of the occurrences of the patterns in input (Archer and Curtin, 2011). Therefore, it is possible that learners in the experiments reported in this dissertation learned the dependencies between the target syllables and the contexts as phonotactic regularities because the input stimuli had enough variability, and learners in Feldman et al. (2011, 2013b) failed to learn the dependencies between the target syllables and the contexts as phonotactic regularities (i.e., the dependencies were learned as parts of the lexical forms), because the input stimuli did not have enough variability.

To test whether the amount of variability in the input plays a crucial role in determining whether learners adopt a phonological interpretation or a lexical interpretation of the context, it would be useful to replicate the non-minimal pair condition of Feldman et al. (2011, 2013b) with more variability in input stimuli. With more variability, I expect that learners would learn the dependencies between the target syllables and the contexts as a kind of complementary distribution and would therefore learn the allophonic relationship between [tɑ] and [tɔ]. Alternatively, it would be useful to replicate the complementary-natural condition of Experiment 1 of this dissertation without the stimulus variability. I expect that learners would interpret the context as lexical context and therefore would fail to learn the allophonic relationship between [ʂ] and [ɕ].

If the amount of variability in input stimuli determines the way the context is

interpreted, gradually increasing the amount of stimulus variability in the input in which target segments are in complementary distribution would induce inverted U-shaped learning. Initial exposure to input with low stimulus variability (e.g., one type of stimuli for each one of the target segments, just like the input used in the non-minimal pair condition of Feldman et al. 2011 and Feldman et al. 2013b) would induce learners to interpret the context as lexical context, and would facilitate the learning of target segments as distinct categories, but subsequent exposure to input with high stimulus variability would induce learners to direct their attention to the complementary distribution and would therefore encourage the learning of an allophonic relationship between the target segments. If these two types of learning happen in succession, learners would initially show good sensitivity to acoustic differences between the target segments, but their sensitivity would decline as they get exposure to input stimuli with more variability.

5.3 Some remaining questions about the context effects hypothesis

In Chapter 3, I explained the results of Experiment 1 and 2 using both articulatory and auditory theories of context effects. Since both articulatory and auditory theories provided the same explanations for the results of Experiments 1 and 2, I did not make any judgments about which theories are better than the others in explaining perceptual biases in the learning of allophony within the context of Experiments 1 and 2. However, when it comes to the question of how robust these theories are in explaining the learning of allophony beyond the results of Experiments 1 and 2, auditory theory (i.e., spectral contrast theory) faces some serious problems. In this section, I will discuss these problems.

5.3.1 Directionality

The first problem is directionality. The pattern of complementary distribution implemented in the input used in the complementary-natural condition of Experiment 1 assumed carryover assimilation. The distribution of the retroflex [ʂ] and alveolopalatal [ɕ] was conditioned by the quality of the preceding vowel: [ʂ] occurring after the high back rounded vowel [u] and [ɕ] occurring after the high front un-

rounded vowel [i]. Therefore, if [ʂ] and [ɕ] are considered to be two variants of a single phoneme, the variation could have arisen as a result of the assimilation of the fricative to the preceding vowel. In this case, both articulatory and auditory theories are able to explain the change in learners' sensitivity to acoustic differences between [ʂ] and [ɕ] after exposure (see Section 3.7.2). However, carryover assimilation is not the only process through which allophony arises; there are a great deal of instances in which allophony arises through anticipatory assimilation.

For example, in Japanese, the distribution of sibilants is conditioned by the quality of the following vowel. While the alveolar sibilants [s] and [(d)z] occur before [a], [e], [o], and [ɯ], the alveolopalatal sibilants [ɕ] and [(d)ʑ] occur before [i].[2] The allophonic relationship between the alveolar sibilants and alveolopalatal sibilants can be described as a result of palatalization; the alveolar sibilants assimilate to the palatality of the following high front unrounded vowel [i]. Moreover, one of the most commonly attested allophonic alternations in natural languages is vowel nasalization, and this predominantly happens as anticipatory assimilation; vowels are nasalized when followed by a nasal consonant. Since spectral contrast theory is all about how the perception of the precursor stimulus modulates the perception of the following stimulus, it has nothing to say about the influence of the following stimulus on the precursor stimulus (Fowler, 2006, pp.163-164). Therefore, spectral contrast theory cannot explain the learning of allophony in general.

5.3.2 Non-spectral information

Another problem of spectral contrast theory is that its explanatory potential is limited to cases in which allophony relies on the spectral properties of the interacting segments. However, there are a great number of cases in which spectral properties are not relevant to allophony. For example, in English, voiceless stops have two allophonic variants, the voiceless aspirated and the voiceless unaspirated. The former type occurs in word-initial syllable onset position as a singleton (e.g., [pʰɪt]), and the latter type occurs in syllable onset position in a cluster following [s] (e.g., [spɪt]). Researchers have argued that the presence or absence of aspiration in voiceless stops in these two contexts can be explained by the timing relation-

[2]The complementary distribution of the alveolar sibilants and alveolopalatal sibilants is seen only in a subset of Japanese lexicon, old Japanese words.

ships between oral and laryngeal gestures (Iverson and Salmons, 1995; Kim, 1970; Kingston, 1990; Löfqvist and Yoshioka, 1981; Yoshioka et al., 1981).

When a voiceless stop is produced in word-initial syllable onset position as a singleton, the vocal folds open for the production of the voiceless stop. Then, the vocal folds start closing after the release of the oral closure in order to be able to produce voicing for the following vowel. Aspiration happens as a consequence of continuous airflow between the release of the oral closure and the onset of the voicing (Kim, 1970). By contrast, when a voiceless stop is produced in syllable onset position in a cluster following [s], the vocal folds open for the production of the voiceless [s]. Then, the vocal folds start closing during the production of [s], and by the time that the oral closure for the following stop is released the vocal folds are already in the setting for the production of the voicing for the following vowel. Therefore, there is not enough time for aspiration to happen between the release of the oral closure and the onset of the voicing (Kim, 1970; Yoshioka et al., 1981; Löfqvist and Yoshioka, 1981). Browman and Goldstein (1986) argue that English has a constraint that allows only one laryngeal gesture in a word-initial (syllable-initial) consonant cluster. The articulatory theories can potentially explain the learning of the allophonic relationship between the voiceless aspirated and voiceless unaspirated stops by assuming that learners have innate knowledge about the complexity in the timing relationships between oral and laryngeal gestures or that learners acquire such knowledge through experience. However, spectral contrast theory cannot explain the learning of the allophony because it has nothing to say about the relative timing of articulatory and acoustic events.

The limitations of spectral contrast theory discussed above indicate that the articulatory theories are more robust in explaining the learning of allophony. However, this does not necessarily mean that spectral contrast effects are irrelevant to the learning of allophony. It is possible that different mechanisms are involved in the learning of different types of allophony, and that spectral contrast effects still play a role in the learning of certain types of allophony, as with the case examined in this dissertation. A better way to test the role of spectral contrast effects in the learning of allophony examined in this dissertation would be to replicate the results of Experiment 1 using non-lingustic contexts.

Studies have demonstrated that listeners show context effects in the perception

of speech sounds even when the sounds are presented in non-linguistic auditory contexts (Holt, 1999, 2005; Holt et al., 2000; Lotto and Kluender, 1998). For example, Lotto and Kluender (1998) reported that English speakers showed context effects in the categorization of a continuum between [da] and [ga] when the stimuli were presented after the frequency modulated (FM) glides that mimicked the F3 trajectories of [l] and [ɹ]. These results have been considered as strong support for spectral contrast effects in speech perception. Therefore, if the learning of allophony examined in this dissertation arises from spectral contrast effects, the same learning effects are expected to be obtained even when the contexts are non-linguistic auditory stimuli. When learners are exposed to the tokens of the retroflex [ʂa] after an FM glide that mimics the F2 trajectory of high back rounded vowel [u] and the tokens of the alveolopalatal [ɕa] after an FM glide that mimics the F2 trajectory of the high front unrounded vowel [i], they should become less sensitive acoustic differences between [ʂa] and [ɕa].

5.4 Future directions

In this dissertation, I presented the first experimental support for the distributional learning of allophony by adults. What we need to test now is how robust the distributional learning of allophony is. In this dissertation, I demonstrated that adult English speakers can learn an allophonic relationship between two novel fricatives, the retroflex [ʂ] and alveolopalatal [ɕ], when they are exposed input in which these fricatives are in a phonetically natural complementary distribution (i.e., the retroflex [ʂ] occurs after the high back vowel [u] and the alveolopalatal [ɕ] occurs after the high front vowel [i]). As I made it clear in Chapter 2, this particular pattern of complementary distribution is artificial in the sense that it is not attested in any natural languages. Therefore, the distributional learning of allophony still needs to be tested with patterns that are attested in natural languages.

Allophony in natural language shows a wide variety of form and complexity. The contexts that determine the distribution include segments, features, and the structural properties of higher-level phonological representations, such as syllable, foot, word, and phrase. Moreover, a phoneme can have more than two allophones. For example, the English alveolar stop /t/ has at least five allophones, and the con-

Table 5.1: Allophones of English /t/

Allophones		Contexts	Examples	
[t^h]	voiceless aspirated	syllable onset	*tie*	[ˈt^haɪ]
[t]	voiceless unaspirated	after [s]	*sty*	[ˈstaɪ]
[t̚]	unreleased	syllable coda	*mat*	[ˈmæt̚]
[ʔ]	glottal stop	before a syllabic nasal	*beaten*	[ˈbiʔn̩]
[ɾ]	tap	between stressed and unstressed vowels	*city*	[ˈsɪɾi]

texts that determine their distributions vary from adjacent segments to prosodic positions (see Table 5.1).[3] /t/ is aspirated when it occurs as a singleton at the onset of a stressed syllable, especially in word initial position. It is unaspirated when it occurs in a consonant cluster following [s]. It is unreleased when it occurs in syllable coda position, especially in word-final position. It is realized as a glottal stop when it is followed by a syllabic nasal consonant. And it is realized as a tap when it occurs between two vowels, especially when the first vowel is stressed and the second vowel is unstressed.

It is also true that allophony is not categorical. First, as Hall (2009) has proposed, phonological relationships are a probabilistic phenomenon rather than a categorical dichotomy between phonemic contrast and allophony. The more the occurrences of two segments are predictable in particular environments, the more allophonic the segments are. Second, the phonetic realization of allophones can be gradient. For example, I discussed in Chapter 1 that English /l/ has two allophones: the light [l] occurring in syllable-initial position and the dark (velarized) [ɬ] occurring in syllable final position. However, Sproat and Fujimura (1993) have demonstrated that the degree of the velarization varies according to the duration of the syllable rhyme; the longer the rhyme is, the greater the degree of velarization is.

The question of how distributional learning can deal with these complexities needs to be explored. For example, this dissertation demonstrated that adults can learn an allophonic relationship between two segments from input in which the

[3] This is not an exhaustive list of the allophones of /t/ and the contexts that determine their distributions. The realization of the allophones also depends on dialect (e.g., Ladefoged and Johnson, 2014).

occurrences of the segments were fully predictable from context. What will happen when learners are exposed to input in which the complementary distribution of two segments is probabilistic (e.g., the target segments occur in mutually exclusive contexts 75% of times but they occur in overlapping contexts 25% of times)? Figure 5.1 shows such a distribution. In this hypothetical input, there are eight sounds taken from a continuum. The frequency distribution of these syllables shows a bimodal shape, implying that the sounds are classified into two segments. The phonotactic distribution of the segments is probabilistic. The instances of each syllable occur in both contexts C1 and C2 but with different frequencies; the syllables that belong to the first category (the distributional peak to the left) occur in C1 25% of time and in C2 75% of time, and the syllables that belong to the second category (the distributional peak to the right) occur in C1 75% of time and in C2 25% of time. Are learners able to learn the probabilistic dependencies between the target segments and their contexts and learn the target segments as probabilistic allophones? If so, how will the learning be reflected in learners' sensitivity to acoustic differences between the target segments?

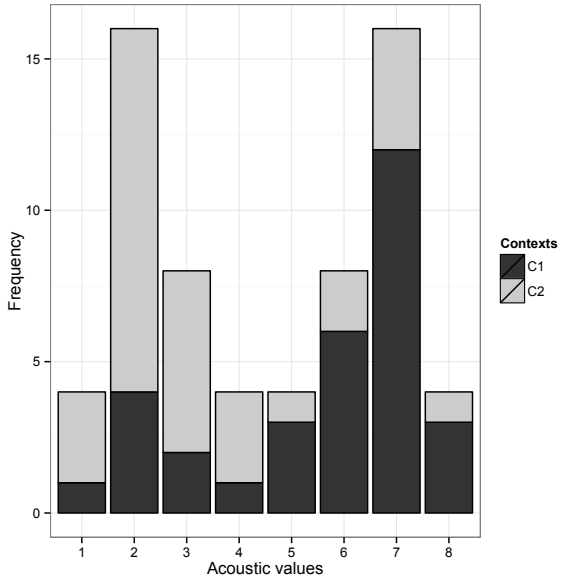


Figure 5.1: Probabilistic distribution

Studies on the distributional learning of syntax have demonstrated that adults are sensitive to probabilistic variation in the grammar of an artificial language (Hudson Kam and Newport, 2005). The learning of allophony depends on the learning of phonotactic regularities. If adults are sensitive to probabilistic variation in the phonotactic regularities, they should be able to learn probabilistic allophones; they will learn two segments as being less allophonic when the segments occur in overlapping contexts 25% of time than when the segments never occur in overlapping contexts. According to Hall (2009)'s information theoretic account of phonological status, listeners's sensitivity to the acoustic differences between allophones is determined by how predictable these allophones are. As the predictability of allophones becomes lower, listeners' sensitivity to the allophonic variation becomes higher. Therefore, compared to learners who are exposed to input in which the occurrences of two segments are either fully predictable or fully unpredictable, those who are exposed to input in which the distribution of the segments is probabilistic should show an intermediate level of sensitivity to acoustic differences between the target segments.[4]

The goal of this dissertation was to understand the mechanisms behind the learning of allophony. This was largely motivated by the question of how infants learn allophony (see Chapter 1). Given the findings that both infants and adults are sensitive to various kinds of distributional information in input—specifically, the phonotactic distribution of segments across contexts—I hypothesized that both infants and adults can learn allophony from the phonotactic distribution of segments in input. However, I only tested the learning of allophony by adults. Therefore, it needs to be tested whether infants can learn the same allophony. There exists a study that already demonstrated that 8.5- and 12-month-old English-learning infants learned an allophonic relationship between two consonants when they were exposed to input in which the consonants occurred in complementary contexts (White et al., 2008). However, we need more experimental studies on the learning of allophony by infants to understand whether infants and adults are using the same

[4]It is worth noting that Hudson Kam and Newport (2005) demonstrated that there is a age difference in the learning of probabilistic variation in artificial language. While adults learn the language with probabilistic variation, children (5- to 7-years-old) regularized the language by generalizing the probabilistically dominant patterns. Therefore, it is possible that there will be a similar age difference in the learning of probabilistic allophony as well.

learning mechanisms or not. In this dissertation, I demonstrated that the learning of allophony by adults is constrained by the phonetic naturalness of the patterns of complementary distribution. Do the same phonetic naturalness constraints apply to the learning of allophony by infants? Some studies have demonstrated that infants' learning of phonological patterns is constrained by the phonetic naturalness of the patterns (e.g., Gerken and Bollt, 2008; White and Sundara, 2014). Therefore, it is very important to understand how the learning of allophony by infants is constrained as well. In order to account for the role of phonetic naturalness in the learning of allophony, I proposed a specific hypothesis about the mechanisms behind the learning of allophony, the context effects hypothesis. If the same phonetic naturalness constraints apply to the learning of allophony by infants, it should be tested whether infants learn the context-dependent perception of sounds through exposure.

5.5 Final remarks

The learning of allophony is a small but important part of phonological acquisition. By studying allophony, we can see how phonology is acquired as a system of knowledge. Human learners have the abilities to learn various elements of phonological systems, such as phonetic categories and phonotactic regularities, from statistical information in input, and the learning is constrained by cognitive factors such as learning biases (both domain-general and language-specific) and perceptual factors such as context effects (both domain-general and language-specific). The learning of allophony is built upon the interplay of these aspects of language learning and speech processing. Therefore, studying the mechanisms behind the learning of allophony is studying the mechanisms involved in the interplay. There is much more work to be done to understand the details of the interplay.

Bibliography

Abramson, A. S. and Lisker, L. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20(3):384–422. → pages 18

Agresti, A. (2002). *Categorical data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, 2nd edition. → pages 109, 110

Agresti, A. (2010). *Analysis of ordinal categorical data*. Wiley, Hoboken, 2nd edition. → pages 110

Allen, J. S. and Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, 106(4):2031–2039. → pages xi, 18

Archer, S. L. and Curtin, S. (2011). Perceiving onset clusters in infancy. *Infant Behavior and Development*, 34(4):534–540. → pages 25, 133

Aslin, R. N. and Pisoni, D. B. (1980). Some developmental processes in speech perception. In Yeni-Komshian, G. H., Kavanagh, J. F., and A, F. C., editors, *Child phonology: Perception*, volume 2, pages 67–96. Academic Press, New York. → pages 62

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321–324. → pages 67

Astheimer, L. B. and Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology*, 80(1):23–34. → pages 8

Astheimer, L. B. and Sanders, L. D. (2011). Predictability affects early perceptual processing of word onsets in continuous speech. *Neuropsychologia*, 49(12):3512–3516. → pages 8

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278. → pages 111

Bateman, N. (2007). *A crosslinguistic investigation of palatalization*. PhD thesis, University of California, San Diego. → pages 52

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). *Parsimonious mixed models*. Manuscript submitted for publication. → pages 111

Beddor, P. S., Harnsberger, J. D., and Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *Journal of Phonetics*, 30(4):591–627. → pages 101

Beddor, P. S. and Strange, W. (1982). Cross-language study of perception of the oral–nasal distinction. *The Journal of the Acoustical Society of America*, 71(6):1551–1561. → pages 6

Best, C. T. (1995). A direct realist view of cross-language speech perception. In Strange, W., editor, *Speech perception and linguistic experience : Issues in cross-language research*, pages 171–204. York Press, Baltimore. → pages 15

Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2):775–794. → pages 15

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):345–360. → pages 15, 16, 39

Bhat, D. (1973). Retroflexion: an areal feature. *Working Papers on Language Universals*, 13:27–67. → pages 52, 53

Bhat, D. (1974). Retroflexion and retraction. *Journal of Phonetics*, 2:233–237. → pages 51

Bhat, D. N. (1978). A general study of palatalization. In Greenberg, J. H., editor, *Universals of human language*, volume 2, pages 47–92. Stanford University Press, Stanford. → pages 52

Blevins, J. (2008). Natural and unnatural sound patterns: A pocket field guide. In Willems, K. and De Cuypere, L., editors, *Naturalness and iconicity in language*, pages 121–148. John Benjamins Publishing, Amsterdam. → pages 68

Boersma, P. and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9):342–345. → pages 42, 44

Boomershine, A., Hall, K. C., Hume, E., and Johnson, K. (2008). The impact of allophony versus contrast on speech perception. In Avery, P., Dresher, B. E., and Rice, K., editors, *Contrasts in phonology: theory, perception, acquisition*, pages 145–171. Mouton de Gruyter, Berlin. → pages 6

Browman, C. P. and Goldstein, L. (1986). Towards an articulatory phonology. *Phonology*, 3(1):219–252. → pages 136

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(2):201–251. → pages 3

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180. → pages 3

Byrd, D. and Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, 24(2):263–282. → pages 69

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, 27(3):345–392. → pages 30, 68, 69

Chambers, K. E., Onishi, K. H., and Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87(2):B69–B77. → pages 2, 25, 70

Chambers, K. E., Onishi, K. H., and Fisher, C. (2010). A vowel is a vowel: generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3):821–828. → pages 26

Chambers, K. E., Onishi, K. H., and Fisher, C. (2011). Representations for phonotactic learning in infancy. *Language Learning and Development*, 7(4):287–308. → pages 25

Chang, Y.-H. S. (2010). *Lip rounding in Taiwan Mandarin retroflex sibilants*. Poster presented at the 84th Annual Meeting of the Linguistic Society of America, Baltimore. → pages 52

Chang, Y.-H. S. (2013). *Variability in cross-dialectal production and perception of contrasting phonemes: the case of the alveolar-retroflex contrast in Beijing and Taiwan Mandarin*. PhD thesis, University of Illinois at Urbana-Champaign. → pages 35, 172

Chao, Y. R. (1948). *Mandarin primer: an intensive course in spoken Chinese*. Harvard University Press, Cambridge. → pages 33, 34, 37

Cheng, C.-C. (1973). *A synchronic phonology of Mandarin Chinese*. Mouton de Gruyter, Berlin. → pages 34

Chiu, C. (2009). Acoustic and auditory comparisons of Polish and Taiwanese Mandarin sibilants. *Canadian Acoustics*, 37(3):142–143. → pages x, 35, 36

Chiu, C. (2010). Attentional weighting of Polish and Taiwanese Mandarin sibilant perception. In Heijl, M., editor, *Proceedings of the 2010 Canadian linguistics association annual conference*. → pages 36, 37, 40

Chomsky, N. and Halle, M. (1968). *The sound pattern of English.* Harper & Row, New York. → pages 3

Christensen, R. H. B. (2015a). Analysis of ordinal data with cumulative link models—estimation with the ordinal package. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/. → pages 109, 110

Christensen, R. H. B. (2015b). ordinal—regression models for ordinal data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/. → pages 110

Christensen, R. H. B. (2015c). A tutorial on fitting cumulative link mixed models with clmm from the ordinal package. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/. → pages 109, 110, 112

Christie Jr., W. M. (1974). Some cues for syllable juncture perception in English. *The Journal of the Acoustical Society of America*, 55(4):819–821. → pages 7

Chung, K. S. (2006). Hypercorrection in Taiwan Mandarin. *Journal of Asian Pacific Communication*, 16(2):197–214. → pages 172

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition*, 25(1):53–69. → pages 7

Clements, G. N. (1985). The geometry of phonological features. *Phonology*, 2(1):225–252. → pages 3

Coady, J. A., Kluender, K. R., and Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *The Journal of the Acoustical Society of America*, 114(4):2225–2235. → pages 91

Connine, C. M., Blasko, D. G., and Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32(2):193–210. → pages 8

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience, Hoboken, 2nd edition. → pages 7

Creel, S. C., Newport, E. L., and Aslin, R. N. (2004). Distant melodies: statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(5):1119–1130. → pages 68

Cristia, A., McGuire, G. L., Seidl, A., and Francis, A. L. (2011a). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, 39(3):388–402. → pages 20, 22

Cristia, A. and Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development*, 4(3):203–227. → pages 25

Cristia, A., Seidl, A., and Francis, A. (2011b). Phonological features in infancy. In Clements, G. N. and Rachid, R., editors, *Where do phonological contrasts come from? Cognitive, physical and developmental bases of distinctive speech categories*, pages 303–326. John Benjamins Publishing Company, Amsterdam. → pages 25

Cristia, A., Seidl, A., and Gerken, L. (2011c). Learning classes of sounds in infancy. In *Proceedings of The 34 th Annual Penn Linguistics Colloquium*, volume 17 of *University of Pennsylvania Working Papers in Linguistics*, page 9. → pages 68

Cutting, J. E. and Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics*, 16(3):564–570. → pages 10

Darcy, I., Peperkamp, S., and Dupoux, E. (2007). Bilinguals play by the rules: Perceptual compensation for assimilation in late L2-learners. In Cole, J. and Hualde, J. I., editors, *Laboratory phonology*, volume 9, pages 411–442. Mouton de Gruyter, Berlin. → pages 17

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., and Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In Kügler, F., Féry, C., and van de Vijver, R., editors, *Variation and gradience in phonetics and phonology*, pages 265–310. Mouton de Gruyter, Berlin. → pages 17

De Boer, B. and Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4):129–134. → pages 20

De Lacy, P. (2004). Markedness conflation in optimality theory. *Phonology*, 21(2):145–199. → pages 69

Delattre, P. C., Berman, A., and Cooper, F. S. (1962). Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Studia Linguistica*, 16(1-2):104–122. → pages 36

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 68(3):843–857. → pages 88

Delgutte, B. (1997). Auditory neural processing of speech. In Hardcastle, W. J. and Laver, J., editors, *The handbook of phonetic sciences*, pages 507–538. Blackwell, Oxford. → pages 88

Delgutte, B. and Kiang, N. Y. (1984). Speech coding in the auditory nerve: I. vowel-like sounds. *The Journal of the Acoustical Society of America*, 75(3):866–878. → pages 88

Dell, G. S., Reed, K. D., Adams, D. R., and Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6):1355–1367. → pages 26

Diehm, E. E. (1998). *Gestures and linguistic function in learning Russian: Production and perception studies of Russian palatalized consonants*. PhD thesis, Ohio State University. → pages 64, 65

Dietrich, C., Swingley, D., and Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences*, 104(41):16027–16031. → pages 14, 24

Dixit, R. P. and Flege, J. E. (1991). Vowel context, rate and loudness effects of linguopalatal contact patterns in Hindi retroflex /ʈ/. *Journal of Phonetics*, 19(2):213–229. → pages 51

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University Press, Oxford. → pages 33, 34

Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, 16(3):513–521. → pages 10

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968):303–306. → pages 9

Endress, A. D. and Mehler, J. (2010). Perceptual constraints in phonotactic learning. *Journal of Experimental Psychology: Human Perception and Performance*, 36(1):235. → pages 70

Farnetani, E. and Recasens, D. (1997). Coarticulation and connected speech processes. In Hardcastle, W. J. and Laver, J., editors, *The handbook of phonetic sciences*, pages 371–404. Blackwell, Oxford. → pages 84

Feldman, N., Myers, E., White, K., Griffiths, T., and Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Boston Uninversity Conference on Language Development*, pages 197–209. → pages 13, 127, 128, 131, 132, 133, 134

Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L. (2013a). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751. → pages 13, 127

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st annual conference of the Cognitive Science Society*, pages 2208–2213. → pages 127, 128

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., and Morgan, J. L. (2013b). Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438. → pages 13, 127, 128, 129, 131, 132, 133, 134

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, 425(6958):614–616. → pages 41

Finn, A. S. and Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, 108(2):477–499. → pages 14

Fiser, J. and Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):458. → pages 67

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In Strange, W., editor, *Speech perception and linguistic experience: Issues in cross-language research*, pages 233–277. York Press, Baltimore. → pages 15

Flemming, E. (2003). The relationship between coronal place and vowel backness. *Phonology*, 20(3):335–373. → pages 51, 52, 53

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1):3–28. → pages 86

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics*, 55(6):597–610. → pages 35

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, 99(3):1730–1741. → pages 86

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, 68(2):161–177. → pages 86, 135

Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech*, 5(4):171–189. → pages 9

Fujisaki, H. and Kawashima, T. (1969). On the modes and mechanisms of speech perception. *Annual Report of the Engineering Research Institute*, 28:67–73. → pages 7

Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. *Annual Report of the Engineering Research Institute*, 29:207–214. → pages 7

Gabay, Y., Dick, F. K., Zevin, J. D., and Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance*, 41(4):1124–1138. → pages 129

Gerken, L. and Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints. *Language Learning and Development*, 4(3):228–248. → pages 68, 141

Gilkerson, J. (2005). Categorical perception of natural and unnatural categories: evidence for innate category boundaries. *UCLA Working Papers in Linguistics*, 13:34–58. → pages 83

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers in Linguistics and the Philosophy of Language*, 6:59–72. → pages 16

Gnanadesikan, A. (1994). The geometry of coronal articulations. In González, M., editor, *Proceedings of the North East Linguistic Society*, volume 24, pages 125–139, Amherst. → pages 52, 53

Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78. → pages 9

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436. → pages 68, 133

Gordon, M., Barthmaier, P., and Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, 32(2):141–174. → pages 34

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r". *Neuropsychologia*, 9(3):317–323. → pages 16

Goudbeek, M., Cutler, A., and Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, 50(2):109–125. → pages 21

Grieser, D. and Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, 25(4):577. → pages 11

Guenther, F. H. and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, 100(2):1111–1121. → pages 20

Guion, S. G. (1996). *Velar palatalization: coarticulation, perception, and sound change*. PhD thesis, University of Texas at Austin. → pages 51

Guion, S. G. (1998). The role of perception in the sound change of velar palatalization. *Phonetica*, 55(1-2):18–52. → pages 51, 52

Guion, S. G., Flege, J. E., Akahane-Yamada, R., and Pruitt, J. C. (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(5):2711–2724. → pages 16

Gulian, M., Escudero, P., and Boersma, P. (2007). Supervision hampers distributional learning of vowel contrasts. In Trouvain, J. and Barry, W. J., editors, *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 1893–1896. → pages 21

Hall, K. C. (2009). *A probabilistic model of phonological relationships from contrast to allophony*. PhD thesis, The Ohio State University. → pages 2, 4, 5, 6, 7, 8, 9, 138, 140

Hall, K. C. (2012). Phonological relationships: a probabilistic model. *McGill Working Papers in Linguistics*, 22(1):1–14. → pages xi, 4, 5

Hall, K. C. (2013a). Documenting phonological change: A comparison of two japanese phonemic splits. In Luo, S., editor, *Proceedings of the 2013 Annual Meeting of the Canadian Linguistic Association*. → pages 5

Hall, K. C. (2013b). A typology of intermediate phonological relationships. *The Linguistic Review*, 30(2):215–275. → pages 4, 5

Hamann, S. (2002). Retroflexion and retraction revised. In Hall, T. A., Pompino-Marschall, B., and Rochon, M., editors, *Papers on phonetics and phonology : The articulation, acoustics and perception of consonants*, volume 28 of *ZAS papers in linguistics*, pages 13–25. Zentrum für Allgemeine Sprachwissenschaft, Sprachtypologie und Universalienforschung, Berlin. → pages 51

Hamann, S. (2003). *The phonetics and phonology of retroflexes*. LOT Press, Utrecht. → pages 51

Hansson, G. Ó. (2010). *Consonant harmony: long-distance interaction in phonology*. UC Publications in Linguistics. University of California Press, Berkeley. → pages 71

Hao, Y.-C. (2012). *The effect of L2 experience on second language acquisition of Mandarin consonants, vowels, and tones*. PhD thesis, Indiana University. → pages x, 37, 38, 39, 100

Harnad, S. (2005). To cognize is to categorize: Cognition is categorization. In Cohen, H. and Lefebvre, C., editors, *Handbook of categorization in cognitive science*, pages 20–45. Elsevier, Amsterdam. → pages 9

Harnsberger, J. D. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics*, 29(3):303–327. → pages 6

Harris, K. S. (1954). Cues for the identification of the fricatives of American English. *The Journal of the Acoustical Society of America*, 26(5):952–952. → pages 36

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1(1):1–7. → pages 36

Hartman, L. M. (1944). The segmental phonemes of the Peking dialect. *Language*, pages 28–42. → pages 34

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, 23(1):65–94. → pages 21, 63

Heinz, J. M. and Stevens, K. N. (1961). On the properties of voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 33(5):589–596. → pages 34

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, 97(5):3099–3111. → pages 18

Hirata, Y., Whitehurst, E., and Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, 121(6):3837–3845. → pages 16

Hohne, E. A. and Jusczyk, P. W. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, 56(6):613–623. → pages 14, 24

Holt, L. L. (1999). *Auditory constraints on speech perception: An examination of spectral contrast*. PhD thesis, University of Wisconsin–Madison. → pages 91, 137

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4):305–312. → pages 137

Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America*, 108(2):710–722. → pages 137

Hu, F. (2008). The three sibilants in standard Chinese. In Sock, R., Fuchs, S., and Yvis, L., editors, *Proceedings of the 8th International Seminar on Speech Production*, pages 105–108. INRIA. → pages 33, 34, 90

Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195. → pages 140

Hughes, G. W. and Halle, M. (1956). Spectral properties of fricative consonants. *The Journal of the Acoustical society of America*, 28(2):303–310. → pages 34

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, 45(3):188. → pages 8

Ingvalson, E. M., Holt, L. L., and McClelland, J. L. (2012). Can native japanese listeners learn to differentiate /r–l/ on the basis of F3 onset frequency? *Bilingualism: Language and Cognition*, 15(02):255–274. → pages 17

Iverson, G. K. and Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, 12(3):369–396. → pages 136

Iverson, P. and Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, 122(5):2842–2854. → pages 16

Iverson, P. and Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, 126(2):866–877. → pages 16

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., and Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87(1):B47–B57. → pages 16

Iverson, P., Pinet, M., and 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. → pages 16

Jakobson, R., Fant, G., and Halle, M. (1951). *Preliminaries to speech analysis. The distinctive features and their correlates*. The MIT Press, Cambridge. → pages 3

Johnson, K. and Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics*, 38(1):127–136. → pages 6

Jones, D. (1950). *The phoneme: Its nature and use*. Cambridge University Press, Cambridge. → pages 2

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263. → pages 34

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., and Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, 32(3):402–420. → pages 25

Jusczyk, P. W., Hohne, E. A., and Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, 61(8):1465–1476. → pages 14, 15

Jusczyk, P. W. and Luce, P. A. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5):630–645. → pages 25

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F., and Smith, L. B. (1977). Categorical perception of nonspeech sounds by 2-month-old infants. *Perception & Psychophysics*, 21(1):50–54. → pages 10

Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3):187–207. → pages 46

Kim, C.-W. (1970). A theory of aspiration. *Phonetica*, 21(2):107–116. → pages 136

Kingston, J. (1990). Articulatory binding. In Kingston, J. and Beckman, M., editors, *Papers in laboratory phonology I: Between the grammar and the physics of speech*, pages 406–434. Cambridge University Press, Cambridge. → pages 136

Kingston, J. (2003). Learning foreign vowels. *Language and Speech*, 46(2-3):295–348. → pages 16

Kirchner, R. M. (1998). *An effort based approach to consonant lenition*. PhD thesis, University of California Los Angeles. → pages 69

Kirkham, N. Z., Slemmer, J. A., and Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, 83(2):B35–B42. → pages 67

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7(312):1–26. → pages 3

Kluender, K. R., Coady, J. A., and Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, 41(1):59–69. → pages 87, 101

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2):93–107. → pages 11, 12

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, 4(6):812–822. → pages 18

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., and Nelson, T. (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(1493):979–1000. → pages 12

Kuhl, P. K. and Iverson, P. (1995). Chapter 4: Linguistic experience and the "perceptual magnet effect,". In Strange, W., editor, *Speech perception and linguistic experience: Issues in cross-language research*, pages 121–154. York Press, Baltimore. → pages 12

Kuhl, P. K. and Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, 63(3):905–917. → pages 10

Kuhl, P. K. and Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, 73(3):1003–1010. → pages 10

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608. → pages 1, 11, 12

Kuo, L.-J. (2009). The role of natural class features in the acquisition of phonotactic regularities. *Journal of Psycholinguistic Research*, 38(2):129–150. → pages 70

Lacerda, F. (1995). The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In Elenuis, K. and Branderud, P., editors, *Proceedings of the 13th International Congress of Phonetic Sciences*, volume 2, pages 140–147. → pages 18

Lacerda, F. (1998). An exemplar-based account of emergent phonetic categories. *Proceedings of the 16th Internatinoal Congresses on Acoustics*, 3:2013–2014. → pages 18

Ladd, D. R. (2014). *Simultaneous structure in phonology*. Oxford University Press, Oxford. → pages 3

Ladefoged, P. and Bhaskararao, P. (1983). Non-quantal aspects of consonant production-a study of retroflex consonants. *Journal of Phonetics*, 11(3):291–302. → pages 51

Ladefoged, P. and Johnson, K. (2014). *A course in phonetics*. Cengage Learning, Stamford, 7th edition. → pages 4, 138

Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*. Wiley-Blackwell, Hoboken. → pages 33, 34

Ladefoged, P. and Wu, Z. (1984). Places of articulation-an investigation of Pekingese fricatives and affricates. *Journal of Phonetics*, 12(3):267–278. → pages 33, 34, 90

Lasky, R. E., Syrdal-Lasky, A., and Klein, R. E. (1975). VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 20(2):215–225. → pages 10

Lee, C.-Y., Zhang, Y., Li, X., Tao, L., and Bond, Z. (2012). Effects of speaker variability and noise on Mandarin fricative identification by native and

non-native listeners. *The Journal of the Acoustical Society of America*, 132(2):1130–1140. → pages 39

Lee, S.-I. (2011). Spectral analysis of Mandarin Chinese sibilant fricatives. In Lee, W.-S. and Zee, E., editors, *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 1178–1181. → pages 35, 42

Lee, W.-S. (2008). Articulation of the coronal sounds in Peking dialect. In Sock, R., Fuchs, S., and Yvis, L., editors, *Proceedings of the 8th International Seminar on Speech Production*, pages 109–122. INRIA. → pages 33, 34, 90

Lee-Kim, S.-I. (2014). Revisiting Mandarin 'apical vowels': An articulatory and acoustic study. *Journal of the International Phonetic Association*, 44(3):261–282. → pages 33, 34, 90

Li, F. (2008). *The phonetic development of voiceless sibilant fricatives in English, Japanese and Mandarin Chinese*. PhD thesis, The Ohio State University. → pages 35

Li, F., Edwards, J., and Beckman, M. (2007). Spectral measures for sibilant fricatives of English, Japanese, and Mandarin Chinese. In Trouvain, J. and Barry, W. J., editors, *Proceedings of the 16th International Congress of Phonetic Sciences*, pages 917–920. → pages 35, 37

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6):431. → pages 9

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358. → pages 9

Liberman, A. M., Harris, K. S., Kinney, J. A., and Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 61(5):379. → pages 9

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36. → pages 86

Liberman, A. M. and Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, 4(5):187–196. → pages 86

Lim, S.-j. and Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, 35(7):1390–1405. → pages 17, 129, 130

Lin, S. S. (2011). *Production and perception of prosodically varying inter-gestural timing in American English laterals.* PhD thesis, The University of Michigan. → pages 4

Lin, Y. (2005). *Learning features and segments from waveforms: A statistical model of early phonological acquisition*. PhD thesis, University of California Los Angeles. → pages 20

Lindblom, B. E. and Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America*, 42(4):830–843. → pages 85, 86, 99, 107

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ vs. /p/ in trochees. *Language and Speech*, 29(1):3–11. → pages 84

Lively, S. E., Logan, J. S., and Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3):1242–1255. → pages 16

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III: Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, 96(4):2076–2087. → pages 16, 17

Löfqvist, A. and Yoshioka, H. (1981). Interarticulator programming in obstruent production. *Phonetica*, 38(1-3):21–34. → pages 136

Logan, J. S., Lively, S. E., and Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, 89(2):874–886. → pages 16

Lotto, A. J. and Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, 60(4):602–619. → pages 87, 101, 137

Lotto, A. J., Kluender, K. R., and Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *The Journal of the Acoustical Society of America*, 102(2):1134–1140. → pages 87, 101

Lotto, A. J., Sato, M., and Diehl, R. L. (2004). Mapping the task for the second language learner: the case of Japanese acquisition of /r/ and /l/. In Slifka, J., Manuel, S., and Matthies, M., editors, *Proceedings of From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, pages C181–C186. → pages 16

Lu, Y.-a. (2011). The psychological reality of phonological representations: The case of Mandarin fricatives. In Zhuo, J.-S., editor, *Proceedings of the 23rd North American Conference on Chinese Linguistics*, volume 1, pages 251–226. → pages 34

Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide*. Lawrence Erlbaum Associate, Inc., Publishers, Mahwah, 2nd edition. → pages 57, 58, 79

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, 28(5):407–412. → pages 85, 86, 87, 88, 99, 107, 116

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r". *Cognition*, 24(3):169–196. → pages 100

Mann, V. A. and Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, 28(3):213–228. → pages 85, 99

Marslen-Wilson, W. and Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):576. → pages 8

Martin, A., Peperkamp, S., and Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124. → pages 127

Maye, J. and Gerken, L. (2000). Learning phonemes without minimal pairs. In Howell, C., Fish, S., and Keith-Lucas, T., editors, *Proceedings of the 24th Annual Boston University Conference on Language Development*, volume 2, pages 522–533, Somerville. Cascadilla Press. → pages 21, 32, 55, 89

Maye, J. and Gerken, L. (2001). Learning phonemes: How far can the input take us. In Domínguez, L. and Johansen, A., editors, *Proceedings of the 25th annual Boston University Conference on Language Development*, volume 1, pages 480–490, Somerville. Cascadilla Press. → pages 21, 23

Maye, J. and Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. In Beachley, B., Brown, A., and Conlin, F., editors, *Proceedings of the 27th annual Boston University Conference on Language Development*, volume 2, pages 508–518, Somerville. Cascadilla Press. → pages 23

Maye, J., Weiss, D. J., and Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11(1):122–134. → pages 20, 23

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111. → pages xi, 1, 2, 20, 21

Maye, J. C. (2000). *Learning speech sound categories from statistical information*. PhD thesis, University of Arizona. → pages 2, 18, 21, 23, 32, 62, 63, 64, 84

McGuire, G. L. (2007a). English listeners' perception of Polish alveopalatal and retroflex voiceless sibilants: A pilot study. *UC Berkeley Phonology Lab Annual Report*, pages 391–415. → pages 40, 44, 90

McGuire, G. L. (2007b). *Phonetic category learning*. PhD thesis, The Ohio State University. → pages 44, 90

McGuire, G. L. (2008). Selective attention and English listeners' perceptual learning of the Polish post-alveolar sibilant contrast. Unpablished manuscript. → pages 40, 90

McMurray, B., Aslin, R. N., and Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, 12(3):369–378. → pages 20

McMurray, B. and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219. → pages 34

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, 18(5):331–340. → pages 16, 100

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics*, 9:283–303. → pages 16

Moore, B. C. and Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The Journal of the Acoustical Society of America*, 74(3):750–753. → pages 43

Moray, N. and Taylor, A. (1958). The effect of redundancy in shadowing one of two dichotic messages. *Language and Speech*, 1(2):102–109. → pages 8

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, 25(1):83–127. → pages 71

Moreton, E. (2012). Inter- and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, 67(1):165–183. → pages 72

Moreton, E. and Pater, J. (2012a). Structure and substance in artificial-phonology learning, part I: Structure. *Language and Linguistics Compass*, 6(11):686–701. → pages 67, 69, 71, 72

Moreton, E. and Pater, J. (2012b). Structure and substance in artificial-phonology learning, part II: Substance. *Language and Linguistics Compass*, 6(11):702–718. → pages 67, 71, 72

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467. → pages 44

Nakatani, L. H. and Dukes, K. D. (1977). Locus of segmental cues for word juncture. *The Journal of the Acoustical Society of America*, 62(3):714–719. → pages 7

Nazzi, T., Bertoncini, J., and Bijeljac-Babic, R. (2009). A perceptual equivalent of the labial-coronal effect in the first year of life. *The Journal of the Acoustical Society of America*, 126(3):1440–1446. → pages 25

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American sign language. *Language Sciences*, 10(1):147–172. → pages 23

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28. → pages 23

Newport, E. L. and Aslin, R. N. (2004). Learning at a distance I: Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2):127–162. → pages 67, 68, 73, 83

Nittrouer, S. (1992). Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics*, 20(3):351–382. → pages 36

Nittrouer, S. (2002). Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, 112(2):711–719. → pages 36

Nittrouer, S. and Miller, M. E. (1997a). Developmental weighting shifts for noise components of fricative-vowel syllables. *The Journal of the Acoustical Society of America*, 102(1):572–580. → pages 36

Nittrouer, S. and Miller, M. E. (1997b). Predicting developmental shifts in perceptual weighting schemes. *The Journal of the Acoustical Society of America*, 101(4):2253–2266. → pages 36

Nittrouer, S. and Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the perception of fricatives by children and adults. *Journal of Speech, Language, and Hearing Research*, 30(3):319–329. → pages 36

Noguchi, M., Chiu, C., Po-Chun, W., and Yamane, N. (2015a). *Uncovering sibilant fricative merger in Taiwan Mandarin: Evidence from ultrasound imaging and acoustics*. Poster presented at Linguistic Society of America 2015 Annual Meeting, Portland. → pages 33, 90, 172

Noguchi, M., Chiu, C., Wei, P.-C., and Yamane, N. (2015b). Contrastive tongue shapes of the three sibilant fricatives in Taiwan Mandarin read speech. *Canadian Acoustics*, 43(3). → pages 172

Noguchi, M. and Hudson Kam, C. L. (2014a). *Learning phonetic categories with phonotactics: The influence of predictability and phonetic naturalness*. Poster presented at The 39th Annual Boston University Conference on Language Development, Boston. → pages iv

Noguchi, M. and Hudson Kam, C. L. (2014b). *Learning sound categories with phonotactics*. Poster presented at The 14th Conference on Laboratory Phonology, Tokyo. → pages iv

Noguchi, M. and Hudson Kam, C. L. (2015a). Categorical perception of post-alveolar sibilants by Taiwan and Beijing Mandarin speakers. *Canadian Acoustics*, 43(3). → pages v, 47, 102

Noguchi, M. and Hudson Kam, C. L. (2015b). *Learning the context-dependent perception of novel speech sounds*. Poster presented at 2015 Annual Meeting on Phonology, Vancouver. → pages iv

Nowak, P. M. (2006). The role of vowel transitions and frication noise in the perception of Polish sibilants. *Journal of Phonetics*, 34(2):139–152. → pages 35

Ong, J. H., Burnham, D., and Escudero, P. (2015). Distributional learning of lexical tones: A comparison of attended vs. unattended listening. *PLOS ONE*, 10(7). → pages 21

Onishi, K. H., Chambers, K. E., and Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, 83(1):B13–B23. → pages 2, 26, 70

Pajak, B. (2012). *Inductive inference in non-native speech processing and learning*. PhD thesis, University of California San Diego. → pages 21, 22, 23, 55, 63

Pajak, B. and Levy, R. (2012). Distributional learning of L2 phonological categories by listeners with different language backgrounds. In Biller, A. K., Chung, E. Y., and Kimball, A. E., editors, *Proceedings of the 36th Boston University Conference on Language Development*, volume 2, pages 400–413, Somerville. Cascadilla Press. → pages 21, 22

Pastore, R. E., Layer, J. K., Morris, C. B., and Logan, R. J. (1988). Temporal order identification for tone/noise stimuli with onset transitions. *Perception & Psychophysics*, 44(3):257–271. → pages 10

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

Peperkamp, S., Le Calvez, R., Nadal, J.-P., and Dupoux, E. (2006a). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41. → pages 2, 4, 24

Peperkamp, S., Pettinato, M., and Dupoux, E. (2003). Allophonic variation and the acquisition of phoneme categories. In Beachley, B., Brown, A., and Conlin, F., editors, *Proceedings of the 27th annual Boston University Conference on Language Development*, volume 2, pages 650–661, Somerville. Cascadilla Press. → pages 6, 7, 26, 27, 32

Peperkamp, S., Skoruppa, K., and Dupoux, E. (2006b). The role of phonetic naturalness in phonological rule acquisition. In Magnitskaia, T. and Colleen, Z., editors, *Proceedings of the 30th annual Boston University Conference on Language Development*, volume 2, pages 464–475, Somerville. Cascadilla Press. → pages 70, 73

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184. → pages 18

Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications, Inc., New York, 2nd edition. → pages 8

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, 46(2-3):115–154. → pages 1, 18

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 13(2):253–260. → pages 7, 55

Polka, L. and Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *The Journal of the Acoustical Society of America*, 100(1):577–592. → pages 11

Polka, L. and Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421. → pages 11

Pons, F., Sabourin, L., Cady, J. C., and Werker, J. F. (2006). *Distributional learning in vowel distinctions by 8-month-old English infants*. Paper presented at The 28th Annual Conference of the Cognitive Science Society, Vancouver. → pages 20

Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences*, 32(1):43–55. → pages 3

Port, R. F. and Leary, A. P. (2005). Against formal phonology. *Language*, 81(4):927–964. → pages 3

Proctor, M., Lu, L. H., Zhu, Y., Goldstein, L., Narayanan, S., et al. (2012). Articulation of Mandarin sibilants: A multi-plane realtime MRI study. In *Proceedings of The 14th Australasian International Conference on Speech Science and Technology*, pages 113–116. → pages 33, 34, 90

Pycha, A., Nowak, P., Shin, E., and Shosted, R. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In Garding, G. and Tsujimura, M., editors, *Proceedings of The 22nd West Coast Conference on Formal Linguistics*, pages 101–114, Somerville. Cascadilla Press. → pages 69, 70

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. → pages 58, 79, 110

Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, 30(3):217–227. → pages 85, 99

Repp, B. H. and Mann, V. A. (1981). Perceptual assessment of fricative–stop coarticulation. *The Journal of the Acoustical Society of America*, 69(4):1154–1163. → pages 85, 99

Rost, G. C. and McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2):339–349. → pages 13, 14, 127

Rost, G. C. and McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, 15(6):608–635. → pages 13, 14, 127

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, 47(1):172–196. → pages 67

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928. → pages 2, 8, 67

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52. → pages 67

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4):606–621. → pages 2, 67

Saffran, J. R. and Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, 39(3):484–494. → pages 25, 70

Schane, S. A., Tranel, B., and Lane, H. (1975). On the psychological reality of a natural rule of syllable structure. *Cognition*, 3(4):351–358. → pages 30, 68

Schneider, W., Eschman, A., and Zuccolotto, A. (2002). *E-Prime: User's guide*. Psychology Software Inc. → pages 55, 79, 107

Schouten, B., Gerrits, E., and van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, 41(1):71–80. → pages 9

Seidl, A. and Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, 1(3-4):289–316. → pages 25, 69

Seidl, A. and Cristia, A. (2012). Infants' learning of phonological status. *Frontiers in psychology*, 3(448):1–10. → pages ii, 2

Seidl, A., Cristia, A., Bernard, A., and Onishi, K. H. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, 5(3):191–202. → pages 2, 14, 24

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., and Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, 115(3):435–443. → pages 129

Seitz, A. R. and Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, 49(21):2604–2610. → pages 129

Sekiyama, K. and Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21(4):427–444. → pages 16

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64. → pages 7

Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of information*. University of Illinois Press, Chicago. → pages 7

Skoruppa, K., Lambrechts, A., and Peperkamp, S. (2011). The role of phonetic distance in the acquisition of phonological alternations. In Lima, S., Mullin, K., and Smith, B., editors, *Proceedings of The 39the Annual Meeting of the North East Linguistic Society*, volume 2, pages 464–475. CreateSpace Independent Publishing Platform. → pages 70, 73

Smith, R. L. (1979). Adaptation, saturation, and physiological masking in single auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 65(1):166–178. → pages 87

Sproat, R. and Fujimura, O. (1993). Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics*, 21(3):291–311. → pages 138

Stager, C. L. and Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640):381–382. → pages 13, 127

Stevens, K. N. and Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, 3(4):215–233. → pages 53

Stevens, K. N., Li, Z., Lee, C.-Y., and Keyser, S. J. (2004). A note on Mandarin fricatives and enhancement. In Fant, G., Fujisaki, H., Cao, J., and Xu, Y., editors, *From traditional phonology to modern speech processing: Festschrift for professor Wu Zongji's 95th birthday*, pages 393–403. Foreign Language Teaching and Research Press, Beijing. → pages 35

Streeter, L. A. (1976). Language perception of 2-mo-old infants shows effects of both innate mechanisms and experience. *Nature*, 259(5538):39–41. → pages 11

Summerfield, Q. (1975). How a full account of segmental perception depends on prosody and vice versa. In Cohen, A. and Nooteboom, S. G., editors, *Structure and process in speech perception*, pages 51–68, Berlin. Springer-Verlag. → pages 85, 99

Sussman, H. M., McCaffrey, H. A., and 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. → pages 35

Svantesson, J.-O. (1986). Acoustic analysis of Chinese fricatives and affricates. *Journal of Chinese Linguistics*, 14(1):53–70. → pages 35

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3617–3632. → pages 127

Takagi, N. (1993). *Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view*. PhD thesis, University of California Irvine. → pages 16

Tees, R. C. and Werker, J. F. (1984). Perceptual flexibility: maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 38(4):579. → pages 15

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56(1):16–34. → pages 13, 127, 128

Thiessen, E. D. (2011a). Domain general constraints on statistical learning. *Child Development*, 82(2):462–470. → pages 67

Thiessen, E. D. (2011b). When variability matters more than meaning: the effect of lexical forms on use of phonemic contrasts. *Developmental Psychology*, 47(5):1448. → pages 13, 127, 128

Toda, M. and Honda, K. (2003). *An MRI-based cross-linguistic study of sibilant fricatives*. Paper presented at 6th International Seminar on Speech Production, Manly. → pages 33, 34, 37, 90

Toscano, J. and McMurray, B. (2010). Cue integration with categories: A statistical approach to cue weighting and combination in speech perception. *Cognitive Science*, 34(3):434–464. → pages 21

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, 47(2):466–472. → pages 11

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12(4):242–248. → pages 8

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *The American Journal of Psychology*, 77(2):206–219. → pages 8

Treisman, A. M. (1965). The effects of redundancy and familiarity on translating and repeating back a foreign and a native language. *British Journal of Psychology*, 56(4):369–379. → pages 8

Trubetzkoy, N. S. (1969). *Principles of phonology.* University of California Press, Berkeley. → pages 1, 2, 4

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno, M., Menyuk, P., and Best, C. T. (1994). Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. In *Proceedings of The 3rd International Conference on Spoken Language Processing*, pages 1695–1698, Yokohama. Acoustical Society of Japan. → pages 11

Twaddell, W. F. (1935). On defining the phoneme. *Language*, 11(1):5–62. → pages 1

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278. → pages 20

Viswanathan, N., Magnuson, J. S., and Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, 36(4):1005–1015. → pages 100

Vlahou, E. L., Protopapas, A., and Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, 141(2):363–381. → pages 129, 130

Wade, T. and Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, 118(4):2618–2633. → pages 129

Wagner, A., Ernestus, M., and Cutler, A. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *The Journal of the Acoustical Society of America*, 120(4):2267–2277. → pages 36

Waters, R. and Wilson, W. (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception & Psychophysics*, 19(4):285–289. → pages 10

Werker, J. F. and Curtin, S. (2005). Primir: A developmental framework of infant speech processing. *Language Learning and Development*, 1(2):197–234. → pages 1

Werker, J. F. and Fennell, C. T. (2004). Listening to sounds versus listening to words: Early steps in word learning. In Hall, D. G. and Waxman, S. R., editors, *Weaving a lexicon*, pages 79–109. The MIT Press, Cambridge. → pages 127

Werker, J. F. and Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, 37(1):35–44. → pages 55

Werker, J. F. and Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 37(2):278. → pages 11

Werker, J. F. and 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. → pages 1, 11

Whalen, D. (1981a). Effects of nonessential cues on the perception of english [s] and [sš]. *The Journal of the Acoustical Society of America*, 69(S1):S94–S94. → pages 36

Whalen, D. H. (1981b). Effects of vocalic formant transitions and vowel quality on the English [s]–[š] boundary. *The Journal of the Acoustical Society of America*, 69(1):275–282. → pages 36

Whalen, D. H. (1991). Perception of the English /s/–/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *The Journal of the Acoustical Society of America*, 90(4):1776–1785. → pages 36

Whalen, D. H., Best, C. T., and Irwin, J. R. (1997). Lexical effects in the perception and production of American English /p/ allophones. *Journal of Phonetics*, 25(4):501–528. → pages 6, 7

White, J. and Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition*, 133(1):85–90. → pages 72, 141

White, K. S., Peperkamp, S., Kirk, C., and Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, 107(1):238–265. → pages 25, 26, 140

Wickens, C. D. (1981). *Processing resources in attention, dual task performance, and workload assessment* (technical report EPL–81–3/ONR–81–3). Engineering-Psychology Research Laboratory, University of Illinois at Urbana-Champaign, Urbana-Champaign. → pages 8

Wilde, L. (1993). Inferring articulatory movements from acoustic properties at fricative-vowel boundaries. *The Journal of the Acoustical Society of America*, 94(3):1881–1881. → pages 34, 35

Wilson, C. (2003). Experimental investigation of phonological naturalness. In Garding, G. and Tsujimura, M., editors, *Proceedings of The 22nd West Coast Conference on Formal Linguistics*, pages 533–546, Somerville. Cascadilla Press. → pages 30, 70, 71, 73

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5):945–982. → pages 30, 51, 71, 72, 73, 83, 132

Yamada, R. A. and Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, 52(4):376–392. → pages 16

Yeung, H. H. and Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113(2):234–243. → pages 127

Yoshida, K. A., Pons, F., Maye, J., and Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, 15(4):420–433. → pages 20, 22

Yoshioka, H., Löfqvist, A., and Hirose, H. (1981). Laryngeal adjustments in the production of consonant clusters and geminates in American English. *The Journal of the Acoustical Society of America*, 70(6):1615–1623. → pages 136

Zygis, M. and Padgett, J. (2010). A perceptual study of Polish fricatives, and its implications for historical sound change. *Journal of Phonetics*, 38(2):207–226. → pages 35

Appendix A

Categorical perception of post-alveolar fricatives by native speakers of Mandarin

In order to find the location of a perceptual boundary between the retroflex [ʂa] and alveolopalatal [ɕa], I tested the categorical perception of the syllables from the 10-step continuum between [ʂa] and [ɕa] by native speakers of Mandarin. Previous studies have demonstrated that there are some differences between the regional varieties of Mandarin with respect to the production of the retroflex sibilants. Specifically, while the phonetic contrast between the dental and retroflex sibilants are maintained in the variety spoken around Beijing (so-called "Standard Mandarin"), the contrast tend to be lost in casual speech in other varieties, including the one spoken in Taiwan (Chang, 2013; Chung, 2006; Noguchi et al., 2015a,b). Since the stimuli used in this study were produced by a speaker from Taiwan, I tested the perception of Mandarin speakers from both Beijing and Taiwan.

A.1 Design

I used an ABX discrimination task as well as an identification task to test the categorical perception of the phonetic contrast between retroflex [ʂ] and alveolopalatal [ɕ]. Stimuli were drawn from the 10-step continuum from retroflex [ʂa] to alve-

olopalatal [ɕa] (see Secttion 2.3.2). In the ABX discrimination task, participants compared two test items that were separated by one step (e.g., step 1 vs. step 3). ISI was 750 ms. In each trial, participants were given a maximum of five seconds to respond, but the trial was terminated whenever they recorded a response. ITI was two seconds. In the identification task, participants were asked to label a single test item either as retroflex or alveolopalatal. In each trial, participants were given a maximum of five seconds to respond, but the trial was terminated whenever they recorded a response. ITI was two seconds.

A.2 Participants

10 Taiwan Mandarin speakers and 7 Beijing Mandarin speakers participated in the study. All were living in or visiting Vancouver at the time of the study. All participants in the Taiwan Mandarin group self-reported living in Taiwan until adolescence. Similarly, all participants in the Beijing group self-reported living in Beijing until adolescence. Participants were paid $5 for their participation.

A.3 Procedure

All participants did the ABX discrimination task first and the identification task second. Instructions were given in Mandarin in written form. For the ABX discrimination task, participants heard three syllables and were asked to decide whether the last syllable was identical to the first one or the second one. There were eight blocks in this task. Each block contained 32 trials (8 ABX triads in 4 different orders). For the identification task, participants heard a single syllable and were asked to identify it as either retroflex [ʂa] or alveolopalatal [ɕa]. Category labels were shown in *Zhuyin* for Taiwan Mandarin speakers and *Pinyin* for Beijing Mandarin speakers. There were eight blocks in the task, each consisting of 10 trials, one per step on the continuum. Trials within a block were presented in random order.

A.4 Results

For the discrimination task, responses to each unique triad were converted into sensitivity scores (d'). Figure A.1 shows the mean d' scores by pair and language group. Participants in both groups show the highest sensitivity for the trials comparing step 5 and step 7. A repeated-measures ANOVA was conducted with d' scores as dependent variable and language group as a between-participant factor and pair as a within-participant factor. The analysis yielded a significant effect of pair [$F(7,105) = 21.45$, $p < 0.001$] but not language group [$F(1,15) = 1.68$, $p = 0.21$]. There was no interaction between the two factors [$F(7,105) = 1.02$, $p = 0.41$].

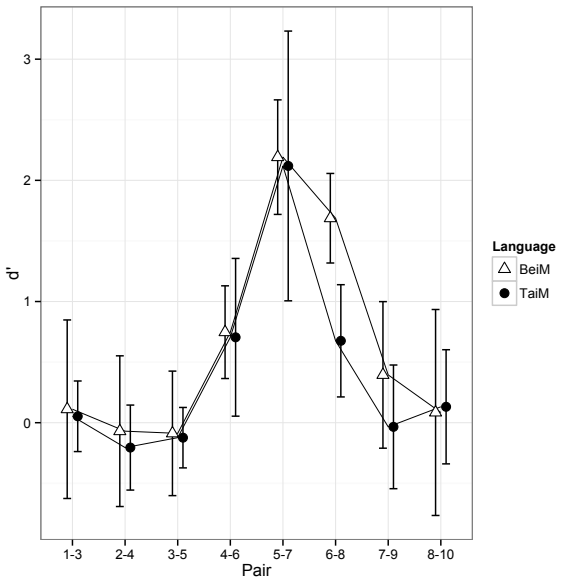


Figure A.1: Mean d' scores (with 95% CI)

For the identification task, the proportion of /ɕa/ responses was calculated for each step on the continuum. Figure A.2 shows the mean proportion of /ɕa/ responses by step and language group. For both groups, the proportion of /ɕa/ responses is close to 0 in steps 1-5, jumps to around .5 at step 6, then increases to close to 1 in steps 7-10, indicating a perceptual boundary between retroflex [ʂa] and alveolopalatal [ɕa] at step 6. Note that both Taiwan and Beijing Mandarin speakers

show the same identification curve.

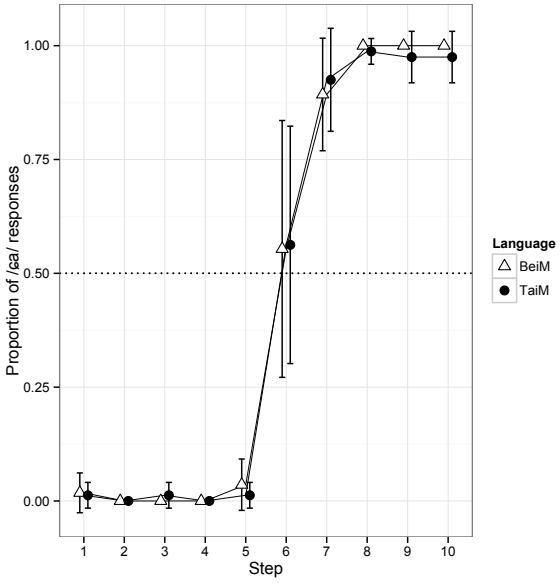


Figure A.2: Proportion of /ɕa/ responses (with 95% CI)

The results of this study show that both Taiwan and Beijing Mandarin speakers perceive the contrast between retroflex [ʂa] and alveolopalatal [ɕa] in a categorical fashion, and there is no difference between the two language groups in where they place the boundary. Both Taiwan and Beijing Mandarin speakers placed the category boundary at step 6.