Attention and salience in lexically-guided perceptual learning

by

Michael McAuliffe

B.A., University of Washington, 2009

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

Doctor of Philosophy

in

THE FACULTY OF ARTS
(Linguistics)

The University Of British Columbia
(Vancouver)

April 2015

Abstract

This dissertation presents three experiments investigating perceptual learning in speech perception under different attention and salience conditions. Listeners across all three experiments were exposed to a speaker's /s/ category that had been modified to sound more like /ʃ/. All listeners showed a perceptual learning effect on a categorization task following exposure, categorizing more of continua from /s/ to /ʃ/ as /s/ than a control group that completed only the categorization task, replicating previous lexically-guided perceptual learning results (e.g., Norris et al., 2003; Reinisch et al., 2013). However, the magnitude of perceptual learning effects differed depending on the linguistic context of the modified /s/ category and the listener's attentional set. The linguistic context of modified categories has previously been shown to affect a wide range of psycholinguistic tasks, such as phoneme identification (Pitt & Samuel, 2006; Borsky et al, 1998) and phoneme restoration (Samuel, 1981). The attentional set of the listener has also been shown to affect how tolerant they are to modified categories (Pitt & Szostak, 2012). This dissertation replicates these studies in the context of lexically-guided perceptual learning, showing that these factors that influence the online processing of speech also influence the longer term adaptation to speakers. In Experiment 1, listeners exposed to the modified category in the middle of words showed larger perceptual learning effects than those exposed to the category at the beginnings of words. However, this difference in magnitude was only observed when listeners were not told that the speaker's /s/ sounds would be ambiguous. When listeners were given these instructions, the difference between word-initial and word-medial exposure was not present. In Experiment 2, the exposure conditions were the same as Experiment 1, but the modified /s/ category was more atypical, sounding more like /ʃ/ than /s/ rather than halfway in between. In this experiment, exposure conditions had no effect, and all participants showed the same size of perceptual learning effect. In Experiment 3, listeners were exposed to the same modified /s/ category as Experiment 1 in word-medial position only, but the words were embedded at the end of sentences. These sentences differed in how strongly they conditioned the final word, with one group of listeners hearing the modified category only in sentences that were predictive of the target word, and the other group only hearing the category in sentences that were not predictive of the target word. The same pattern as Experiment 2 is present for Experiment 3, with no significant differences in exposure conditions. These results are framed in a predictive coding framework (Clark, 2013), which is a hierarchical Bayesian model of the brain. Predictions from higher levels are sent to lower ones, and mismatches between these predictions and the lower level's input cause an error signal to propagate back to the higher levels. The expectations for future stimuli are then tuned based on this error signal. The model captures perceptual learning results well as a whole, but the mechanism for attention in the framework is a simple gain based one, where attention gives increased weight to the error signals. Such a mechanism would predict that attention would result in increased perceptual learning, but that prediction is not supported by results of this dissertation's experiments. Instead, increased attention to low-level perception reduces the amount of perceptual learning observed. Increased attention to perception can be induced in different ways either through explicit instructions, linguistic prominence, or the salience of the atypicality of the category. I propose a mechanism of attention in the predictive coding framework that inhibits propagation of error signals beyond the level to which attention is oriented. Attention focused on perception will inhibit the propagation of error signals to higher levels, limiting the updating of expectations at higher levels and leading to reduced perceptual learning. Perception-oriented tasks, such as psychophysical visual perceptual learning or visually-guided perceptual learning in speech perception, show a lack of generalization, with effects limited to exposure items (Gilbert et al., 2001; Reinisch et al., 2014), but comprehension-oriented tasks, such as lexically-guided perceptual learning, are more likely to show generalization to new items (Norris et al., 2003; Reinisch et al., 2013). The propagation-inhibiting mechanism for attention can account for the degrees of generalization reported across the literature.

Revision: May 2, 2015

Preface

At University of British Columbia (UBC), a preface may be required. Be sure to check the Graduate and Postdoctoral Studies (GPS) guidelines as they may have specific content to be included.

Contents

List of Tables

List of Figures

Glossary

This glossary uses the handy `acroynym` package to automatically maintain the glossary. It uses the package's `printonlyused` option to include only those acronyms explicitly referenced in the LaTeX source.

GPS	Graduate and Postdoctoral Studies

Acknowledgments

Thank those people who helped you.

Molly!

Kathleen!

530 class!

Jobie!

Jamie!

Michelle!

Don’t forget your parents or loved ones.

Mom!

Dad!

Laura!

You may wish to acknowledge your funding sources.

Chapter 1

Introduction

Listeners of a language are faced with a large degree of phonetic variability when interacting with their fellow language users. Speakers can have different sizes, different genders, and different backgrounds that make speech sound categories, at first blush, overlapping in distribution and hard to separate in acoustic dimensions. In addition to properties of the speaker varying, a listener's attention or goals in an interaction can vary, such as paying attention to a non-native speaker or being distracted by planning upcoming utterances or by another task entirely. Despite variability on the part of both the listener and the speaker, listeners can interpret disparate and variable productions as belonging to a single word type or sound category, a phenomenon referred to as perceptual constancy in categorization studies [30, 56] and as recognition equivalence in word recognition tasks [58]. One of the processes for achieving this constancy is perceptual learning, whereby perceivers update their representation of a category based on contextual factors.

In the speech perception literature, perceptual learning refers to the updating of distributions corresponding to a sound category, such as /s/ or /ʃ/, following exposure to speech with a modification to the sound category's distribution [40]. Exposure to a speaker affects a listener's perceptual system for that speaker, even after only a few tokens [29, 60] or within the span of a single utterance [31]. Perceptual learning is not limited to single speakers, and exposure to multiple speakers sharing a non-native accent increases the intelligibility of new speakers with that accent [8].

Generalization to novel contexts has been a locus of investigation within the speech-focused perceptual learning in speech literature. In general, exposure to modified sound categories, such as /s/ or /ʃ/, generalizes well to other words and nonwords when the modified exposure tokens are embedded in real words [40, 49]. This paradigm is referred to as lexically-guided perceptual learning, as listeners are exposed to the modified sounds in the context of real words in a lexical decision task.

On the other hand, generalization appears to be more limited when perceptual learning is induced through visually-guided paradigms in which listeners are exposed to a modified category through matching it to an unambiguous video signal. The perceptual learning exhibited from these visually-guided experiments is only found to influence the specific nonword that perceivers are exposed to, and not other similar nonwords [50]. This kind of exposure-specificity effect has been widely reported in perceptual learning studies in the psychophysics literature [20].

Why, then, does lexically-guided perceptual learning produce such generalization? Here I posit that these results can be understood by considering the attentional set exploited in the exposure phase. An attentional set is a strategy that is employed by perceivers to prioritize certain aspects of stimuli, such as color or shape in the visual domain. In the speech perception literature, two broad attentional sets have been posited under varying names [14, 43]. The first is a comprehension-oriented, or diffuse, attentional set, which is likely the most similar to normal language use. In this attentional set, listeners are focused on comprehending the intended message of the speech, and a comprehension set is promoted by tasks that focus on word identity and word recognition. The comprehension-oriented attentional set would be elicited in the lexically-guided perceptual learning paradigms through their use of lexical decision tasks and the embedding of modified sound categories in word tokens. A second kind of attentional set is a perception-oriented, or focused, attentional set, where a listener is focused more on the low-level, signal properties of the speech rather than the message. The perception-oriented attentional set is promoted by tasks such as phoneme or syllable monitoring or mispronunciation detection. The tasks used in visually-guided perceptual learning and perceptual learning within the psychophysics literature would elicit this attentional set, due to their focus on stimuli that are devoid of linguistic meaning.

The hypothesis of this dissertation is that perceivers who adopt a more comprehension-oriented attentional set will show more generalization than those who adopt a more perception-oriented attentional set.

To test this hypothesis, I use a lexically-guided perceptual learning paradigm to expose listeners to an /s/ category modified to sound more like /ʃ/. Groups of participants differ in whether comprehension-oriented or perception-oriented attentional sets are favored when processing the modified /s/ category based on experimental manipulations. The favoring of attentional sets are implemented in four ways across the three experiments presented in this dissertation. The first two manipulations are linguistic in nature, and the other two are instruction-based and stimuli-based.

The first linguistic manipulation that affects attentional sets is the position of the modified /s/ in the exposure words. Some groups of participants are exposed to the modified sound only at the beginnings of words, such as *silver* or *settlement*, and other groups are exposed to the category only in the middle of words, such as carousel or *fossil*. Word-initial exposure is predicted to promote a more perception-oriented attentional set, as that position is important in linguistic theory [3] and lexical information exhibits less of an effect on that position [46]. In contrast, word-medial exposure is predicted to promote a more comprehension-oriented attentional set, as the identity of the word will be largely known by the time the modified sound is processed. Previous work has shown that lexical information exerts greater influence the later in the word a given sound is [46]. Perception is more critical earlier in the word than later, as the set of possible words shrinks as incoming information progressively narrows the possible words. Experiments 1 and 2 use this manipulation in Chapter 2, and further background is given in Section 1.2.1 of this chapter.

The second linguistic manipulation employed is the context that a word appears in. In Experiments 1 and 2, participants are exposed to the modified sound category in words in isolation, as in previous work [40]. However, in Experiment 3, the words containing the sound category have been embedded in sentences that are either predictive or unpredictive of the target word. Using sentence frames is predicted to promote comprehension-oriented attentional sets more than words in isolation; further background on the use of the sentence frames is given in Section 1.2.2.

Participants in all three experiments receive the same general instructions for the exposure task, but groups of participants in each experiment receive additional instructions about the nature of the /s/ category, following previous studies [43]. Without any additional instructions, the task is predicted to promote a comprehension-oriented attentional set, and the instructions about /s/ are expected to promote a perception-oriented attentional set. Section 1.3 contains background on attention and instructions.

The stimuli-based manipulation is the typicality of the modified /s/ category – Experiments 1 and 2 differ in this respect. In line with previous work [40], participants in Experiment 1 are exposed to a modified category halfway between /s/ and /ʃ/. Participants in Experiment 2 are exposed to an even more atypical /s/ – the modified fricative is more /ʃ/-like than /s/-like. A more atypical category is predicted to promote a perception-oriented attentional set than a more typical category because its physical salience is predicted to be more perceptually salient.

Perceptual salience is relevant in this dissertation only in so far as it promotes a perception-oriented attentional set. Salience is a widely-used and poorly-defined term across literatures, and for the purposes of this dissertation I adopt the following definition: an element is salient if it is unpredictable from context or easily distinguishable from other possible elements. The sound category that is being learned in this dissertation already has salient signal properties (e.g., in the form of high frequency, relatively high amplitude, aperiodic noise). Therefore, increasing salience of the modified /s/ category is a function of embedding it a linguistic position with little conditioning context. Increasing the acoustic distance of the modified category to the typical distribution of /s/, and in turn increasing the distinguishability, will also increase the salience of the category and promote a more perception-oriented attentional set.

The results of these experiments will be contextualized in the existing perceptual learning literature and in models of perceptual learning. Section 1.1 reviews the perceptual learning literature in greater depth, as well as conceptual models that have been proposed for it [11, 25]. Broadly speaking, perceptual learning is well accounted for by a Bayesian model, where differences between expectations and observed input cause an error signal which updates future expectations, leading to perceptual learning. The results of the experiments in this dissertation support a more nuanced attention mechanism, however, than that proposed in Clark [11]. The mechanism for attention proposed by Clark is gain-based such that increased attention increases the weight of error signals, thus increased attention should lead to increased perceptual learning. As reviewed in Section 1.1, generalization of perceptual learning is dependent on the task and the attentional set being employed. An expanded attention mechanism is therefore proposed to unify the results of perceptual learning studies across the psychophysics and speech perception literatures, encompassing the results of the current experiments.

The proposed mechanism inhibits error propagation beyond the level where attention is directed. In a perception-oriented attentional set, perceptual learning is hypothesized to occur at the initial perception where encoding is more fine-grained, inhibiting generalization. Comprehension-oriented attentional sets are predicted to allow for greater error propagation and abstraction, leading to greater levels of generalization.

The structure of the thesis is as follows. This chapter will review the recent literature on perceptual learning in speech perception (Section 1.1), as well as literature on linguistic expectations (Section 1.2), attention (Section 1.3), and category typicality (Section 1.4) as they relate to perceptual learning and to the three experiments of this dissertation. Chapter 2 will detail two experiments using a lexically-guided perceptual learning paradigm, each with different conditions for levels of lexical bias and attention. The two experiments differ in the acoustic properties of the exposure tokens, with the first experiment using a slightly atypical /s/ category that is halfway between /s/ and /ʃ/, and the second experiment using a more atypical /s/ category that is more /ʃ/-like than /s/-like. Chapter 3 details an experiment using a novel perceptual learning paradigm that manipulates an additional linguistic factor, namely semantic predictability, to increase the linguistic expectations during exposure. Finally, Chapter 4 summarizes the results and places them in a theoretical framework. The perceptual learning literature has generally used consistent processing conditions to elicit perceptual learning effects, and a goal of this dissertation is to examine the robustness and degree of perceptual learning across conditions that promote the two primary attentional sets.

1.1 Perceptual learning

Perceptual learning is a well established phenomenon in the psychology and psychophysics literature. Training can improve a perceiver's ability to discriminate in many disparate modalities (e.g., visual acuity, somatosensory spatial resolution, weight estimation, and discrimination of hue and acoustic pitch [20, for review]). In the psychophysics literature, perceptual learning is an improvement in a perceiver's ability to judge the physical characteristics of objects in the world through training that assumes attention on the task, but doesn't require reinforcement, correction or reward. This definition of perceptual learning corresponds more to what is termed "selective adaptation" in the speech perception literature rather than what is termed "perceptual learning", "perceptual adaptation" or "perceptual recalibration." In speech perception, selective adaptation is the phenomenon where listeners that are exposed repeatedly to a narrow distribution of a sound category, narrow their own perceptual category, resulting in a change in variance of the category, but not of the mean of the category (along some acoustic-phonetic dimension) [15, 52, 60]. Perceptual learning or recalibration in the speech perception literature is a more broad updating of perceptual categories, either in mean or variance [40, 60].

Norris et al. [40] began the recent set of investigations into lexically-guided perceptual learning in speech. Norris and colleagues exposed one group of Dutch listeners to a fricative halfway between /s/ and /f/ at the ends of words like *olif* ”olive” and *radijs* ”radish”, while exposing another group to the ambiguous fricative at the ends of nonwords, like *blif* and *blis*. Following exposure, both groups of listeners were tested on their categorization of a fricative continuum from 100% /s/ to 100% /f/. Listeners exposed to the ambiguous fricative at the end of words shifted their categorization behaviour, while those exposed to the same sounds at the end of nonwords did not. The exposure using words was further differentiated by the bias introduced by the words. That is, half the tokens ending in the ambiguous fricative formed a word if the fricative was interpreted as /s/ but not if it was interpreted as /f/, and the others were the reverse. Listeners exposed only to the /s/-biased tokens categorized more of the /f/-/s/ continuum as /s/, and listeners exposed to /f/-biased tokens categorized more of the continuum as /f/. The ambiguous fricative was associated with either /s/ or /f/ according to the bias induced by the word, which led to an expanded category for that fricative at the expense of the other category. These results crucially show that perceptual categories in speech are malleable to new exposure, and that the linguistic system of the listener facilitates generalization to that category in new forms and contexts.

In addition to lexically-guided perceptual learning, unambiguous visual cues to sound identity can cause perceptual learning as well; this is referred to as perceptual recalibration. In Bertelson et al. [4], an auditory continuum from /aba/ to /ada/ was synthesized and paired with a video of a speaker producing /aba/ and a video of /ada/. Participants first completed a pretest that identified the maximally ambiguous step of the /aba/-/ada/ auditory continuum. In eight blocks, participants were randomly exposed to the ambiguous auditory token paired with video for /aba/ or with the video for /ada/. Following each block, they completed a short categorization test. Participants showed perceptual learning effects, such that they were more likely to respond with /aba/ if they had been exposed to the video of /aba/ paired with the ambiguous token in the preceding block, and vice versa for /ada/.

van Linden and Vroomen [59] compared the perceptual recalibration effects from the visually-guided perceptual learning paradigm [4] to the standard lexically-guided perceptual learning paradigm [40]. Speech-read recalibration and perceptual learning effects had comparable sizes, lasted equally as long, were enhanced when presented with a contrasting sound, and were both unaffected by periods of silence between exposure and categorization. The effects for both lexically- and visually-guided paradigms did not last through prolonged testing in the vanLinden and colleagues study, however, unlike other work [17, 26]. In Kraljic and Samuel [26] and Eisner and McQueen [17], lexically-guided perceptual learning effects persisted through intermediate tasks, as long as the task did not involve any contradictory evidence of the trait learned.

Visually-guided perceptual learning in speech perception has been modeled using a Bayesian belief updating framework [25]. A more general Bayesian framework for perception and action in the brain is the predictive coding model [11]. In Bayesian belief updating, the model categorizes the incoming stimuli based on multimodal cues, and then updates the distribution to reflect that categorization. This updated conditional distribution is then used for future categorizations in an iterative process. Kleinschmidt and Jaeger [25] effectively model the results of the behavioural study in Vroomen et al. [60] in a Bayesian framework, with models fit to each participant capturing the perceptual recalibration and selective adaptation shown by the participants over the course of the experiment. A similar, but more broad, framework is that of predictive coding [11]. This framework uses a hierarchical generative model that aims to minimize prediction error between bottom-up sensory inputs and top-down expectations. Mismatches between the top-down expectations and the bottom-up signals generate error signals that are used to modify future expectations. Perceptual learning then is the result of modifying expectations to match learned input.

Perceptual learning in the psychophysics literature has shown a large degree of exposure-specificity, where observers only show learning effects on the same or very similar stimuli as those they were trained on. As such, perceptual learning has been argued to reside or affect the early sensory pathways, where stimuli are represented with the greatest detail [21]. Perceptual learning in speech perception has also shown a large degree of exposure-specificity, where participants do not generalize cues across speech sounds [50] or across speakers unless the sounds are sufficiently similar across exposure and testing [16, 26, 27, 48]. Crucially, lexically-guided perceptual learning in speech has shown a greater degree of generalization than would be expected from a purely psychophysical standpoint. The testing stimuli are in many ways quite different from the exposure stimuli, with participants exposed to multisyllabic words ending in an ambiguous sound and tested on monosyllabic words [49] and nonwords [26, 40], though exposure-specificity has been found when exposure and testing use different positional allophones [38].

Why is lexically-guided perceptual learning more context-general? The experiments performed in this dissertation provide evidence that this context-generality is the result of a listener's attentional set, which can be influenced by linguistic, attentional and stimulus properties. A comprehension-oriented attentional set, where a listener's goal is to understand the meaning of speech, promotes generalization and leads to greater perceptual learning. A purely perception-oriented attentional set, where a listener's goal is to perceive specific qualities of a signal, does not promote generalization. The experiments in this dissertation use the lexically-guided perceptual learning paradigm, which uses tasks oriented towards comprehension, so generalization is to be expected in general, but the more perception-oriented the attentional set, the less perceptual learning should be observed. In terms of a Bayesian framework with error propagation, a more perception-oriented attentional set may keep error propagation more local, resulting in the exposure-specificity seen more in the psychophysics literature. A more comprehension-oriented attentional set would propagate errors farther upward in the hierarchy of expectations. In both cases, errors would propagate to where attention is focused, but larger updating associated with farther error propagation would lead to larger observed perceptual learning effects. These attentional sets will be explored in more detail in Section 1.3 following an examination of the linguistic factors that will be manipulated in the experiments in Chapters 2 and 3, namely lexical bias and semantic predictability; both of which are hypothesized to aid the listener in adopting comprehension-oriented attentional sets.

1.2 Linguistic expectations and perceptual learning

Two linguistic manipulations on attentional sets are lexical bias, as a function of the position of the modified category in the word, and semantic predictability. Chapter 2 presents two experiments using a standard lexically-guided perceptual learning paradigm, which uses lexical bias as the means to link an ambiguous sound to an unambiguous category. In Chapter 3, a novel, sententially-guided perceptual learning paradigm is used to further promote use of comprehension-oriented attentional sets.

1.2.1 Lexical bias

Lexical bias is the primary way through which perceptual learning is induced in the experimental speech perception literature. Lexical bias, also known as the Ganong Effect, refers to the tendency for listeners to interpret a speaker's (noncanonical) production as a particular, meaningful word rather than a nonsense word. For instance, given a continuum from a nonword like *dask* to word like *task* that differs only in the initial sound, listeners in general are more likely to interpret any step along the continuum as the word endpoint rather than the nonword endpoint compared to a continuum where neither endpoint is a word [19]. This bias is exploited in perceptual learning studies to allow for noncanonical, ambiguous productions of a sound to be readily linked to pre-existing sound categories. Given that ambiguous productions must be associated with a word to induce lexically-guided perceptual learning [40], differing degrees of lexical bias could lead to differing degrees of perceptual learning, a prediction which is tested in Chapter 2. The stronger the lexical bias, the stronger the link will be between the ambiguous sound and the general sound category.

Lexical bias effects have also been found for reaction time in phoneme detection tasks, where sounds are detected faster in words than in nonwords. However, such lexical effects are dependent on the attentional set being employed. If the stimuli are sufficiently repetitive, such as all having the same CV shape, the lexical bias effects disappear [14]. The monotony or variation of filler items is sufficient to bias listeners towards one attentional set or the other. Increasing cognitive load has been found to increase lexical bias effects as well. Performing a hard task concurrent to a phoneme categorization task results in a greater weighting for lexical information, due to impoverished auditory encoding [35]. Both of these tasks are perception-oriented as well, and the prevalence of effects from comprehension-oriented attentional sets speaks to its prominent role in ordinary language use.

Aspects of the stimuli items themselves have a large impact on the size of lexical bias effects. Longer words, in number syllables, show stronger lexical bias than shorter words [46]. Continua formed using trisyllabic words, such as *establish* and *malpractice*, were found to show consistently larger lexical bias effects than monosyllabic words, such as *kiss* and *fish*. Pitt and Samuel [46] also found that lexical bias from trisyllabic words was robust across experimental conditions (e.g., compressing the durations by up to 30%), but lexical bias from monosyllabic words was more fragile and condition dependent. The lexical bias effects shown by monosyllabic words only approached those of trisyllabic words when the participants were told to keep response times within a certain margin and given feedback when the response time fell outside the desired range. The reaction time monitoring could have added a greater cognitive load for participants, which has been shown to increase lexical bias effects [35]. They argue that longer words exert stronger lexical bias due to both the greater bottom-up information present in longer words and the greater lexical competition for shorter words.

Different positions within words of a given length have stronger or weaker lexical bias effects. Pitt and Szostak [43] used a lexical decision task with a continuum of fricatives from /s/ to /ʃ/ embedded in words differing in the position of a sibilant. They found that ambiguous fricatives later in the word, such as *establish* or *embarrass*, show greater lexical bias effects than the same ambiguous fricatives embedded earlier in the word, such as *serenade* or *chandelier*. In the experiments and meta-analysis of phoneme identification results presented in Pitt and Samuel [45], they found that for monosyllabic word frames token-final targets produce more robust lexical bias effects than token-initial targets. Considering word length and position in the word, lexical bias appears to be strengthened over the course of the word. As a listener hears more evidence for a particular word, their expectations for hearing the rest of that word increase.

One final research paradigm that has investigated lexical biases is phoneme restoration tasks [51]. In this paradigm, listeners hear words with noise added to or replacing sounds and are asked to identify which they had heard for each trial. Lower sensitivity to noise addition versus noise replacement and increased bias for responding that noise has been added is indicative of phoneme restoration – that is, listeners are perceiving sounds not physically present in the signal. Samuel [51] identified several factors that increase the likelihood of the phoneme restoration effect. In the lexical domain, words are more likely than nonwords to have phoneme restorations, and this discrepancy strengthens when listeners are primed with a form without noise before the trial. More frequent words are also more likely to exhibit phoneme restoration effects, and longer words also show greater phoneme restoration effects. Position of the sound in the word also influences listeners' decisions, with non-initial positions showing greater phoneme restoration effects. The other influences on phoneme restoration discussed in Samuel [51], namely the signal properties and sentential context will be discussed in subsequent sections.

Sounds at the beginnings of words show more fragile lexical bias effects than those later in the word. Sounds at the beginning of words require more accurate perception than sounds later in words, because they are used to narrow the set of potential words and no particular expectations are available before the word begins. As such, perception-oriented attentional sets are favored in the processing of word-initial sounds. Lexically-guided perceptual learning typically maximizes the lexical bias of the modified category by placing it at the end of the word [40], but perceptual learning has been found when the ambiguous stimuli is embedded earlier in the word (e.g., the onset of the final syllable [26, 28, 29] or the onset of the first syllable [10]). However, in light of the findings in Pitt and Szostak [43], we expect less lexical bias when the ambiguous sounds occur earlier in the word, and therefore, I hypothesize, lower endorsement rates and smaller perceptual learning effect sizes for word-initial manipulations. Experiments 1 and 2 in Chapter 2 implement this manipulation.

1.2.2 Semantic predictability

The second type of linguistic expectation manipulation used in this dissertation is semantic predictability [22]. Sentences are semantically predictable when they contain words prior to the final word that point almost definitively to the identity of that final word. For instance, the sentence fragment *The cow gave birth to the...* from Kalikow et al. [22] is almost guaranteed to be completed with the word *calf*. On the other hand, a fragment like *She is glad Jane called about the...* is far from having a guaranteed completion beyond being a noun.

Words that are predictable from context are temporally and spectrally reduced compared to words that less predictable [12, 53]. Despite this acoustic reduction, high predictability sentences are generally more intelligible. Sentences that form a semantically coherent whole have higher word identification rates across varying signal-to-noise ratios [22] in both children and adults [18], and across native monolingual and early bilingual listeners, but not late bilingual listeners [36]. Highly predictable sentences are more intelligible to native listeners in noise, even when signal enhancements are not made, though non-native listeners require both signal enhancements and high predictability together to see any benefit [7]. However, when words at the ends of predictive sentences are excised from their context, they tend to be less intelligible than words excised from non-predictive contexts [34].

Semantic predictability has been found to have similar effects on phoneme categorization as lexical bias [6, 13]. In those studies, a continuum from one word to another, such as *coat* to *goat*, was embedded in a sentence frame that semantically cohered with either one of the endpoints or the other. Participants showed a shift in the category boundary based on the sentence frame. If the sentence frame cued the voiced stop, more of the continuum was subsequently categorized as the voiced stop and vice versa for the voiceless stop.

In the phoneme restoration paradigm, higher semantic predictability has been found to bias listeners toward interpreting the stimuli as intact with added noise rather than replaced with noise [51]. This increased bias towards interpreting the stimuli as an intact word was also coupled with an increase in sensitivity between the two types of stimuli, which Samuel [51] suggests is the result of a lower cognitive load in predictable contexts. Later work has suggested that in cases of lower cognitive load, greater phonetic encoding is available [see also 35].

To summarize, the literature on semantic predictability has shown largely similar effects as lexical bias in terms of how sounds are categorized and restored. From this, I hypothesize that increasing the expectations for a word through semantic predictability will promote a comprehension-oriented attentional set, as perception of the modified sound category will not be strictly necessary for comprehension. Listeners that are exposed to an /s/ category that is more /ʃ/-like only in words that are highly predictable from context are therefore predicted to show larger perceptual learning effects than listeners exposed to the same category only in words that are unpredictable from context. However, there may be an upper limit for listener expectations when both semantic predictability and lexical bias are high, as committing too much to a particular expectation could lead to garden path phenomena [33] or other misunderstandings. The effect of semantic predictability on perceptual learning is explicitly tested in Chapter 3.

1.3 Attention and perceptual learning

Attention is a large topic of research in its own right, and this section only reviews literature that is directly relevant to perceptual learning and this thesis. Attention has been found to have a role on perceptual learning in the psychophysics literature, indeed Gibson [20] identifies it as the sole prerequisite to perceptual learning. As an example, Ahissar and Hochstein [1] found that, in general, attending to *global* features for detection (i.e., discriminating different orientations of arrays of lines) does not make participants better at using *local* features for detection (i.e., detection of a singleton that differs in angle in the same arrays of lines), and vice versa. Perceptual learning, in this case, in this visual domain was limited to the task and stimuli on which a given participant was trained.

Attentional sets can refer to the strategies that the perceiver uses to perform a task. The attentional sets widely used in the visual perception literature do not align completely with the notion of perception-oriented and comprehension-oriented attentional sets used here. For instance, in a visual search task, colour, orientation, motion and size are the predominant strategies [61]. However, some parallels are present. The two broad categories in the visual perception literatures are *focused* and *diffuse* attentional sets. Focused sets direct attention to components of the sensory input, perceiving the trees instead of the forest. Diffuse sets direct attention to global properties of the sensory input, perceiving the forest instead of the trees. In the visual search literature, an example of a focused attentional set is the "feature search mode", which gives priority to a single feature, such as the colour of the target. An example of a diffuse attentional set is the "singleton detection mode", which gives priority to any salient features [2]. Perception-oriented and comprehension-oriented attentional sets have also been referred to as focused and diffuse attentional sets in recent speech perception work. In Pitt and Szostak [43], a diffuse attentional set is where primary attention is on detecting words from nonwords in a lexical decision task, and a focused attentional set is where instructions direct participants' attention to a potentially misleading sound. Attentional set selection is not necessarily optimal on the part of the perceiver, and it has been shown to be biased based on experience, with the amount of training performed influencing the length of time that perceivers will continue to use non-optimal sets after the task has changed [32].

Listeners can employ different attentional sets depending on the nature of the task, as well as other processing considerations. For instance, listeners can attend to particular syllables or sounds in syllable- or phoneme-monitoring tasks [39, and others], and even particular linguistically relevant temporal positions [44]. However, even in these low-level, signal based tasks, lexical properties of the signal can exhibit some influence if the stimuli are not monotonous enough to disengage comprehension [14]. Additionally, when performing a phoneme categorization task under higher cognitive load, such as performing a more difficult concurrent task, listeners show increased lexical bias effects [35]. Variation in speech in general seems to lead towards a more diffuse, comprehension-oriented attentional set, which is likely the default attentional set employed in everyday use of language, where the goal is firmly more comprehension-based than low-level perception-based. In comprehension-oriented tasks, such as a lexical decision tasks, explicit instructions can promote a more perception-oriented attentional set. When listeners are told that the speaker's /s/ is ambiguous and to listen carefully to ensure correct responses, they are less tolerant of noncanonical productions across all positions in the word [43]. That is, listeners whose attention is directed to the speaker's sibilants are less likely to accept the modified production as a word than listeners given no particular instructions about the sibilants. While the primary task has a large influence on the type of attentional set adopted, other instructions and aspects about the stimuli can shift the listener's attentional set toward another one.

Attentional sets have not been manipulated in previous work on perceptual learning in speech perception, but some work has been done on how individual differences in attention control can impact perceptual learning. Scharenborg et al. [55] presents a perceptual learning study of older Dutch listeners in the model of Norris et al. [40]. In addition to the exposure and test phases, these older listeners completed tests for hearing loss, selective attention and attention-switching control. They found no evidence that perceptual learning was influenced by listeners' hearing loss or selective attention abilities, but they did find a significant relationship between a listener's attention-switching control and their perceptual learning. Listeners with worse attention-switching control showed greater perceptual learning effects, which the authors ascribed to an increased reliance on lexical information. Older listeners were shown to have smaller perceptual learning effects compared to younger listeners, but the differences were most prominent directly following exposure [54]. Younger listeners initially had a larger perceptual learning effect in the first block of testing, but the effect lessened over the subsequent blocks. Older listeners showed smaller initial perceptual learning effects, but no such decay. Scharenborg and Janse [54] also found that participants who endorsed more of the target items as words in the exposure phase showed significantly larger perceptual learning effects in the testing phase.

There is evidence that attentional sets in the visual domain become entrenched over time [32], but the fact that attention-switching control in older adults was a significant predictor of the size of perceptual learning effects [55] suggests that listeners are indeed switching between sets through an experiment. The lexical decision task is oriented toward comprehension, so the primary attentional set is likely to be a diffuse one relying more on lexical information than acoustic. Participants with worse attention-switching control would have adopted this attentional set for more of the exposure than those with better attention-switching control, and it is precisely those with worse attention-switching control that showed the larger perceptual learning effects. The ability to switch attention to a more perception-oriented set could have allowed those listeners prevent over-generalization, leading to a smaller perceptual learning effect for participants with better attention-switching control.

As stated above, Bayesian models account well for perceptual learning experiments. Attentional sets are crucial to the hypothesis tested in this dissertation, but they do not play a role in the conceptual and computational models of perceptual learning. The predictive coding framework [11] provides a gain-based attentional mechanism. In this model, attention causes greater weight to be attached to error signals from mismatched expectations and sensory input, increasing their weight and their effect on future expectations. However, as noted by Block and Siegel [5], this view of attention does not capture the full range of experimental results. For instance, in a texture segregation task, spatial attention to the periphery improves performance where spatial resolution is poor, but attention to central locations, where spatial resolution is high, actually harms performance [62]. This detrimental effect is an instance of missing the forest for the trees, as spatial resolution increased too much in the central locations to perceive the larger texture. The attentional mechanism that I propose for the predictive coding framework is one where attention is not simply gain, but is also a mechanism that affects where errors must be resolved and expectations updated. Attending to perception rather than comprehension should only update expectations about perception of that individual instance. The lower sensory levels is where stimuli are represented with the greatest degree of detail [21]. Perceptual learning at these lower levels should be more exposure specific and less generalized than any learning that propagates to higher representational levels. In contrast, according to the mechanism proposed in Clark [11], any increases in attention, perception-oriented or otherwise, is predicted to lead to greater perceptual learning.

In this dissertation, focusing a listener's attention on the signal through explicit instructions is hypothesized to lead to smaller perceptual learning effects. In a focused perception-oriented attentional set, comprehension will be de-emphasized and expectations will be updated at a lower level, leading to a more exposure-specific learning that will not generalize to novel tokens as readily. The work on phoneme restoration suggests that higher predictability sentences have a lower cognitive load than unpredictable sentences [51], which may lead to greater encoding for the sounds to be learned, leading to larger perceptual learning effects. However, if fewer attentional resources are required for comprehending the word, there could be more attention on perception in the predictable sentences, leading to a reduced perceptual effect.

1.4 Category typicality and perceptual learning

A primary finding across the perceptual learning literature is that learning effects are only found on testing items that are similar in some sense to the exposure items. In the most extreme instance, perceptual learning is only found on the exact same items as exposure [50], but most commonly, perceptual learning is limited to items produced by the same speaker as exposure [40, 49]. However, a less studied question is what properties of the exposure items cause different degrees of perceptual learning.

Variability is a fundamental property of the speech signal, so sound categories must have some variance associated with them, and certain contexts can have increased degrees of variability. For example, Kraljic et al. [28] exposed participants to ambiguous sibilants between /s/ and /ʃ/ in two different contexts. In one, the ambiguous sibilants were intervocalic, and in the other, they occurred as part of a /str/ cluster in English words. Participants exposed to the ambiguous sound intervocalically showed a perceptual learning effect, while those exposed to the sibilants in /str/ environments did not. The sibilant in /str/ often surfaces closer to [ʃ] in many varieties of English, due to coarticulatory effects from the other consonants in the cluster, but the coarticulatory effects for merging /s/ and /ʃ/ are much weaker in intervocalic position. They argue that the interpretation of the ambiguous sound is done in context of the surrounding sounds, and only when the pronunciation variant is unexplainable from context is the variant learned and attributed to the speaker [see also 29]. In other words, a more /ʃ/-like /s/ category is typical in the context of the /str/ clusters, but is atypical in intervocalic position. Interestingly, given the lack of learning present in in the /str/ context, some degree of salience seems to be required to trigger perceptual learning.

Sumner [57] investigated whether differences in presentation order in a perceptual learning experiment led to different learning effects of the /b/-/p/ category boundary for a French-accented English talker. The presentation order manipulations that showed the greatest perceptual learning effects occurred when the tokens started out more closely matching what an English-speaking listener would expect for the categories (English-like voice onset time for /b/ and /p/) and shifted over the course of the experiment to what the speaker's actual categories were (French-like voice onset time for /b/ and /p/). Note that this is despite the fact that such a presentation order is not anything like what a listener would normally encounter when interacting with a non-native speaker of English. The condition that mirrored the more normal course of non-native speaker pronunciation changes, starting as more French-like and ending as more English-like, did not produce significantly different behaviour than control participants who only completed the categorization task. These results suggest that listeners constantly update their category following each successive input, rather than only relying on initial impressions [contra 29]. This finding is mirrored in Vroomen et al. [60], where participants initially expand their category in response to a single, repeated modified input, but then entrench that category as subsequent input is the same. The data in Vroomen et al. [60] is modeled using a Bayesian framework with constant updating of beliefs. However, in Sumner [57], the constantly shifting condition also had more perceptual learning than a random order of the same stimuli, suggesting that small differences in expectations and observed input induce greater updating than large differences. The bias towards small differences is better captured by the exemplar model proposed by Pierrehumbert [42], where only input similar to the learned distribution is used for updating that distribution, and input too ambiguous given learned distributions is discarded.

Similarity surfaces as an important factor in the phoneme restoration literature as well. Samuel [51] found that the likelihood of restoring a sound increases when said sound is acoustically similar to the noise replacing it. When the replacement noise is white noise, fricatives and stops are more likely to be restored than vowels and liquids. Simply, acoustic signals that better match expectations are thus less likely to be noticed as atypical.

In this dissertation, the degree of typicality of the modified category is manipulated across Experiments 1 and 2. In one case, the /s/ category for the speaker is maximally ambiguous between /s/ and /ʃ/, but in the other, the category is more like /ʃ/ than /s/. The maximally ambiguous category is hypothesized to be less salient than the more /ʃ/-like /s/ category. This lessened salience will result in greater use of comprehension-oriented attentional sets. I hypothesize that the more /ʃ/-like category will shift listeners' attentional sets to be more perception-oriented due to their greater atypicality, which will lead to lower perceptual learning. As such, perceptual learning effects are predicted to be greater when the modified /s/ category is more ambiguous between the two categories than when it is closer to the expected /ʃ/ category.

1.5 Current contribution

Perceptual learning in speech perception generalizes to new forms and contexts far more than would be expected from a purely psychophysical perspective [21, 40]. The two paradigms promote different attentional sets, with speech perception more focused on comprehension and psychophysics tasks more focused on perception. Indeed, visually-guided perceptual learning in speech perception, with its emphasis on perception, shows largely similar exposure-specificity effects as the psychophysics findings [50]. This dissertation expands the existing literature by modifying the exposure tasks to promote comprehension- or perception-oriented attentional sets. Perceptual learning effects are hypothesized to be smaller in the conditions that promote perception-oriented attentional sets, as perception exposure tasks have shown greater exposure-specificity effects than comprehension exposure tasks.

Chapter 2

Lexical decision

2.1 Motivation

The experiments in this chapter implement a standard lexically-guided perceptual learning experiment with exposure to a modified /s/ category during a lexical decision task. Because the exposure task is one of word recognition, participants will default to a comprehension-oriented attentional set. The manipulations in the two experiments presented here will encourage the participants to adopt a more perception-oriented attentional set. The first manipulation will be through the position of the modified /s/ category in the exposure tokens. Lexical bias effects increase as the length of the word increases and as the word unfolds [43, 46]. The second manipulation will be through explicit instructions about the modified /s/category. Such instructions have been shown to reduce lexical bias effects in lexical decision tasks [43]. Both of these manipulations will be present in Experiments 1 and 2.

The experiments differ in how typical the modified /s/ category. Studies have reported greater perceptual learning when ambiguous stimuli are closer to the distribution expected by a listener than when the ambiguous stimuli are farther away from expected distributions [57]. Words containing stimuli farther away from the target production are in general less likely to be endorsed as words, but similar effects of attention were found across word position [43]. Experiment 2 contains the same manipulations to attention and lexical bias as Experiment 1, but with ambiguous stimuli farther from the target production than those used in Experiment 1.

The hypothesis tested in this dissertation predicts that greater perceptual learning effects should be observed when no attention is directed to the ambiguous sounds and when lexical bias is maximized. In that instance, participants should use a comprehension-oriented attentional set. If selective attention is directed to the ambiguous sounds, a more perception-oriented attentional set should be adopted. Likewise, if the ambiguous sound is in a linguistically salient position with little to no lexical bias, a listener's attention should be drawn to the ambiguous sound, causing an adoption of a more perception-oriented attentional set. Finally, if the ambiguous sounds are more atypical, they should be more salient to listeners regardless of lexical bias, leading to a more perceptual oriented attentional set. Regardless of the cause, adopting a perception-oriented attentional set is predicted to inhibit a generalized perceptual learning effect.

2.2 Experiment 1

In this experiments, listeners will be exposed to ambiguous productions of words containing a single instance of /s/, where the /s/ has been modified to sound more like /ʃ/ in a lexical decision task. In one group, the critical words will have an /s/ in word-initial position, like in *cement*, with no /ʃ/ neighbour, like *shement*. In the other group, the critical words will have an /s/ in word-medial position, like in *tassel*, with no /ʃ/ neighbour like *tashel*. In addition, half of each group will be given instructions that the speaker has an ambiguous /s/ and to listen carefully, following Pitt and Szostak [43].

Given the difference in word response rates depending on position in the word in Pitt and Szostak [43], we would predict that listeners exposed to ambiguous sounds earlier in words would be less likely to accept these productions as words as compared to listeners exposed to ambiguous sounds later in words. In addition, given the reliance of perceptual learning on lexical scaffolding, this lower acceptance rate for the former group should lead to a smaller perceptual learning

Table 2.1: Mean and standard deviations for frequencies (log frequency per million words in SUBTLEXus) and number of syllables of each item type

Item type	Frequency	Number of syllables
Filler words	1.81 (1.05)	2.4 (0.55)
/s/ Word-initial	1.69 (0.85)	2.4 (0.59)
/s/ Word-medial	1.75 (1.11)	2.3 (0.47)
/ʃ/ Word-initial	2.01 (1.17)	2.3 (0.48)
/ʃ/ Word-medial	1.60 (1.12)	2.4 (0.69)

Table 2.2: Frequencies (log frequency per million words in SUBTLEXus) of words used in categorization continua

Continuum	/s/-word frequency	/ʃ/-word frequency
sack-shack	1.11	0.75
sigh-shy	0.53	1.26
sin-shin	1.20	0.48
sock-shock	0.95	1.46

effect as compared to the latter group.

2.2.1 Methodology

Participants

A total of 173 participants from the UBC population completed the experiment and were compensated with either $10 CAD or course credit. Of these, 77 were non-native speakers of English and two native speakers of English reported speech or hearing disorders, and their results were not included in the analyses, resulting in data from 94 participants. Twenty additional native English speakers participated in a pretest to determine the most ambiguous sounds. Twenty five other native speakers of English participated for course credit in a control experiment.

Materials

One hundred and twenty English words and 100 nonwords that were phonologically legal in English were used as exposure materials. The set of words consisted of 40 critical items, 20 control items and 60 filler words. Half of the critical items had an /s/ in the onset of the first syllable (Word-initial) and half had an /s/ in the onset of the final syllable (Word-medial). All critical tokens formed nonwords if their /s/ was replaced with /ʃ/. Half the control items had an /ʃ/ in the onset of the first syllable and half had an /ʃ/ in the onset of the final syllable. Each critical item and control item contained just the one sibilant, with no other /s z ʃ ʒ ʧ ʤ/. Filler words and nonwords did not contain any sibilants. Frequencies and number of syllables across item types are in Table 2.1

Four monosyllabic minimal pairs of voiceless sibilants were selected as test items for categorization (*sack-shack*, *sigh-shy*, *sin-shin*, and *sock-shock*). Two of the pairs had a higher log frequency per million words (LFPM) from SUBTLEXus [9] for the /s/ word, and two had higher LFPM for the /ʃ/ word, as shown in Table 2.2.

All words and nonwords were recorded by a male Vancouver English speaker in quiet room. Critical words for the exposure phase were recorded in pairs, once normally and once with the sibilant swapped forming a nonword. The speaker was instructed to produce both forms with comparable speech rate, speech style and prosody.

For each critical item, the word and nonword versions were morphed together in an 11-step continuum (0%-100% of the nonword /ʃ/ recording, in steps of 10%) using STRAIGHT [23] in Matlab. Prior to morphing, the word and nonword versions were time aligned based on acoustic landmarks, such as stop bursts, onset of F2, nasalization or frication, etc.

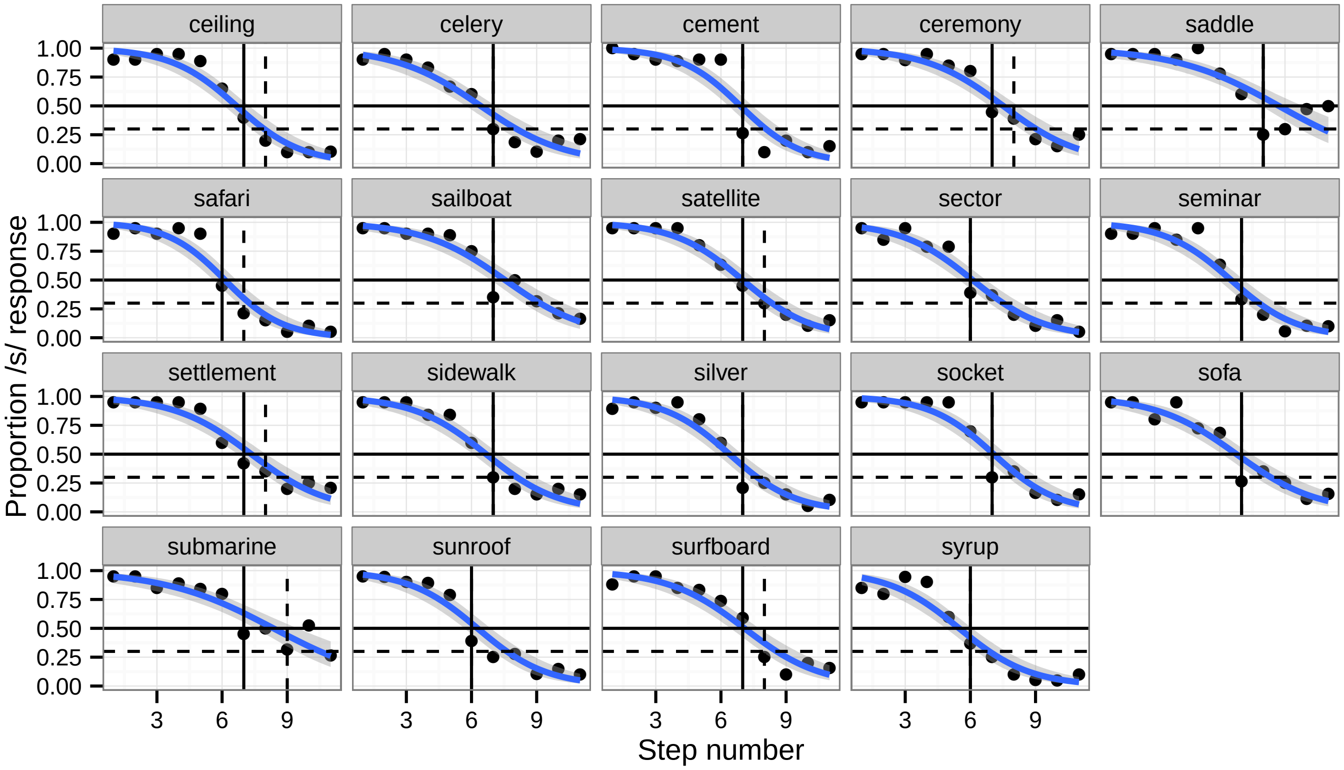


Figure 2.1: Proportion of word-responses for /s/-initial exposure words. Solid lines represent Experiment 1 selection criteria (50% word-response rate) and dashed lines represent Experiment 2 selection criteria (30% word-response rate). Dots are averaged word-response across subjects, and the blue line is a binomial model constructed from the responses.

All control items and filler words were processed and resynthesized by STRAIGHT to ensure a consistent quality across stimulus items.

Pretest

To determine which step of each continua would be used in exposure, a phonetic categorization experiment was conducted. Participants were presented with each step of each exposure word-nonword continuum and each categorization minimal pair continuum, resulting in 495 trials (40 exposure words plus five minimal pairs by 11 steps). The experiment was implemented in E-prime [47]. As half of the critical items had a sibilant in the middle of the word (onset of the final syllable), participants were asked to respond with word or non word rather than asking for the identity of the ambiguous sound, as in previous research [49].

The proportion of /s/-responses (or word responses for exposure items) at each step of each continuum was calculated and the most ambiguous step chosen. The threshold for the ambiguous step for this experiment was when the percentage of /s/-response dropped near 50%. The list of steps chosen for Word-initial target stimuli is in Table 2.3 and in Table 2.4 for Word-medial target stimuli. For the minimal pairs, six steps surrounding the 50% cross over point were selected for use in the phonetic categorization task. Due to experimenter error, the continuum for *seedling* was not included in the stimuli, so the chosen step was the average chosen step for the /s/-initial words. The average step chosen for Word-initial /s/ words was 6.8 ($SD = 0.5$), and for Word-medial /s/ words the average step was 7.7 ($SD = 0.8$).

To visualize the effect of morphing on the acoustics of the sibilants, a multidimensional scaled plot of acoustic distance was constructed. Using the `python-acoustic-similarity` package [37], sibilants were transformed into arrays of mel-frequency cepstrum coefficients (MFCC), and then pairwise distances were computed via dynamic time

Table 2.3: Step chosen for each Word-initial stimulus in Experiment 1 and the proportion /s/ response in the pretest

Word	Step chosen	Proportion /s/ response
ceiling	7	0.40
celery	7	0.30
cement	7	0.26
ceremony	7	0.44
saddle	8	0.25
safari	6	0.45
sailboat	7	0.35
satellite	7	0.45
sector	6	0.39
seminar	7	0.33
settlement	7	0.42
sidewalk	7	0.30
silver	7	0.21
socket	7	0.30
sofa	7	0.26
submarine	7	0.45
sunroof	6	0.39
surfboard	7	0.59
syrup	6	0.37
Average	6.8	0.36

Table 2.4: Step chosen for each Word-medial stimulus in Experiment 1 and the proporation /s/ response in the pretest

Word	Step chosen	Proportion /s/ response
carousel	7	0.45
castle	7	0.50
concert	7	0.53
croissant	7	0.42
currency	7	0.58
cursor	8	0.53
curtsy	8	0.40
dancer	7	0.45
dinosaur	7	0.50
faucet	7	0.45
fossil	8	0.30
galaxy	9	0.47
medicine	8	0.55
missile	10	0.30
monsoon	8	0.42
pencil	7	0.45
pharmacy	8	0.42
tassel	8	0.35
taxi	8	0.50
whistle	7	0.58
Average	7.7	0.45

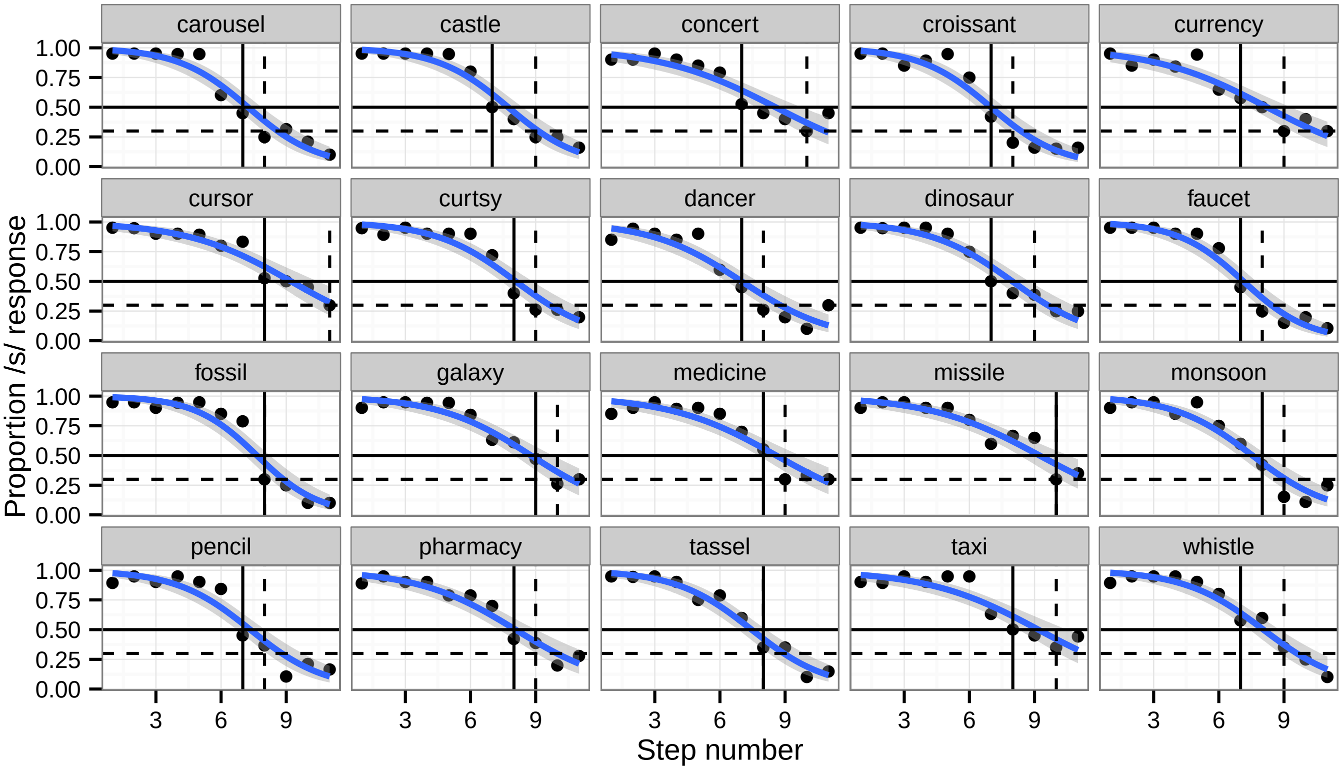


Figure 2.2: Proportion of word-responses for /s/-final exposure words. Solid lines represent Experiment 1 selection criteria (50% word-response rate) and dashed lines represent Experiment 2 selection criteria (30% word-response rate). Dots are averaged word-response across subjects, and the blue line is a binomial model constructed from the responses

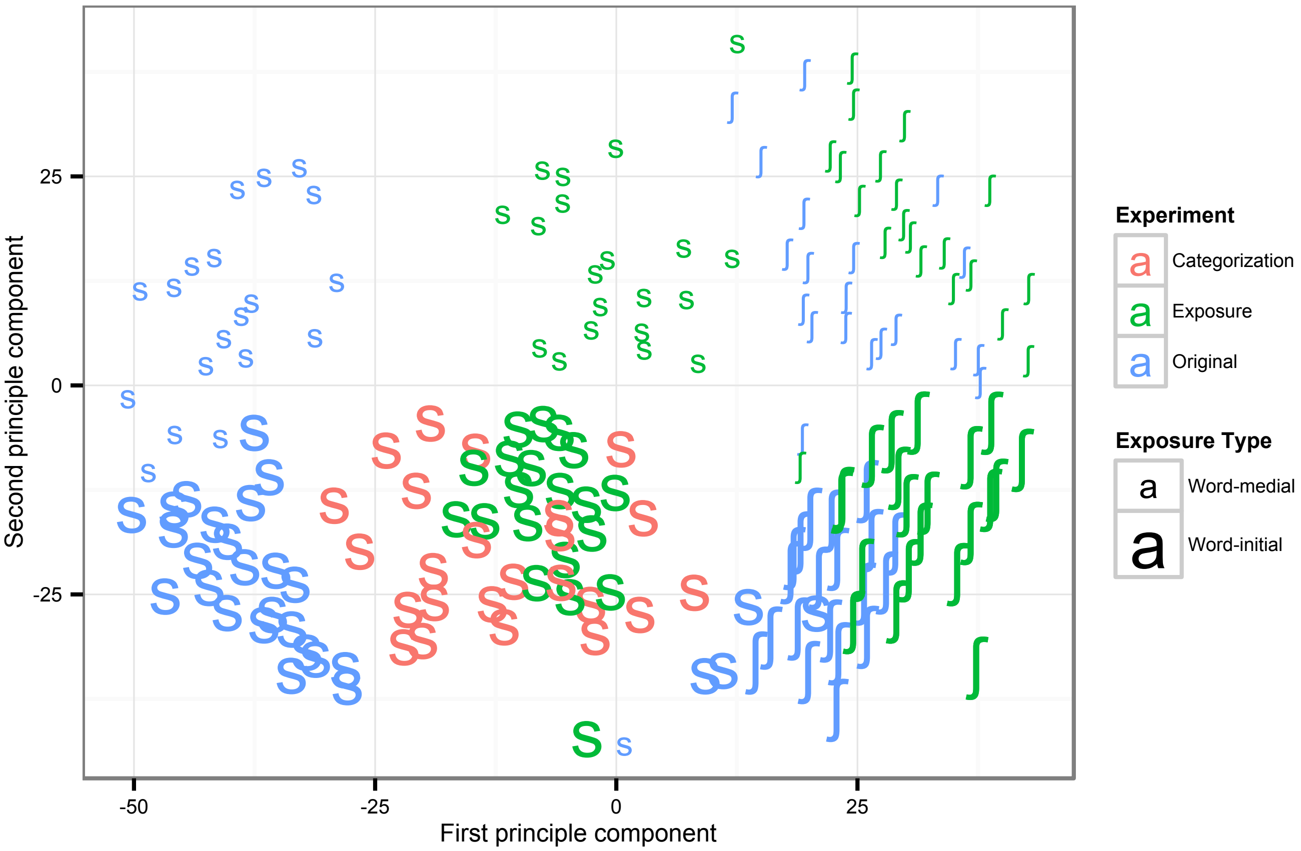
warping to create a distance matrix of the sibilant productions. This distance matrix was then multidimensionally scaled to produce Figure 2.3. As seen there, the original, unsynthesized productions (in blue) form four quadrants based on the two principle components of the distance matrix. The first principle component is associated with the centroid frequency of the sibilant, separating /s/ tokens from /ʃ/. The second principle component separates out the word-medial sibilants (in smaller font) from the word-initial sibilants (in larger font), largely from the different coarticulatory effects between the two positions. The categorization tokens (all word-initial) predictably occupy the space between the word-initial /s/ tokens and the word-initial /ʃ/ tokens. The exposure tokens pattern as expected. Exposure /ʃ/ tokens are overlapping with the original distributions for /ʃ/ tokens. Exposure /s/ tokens are in between /s/ and /ʃ/, though word-medial /s/ tokens are closer to the original /ʃ/ distribution, reflecting the difference in average stimuli step chosen in Tables 2.3 and 2.4.

Procedure

Participants in the experimental conditions completed an exposure task and a categorization task. The exposure task was a lexical decision task, where participants heard auditory stimuli and were instructed to respond with either ”word” if they thought what they heard was a word or ”nonword” if they didn’t think it was a word. The buttons corresponding to ”word” and ”nonword” were counterbalanced across participants. Trial order was pseudorandom, with no critical or control items appearing in the first six trials, and no critical or control trials in a row, but random otherwise, following Reinisch et al. [49].

In the categorization task, participants heard an auditory stimulus and had to categorize it as one of two words, differing only in the onset sibilant, i.e. *sin* or *shin*. The buttons corresponding to the words were counterbalanced across participants. The six most ambiguous steps of the minimal pair continua were used with seven repetitions each, giving a

Figure 2.3: Multidimensional scaling of the acoustic distances between the sibilants of original productions, categorization tokens and the exposure tokens in Experiment 1.



total of 168 trials. Participants were instructed that there would two tasks in the experiment, and both tasks were explained at the beginning to remove experimenter interaction between exposure and categorization.

Participants were assigned to one of four conditions, plus the control experiment. Two of the conditions exposed participants to only critical items that began with /s/ (Word-initial condition), and the other two exposed them to only critical items that had an /s/ in the onset of the final syllable (Word-medial condition), giving a consistent 200 trials in all exposure phases with control and filler items shared across all participants. Participants in half the conditions received additional instructions that the speaker's "s" sounds were sometimes ambiguous, and to listen carefully to ensure correct responses in the lexical decision. Participants assigned to the control experiment completed only the categorization task.

2.2.2 Results

Control experiment

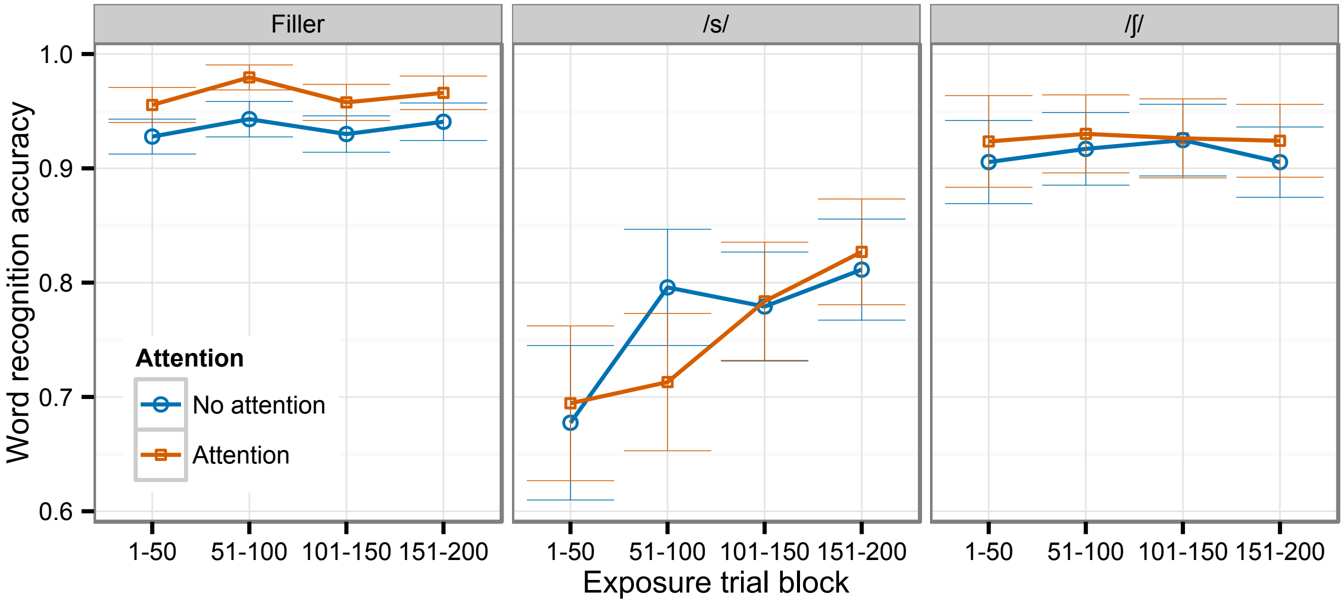
Responses with reaction times less than 200 ms or greater than 2500 ms were excluded from analyses. A logistic mixed effects models was fit with Subject and Continuum as random effects and Step as a fixed effect with by-Subject and by-Continuum random slopes for Step. The intercept was not significant ($\beta = 0.43, SE = 0.29, z = 1.5, p = 0.13$), and Step was significant ($\beta = -2.61, SE = 0.28, z = -9.1, p < 0.01$).

Exposure

Trials with nonword stimuli and responses faster than 200 ms or slower than 2500 ms were excluded from analysis. Performance on the exposure task was high overall, with accuracy on filler trials averaging 92%. A logistic mixed-effects model with accuracy as the dependent variable was fit with fixed effects for Trial (0-200), Trial Type (Filler, /s/, and /ʃ/), Attention (No Attention and Attention), Exposure Type (Word-Initial and Word-Medial) and their interactions. Trial Type was coded using treatment (dummy) coding, with Filler as the reference level. Deviance contrast coding was used for Exposure Type (Word-initial = 0.5, Word-medial = -0.5) and Attention (No attention = 0.5, Attention = -0.5). The random effect structure was as maximally specified as possible with random intercepts for Subject and Word, and by-Subject random slopes for Trial, Trial Type and their interactions and by-Word random slopes for Attention and Exposure Type and their interactions.

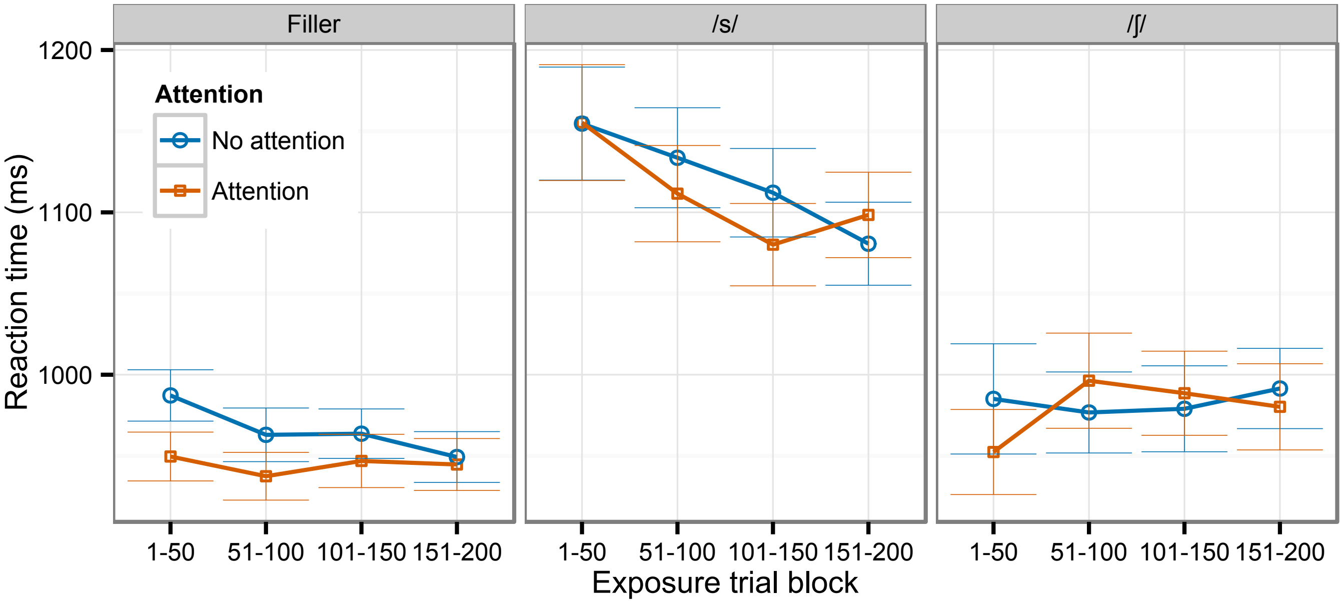
A significant fixed effect was found for Trial Type of /s/ versus Filler ($\beta = -2.13, SE = 0.31, z = -6.8, p < 0.01$), as participants were less likely to recognize words containing the modified /s/ category as compared to filler words. However, there was a significant interaction between Trial and Trial Type of /s/ versus Filler ($\beta = -0.45, SE = 0.14, z = 3.1, p = 0.01$), so the differences in accuracy between words with /s/ and filler words diminished over time. There was also a significant main effect of Attention ($\beta = -0.57, SE = 0.28, z = -2.0, p = 0.04$), indicating that participants were more accurate at identifying words in the Attention condition as compared to the No Attention condition. However, there was a marginal interaction between Attention and Trial Type of /s/ versus Filler ($\beta = 0.72, SE = 0.39, z = 1.8, p = 0.06$), suggesting that attention only increased accuracy for words not containing the modified /s/ category. Figure 2.8 shows within-subject mean accuracy across exposure, with Trial in four blocks.

Figure 2.4: Within-subject mean accuracy in the exposure phase of Experiment 1, separated out by Trial Type (Filler, /s/, and /ʃ/). Error bars represent 95% confidence intervals.



A linear mixed-effects with logarithmically-transformed reaction time as the dependent variable was fit with identical fixed effect and random effect structure as the logistic model for accuracy. Significant effects were found for Trial Type of /s/ versus Filler ($\beta = 0.71, SE = 0.07, t = 10.8$) and the interaction between Trial and Trial Type of /s/ versus Filler ($\beta = -0.08, SE = 0.02, t = -3.1$). These effects follow the pattern found in the accuracy model, where participants begin with slower reaction times to words with /s/ than filler words, but over time, this difference lessens. Figure 2.9 shows within-subject mean reaction time across exposure, with Trial in four blocks.

Figure 2.5: Within-subject mean reaction time in the exposure phase of Experiment 1, separated out by Trial Type (Filler, /s/, and /ʃ/). Error bars represent 95% confidence intervals.



Categorization

Responses with reaction times less than 200 ms or greater than 2500 ms were excluded from analyses. Two participants were excluded because their initial estimated crossover point for the continuum lay outside of the 6 steps presented. A logistic mixed effects model was constructed with Subject and Continuum as random effects and a by-Subject random slope for Step and by-Continuum random slopes for Step, Attention and Exposure Type and their interactions. Fixed effects for the model were Step, Exposure Type, Attention and their interactions. Deviance contrast coding was used for Exposure Type (Word-initial = 0.5, Word-medial = -0.5) and Attention (No attention = 0.5, Attention = -0.5). An /s/ response was coded as 1 and an /ʃ/ response as 0.

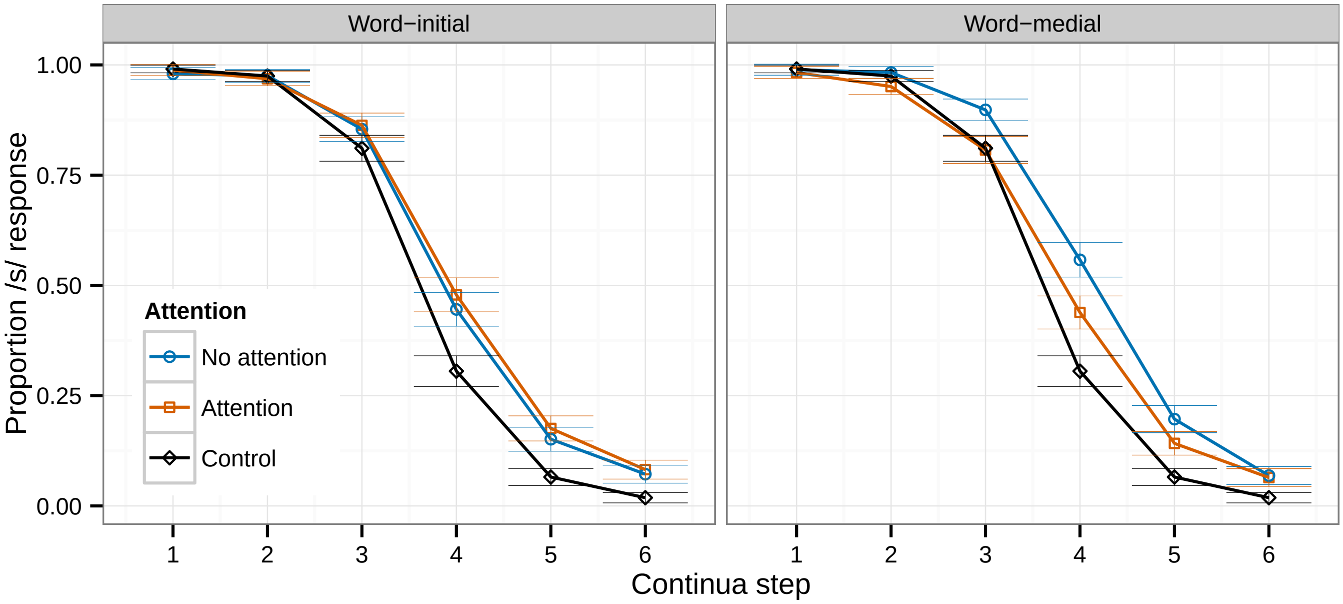
There was a significant effect for the intercept ($\beta = 0.76, SE = 0.22, z = 3.3, p < 0.01$), indicating that participants categorized more of the continua as /s/ in general. There was also a significant main effect of Step ($\beta = -2.14, SE = 0.15, z = -14.2, p < 0.01$), and a significant interaction between Exposure Type and Attention ($\beta = -0.93, SE = 0.45, z = -2.04, p = 0.04$). The results are visualized Figure 2.6. When exposed to ambiguous /s/ tokens in the first syllables of words, participants show a general expansion of the /s/ category, but no differences in behaviour if they are warned about ambiguous /s/ productions. However, when the exposure is to ambiguous /s/ tokens later in the words, we can see differences in behaviour beyond the general /s/ category expansion. Participants not warned of the speaker's ambiguous tokens categorized more of the continua as /s/ than those who were warned of the speaker's ambiguous /s/ productions.

2.2.3 Discussion

The condition that showed the largest perceptual learning effect was the one most biased toward a comprehension-oriented attentional set. Participants exposed to the modified /s/ category in the middle of words and with no explicit instructions about it had larger perceptual learning effects than any of the other conditions. The other conditions showed roughly equivalent sizes of perceptual learning, suggesting that there was not a compounding effect of explicit attention and word position. That is, the comprehension-oriented nature of the primary task still exerts an effect on attentional set selection.

The findings of this experiment do not support the predictions of a purely gain-based mechanism for attention, such as the one posited by Clark [11]. If attention functioned as a gain mechanism, increasing the weight of error signals

Figure 2.6: Proportion /s/ response along the 6 step continua as a function of Exposure Type and Attention in Experiment 1. Error bars represent 95% confidence intervals.



generated by mismatches between expectations and incoming signals, we should expect to see greater perceptual learning when listeners were instructed to attend to the speaker's /s/ sounds. Instead, the opposite occurred. Participants told to attend to the speaker's /s/ sounds showed smaller perceptual learning effects. This outcome may be determined by the nature of the instructions, which suggested that the ambiguity could harm accuracy, but different valences of attention are not predicted to behave any differently from one another in a gain mechanism.

In addition to the perceptual learning effects of the categorization phase, the exposure phase also demonstrates learning. In the initial trials, words with /s/ in them are responded to slower and less accurately, but over the course of exposure, both reaction times and accuracy approaches the other kinds of words. Interestingly, only the attention manipulation had an effect on exposure performance, with participants attending to the /s/ category responding more accurately overall. Exposure Type did not significantly influence accuracy or reaction time in the exposure.

One possible consideration is that the perceptual learning in this study might be showing exposure-specificity effects. However, as shown in Figure 2.3, Word-medial exposure tokens are acoustically farther from the categorization tokens than the Word-initial exposure tokens. The modified /s/ category used in this study aimed to be perfectly ambiguous between /s/ and /ʃ/. Many types of methods have been used for synthesizing the exposure items for perceptual learning experiments, such as selecting the maximally ambiguous mixture of extracted /s/ and /ʃ/ as judged by experimenters [26] or splicing the maximally ambiguous sound from a nonword continuum into the exposure items [40]. Perhaps the perceptual learning from the Word-initial exposure tokens could be enhanced if their /s/ category was directly taken from the categorization stimuli.

Reinisch et al. [49] used a similar method for exposure stimuli selection, but generated a more typical category by using 70% word response rate as the cutoff. Perhaps as a result, they report word endorsement rates that consistently exceeded 85%. In contrast, this experiment used 50% as the threshold and had correspondingly lower word endorsement rates (mean = 76%, sd = 22%). Interestingly, despite the lower word endorsement rates and the less canonical stimuli used, perceptual learning effects remained robust, which raises the question, can perceptual learning occur from a modified category even more atypical than the one used in this experiment? More atypical categories should be more salient and induce more a perception-oriented attentional set, and therefore result in smaller perceptual learning effects.

Table 2.5: Step chosen for each Word-initial stimulus in Experiment 2 and the proporation /s/ response in the pretest

Word	Step chosen	Proportion /s/ response
ceiling	8	0.20
celery	7	0.30
cement	7	0.26
ceremony	8	0.39
saddle	8	0.25
safari	7	0.21
sailboat	7	0.35
satellite	8	0.30
sector	6	0.39
seminar	7	0.33
settlement	8	0.35
sidewalk	7	0.30
silver	7	0.21
socket	7	0.30
sofa	7	0.26
submarine	9	0.32
sunroof	6	0.39
surfboard	8	0.25
syrup	6	0.37
Average	7.3	0.30

2.3 Experiment 2

Experiment 2 follows up on Experiment 1 by using stimuli that are farther from the canonical productions of the critical exposure tokens containing /s/.

2.3.1 Methodology

Participants

A total of 127 participants from the UBC population completed the experiment and were compensated with either $10 CAD or course credit. Of these, 31 were non-native speakers of English and their results were not included in the analyses, resulting in data from 96 participants.

Materials

Experiment 2 used the same items as Experiment 1, except that the step along the /s/-/ʃ/ continua chosen as the ambiguous sound had a different threshold. For this experiment, 30% identification as the /s/ word was used the threshold. The average step chosen for /s/-initial words was 7.3 ($SD = 0.8$), and for /s/-medial words the average step was 8.9 ($SD = 0.9$). The list of steps chosen for Word-initial and Word-medial target stimuli are in Tables 2.5 and 2.6, respectively.

Looking at the multidimensional scaling plot of the sibilant tokens involved Experiment 2 reveals similar patterns to that of Experiment 1. The axes remain the same as before, with the first principle component corresponding to differences between sibilants, and the second principle component corresponding to differences in word position. The original productions, categorization tokens and /ʃ/ tokens in Figure 2.7 are identical to those shown in Figure 2.3, but the exposure token distribution is shifted towards the /ʃ/ distribution, as expected. In the Word-medial position, the distributions of /s/ and /ʃ/ are close to overlapping, and in the Word-initial position, they are still separated, but closer than in Experiment 1.

Table 2.6: Step chosen for each Word-medial stimulus in Experiment 2 and the proporation /s/ response in the pretest

Word	Step chosen	Proportion /s/ response
carousel	8	0.25
castle	9	0.25
concert	10	0.30
croissant	8	0.20
currency	9	0.30
cursor	11	0.30
curtsy	9	0.26
dancer	8	0.26
dinosaur	9	0.39
faucet	8	0.25
fossil	8	0.30
galaxy	10	0.26
medicine	9	0.30
missile	10	0.30
monsoon	9	0.15
pencil	8	0.37
pharmacy	9	0.39
tassel	8	0.35
taxi	10	0.35
whistle	9	0.35
Average	8.9	0.29

Procedure

The procedure and instructions were identical between Experiment 1 and Experiment 2.

2.3.2 Results

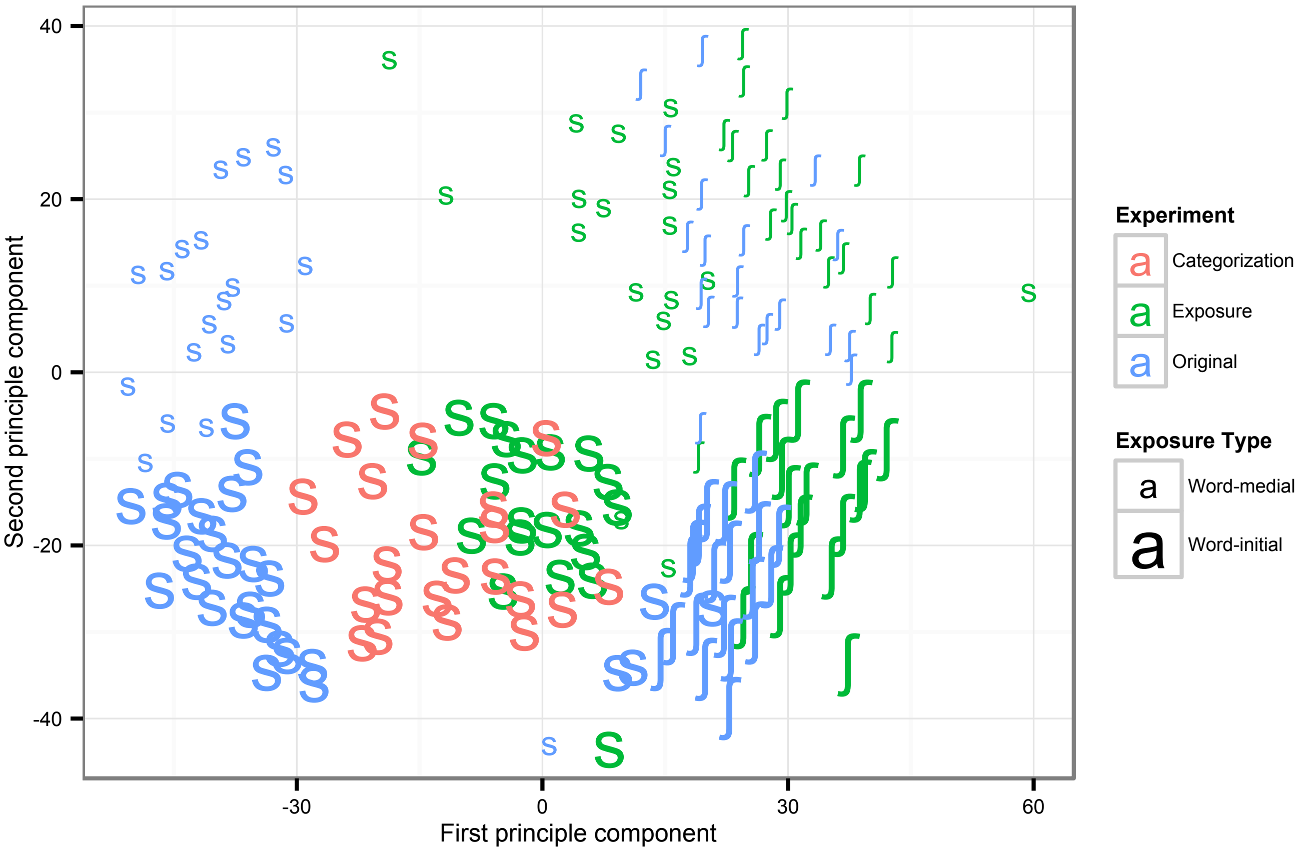
Exposure

Trials with nonword stimuli and responses faster than 200 ms or slower than 2500 ms were excluded from analysis. Performance on the exposure task was as high for Experiment 1, with accuracy on filler trials averaging 92%. A logistic mixed-effects model with accuracy as the dependent variable and a linear mixed-effects model with logarithmically-transformed reaction time as the dependent variable were fit with identical specifications as Experiment 1.

In the logistic mixed-effects model of accuracy, a significant fixed effect was found for Trial Type of /s/ versus Filler ($\beta = -3.56, SE = 0.31, z = -11.4, p < 0.01$), with participants less likely to respond that an item was a word if it contained the modified /s/ category. There was a significant interaction between Trial and Trial Type of /s/ versus Filler ($\beta = -0.29, SE = 0.11, z = 2.5, p < 0.01$) and between Trial and Trial Type of /ʃ/ versus Filler ($\beta = -0.33, SE = 0.16, z = -2.0, p = 0.04$). These interactions indicate that participants became more likely to endorse words with modified /s/ productions over time, but also became less accurate on words containing /ʃ/. Figure 2.8 shows within-subject mean accuracy across exposure, with Trial in four blocks.

In the linear mixed-effects model of reaction time, significant effects were found for Trial Type of /s/ versus Filler ($\beta = 0.94, SE = 0.07, t = 14.4$), indicating, as in Experiment 1, that reaction times were slower for words containing the modified /s/ category. Also significant was the interaction between Trial and Trial Type of /ʃ/ versus Filler ($\beta = 0.07, SE = 0.02, t = 3.4$), but not the interaction between Trial and Trial Type of /s/ versus Filler ($\beta = -0.02, SE = 0.02, t = -0.8$), indicating that reaction time remained stable for words containing the modified /s/ category, but lengthened for words

Figure 2.7: Multidimensional scaling of the acoustic distances between the sibilants of original productions, categorization tokens and the exposure tokens in Experiment 2.



containing the /ʃ/ control. There was a marginal effect for Trial Type of /ʃ/ versus Filler ($\beta = 0.14, SE = 0.07, t = 1.9$), indicating that words with /ʃ/ tended to be responded to more slowly than filler times, and a marginal effect of Exposure Type ($\beta = 0.17, SE = 0.09, t = 1.9$), indicating that words in the Word-medial condition tended to be responded to faster. Figure 2.9 shows within-subject mean reaction time across exposure, with Trial in four blocks.

Categorization

Responses with reaction times less than 200 ms or greater than 2500 ms were excluded from analyses. Two participants were excluded because their initial estimated cross over point for the continuum lay outside of the 6 steps presented. A logistic mixed effects model was constructed with identical specification as Experiment 1.

There was a significant effect for the Intercept ($\beta = 0.60, SE = 0.26, z = 2.3, p = 0.02$), indicating that participants categorized more of the continua as /s/ in general. There was also a significant main effect of Step ($\beta = -2.51, SE = 0.19, z = -13.1, p < 0.01$). Unlike in Experiment 1, there were no other significant effects, suggesting that participants across group had similar perceptual learning effects.

2.4 Grouped results across experiments

To see what degree the stimuli used had an effect on perceptual learning, the data from Experiment 1 and Experiment 2 were pooled and analyzed identically as above, but with Experiment and its interactions as fixed effects. In the lo-

Figure 2.8: Within-subject mean accuracy in the exposure phase of Experiment 2, separated out by Trial Type (Filler, /s/, and /ʃ/). Error bars represent 95% confidence intervals.

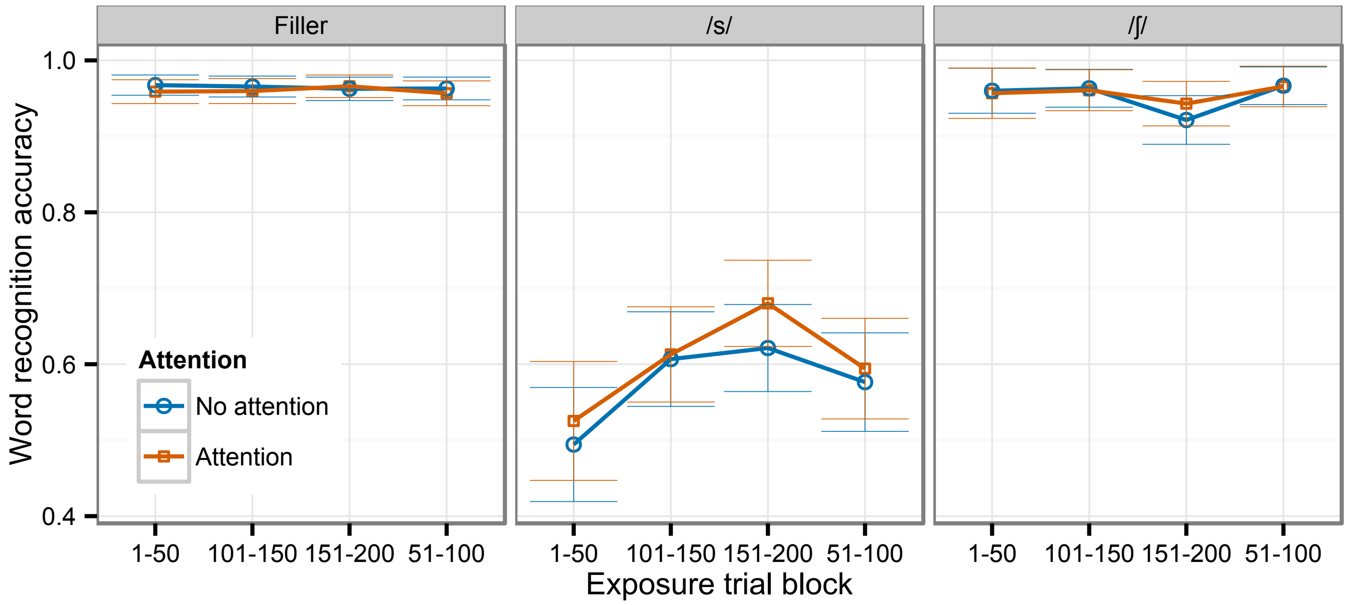
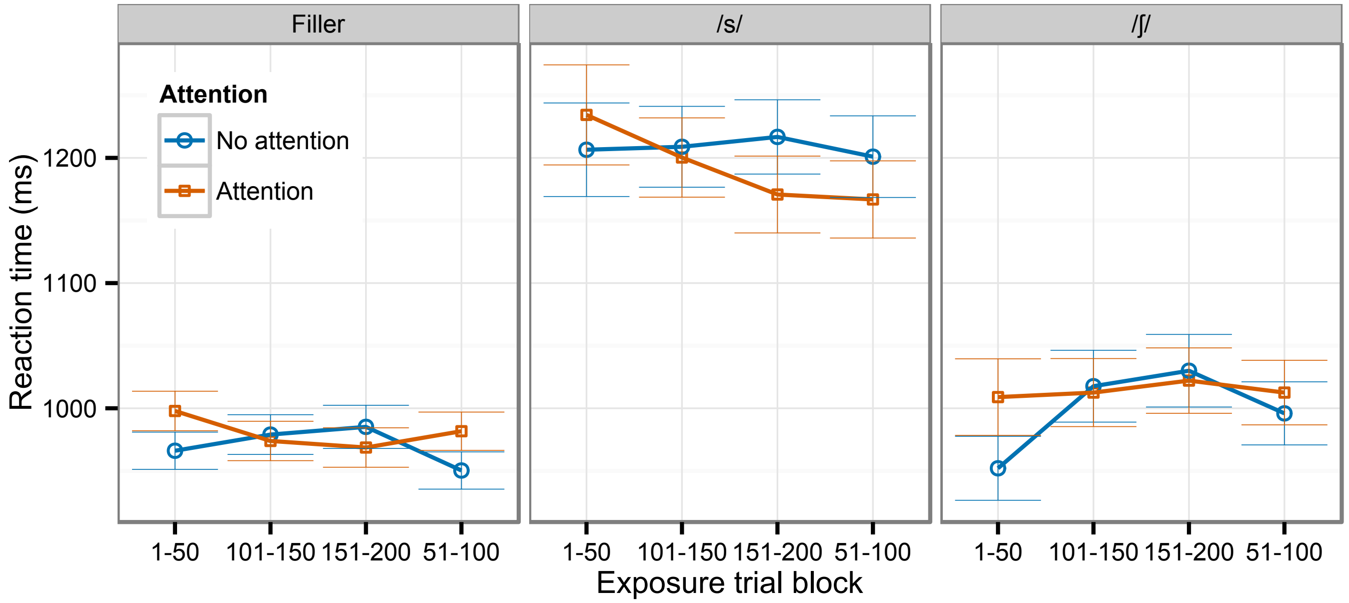
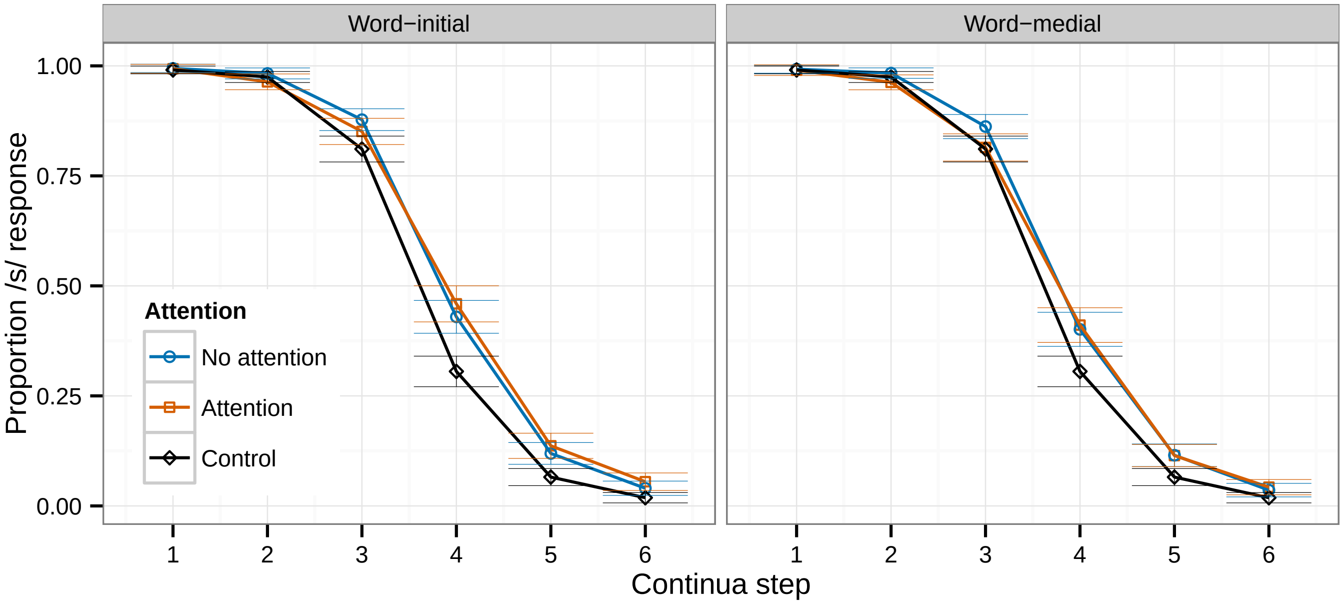


Figure 2.9: Within-subject mean reaction time in the exposure phase of Experiment 2, separated out by Trial Type (Filler, /s/, and /ʃ/). Error bars represent 95% confidence intervals.



gistic mixed effects model, there was significant main effects for Intercept ($\beta = 1.00, SE = 0.36, z = 2.7, p < 0.01$) and Step ($\beta = -2.64, SE = 0.21, z = -12.1, p < 0.01$), and a significant two-way interaction between Experiment and Step

Figure 2.10: Proportion /s/ response along the 6 step continua as a function of Exposure Type and Attention in Experiment 2. Error bars represent 95% confidence intervals.



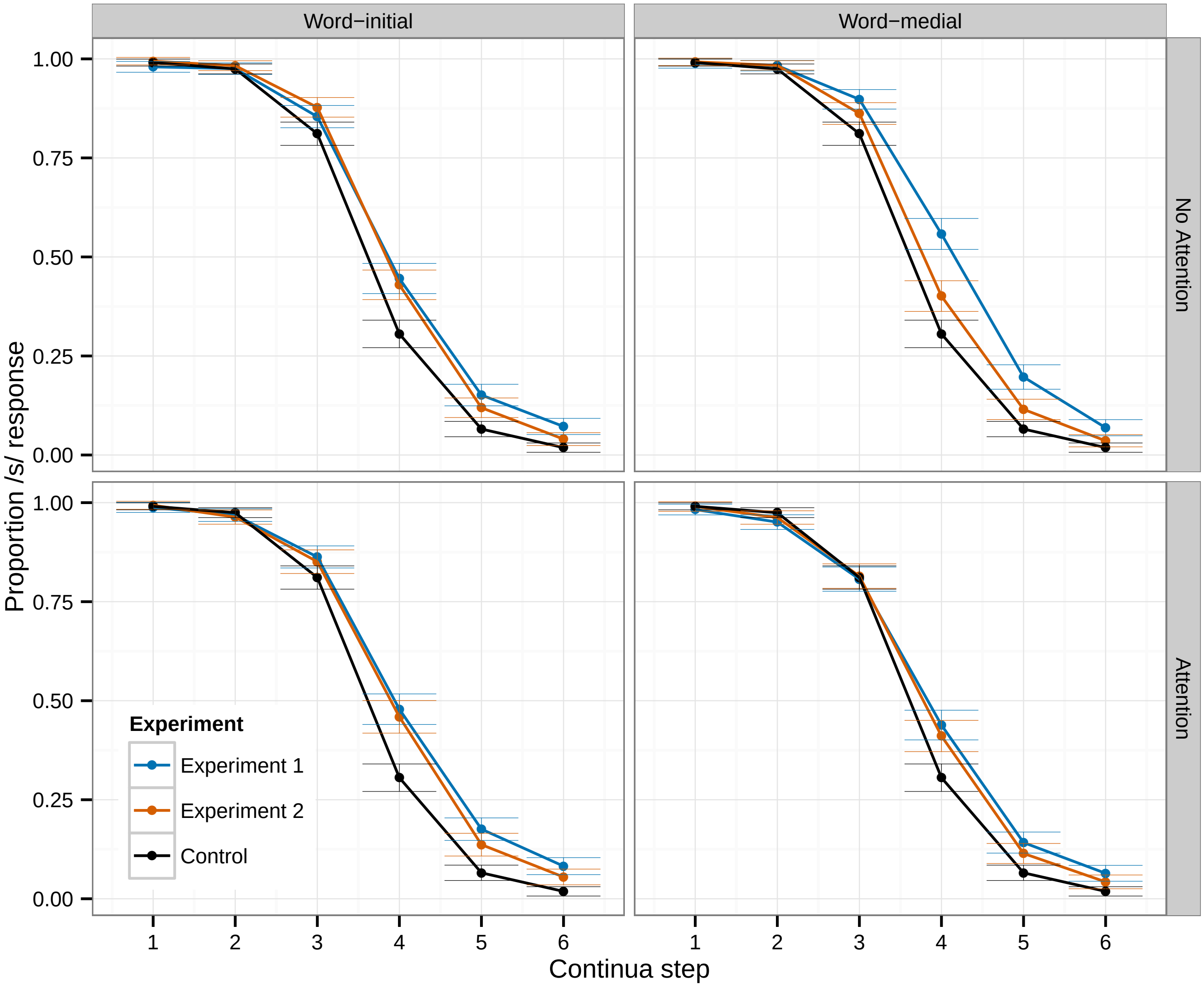
($\beta = 0.51, SE = 0.20, z = 2.5, p = 0.01$), and a marginal four-way interaction between Step, Exposure Type, Attention and Experiment ($\beta = 0.73, SE = 0.42, z = 1.7, p = 0.08$). These results can be seen in Figure 2.11. The four-way interaction can be seen in S-Final/No Attention conditions across the two experiments, where Experiment 1 has a significant difference between the Attention and No Attention condition, but Experiment 2 does not. The two-way interaction between Experiment and Step and the lack of a main effect for Experiment potentially suggests that while the category boundary was not significantly different across experiments, the slope of the categorization function was.

As an individual predictor of participants' performance the proportion critical word endorsements were compared to estimated points of the categorization continua where participants' perception switch from predominantly /s/ to predominantly /ʃ/. The crossover point was determined from the Subject random intercept and the by-Subject random slope of Step in a simple model containing only those random effects, similar by-Continuum random effects, and a fixed effect for Step [24]. An ANOVA with crossover point as the independent variable and word endorsement rate, Exposure Type, Attention, Experiment and their interactions, found only a main effect of word endorsement rate ($F(1,174) = 23.3, p < 0.01$), and a marginal interaction between word endorsement rate and Experiment ($F(1,174) = 3.5, p = 0.06$). A scatter plot of word endorsement rate and crossover point is shown in Figure 2.12. In general there is a positive correlation between word endorsement rate in the exposure phase and the crossover point from /s/ to /ʃ/ in the categorization phase. Participants in the Experiment 1, who were exposed to more typical /s/ stimuli showed a stronger correlation across conditions than participants in Experiment 2, who were exposed to a more atypical /s/ category.

In Experiment 1, the strongest correlations between word endorsement rate and crossover point are seen in the Word-initial conditions (Attention: $r = 0.46, t(22) = 2.4, p = 0.02$; No Attention: $r = 0.45, t(22) = 2.4, p = 0.02$), with the next strongest, and more marginal, correlation in Word-final/Attention condition ($r = 0.39, t(23) = 2.0, p = 0.06$). The condition for which the most perceptual was observed actually has the weakest correlation ($r = 0.32, t(21) = 1.5, p = 0.13$).

In Experiment 2, the strongest correlation is in the Word-initial/Attention condition ($r = 0.40, t(23) = 2.1, p = 0.05$), with two trending correlations for the Word-medial conditions (Attention: $r = 0.33, t(20) = 1.6, p = 0.12$; No Attention: $r = 0.27, t(22) = 1.3, p = 0.20$). Finally, the correlation for the Word-initial/Attention condition is not significant ($r = -0.05, t(20) = -0.2, p = 0.82$).

Figure 2.11: Proportion /s/ response along the 6 step continua as a function of Exposure Type and Attention in Experiment 1 and Experiment 2. Error bars represent 95% confidence intervals.

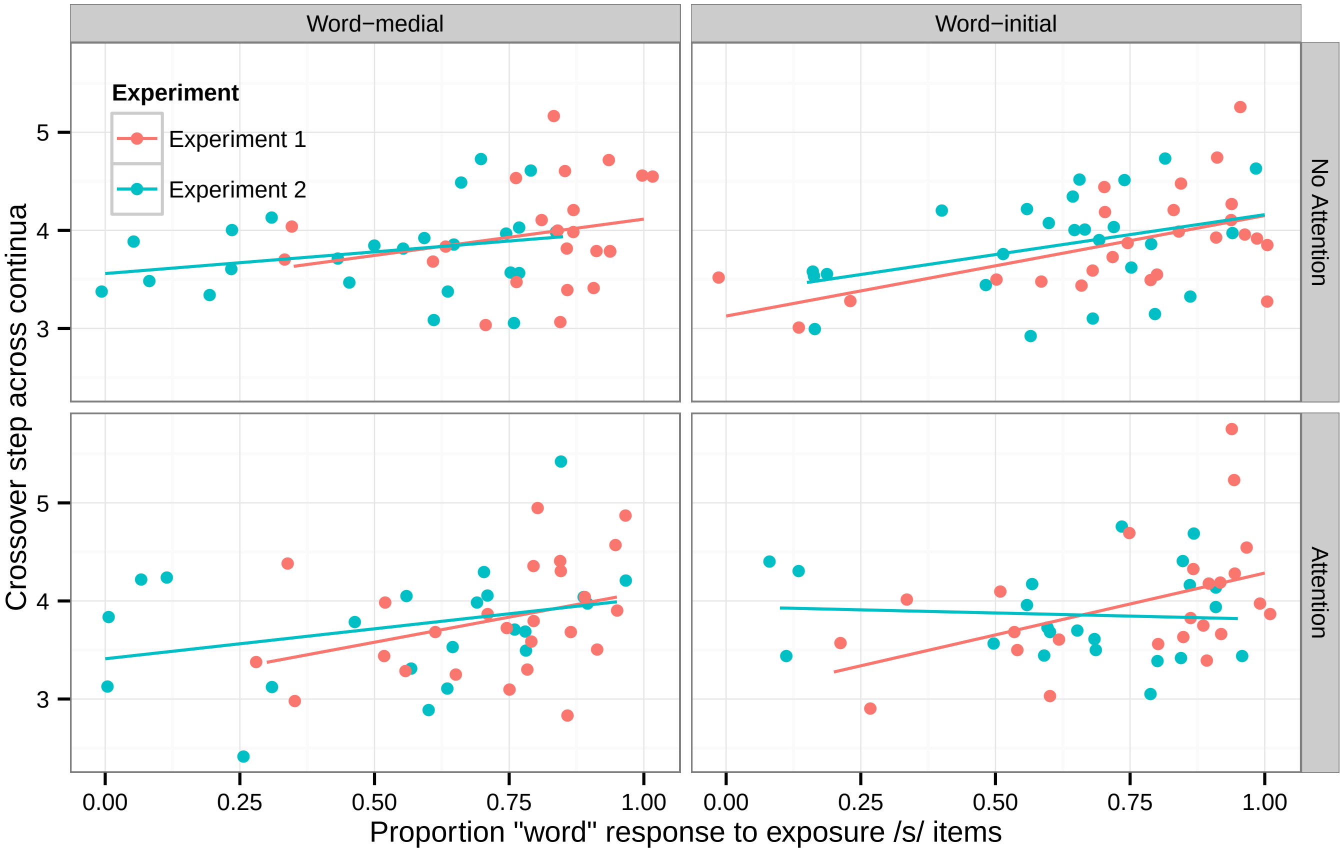


2.5 General discussion

Perceptual learning was found across all experimental conditions, as in other perceptual learning tasks that employed a lexical decision task as the exposure. However, the condition with the most bias toward a comprehension-oriented attentional set showed larger perceptual learning effects than all other conditions. Those conditions contained different kinds of manipulations all aimed at inducing a more perception-oriented attentional set. These manipulations did not have additive effects on perceptual learning, but rather, the presence of one or more was enough to shift participants into a more perception-oriented attentional set.

The perceptual learning effects found in Experiment 1 and 2 appear to fall into two categories, aligning with either comprehension-oriented or perception-oriented attentional sets. For most of exposure conditions, participants showed a small perceptual learning effect of similar magnitude. These conditions perception-oriented attentional sets were likely to be recruited, such as when attention was explicitly drawn to the ambiguous sounds, or the ambiguous sounds were

Figure 2.12: Correlation of crossover point in categorization with the proportion of word responses to critical items containing an ambiguous /s/ token in Experiments 1 and 2.



at the beginnings of words, or when the ambiguous sounds were the least typical of /s/. The participants exposed to the ambiguous sounds with the most lexical bias and with no attention drawn to them, however, showed a significantly stronger perceptual learning effect. This condition is the one where comprehension-oriented attentional sets were predicted to dominate. The perception-oriented attentional sets exhibit less generalization, like what is seen in psychophysics literature and in visually-guided perceptual learning in speech perception [50].

Compared to Experiment 1, Experiment 2 had a weaker, yet still significant, correlation between critical word endorsement rates and crossover boundary points, suggesting that although the stimuli used in Experiment 2 were farther from the canonical production, they did not shift the category boundary as much as the stimuli in Experiment 1. While neither attention or position of the ambiguous sound in the word had an effect on the correlation, the distance from the canonical production did. This potentially suggests that the degree to which a category is shifted is inversely related to its distance. Or perhaps more likely, the distance is inversely related to its goodness or plausibility that the ambiguous sound was indeed meant to be an /s/.

The correlation between word response rate in the exposure phase and the category boundary in categorization phase across both experiments raises two possible explanations. In a causal interpretation between exposure and categorization, as each ambiguous sound is processed and errors propagate, the distribution for that category (for that particular speaker) is updated. Participants who processed more of the ambiguous sound as an /s/ updated their perceptual category for /s/ more. This explanation fits within a Bayesian model of the brain [11] or a neo-generative model of spoken language processing [41]. A non-causal story is also plausible: the correlation may reveal individual differences on the part of the participants, where some participants are more adaptable or tolerant of variability than others, leading to greater degrees of perceptual adaptation. Individual differences in attention-switching control have previously been found to affect perceptual learning [55], which supports a non-causal interpretation as well.

However, an important note is that the condition that saw the largest perceptual learning did not have the largest correlation between word response rate and crossover point. Both the causal and non-causal explanations above would require a strong correlation for that condition. As that is precisely the condition where comprehension-oriented attentional sets are most promoted, it may be the case that only a few tokens have to be endorsed in such an attentional set to cause expectations for future tokens to be updated. In contrast, perception-oriented attentional sets only allow for low level sensory expectations to be updated, which would then be reinforced over time and perhaps propagated more slowly to higher sensory levels.

As mentioned in the discussion for Experiment 1, the findings do not support a simple gain mechanism for attention as proposed in Clark [11]. In Yeshurun and Carrasco [62], attention to areas with finer spatial resolution caused observers to miss larger patterns. If attention simply boosted the error signal, attention of all kinds should always be beneficial to perceptual learning. The findings of these two experiments supports a larger role for attention in a predictive coding framework, as noted in Block and Siegel [5]. The propagation-limiting mechanism attention proposed earlier explains both the findings in the visual domain and the current findings. Attention to a level of representation causes errors between expectations and observed signals to be resolved and updated at that level. If attention is more oriented towards comprehension, either of language or other patterns in the world, errors can be propagated higher before updating expectations.

The manipulations in the two experiments in this chapter were ones that induced perception-oriented attentional sets. The task itself focuses on recognition, and a comprehension-oriented attentional set is likely more helpful for this task than a perception-oriented one. However, participants likely switch between the two throughout the course of the experiment. To fully examine the use of perception- and comprehension-oriented attentional sets in perceptual learning, manipulations that induce comprehension-oriented attentional sets are necessary. In the following chapter, such a manipulation is implemented through increasing linguistic expectations from semantic predictability.

Chapter 3

Cross-modal word identification

3.1 Motivation

In Chapter 2, listeners showed greater perceptual learning effects when the lexical bias was greater. However, this relationship was not found when attention was drawn to the ambiguous sound and when the ambiguous sound was farther from the canonical production. The experiment in this chapter seeks to increase the linguistic expectations for the words containing the ambiguous productions as a way to induce more comprehension-oriented attentional sets, even in the face of explicit instructions promoting perception-oriented attentional sets. To this end, semantic predictability is used to boost linguistic expectations in conjunction with lexical bias.

Semantic predictability in production studies has been found to have an effect on vowel realization and duration independent of lexical factors such as frequency and neighbourhood density [12, 53]. In speech perception work, semantic predictability has been found to have an effect on phoneme categorization similar to that of lexical bias [6], and increased intelligibility in noise, particularly for native speakers [7, 18, 22, 36, and others]. In phoneme restoration, semantic predictability both increased the bias to respond that a word intact and also increased the sensitivity to noise addition versus noise replacement [51].

As in previous work, two levels of semantic predictability are used in this experiment. Highly predictable sentences are ones where the final word is constrained to only a few words. Unpredictable sentences are ones where the possible completions are numerous, and any completions given are more guesses than anything else. All target words and the threshold for ambiguity are taken from Experiment 1 and have the highest lexical bias available. When a listener is exposed to these words in unpredictable sentences, they should show perceptual learning effects equivalent to those participants in the Word-medial condition of Experiment 1.

The exposure task in this experiment differs from lexical decision task used in the previous chapter. Instead, participants identify the picture that matches the final word in the sentence, and all trials involve a real word. The attentional set for this task, where participants make a decision about the identity of a specific word, may be different from the previous experiment, where they make a decision about the lexicality of a word, but the decision is about the target word, and comprehension is critical in both instances.

If linguistic expectations are cumulative, we should expect to see a larger perceptual learning effect for participants exposed to ambiguous sounds in words that are highly predictable, due to greater use of comprehension-oriented attentional sets. Additionally, if perception-oriented attentional sets only require one trigger to counteract the effects of comprehension-oriented attentional sets, we should see a similar size of perceptual learning in this experiment in the attention conditions as in the previous experiments. Previous research has found an increased sensitivity to signal properties in semantically predictable sentences [51], which could lead to one of two outcomes. Given that bias to restore the phoneme was greater in these conditions as well, we might see greater perceptual learning effects, as the link between the word and the production would be enhanced, but the signal would be encoded more accurately from the lower cognitive load. On the other hand, the lower cognitive load could lead to more attention available to notice the deviant production, and thus adopt a perception-oriented attentional set. Also relevant are the findings of Kraljic et al. [28], where no perceptual learning was observed when the modified sound was in a /str/ context. In that context, the modified sound was not salient, and no learning was observed. If participants do not pay attention to variation in the final word in a predictable sentence, because they tend to be acoustically and temporally reduced, perceptual learning might also be reduced or absent.

3.2 Methodology

3.2.1 Participants

One hundred native speakers of English participated in the experiment and were compensated with either $10 CAD or course credit. They were recruited from the UBC student population. Twenty additional native English speakers participated in a pretest to determine sentence predictability, and 10 other native English speakers participated in a picture naming pretest. Participants were native North American English speakers with no reported speech or hearing disorders.

3.2.2 Materials

One hundred and twenty sentences were used as exposure materials. The set of sentences consisted of 40 critical sentences, 20 control sentences and 60 filler sentences. The critical sentences ended in one of 20 of the critical words in Experiments 1 and 2 that had an /s/ in the onset of the final syllable. The 20 control sentences ended in the 20 control items used in Experiments 1 and 2, and the 60 filler sentences ended in the 60 filler words in Experiments 1 and 2. Half of all sentences were written to be predictive of the final word, and the other half were written to be unpredictive of the final word. Unlike previous studies using sentence or semantic predictability [22], Unpredictive sentences were written with the final word in mind with a variety of sentence structures, and the final words were plausible objects of lexical verbs and prepositions. A full list of words and their contexts can be found in the appendix. Aside from the sibilants in the critical and control words, the sentences contained no sibilants (/s z ʃ ʒ ʧ ʤ/). The same minimal pairs for phonetic categorization as in Experiments 1 and 2 were used.

Sentences were recorded by the same male Vancouver English speaker used in Experiments 1 and 2. Critical sentences were recorded in pairs, with one normal production and then a production of the same sentence with the /s/ in the final word replaced with an /ʃ/. The speaker was instructed to produce both sentences with comparable speech rate, speech style and prosody.

As in Experiments 1 and 2, the critical items were morphed together into an 11-step continuum using STRAIGHT [23]; however, only the final word in sentence was morphed. For all steps, the preceding words in the sentence were kept as the natural production to minimize artifacts of the morphing algorithm. The control and filler items were also processed and resynthesized to ensure consistent quality. The ambiguous point selection was based on the pretest performed for Experiment 1 and 2 exposure items. The ambiguous steps of the continua chosen corresponded to the 50% cross over point in Experiment 2.

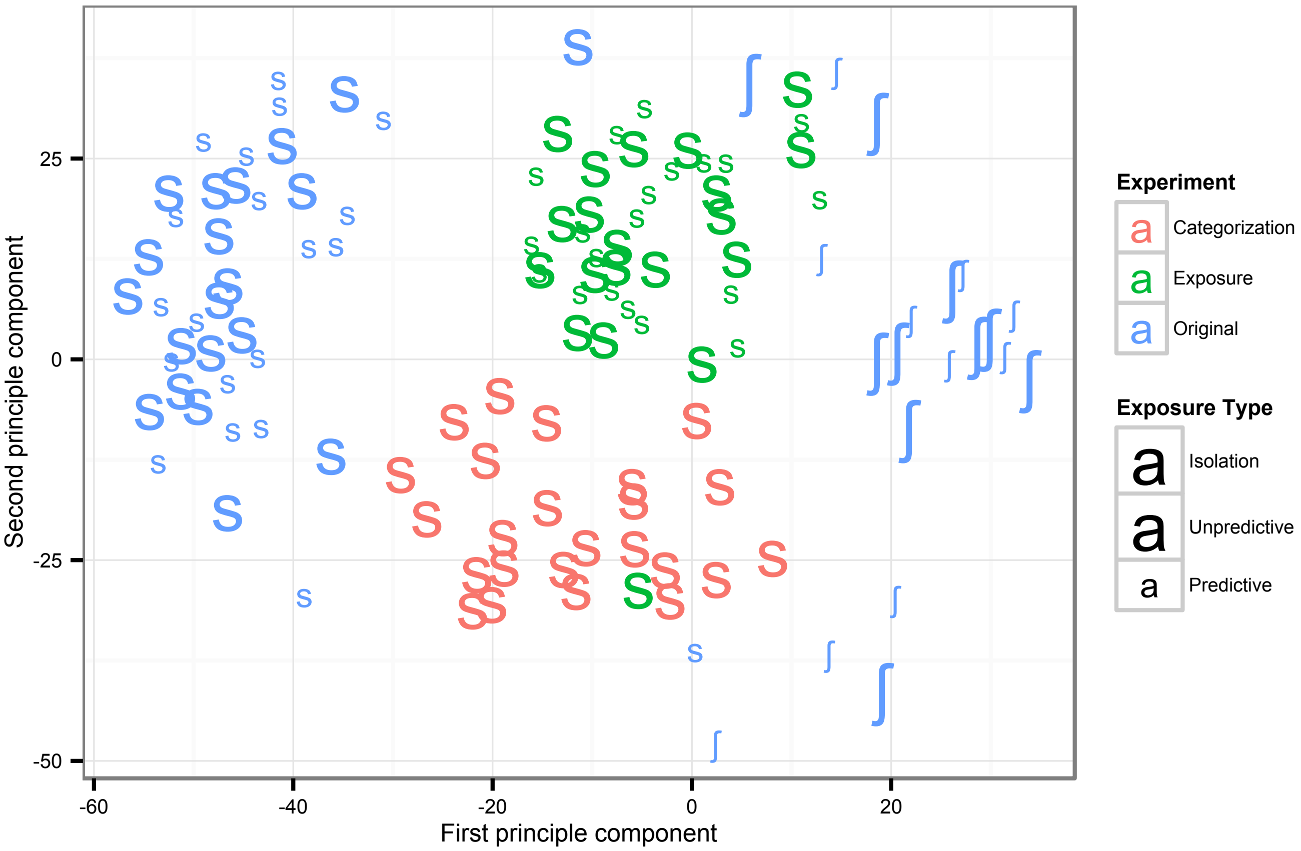
As done for Experiments 1 and 2, acoustic distances between exposure tokens, categorization tokens, and their original productions were multidimensionally scaled. In Figure 3.1, the original productions are separated again by the first principle component, which corresponds to the centroid frequency of the sibilant. The categorization tokens are predictably in between the original productions, and offset in the second principle component, due to their different position in the word. The exposure tokens for Experiment 3 fit in between the original productions and the categorization tokens.

Pictures of 200 words, with 100 pictures for the final word of the sentences and 100 for distractors, were selected in two steps. First, a research assistant selected five images from a Google image search of the word, and then a single image representing that word was selected from amongst the five by me. To ensure consistent behaviour in E-Prime [47], pictures were resized to fit within a 400x400 area with a resolution of 72x72 DPI and converted to bitmap format. Additionally, any transparent backgrounds in the pictures were converted to plain white backgrounds.

3.2.3 Pretest

The same twenty participants that completed the lexical decision continua pre-test also completed a sentence predictability task before the phonetic categorization task described in Experiment 1. Participants were compensated with $10 CAD for

Figure 3.1: Multidimensional scaling of the acoustic distances between the sibilants of original productions, categorization tokens and the exposure tokens in Experiment 3.



both tasks, and were native North American English speakers with no reported speech, language or hearing disorders. In this task, participants were presented with sentence fragments that were lacking in the final word. They were instructed to type in the word that came to mind when reading the fragment, and to enter any additional words that came to mind that would also complete the sentence. There was no time limit for entry and participants were shown an example with the fragment “The boat sailed across the...” and the possible completions “bay, ocean, lake, river”. Responses were collected in E-Prime [47], and were sanitized by removing miscellaneous keystrokes recorded by E-Prime, spell checking, and standardizing variant spellings and plural forms.

The measure used for determining rewriting of sentences was the proportion of participants that included the target word in their responses. For predictive sentences, the mean proportion was 0.49 (range 0-0.95) and for unpredictive sentences, the mean proportion was 0.03 (range 0-0.45). Predictive sentences that had target response proportions of 20% or less were rewritten. The predictive sentences for *auction*, *brochure*, *carousel*, *cashier*, *cockpit*, *concert*, *cowboy*, *currency*, *cursor*, *cushion*, *dryer*, *graffiti*, and *missile* were rewritten to remove any ambiguities.

Five volunteers from the Speech in Context lab participated in another pretest to determine how suitable the pictures were at representing their associated word. All participants were native speakers of North American English, with reported corrected-to-normal vision. Participants were presented with a single image in the middle of the screen. Their task was to type the word that first came to mind, and any other words that described the picture equally well. There was no time limit and presentation of the pictures was self-paced. Responses were sanitized as above.

Pictures were replaced if 20% or less of the participants (1 of 5) responded with the target word and the responses were semantically unrelated to the target word. Five pictures were replaced, *toothpick* and *falafel* with clearer pictures and *ukulele*, *earmuff* and *earplug* were replaced with *rollerblader*, *anchor* and *bedroom*. All five replacements were for distractor words.

3.2.4 Procedure

As in Experiments 1 and 2, participants completed an exposure task and a categorization task. For the exposure task, participants heard a sentence via headphones for each trial. Immediately following the auditory presentation, they were presented with two pictures on the screen. Their task was to select the picture on the screen that corresponded to the final word in the sentence they heard. As in Experiments 1 and 2, the order was pseudorandom, with the same constraints.

Participants were assigned to one of four groups of 25 participants. In the exposure phase, half of the participants were exposed to a modified /s/ sound only in Predictive sentences and half were exposed to it only in Unpredictive sentences. Half of all participants were told that the speaker's production of "s" was sometimes ambiguous, and to listen carefully to ensure correct responses.

Following the exposure task, participants completed the same categorization task described in Experiments 1 and 2.

3.3 Results

3.3.1 Exposure

Performance in the task was high, with accuracy near perfect across all subjects (mean accuracy = 99.5%, sd = 0.8%). Due to the extreme ceiling effects, a logistic mixed-effects model of accuracy was not constructed. A linear mixed effects model for logarithmically-transformed reaction time was constructed with a similar structure as in Experiments 1 and 2. Fixed effects were Trial (0-100), Trial Type (Filler, /s/, and /ʃ/), Attention (No Attention and Attention), Predictability (Unpredictive and Predictive), and their interactions. By-Subject and by-Word random effect structure was as maximal as permitted by the data, with by-Subject random slopes for Trial, Trial Type, and Predictability and their interactions, and by-Word random slopes for Attention, Predictability and their interaction. Trial Type was coded using treatment (dummy) coding, with Filler as the reference level. Deviance contrast coding was used for Predictability (Unpredictive = 0.5, Predictive = -0.5) and Attention (No attention = 0.5, Attention = -0.5).

A significant effect was found for Trial ($\beta = -0.20, SE = 0.01, t = -11.0$), indicating that reaction time became faster over the course of the experiment. There was a significant effect for Trial Type of /ʃ/ versus Filler ($\beta = 0.19, SE = 0.09, t = 2.1$), but not for /s/ versus Filler ($\beta = 0.11, SE = 0.09, t = 1.3$), suggesting that words with /ʃ/ in them were responded to more slowly than filler words or those with /s/ in them. There was a significant interaction between Trial and Trial Type of /s/ versus Filler ($\beta = 0.05, SE = 0.02, t = 2.4$) and between Trial and Trial Type of /ʃ/ versus Filler ($\beta = 0.05, SE = 0.02, t = 2.9$), indicating that reaction time for words with /s/ or /ʃ/ in them did not become as fast as those for filler words.

3.3.2 Categorization

Responses with reaction times less than 200 ms or greater than 2500 ms were excluded from analyses. A logistic mixed effects model was constructed with Subject and Continua as random effects and continua Step as random slopes, with 0 coded as a /ʃ/ response and 1 as a /s/ response. Fixed effects for the model were Step, Exposure Type, Attention and their interactions, with deviance coding used for contrasts for Exposure Type (Unpredictive = 0.5, Predictive = -0.5) and Attention (No attention = 0.5, Attention = -0.5).

As in the previous experiments, there was a significant effect of the intercept ($\beta = 0.52, SE = 0.20, z = 2.6, p < 0.01$) and of Step ($\beta = -2.49, SE = 0.19, z = -12.7, p < 0.01$). Exposure Type ($\beta = 0.23, SE = 0.23, z = 0.97, p = 0.33$), Attention ($\beta = 0.30, SE = 0.21, z = 1.4, p = 0.15$), and their interaction ($\beta = 0.38, SE = 0.44, z = 0.9, p = 0.39$) are all not significant, despite the visible differences in Figure 3.4.

As an additional comparison, the data from this experiment was combined with the subset of participants in Experiment 1 who were exposed to the same set of words (the word-medial condition). Exposure Type was recoded as a three-level factor, using treatment (dummy) contrast coding, with the Experiment 1 exposure (Isolation) as the refer-

Figure 3.2: Within-subject mean reaction time in the exposure phase of Experiment 3, separated out by Trial Type (Filler, /s/, and /ʃ/). Error bars represent 95% confidence intervals.

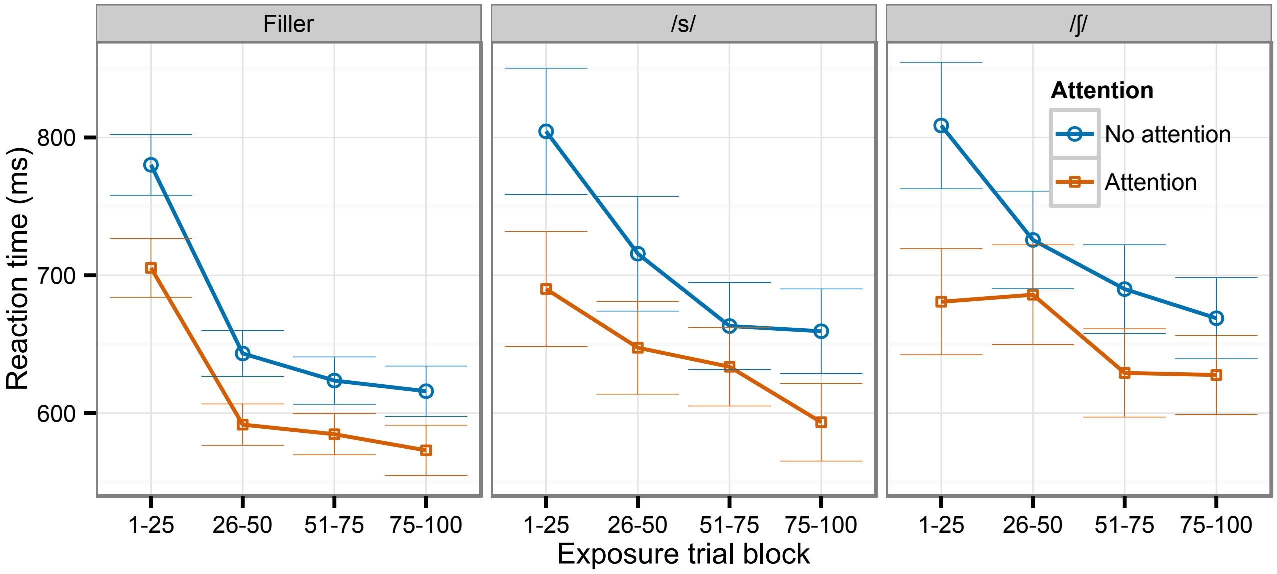
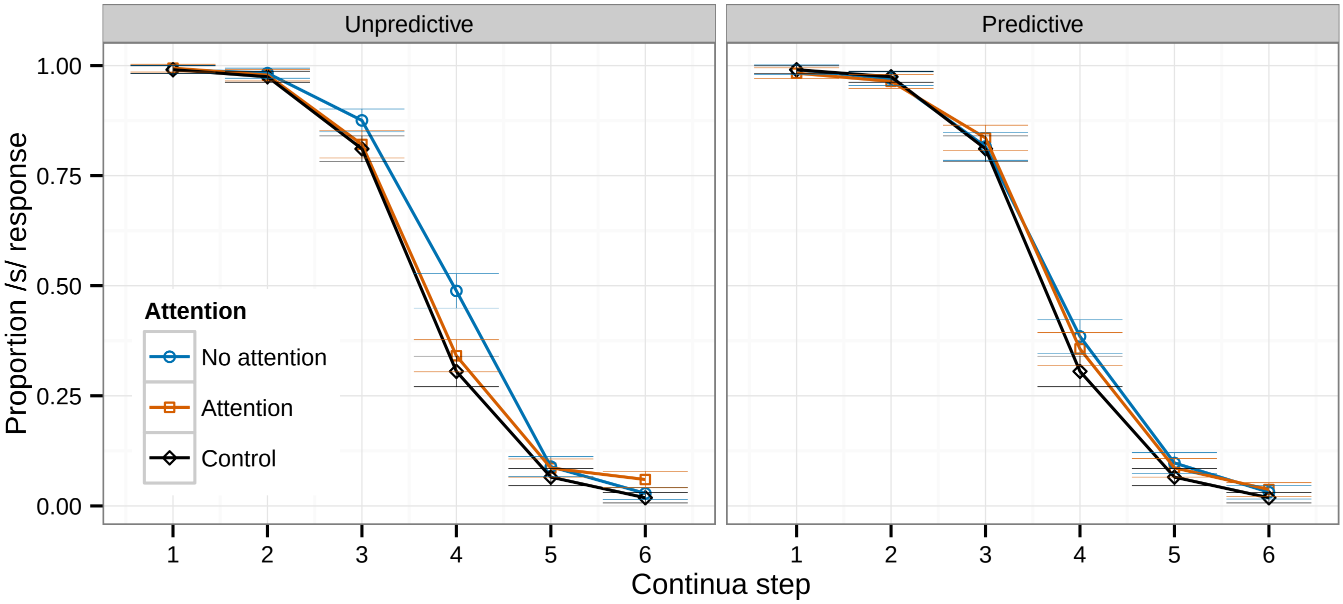
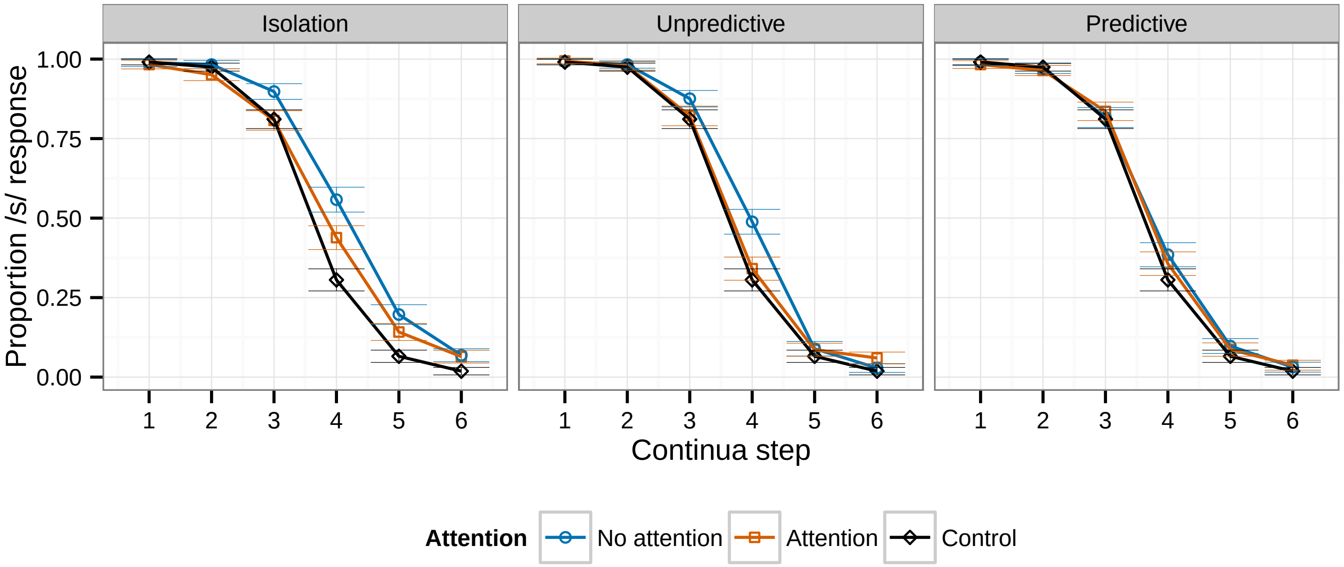


Figure 3.3: Proportion /s/ response along the 6 step continua as a function of Exposure Type and Attention in Experiment 3. Error bars represent 95% confidence intervals.



ence level. An identically specified logistic mixed effects model was fit to this set data as to the initial data. In this model, there was a significant effect of Attention ($\beta = 0.74, SE = 0.32, z = 2.2, p = 0.02$), such that participants in Attention conditions were less likely to categorize the continua steps as /s/. Additionally, Exposure Type had a marginal

Figure 3.4: Proportion /s/ response along the 6 step continua as a function of Exposure Type and Attention in Experiment 3 and the word-medial condition of Experiment 1. Error bars represent 95% confidence intervals.



effect of Predictive compared to Isolation ($\beta = -0.43, SE = 0.23, z = -1.9, p = 0.05$) and two significant interactions, one between Step and Unpredictive ($\beta = -0.42, SE = 0.17, z = -2.4, p = 0.01$) and one between Step and Predictive ($\beta = -0.32, SE = 0.14, z = -2.2, p = 0.02$). These marginal effect of Predictive indicates that participants in that condition were less likely categorize the continua as /s/ overall, and the interactions between the sentential conditions and Step indicate that the slope of categorization functions, as seen in Figure 3.4, were significantly steeper than in the Isolation condition. Attention and Exposure Type did not significantly interact.

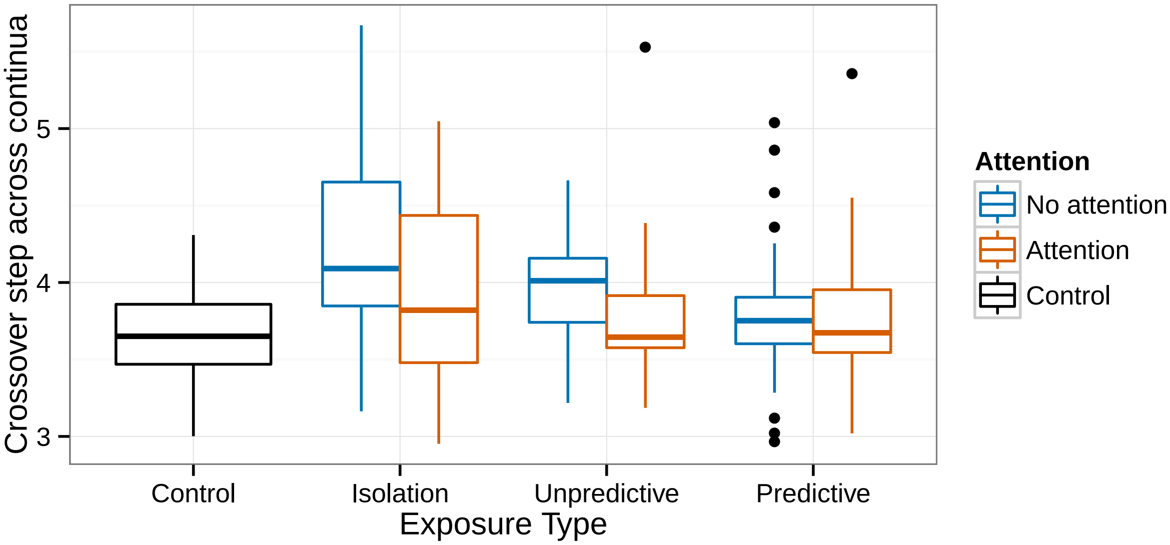
Two additional models were run with the reference level for Exposure Type as Predictive and Unpredictive. In the model with Predictive as reference, the intercept is no longer significant ($\beta = 0.44, SE = 0.28, z = 1.5, p = 0.12$), indicating perceptual learning was not robustly present in participants in the Predictive condition. The difference between the Predictive condition and the Isolation condition remains ($\beta = 0.77, SE = 0.32, z = 2.4, p = 0.01$), and, as above, the difference between Predictive and Unpredictive is not significant ($\beta = 0.42, SE = 0.31, z = 1.3, p = 0.17$).

3.4 Discussion

One key finding of the current experiment is that modified categories embedded in words in meaningful sentences produce less perceptual learning than words in isolation. Particularly, as the predictability of the context increases, the amount of perceptual learning decreases. In fact, participants exposed to a modified category only in predictive words had a similar boundary as the control experiment with no exposure phase at all.

Figure 3.5 shows the distributions of cross over points across the word-final conditions of Experiment 1 (labeled as Isolation) and all conditions of Experiment 3, with the distribution of the control experiment for comparison. Across all exposure types, there is a clear tendency for speakers in the Attention conditions to have smaller perceptual learning effects than their No attention counterparts. The distribution of the crossover points of participants from Experiment 1 have a larger interquartile range than those involving sentential contexts, but outliers are most prevalent in the Predictive condition with No attention, with a relatively narrow interquartile range. This abundance of outliers is likely due to a relative lack of constraint on the attentional set required by the task. In all other conditions, participants are guided toward perception-oriented attentional sets, either through a lack of contextual cues or through explicit instructions. As such, participants may naturally be inclined towards one attentional set or another.

Figure 3.5: Distribution of cross over points for each participant across comparable exposure tokens in Experiments 1 and 3.



There are several potential explanations for the lack of perceptual learning observed in the participants who heard the modified sound category only in predictable sentences. The first is that when the sentence is predictive of the final word, participants simply did not use the actual production of the final word for their response. That is, because the task was to identify the correct picture as fast as possible, the available information earlier in the sentence provided sufficient information to render a decision, so processing the final word was not necessary. An experiment that would shed light on whether this explanation is valid would be one that looked at differences in form encoding. For instance, if the form encoding is impoverished, long-term form priming would be affected, but long-term semantic priming would not be. For unpredictive sentences, both long-term form priming and semantic priming should be to be as robust as such priming for words in isolation.

Another potential explanation for the lack of perceptual learning effects observed is that predictive sentences are lower cognitive load than unpredictive sentences, as put forth by Samuel [51], where it was found that listeners are more sensitive to signal properties of words in predictive contexts. In line with that reasoning, the reduced cognitive load in predictive sentences would allow listeners to attend to the signal more easily, reducing the perceptual learning, as in the previous two experiments. However, perceptual learning effects were still observed in those experiments, whereas it appears to be absent entirely for the participants exposed to the sound category in predictive sentences.

One final potential explanation, related to the first, is that high predictability sentences are a “known” source of variation, much like a consonantal context like /str/. Recall that in Kraljic et al. [28], no perceptual learning was observed when the modified /s/ was in cluster where /s/ is typically realized closer to /ʃ/. Similarly, the modified category in this experiment may have been within in the range of variation for /s/ in predictable speech. Acoustic realizations of words in high predictability contexts are temporally and spectrally reduced [53], yet still highly intelligible in context [22]. Listeners, then, would not require a high degree of acoustic match a token in a predictable environment and their stored lexical entry.

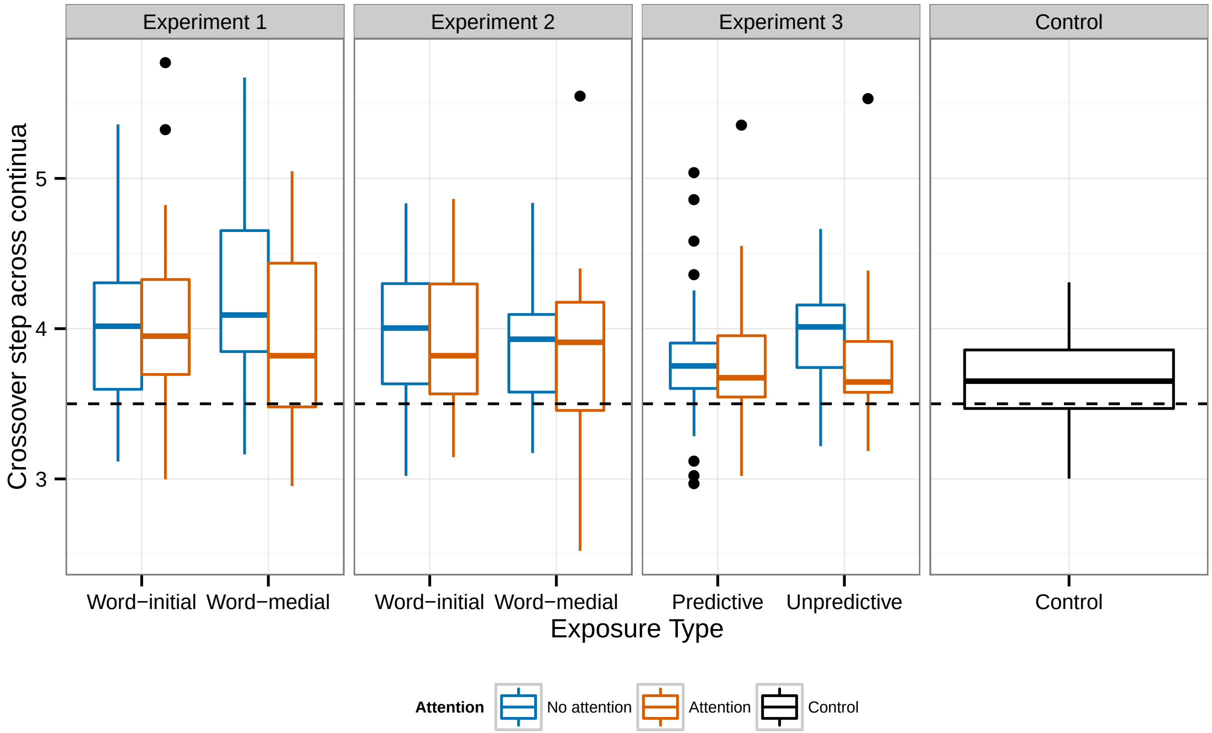
Many of the visually distinctive effects, as shown in the graphs, did not come out as significant in the mixed-effects models. Sentences may elicit more variable behaviour from participants than words in isolation, and this variation would have exceeded the tolerance for the models. The effect sizes may also have been smaller for sentential exposure, so the visual effects ended up only trending toward significance.

Chapter 4

Discussion and conclusions

This dissertation has examined the influence of attentional sets on perceptual learning in speech perception. Through explicit instructions and experimental manipulations, listeners were induced towards a more comprehension-oriented attentional set or a more perception-oriented attentional set. Those participants in conditions biased towards comprehension showed larger perceptual learning effects than those in conditions biased more towards perception. I adopted a predictive coding framework [11] that provides an intuitive model of perceptual learning and has been computationally implemented [25]. However, the attentional mechanism in the predictive coding framework does not work well for some instances of visual attention [5], or for the current results. I propose a new attentional mechanism for predictive coding, one in which attention inhibits error propagation beyond the level to which attention is directed. Such a mechanism explains both the previous findings and the current results.

Figure 4.1: Distribution of crossover points for each participant across all experiments. Dashed line represents the midpoint of the continua.



4.1 Attentional control of perceptual learning

The findings of Experiment 1 support the hypothesis that comprehension-oriented attentional sets produce larger perceptual learning effects than perception-oriented attentional sets. Although all participants showed perceptual learning effects, those exposed to the ambiguous sound with increased lexical bias only showed larger perceptual learning effects when the instructions about the speaker's ambiguous sound were withheld. Attention on the ambiguous sound equalized the perceptual learning effects across lexical bias. However, in Experiment 2, there is no such effect of attention, suggesting that the ambiguous sounds that were farther away from the canonical production induced a more perception-oriented attentional set.

One question raised by the current results is whether perception-oriented attentional sets always result in decreased perceptual learning. The instructions used to focus the listener's attentional set framed the ambiguity in a negative way, with listeners being cautioned to listen carefully to ensure they made the correct decision. If the attention were directed to the ambiguous sound by framing the ambiguity in a positive way, would we still see the same pattern of results? The current mechanism would predict that attention of any kind to signal properties would block the propagation of errors, reducing perceptual learning. This prediction will be tested in future work.

4.2 Specificity versus generalization in perceptual learning

The results of this dissertation speak to dichotomy between specificity and generalization found in perceptual learning work. In Experiment 1, greater perceptual learning was shown by participants exposed to ambiguous sounds later in the words, not to those at the beginnings of words. And yet, the testing continua consisted of stimuli with the sibilant at the beginnings of words, which are more similar to the words beginning with the ambiguous sound. In this sense, there was not only a lack of exposure-specificity, but a level of abstraction that is usually not assumed in perceptual learning studies. Most lexically-guided perceptual learning studies attempt to make the exposure tokens and the categorization as similar, and in some cases, the same, in order to maximize exposure-specificity effects. The results of this dissertation show that not only do listeners show perceptual learning effects from stimuli with large degrees of coarticulation (i.e., in the middle of the word) to stimuli without as much coarticulation, in some cases, perceptual learning is enhanced in precisely those cases.

The effect of attentional set manipulations in Experiments 1 and 2 suggest that when listeners adopt more perceptually-oriented attentional sets, even within tasks that are oriented toward comprehension, generalization of perceptual learning to new forms is inhibited. While visually-guided perceptual learning shows comparable effects to lexically-guided perceptual learning [59], lexically-guided perceptual learning is more likely to be expanded to new contexts [28, 40], though with some restrictions [38], than visually-guided perceptual learning [50]. Visually-guided and psychophysical perceptual learning paradigms often have highly repetitive stimuli with little variation. Both of these aspects add to the monotony and the likelihood of perception-oriented attentional sets [14]. Lexically-guided perceptual learning, on the other hand, requires very few instances to affect the perceptual system, usually around 20 ambiguous tokens within 200 trials, but as few as 10 ambiguous tokens have been shown to have comparable effects [29]. Following from the proposed attention mechanism, tokens heard under a comprehension-oriented attentional set should have a relatively large effect on the perceptual system as compared to tokens heard under a perception-oriented attentional set. This prediction is borne out by the lack of correlation between word endorsement rate and crossover point in Experiment 1 in the Word-final/No attention condition. This condition is predicted to have the most comprehension-oriented attentional set of the conditions, and here is the only instance in Experiment 1 where a strong correlation between word endorsement rate and crossover point is not present. Participants in this condition have relatively high crossover points that do not depend as much on the sheer number of tokens endorsed.

4.3 Effect of increased linguistic expectations

In comparing Experiments 1 and 3, increasing linguistic expectations does not increase perceptual learning. While the Isolation condition (Word-medial condition in Experiment 1) was not significantly different from the Unpredictive condition of Experiment 3, the trend was reduced perceptual learning when the modified sound category was embedded in a sentential context. The Predictive condition showed significantly less perceptual learning than the Isolation condition, but this could be driven by the wide variability found in that condition, as shown in Figure 4.1. Sentence processing may be too comprehension-oriented, and a balance between comprehension and perception is necessary for perceptual learning to be generalized.

Several studies have shown perceptual learning benefits for multiple talkers of an accent using sentential stimuli [8, and others]. Clearly, perceptual learning of accents is possible through hearing sentences, but the variability involved in those tasks reaches far beyond that involved here. The speaker producing the sentences in this dissertation is a native English speaker. Even with the synthesis applied to the sound files, he is likely more intelligible than the speakers in studies involving foreign accents. The ease of comprehension of the speaker in this study might actually inhibit perceptual learning in sentences, because listeners can leverage so much of their perceptual experience with other speakers of the local dialect.

4.4 Category atypicality

In Experiment 2, there was no effect of explicit instructions or lexical bias on perceptual learning, with a stable perceptual learning effect present for all listeners. There are two potential, non-exclusive explanations for the lack of effects. As stated above, the increased distance to the canonical production drew the listener's attention to the ambiguous productions, resulting in a perception-oriented attentional set. The second potential explanation is that the productions farther from canonical produce a weaker effect on the updating of a listener's categories, as predicted from the neo-generative model in [41]. This explanation is supported in part by the weaker correlation between word endorsement rate and crossover point found in Experiment 2, and the findings of Sumner [57] where the most perceptual learning was found when the categories begin like the listener expects and gradually shift toward the speaker's actual boundaries over the course of exposure. This explanation could be tested straightforwardly by implementing the same gradual shift paradigm used in Sumner [57] with the manipulations used in this dissertation.

An interesting extension to the current findings would be to observe the perceptual learning effects in a cognitive load paradigm. Speech perception under cognitive load has shown to have greater reliance on lexical information due to weaker initial encoding of the signal [35]. Following exposure to a modified ambiguous category, we might expect to see less perceptual learning if the exposure was accompanied by high cognitive load. However, Scharenborg et al. [55] found that hearing loss of older participants did not significantly influence their perceptual learning. Therefore it may be that perceptual learning would likewise be similar across cognitive loads. Higher cognitive loads, however, might allow for more atypical ambiguous stimuli to be learned, due to the increased reliance on lexical information during initial encoding.

It is important to bear in mind that what is typical in one context is not necessarily typical in another. The methodology employed for Experiment 3 drew on the assumptions that variance would common between the Experiments. However, it may be that the perfectly ambiguous /s/ category in Experiment 3 was within the range of variation when a word is predicted from context. In this case, had the category atypicality been more like that of Experiment 2, we may have actually seen more of an effect, perhaps back to the level of Experiment 1.

4.5 Conclusion

This dissertation investigated the influences of attention and linguistic salience on perceptual learning in speech perception. Perceptual learning was modulated by the attentional set of the listener. Decreased linguistic salience by increasing linguistic expectations induced more comprehension-oriented attentional sets, resulting in larger perceptual learning effects. Inducing perception-oriented attentional sets through explicit instructions or making the modified sound category more atypical resulted in smaller perceptual learning effects. These results support a greater role of attention than previously assumed in predictive coding frameworks, such as the proposed propagation-blocking mechanism. Finally, these results suggest that the degree to which listeners perceptually adapt to a new speaker is under their control to a degree, but given the robust perceptual learning effects across all conditions, perceptual learning is a largely automatic process.

Bibliography

[1] M. Ahissar and S. Hochstein. Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5718–22, 1993. ISSN 0027-8424. doi:10.1073/pnas.90.12.5718. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=46793&tool=pmcentrez&rendertype=abstract. → pages 7

[2] W. F. Bacon and H. E. Egeth. Overriding stimulus-driven attentional capture. *Perception & psychophysics*, 55(5): 485–496, 1994. ISSN 0031-5117. → pages 7

[3] J. N. Beckman. *Positional Faithfulness*. PhD thesis, University of Massachusetts Amherst, 1998. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470756171.ch16/summary. → pages 2

[4] P. Bertelson, J. Vroomen, and B. De Gelder. Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect. *Psychological Science*, 14(6):592–597, 2003. → pages 3, 4

[5] N. Block and S. Siegel. Attention and perceptual adaptation. *The Behavioral and brain sciences*, 36(3):205–6, 2013. ISSN 1469-1825. URL http://www.ncbi.nlm.nih.gov/pubmed/23663745. → pages 8, 26, 34

[6] S. Borsky, B. Tuller, and L. P. Shapiro. ”How to milk a coat:” the effects of semantic and acoustic information on phoneme categorization. *Journal of the Acoustical Society of America*, 103(5 Pt 1):2670–2676, 1998. URL http://www.ncbi.nlm.nih.gov/pubmed/9604360. → pages 6, 27

[7] A. R. Bradlow and J. a. Alexander. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *Journal of the Acoustical Society of America*, 121(4):2339–2349, 2007. ISSN 00014966. doi:10.1121/1.2642103. URL http://link.aip.org/link/JASMAN/v121/i4/p2339/s1&Agg=doi. → pages 6, 27

[8] A. R. Bradlow and T. Bent. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729, 2008. → pages 1, 36

[9] M. Brysbaert and B. New. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990, 2009. ISSN 1554-3528. → pages 11

[10] E. Clare. Applying phonological knowledge to phonetic accommodation, 2014. → pages 6

[11] A. Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences*, 36(3):181–204, 2013. ISSN 1469-1825. URL http://www.ncbi.nlm.nih.gov/pubmed/23663408. → pages 2, 4, 8, 17, 25, 26, 34

[12] C. G. Clopper and J. B. Pierrehumbert. Effects of semantic predictability and regional dialect on vowel space reduction. *Journal of the Acoustical Society of America*, 124(3):1682–1688, Sept. 2008. ISSN 1520-8524. doi:10.1121/1.2953322. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2676620&tool=pmcentrez&rendertype=abstract. → pages 6, 27

[13] C. Connine. Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, 538:527–538, 1987. URL http://www.sciencedirect.com/science/article/pii/0749596X87901380. → pages 6

[14] A. Cutler, J. Mehler, D. Norris, and J. Segui. Phoneme identification and the lexicon, 1987. ISSN 00100285. → pages 1, 5, 7, 35

[15] P. D. Eimas, W. E. Cooper, and J. D. Corbit. Some properties of linguistic feature detectors, 1973. ISSN 0031-5117. → pages 3

[16] F. Eisner and J. M. McQueen. The specificity of perceptual learning in speech processing. *Perception & psychophysics*, 67(2):224–238, 2005. → pages 4

[17] F. Eisner and J. M. McQueen. Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, 119(4):1950, 2006. ISSN 00014966. doi:10.1121/1.2178721. URL http://scitation.aip.org/content/asa/journal/jasa/119/4/10.1121/1.2178721. → pages 4

[18] M. Fallon, S. E. Trehub, and B. A. Schneider. Children's use of semantic cues in degraded listening environments. *The Journal of the Acoustical Society of America*, 111(5 Pt 1):2242–2249, 2002. ISSN 00014966. → pages 6, 27

[19] W. F. Ganong. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110–125, 1980. URL http://www.ncbi.nlm.nih.gov/pubmed/6444985. → pages 5

[20] E. J. Gibson. Improvement in perceptual judgments as a function of controlled practice or training. *Psychological bulletin*, 50(6):401–431, 1953. → pages 1, 3, 7

[21] C. Gilbert, M. Sigman, and R. Crist. The neural basis of perceptual learning. *Neuron*, 31:681–697, 2001. ISSN 0896-6273. doi:10.1016/S0896-6273(01)00424-X. URL http://www.sciencedirect.com/science/article/pii/S089662730100424X. → pages 4, 8, 9

[22] D. Kalikow, K. Stevens, and L. Elliott. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. ... *Journal of the Acoustical Society of* ..., 61(5), 1977. URL http://link.aip.org/link/?JASMAN/61/1337/1. → pages 6, 27, 28, 33

[23] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 3933–3936, 2008. → pages 11, 28

[24] F. Kleber, J. Harrington, and U. Reubold. The Relationship between the Perception and Production of Coarticulation during a Sound Change in Progress, 2012. ISSN 0023-8309. → pages 23

[25] D. Kleinschmidt and T. F. Jaeger. A Bayesian belief updating model of phonetic recalibration and selective adaptation. *Science*, (June):10–19, 2011. URL http://www.aclweb.org/anthology-new/W/W11/W11-06.pdf#page=20. → pages 2, 4, 34

[26] T. Kraljic and A. G. Samuel. Perceptual learning for speech: Is there a return to normal? *Cognitive psychology*, 51 (2):141–78, Sept. 2005. ISSN 0010-0285. doi:10.1016/j.cogpsych.2005.05.001. URL http://www.ncbi.nlm.nih.gov/pubmed/16095588. → pages 4, 6, 18

[27] T. Kraljic and A. G. Samuel. Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56 (1):1–15, Jan. 2007. ISSN 0749596X. doi:10.1016/j.jml.2006.07.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0749596X06000842. → pages 4

[28] T. Kraljic, S. E. Brennan, and A. G. Samuel. Accommodating variation: dialects, idiolects, and speech processing. *Cognition*, 107(1):54–81, Apr. 2008. ISSN 0010-0277. doi:10.1016/j.cognition.2007.07.013. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2375975&tool=pmcentrez&rendertype=abstract. → pages 6, 8, 27, 33, 35

[29] T. Kraljic, A. G. Samuel, and S. E. Brennan. First impressions and last resorts: how listeners adjust to speaker variability. *Psychological science*, 19(4):332–8, Apr. 2008. ISSN 0956-7976. doi:10.1111/j.1467-9280.2008.02090.x. URL http://www.ncbi.nlm.nih.gov/pubmed/18399885. → pages 1, 6, 9, 35

[30] P. K. Kuhl. Speech perception in early infancy: perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, 66(6):1668–1679, 1979. ISSN 00014966. → pages 1

[31] P. Ladefoged and D. E. Broadbent. Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29(1):98–104, 1957. ISSN 00014966. URL http://scitation.aip.org/content/asa/journal/jasa/29/1/10.1121/1.1908694. → pages 1

[32] A. B. Leber and H. E. Egeth. Attention on autopilot: Past experience and attentional set, 2006. ISSN 1350-6285. → pages 7, 8

[33] R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008. → pages 6

[34] P. Lieberman. Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech. *Language and Speech*, 6(3):172 –187, 1963. URL http://las.sagepub.com/content/6/3/172.abstract. → pages 6

[35] S. L. Mattys and L. Wiget. Effects of cognitive load on speech recognition. *Journal of Memory and Language*, 65 (2):145–160, 2011. → pages 5, 6, 7, 36

[36] L. H. Mayo, M. Florentine, and S. Buus. Age of second-language acquisition and perception of speech in noise. *Journal of speech, language, and hearing research : JSLHR*, 40(3):686–693, 1997. ISSN 1092-4388. → pages 6, 27

[37] M. Mcauliffe. python-acoustic-similarity, 2015. URL https://github.com/mmcauliffe/python-acoustic-similarity. → pages 12

[38] H. Mitterer, O. Scharenborg, and J. M. McQueen. Phonological abstraction without phonemes in speech perception. *Cognition*, 129(2):356–361, 2013. → pages 4, 35

[39] D. Norris and A. Cutler. The relative accessibility of phonemes and syllables. *Perception & psychophysics*, 43(6): 541–550, 1988. ISSN 0031-5117. → pages 7

[40] D. Norris, J. M. McQueen, and A. Cutler. Perceptual learning in speech, 2003. → pages 1, 2, 3, 4, 5, 6, 7, 8, 9, 18, 35

[41] J. Pierrehumbert. Word-specific phonetics. *Laboratory phonology*, 2002. URL http://books.google.com/books?hl=en&lr=&id=N5quZxfmpswC&oi=fnd&pg=PA101&dq=Word-specific+phonetics&ots=3kCkpgcbgI&sig=kMSAYjSD1UDqQr8VOpP0cst4LDE. → pages 25, 36

[42] J. B. Pierrehumbert. Exemplar dynamics: Word frequency, lenition and contrast. *Frequency and the emergence of linguistic structure*, pages 137–158, 2001. ISSN 01677373. → pages 9

[43] M. Pitt and C. Szostak. A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation. *Language and Cognitive Processes*, (April 2013):37–41, 2012. URL http://www.tandfonline.com/doi/abs/10.1080/01690965.2011.619370. → pages 1, 2, 5, 6, 7, 10

[44] M. A. Pitt and A. G. Samuel. Attentional allocation during speech perception: How fine is the focus? *Journal of Memory & Language*, 29(5):611–632, 1990. ISSN 0749596X. → pages 7

[45] M. A. Pitt and A. G. Samuel. An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of experimental psychology. Human perception and performance*, 19(4):699–725, 1993. ISSN 0096-1523. → pages 5

[46] M. A. Pitt and A. G. Samuel. Word length and lexical activation: longer is better. *Journal of experimental psychology. Human perception and performance*, 32(5):1120–1135, 2006. ISSN 0096-1523. → pages 2, 5, 10

[47] I. Psychology Software Tools. E-Prime, 2012. URL http://www.pstnet.com. → pages 12, 28, 29

[48] E. Reinisch and L. L. Holt. Lexically Guided Phonetic Retuning of Foreign-Accented Speech and Its Generalization. *Journal of experimental psychology. Human perception and performance*, 40(2):539–555, 2013. ISSN 1939-1277. URL http://www.ncbi.nlm.nih.gov/pubmed/24059846. → pages 4

[49] E. Reinisch, A. Weber, and H. Mitterer. Listeners retune phoneme categories across languages. *Journal of experimental psychology. Human perception and performance*, 39(1):75–86, Feb. 2013. ISSN 1939-1277. doi:10.1037/a0027979. URL http://www.ncbi.nlm.nih.gov/pubmed/22545600. → pages 1, 4, 8, 12, 14, 18

[50] E. Reinisch, D. R. Wozny, H. Mitterer, and L. L. Holt. Phonetic category recalibration: What are the categories? *Journal of Phonetics*, 45:91–105, 2014. ISSN 00954470. doi:10.1016/j.wocn.2014.04.002. URL http://linkinghub.elsevier.com/retrieve/pii/S009544701400045X. → pages 1, 4, 8, 9, 25, 35

[51] A. G. Samuel. Phonemic restoration: insights from a new methodology. *Journal of experimental psychology. General*, 110(4):474–494, 1981. ISSN 0096-3445. → pages 5, 6, 8, 9, 27, 33

[52] A. G. Samuel. Red herring detectors and speech perception: in defense of selective adaptation. *Cognitive psychology*, 18(4):452–499, 1986. ISSN 00100285. → pages 3

[53] R. Scarborough. Lexical and contextual predictability: Confluent effects on the production of vowels. *Laboratory phonology*, pages 575–604, 2010. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Lexical+and+contextual+predictability:+confluent+effects+on+the+production+of+vowels#0. → pages 6, 27, 33

[54] O. Scharenborg and E. Janse. Comparing lexically guided perceptual learning in younger and older listeners. *Attention, perception & psychophysics*, 75(3):525–36, 2013. ISSN 1943-393X. URL http://www.ncbi.nlm.nih.gov/pubmed/23354594. → pages 7, 8

[55] O. Scharenborg, A. Weber, and E. Janse. The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention, Perception, & Psychophysics*, pages 1–15, 2014. → pages 7, 8, 26, 36

[56] D. Shankweiler, W. Strange, and R. R. Verbrugge. Speech and the problem of perceptual constancy. *Perceiving, acting, and knowing: Toward an ecological psychology*, pages 315–345, 1977. → pages 1

[57] M. Sumner. The role of variation in the perception of accented speech. *Cognition*, 119(1):131–136, 2011. ISSN 0010-0277. doi:10.1016/j.cognition.2010.10.018. URL http://dx.doi.org/10.1016/j.cognition.2010.10.018. → pages 9, 10, 36

[58] M. Sumner and R. Kataoka. Effects of phonetically-cued talker variation on semantic encoding. *The Journal of the Acoustical Society of America*, 134(6):EL485, Dec. 2013. ISSN 1520-8524. doi:10.1121/1.4826151. URL http://www.ncbi.nlm.nih.gov/pubmed/25669293. → pages 1

[59] S. van Linden and J. Vroomen. Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of experimental psychology. Human perception and performance*, 33(6):1483–1494, 2007. ISSN 0096-1523. → pages 4, 35

[60] J. Vroomen, S. van Linden, B. de Gelder, and P. Bertelson. Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(3):572–577, 2007. → pages 1, 3, 4, 9

[61] J. M. Wolfe and T. S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature reviews. Neuroscience*, 5(6):495–501, 2004. → pages 7

[62] Y. Yeshurun and M. Carrasco. Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, 396(6706):72–75, 1998. ISSN 0028-0836. → pages 8, 26