ABSTRACT

Title of Dissertation: THE LEARNING AND USAGE OF SECOND LANGUAGE SPEECH SOUNDS: A COMPUTATIONAL AND NEURAL APPROACH

Craig Thorburn
Doctor of Philosophy, 2023

Dissertation Directed by: Associate Professor Naomi Feldman
Department of Linguistics

Language learners need to map a continuous, multidimensional acoustic signal to discrete abstract speech categories. The complexity of this mapping poses a difficult learning problem, particularly for second language learners who struggle to acquire the speech sounds of a non-native language, and almost never reach native-like ability. A common example used to illustrate this phenomenon is the distinction between /r/ and /l/ (Goto, 1971). While these sounds are distinct in English and native English speakers easily distinguish the two sounds, native Japanese speakers find this difficult, as the sounds are not contrastive in their language. Even with much explicit training, Japanese speakers do not seem to be able to reach native-like ability. (Logan, Lively, & Pisoni, 1991; Lively, Logan, & Pisoni, 1993).

In this dissertation, I closely explore the mechanisms and computations that underlie effective second-language speech sound learning. I study a case of particularly effective learning—a video game paradigm where non-native speech sounds have functional significance (Lim &

Holt, 2011). I discuss the relationship with a Dual Systems Model of auditory category learning and extend this model, bringing it together with the idea of perceptual space learning from infant phonetic learning. In doing this, I describe why different category types are better learned in different experimental paradigms and when different neural circuits are engaged. I propose a novel split where different learning systems are able to update different stages of the acoustic-phonetic mapping from speech to abstract categories. To do this I formalize the video game paradigm computationally and implement a deep reinforcement learning network to map between environmental input and actions.

In addition, I study how these categories could be used during online processing through an MEG study where second-language learners of English listen to continuous naturalistic speech. I show that despite the challenges of speech sound learning, second language listeners are able to predict upcoming material integrating different levels of contextual information and show similar responses to native English speakers. I discuss the implications of these findings and how the could be integrated with literature on the nature of speech representation in a second language.

THE LEARNING AND USAGE OF SECOND LANGUAGE SPEECH SOUNDS: A COMPUTATIONAL AND NEURAL APPROACH

by

Craig Thorburn

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Associate Professor Naomi Feldman, Chair
Associate Professor Ellen Lau
Professor William Idsardi
Professor Philip Resnik
Professor Jonathan Simon, Dean's Representative

© Copyright by
Craig Thorburn
2023

Preface

The work in this dissertation is highly collaborative. Chapters 3 and 4 report joint work with Naomi Feldman and Ellen Lau, and were supported by NSF Grant BCS-2120834 . As of submission, a version of Chapter 3 is currently under review as Thorburn, Lau & Feldman (under review). Chapter 5 reports joint work with Jonathan Simon, Ellen Lau, Dushyanthi Karunathilake and London Dixon and was supported by NIDCD Grant NIH-R01-DC019394, NSF Grant SMA-1734892, and a Seed Grant from the UMD Brain and Behavior Institute.

Acknowledgments

Writing a dissertation and completing a PhD is impossible to do alone, and I have been incredibly lucky to work with and be supported by many amazing people throughout my time at the University Of Maryland. I thank everyone who has helped me throughout the last five years—even the smallest nods of encouragement have made a big difference and short conversations have led to interesting ideas and thoughts which have in some form shaped the work as I present here.

First, I need to thank my amazing advisors, Naomi Feldman and Ellen Lau. My weekly conversations with Naomi have particularly formed me into the scientist that I am today. Naomi has an incredible ability to ask tough questions that are difficult to answer in the moment but after thinking about them more, have pushed me to refine my ideas and theories. You have always been very encouraging for whatever direction I've wanted to take my research and embrace and push some of my more adventurous ideas.

Ellen is such a joy to have as an advisor and has been a huge source of encouragement and support during my PhD. I have enjoyed our wide-ranging discussions about the field and how science is conducted just as much as your thought-provoking questions about the deeper ramifications of the questions I am asking. You have always reminded me to not shy away from challenging the greater assumptions of the field and pushing the boundaries with my studies and theories. I love the energy you brings to meetings, often reminding me of how much fun science can be. I am so lucky to have worked with these two amazing mentors, and they will both have a

long-lasting impact both personally and on my scientific career.

The other members of my committee have provided a lot of very valuable feedback and advice as this work has progressed. Bill Idsardi has been an incredible repository of knowledge of speech and has helped refine many of the ideas in this dissertation through helpful feedback. As chair of the department you also provided much support, particularly to international students, during the pandemic, of which I am very thankful for. Jonathan Simon has proven a great resource for MEG analysis techniques, particularly when I came to this area later in my graduate career. Philip Resnik was one of the first people to inspire me to combine computational work with neuroscience techniques through discussions early in my graduate career, which started me along many of the paths that resulted in this work. I still remember fondly debates in seminars in my first two years which formed many of my thoughts on the combination of these methods to this this day.

I am thankful for the many interesting discussions with the other students and postdocs who work on computational models and speech at UMD. Thomas Schatz provided great support in my first computational project five years ago and set me up with a solid foundation to perform much of the work presented here. Nika Jurov, Leslie Li and Joselyn Rodriguez have been fantastic confidants both as friends, and in developing theories and discussing my many outlandish ideas, and I very much hope we will work together in the future. I have also benefited greatly from many other discussions and feedback from others including Jackie Nelligan, Kasia Hitczenko, Nathalie Fernandez Echeverri, Grace Brown, Cassidy Henry, Sathvik Nair and Zara Harmon.

On the neuroimaging side, I have enjoyed working with London Dixon and Dushyanthi Karunathilake on our Orchard project. Dushyanthi has had a wealth of knowledge on analyzing this data and has been very patient as I have learned the ropes. London has been a great friend

and sounding board as I have ventured more into sentence processing than I might be used to. Christian has provided assistance from afar several times and has left a wealth of scripts and programs which have streamlined much of my recent work.

I have been fortunate to work with two talented undergraduates. Saahiti Potluri shows such enthusiasm for learning and has brought many interesting perspectives to the table during our last two years working together. Xiaoyu Yang has always been ready to step in at a moment's notice to help out and has been a great help in running MEG participants.

Ciaran Stone has been invaluable as lab manager of the MEG lab, showing extensive knowledge about the system and taking much time to train and carefully walk through everything. This includes diagnosing any and all unusual artifacts observed in my data—even when that artifact turns out to be the participant's heart rate.

I have had the privilege of working with many great teachers during my time, particularly Tonia Bleam and Peggy Antonisse. I enjoyed many semesters as a teaching assistant for introductory phonology with Peggy, an experience that not only allowed to develop my teaching, but also helped refine many ideas that relate theoretical work to the research I present here. In addition, my time spent doing outreach with Kathleen Oppenheimer, Lauren Salig and Erika Exton has been very rewarding and always provided a light-hearted and well-needed break from research. I've also enjoyed working in other mentoring roles including PULSAR with Andrea Zukowski and Tess Wodd and the Language Science Station and Planet Word with Charlotte Vaughn and Hannah Mechtenberg.

I have greatly appreciated that in this department everyone is excited to talk about everyone's research and I have enjoyed classes and/or conversations with Colin Phillips, Jeff Lidz, Juan Uriagereka, Alexander Williams, Masha Polinsky and Howard Lasnik. I'm not sure I could

be in a more different area of linguistics to Howard, but I've particularly valued my many interactions with you with your friendly presence and incredible teaching. I'm honored to say that I've taken both a syntax class with you as a famed linguist and Scottish dancing with you as a prolific dance instructor. Outside the department, I've had great discussions and classes with many other faculty—in particular I loved two classes with Daniel Butts, which prompted me to think more deeply about the more implementational details of neural circuitry.

Kim Kwok, as ever, has been an incredible resource in the department and has helped me navigate the bureaucracy of university life. You have made this entire experience much easier with your problem-solving and wonderful presence around the office. The support of the Maryland Language Science Center with Shevaun Lewis and Caitlin Eaves has also been important in many of my interdisciplinary connections as well as the outreach work that I have done, as well as the staff in UMIACS.

On a more personal note, I am happy to leave this department with many friends, often forged through the struggles of research. Our cohort of Masato Nakamura, Cassidy Henry and I may be small, but I still fondly remember our insomnia cookie runs in our first year as we worked on problems in the office together. Masato, we may enjoy whiskeys of different origins, but it's been fun to go through the last few years with you.

I am grateful to the many other researchers and graduate students that I have been able to share my time at UMD including Max Papillon, Phoebe Gaston, Paulina Lyskawa, Aaron Dolina, Anouk Dieuleveut, Mina Hirzel, Rodrigo Ranero, Tyler Knowlton, Sigwan Thivierge, Adam Liter, Hanna Muller, Jackie Nelligan, Yu'an Yang, Maša Beslin, Alex Krauska, Jessica Mendes, Polina Pleshak, Imane Bou-Saboun, Clara Cuonzo, Rosa Lee, Luisa Seguin, Jack Ying, Xinchi Yu, Jingyi Chen, Fedor Golosov, Zulfiyya Aghakishiyeva, Katherine Howitt, Aura Cruz Heredia,

Laurel Whitfield and Jad Wehbe. While sadly the pandemic cut our time short for those who started before me and delayed meeting for those who have started since, I am incredibly happy to have met so many friends along the way. You have all made this department a special place to be for the last five years.

There are a few people who deserve special mention here. Alex has been a good friend through the last several years, from organizing trivia nights through the pandemic to always being down for a good cocktail since. I am so grateful for your constant willingness to take a break to simply chat and spend time together. Every time I knock on Nika's office door she is more than happy to spend an hour talking about some study we've just heard about or life in general. While this might not be the most useful for productivity, those conversations are always very fun—thank you for letting me distract you so much. I've enjoyed our many happy hours together—whether for fun or to talk about reinforcement learning. Masa, Katherine, and Luisa I've been so happy for our time together over the last few years. Your encouragement has been particularly meaningful during the last stretch of my time here and I am grateful to have many great memories from the end of my degree.

I would also not have made it through the last five years without the support and fun times with so many friends outside academia, both near and far. Rima and Richard, your friendship and support as always means so much—both our calls and fun weekends together have helped me through many times. I'm so happy I've gotten to see you so much over the last five years, although it never seems to be enough. Similarly, Zoom calls with the rest of our 'Squid Squad' of Katy, Rachel, Tessa, Claire and Ayesha have been particularly meaningful during long periods of quarantine and I'm so glad to stay good friends. For all my friends here that I have met along the way - Sarah, Mary, Santi, Zack, Fran, Camille, Rob, Rosette, David, Margaret, Ben, Beau,

Theresa, and Fiona I am glad to have so many fun memories from my time in DC. Knowing Bill, Betsy, Ellen and Rich has also been very meaningful in letting me to have some of Scotland in my life through dancing—those Wednesday nights were one of the highlights of my first year.

My parents—Fiona and Derek —you have always provided so much unconditional support in everything that I do. Even when I ramble about Linguistics you are always happy to listen. Your love and support always means so much and I could not have done any of this without you.

And finally Emily - from reading through my paper drafts, hearing me talk about the /r/ and /l/ sounds more than anyone ever needs to, or simply providing love and support through the difficult times, you have stood by me through it all. I've been so glad to have you by my side when I've been at a loss and celebrate with you through the good times. As much as you might disagree, I really couldn't have done this without you. Thank you for everything.

Table of Contents

List of Tables

List of Figures

Chapter 1: Speech Sound Learning

The ability to form categories is a facet of unique importance to human cognition. Whether sorting pets into the categories of 'cat' or 'dog', furniture into categories of a 'table' or 'chair' or basic visual input into 'square' or 'circle', humans are continually forming boundaries between different pieces of information. One case where category mapping is particularly important for communication is the ability to sort sounds into distinct phonetic categories.

Phonetic categories[1] are the building blocks of larger units of linguistic meaning and one of the most fundamental discrete units used in language comprehension. The difference between the category /r/ and /l/ is the difference between the word 'rock' and the word 'lock' and the ability to distinguish between these two categories is important in differentiating these words. Language learners need to map a continuous, multidimensional acoustic signal onto these discrete abstract speech categories and the complexity of this mapping poses a difficult learning problem.

Infants as young as six months of age demonstrate some specialization for this knowledge in their native language —- seemingly losing the ability to differentiate between soundsthat are not part of any languages they hear while developing (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). As adults, however, phonetic categories are one of the most challenging features

[1] In this dissertation I do not make a distinction between phones and phonemes. The categories in this dissertation are considered to have some status as an underlying representation in the language and could be considered phonemes, however since I am discussing an abstract representation of speech sounds more generally, without mention of their specific status in phonological theory, I will refer to these as 'phonetic categories' throughout this dissertation.

of a second language to grasp and even with extensive training, it is unlikely that adult language learners will ever reach native-like ability at phonetic category discrimination and categorization.

In this dissertation, I closely explore the mechanisms and computations that underlie effective second-language speech sound learning and how speech category representations can be used in online processing and my contribution to the wider literature is twofold. First, I extend the Dual Stream Model of auditory category learning, bringing this literature together with the idea of perceptual space learning from infant phonetic learning. In doing so, I describe why different category types are better learned in different experimental paradigms and when different neural circuits are engaged. I use computational models where reinforcement learning represents learning in an implicit paradigm and supervised learning represents learning in an explicit paradigm.

Second, this dissertation considers how the representations that learners have of a second language can be used to predict upcoming input, specifically using information at various context levels. I show that listeners do integrate information at different context levels when anticipating the next phoneme. These results are particularly novel considering previous evidence showing that in controlled settings, second-language learners do not predict upcoming input to the same extent as native listeners. I discuss the implications of my analysis and that my results may in fact be complementary to previous findings.

In this first chapter, I discuss what typical second-language speech sound categorization looks like and the potential methods of learning that have been proposed. I focus particularly on one of these methods which appears to be particularly effective for learning speech categories as an adult—- implicit learning through a video game paradigm (Wade & Holt, 2005; Gabay, Dick, Zevin, & Holt, 2015; Lim & Holt, 2011). Improved performance for the recognition of categories

between the start and end of this video game paradigm is observed for various types of stimuli — both synthesized noise (Lim, Fiez, & Holt, 2019) and speech tokens (Lim & Holt, 2011). In the latter experiment, adults show as much improvement in the discrimination of non-native speech sounds after 2 hours of video game training as they do during 10 hours of explicit training over a period of 2-4 weeks (Logan et al., 1991). A clue to the effectiveness of the paradigm could lie in its engagement of reward-based learning mechanisms and in Chapter 2 I discuss a specific implementation of these algorithms— reinforcement learning.

I use this information to present an explicit computational model of reinforcement learning for this paradigm in Chapter 3 and discuss the implications of these results as they apply to second-language speech sound learning. This is one of the first computational models of speech sound learning in a video game paradigm and helps to illustrate why this paradigm is so effective. I provide evidence that the nature of the algorithms used when learning through this paradigm do not provide a boost in performance, instead suggesting that other properties of the system such as the involvement of different neural circuitry or the specific knowledge that each system is able to update are the cause of the effectiveness of the paradigm.

I explore a Dual Learning Systems Model of auditory category learning in Chapter 4 and propose a novel theory of differences between two auditory learning systems inspired by modeling work and recent findings in infant phonetic learning literature. I suggest that prior literature that proposes two systems is consistent with a theory where each learning system can influence separate stages of the acoustic-phonetic mapping of speech sounds, explaining why there is a split in the types of category boundaries that are best learned by each of the systems. I extend the presented reinforcement algorithm and conduct an experiment to test this proposal. To my knowledge, while prior models have described how different systems show a split in the nature

of categories they preferably learn, few theories have proposed why this split exists, particularly within speech and auditory category learning.

In Chapter 5, I investigate how second-language learners could use their speech representations during online speech processing and whether the categories that are learned during second-language acquisition affect how these listeners predict upcoming information at different context levels. I show that second-language learners may in fact have similar neural responses to prediction as native speakers and integrate different levels of contextual information when listening to naturalistic continuous speech. I present a new corpus where second-language learners of English listen to an audiobook while neural responses are recorded using MEG. Finally, I discuss future directions for this work and ways in which neural studies of sentence processing in second-language learners can be integrated with studies on phonetic category acquisition.

1.1 Second Language Speech Sound Learning

Before discussing the video game paradigm for speech sound learning and its integration with a Dual Stream Model of auditory category learning, I first provide some background on second language speech sound representation and learning mode generally. I introduce work on /r/-/l/ discrimination for native Japanese speakers—an example of a challenging contrast that is difficult to learn later in life and I will use throughout the dissertation; before discussing work that addresses the outcomes of various learning strategies. My proposal also has some links to two prior models of second language speech sound learning as well as infant phonetic learning and the concept of a perceptual space which are also introduced here.

1.1.1 The /r/-/l/ Contrast

One of the most cited examples of challenges of adult learning of speech sound categories in a second language comes from Goto (1971), who show that Japanese adults struggle to differentiate between /r/ and /l/ sounds in English. These two sounds are important to distinguish in English as they create 'minimal pairs' — pairs of words with different meanings where the only difference between the words is the distinction between two sounds. When Japanese speakers are presented with two words that form a minimal pair between /r/ and /l/ such as 'rock' and 'lock', they are significantly worse than native English speakers at identifying the correct word. The primary auditory dimension along which the two sounds differ in American English is the third fundamental frequency produced by the vocal tract during production (F3). Where English has two distinct phonemic categories along this dimension — /r/ and /l/, Japanese only has one sound and the dimension is not required to distinguish between any consonants[2].

For the /r/-/l/ contrast, English speakers show good discrimination of speech sounds along the F3 continuum when they cross a category boundary but decreased discrimination for sounds that are within a category. However Japanese native speakers—because they do not have two categories along this continuum—are only slightly above chance in discrimination for all pairs of sounds across an F3 speech continuum. When the F3 dimension itself is isolated from a word and the sound does not sound like speech, Japanese and American English native speakers perform

[2]A note on terminology: In this dissertation, I mostly use the terms 'second-language learner' and 'non-native listener' interchangeably to mean one who is learning and listening to a language as an adult, which they did not hear from birth. The concept of 'nativeness' is very complex and recent work has suggested that classifying speakers into 'native' and 'non-native' groups may lose a lot of complexity in language knowledge (Weissler et al., 2023). These are important points and definitely warrant further consideration, however for simplicity in this dissertation I use 'second-language learner' to refer to someone who has not been exposed to the 'second language' as an infant. Specifically in my experiment, I use age 12 as an arbitrary cutoff, although the studies discussed in this dissertation may use different criteria to classify these speakers.

equally in discrimination of pairs along the continuum, suggesting that these effects are isolated to the speech perception system (Miyawaki et al., 1975).

It has been suggested that Japanese adults have difficulty with this particular contrast because they are overly sensitive to phonetic cues which do not indicate differences between /r/ and /l/ (Iverson et al., 2003; Iverson, Hazan, & Bannister, 2005)—in other words, they are not correctly weighting F3 above all other cues to be able to successfully differentiate these categories. One can measure discrimination across varying F2 and F3 cues for participants and create a space that demonstrates the perceptual acuity across these two dimensions, where sound pairs that are less likely to be discriminated successfully are grouped closer together. Japanese speakers seem to have more perceptual acuity along the F2 dimension than the F3 dimension when compared with English speakers (Iverson et al., 2003) and they may be attempting to discriminate the /r/ and /l/ sounds along this dimension rather than using F3 as native English speakers do. Variability in the stimuli presented can have a large effect on categorization, as well as the exact category options presented (Yamada & Tohkura, 1992)—for example along this continuum, native Japanese speakers will sometimes categorize some sounds as /w/. In this work I will focus on the case where /r/ and /l/ are the only possible responses.

Production ability is also impacted by native language and native Japanese speakers of English have a wider variation in production of /r/ and /l/ sounds. Production of these sounds improves with exposure (for example, living in an English-speaking country), although even with improved production ability, there is still variability in perception, suggesting that the acquisition of production may come before perception (Yamada, Strange, Magnuson, Pruitt, & Clarke, 1994). Length of residency in an English-speaking country does not correlate with discrimination/identification ability (Aoyama & Flege, 2011), however this may not be a good measure of

proficiency of a language (Park, Solon, Dehghan-Chaleshtori, & Ghanbar, 2022).

Difficulty with contrasts in a second language is well documented across a variety of different sounds and languages including other contrasts that can be challenging for Japanese speakers such as /s/-/θ/ and /b/-/v/ (Guion, Flege, Akahane-Yamada, & Pruitt, 2000); contrasts that can be challenging for native English speakers such as French /y/-/u/ (Levy, 2009), Korean stop consonants (Holliday, 2015), Mandarin /ʂ/ - /ç/ (Noguchi & Kam, 2018), and Nthlakampx uvular and velar ejectives (Werker & Tees, 1984; Werker, Gilbert, Humphrey, & Tees, 1981); as well as many more (Broersma, 2005)

As a running example and given the extensive background on this one particular contrast, much of the discussion in this dissertation will revolve around the /r/-/l/ contrast. In Chapter 4, I will discuss using alternative cues to signal the /b/-/p/ contrast for native English speakers. Although these examples will be used throughout, the ideas and theories presented in this dissertation are widely applicable and illustrate how the perceptual system functions more generally.

1.1.2 Second Language Phonetic Training

Training can help second language learners increase in categorization ability although they often do not reach the performance level of native speakers. Japanese participants can be trained on /r/ vs /l/ sounds in a discrimination task, where they hear two tokens and are asked to judge if the tokens are two of the same sound or different sounds before receiving explicit feedback about whether the sound was same or different (Strange & Dittmann, 1984). Participants improve in discrimination of these speech sounds after training on synthetic stimuli, however, this performance boost does not transfer to natural speech.

A variety of learning paradigms or adjustments to presented stimuli have been explored to try and show increased improvement for second language categorization. Presenting a large variety of tokens, instead of just a single or few exemplars of a particular category has been shown to improve performance, particularly when the tokens presented present a wide variety of the perceptual space. This technique—dubbed High Variability Training—has proven effective in numerous studies and was first explored by Logan et al. (1991) where Japanese speakers are trained with a large variety of tokens in each category—a range of natural stimuli including the sounds presented within numerous phonetic environments. While this does result in some small boost in discrimination, this improvement is still not to the level of native listeners. Results from this form of training have been replicated (Lively et al., 1993), extended to production ability (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997) and improvement is retained after 3 and 6 months have passed, with accuracy only decreasing slightly (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994). Shinohara and Iverson (2018) show that both discrimination and identification training paradigms benefit from high variability in stimuli. Moreover, the improvement compared to more typical training paradigms is observed even for L2 learners that have already had extensive exposure to the second language. This suggests that high variability training is doing something more than just mere exposure to the sounds, and presenting varying categories allows participants to better update their perceptual representations. Despite this extensive work, it is incredibly uncommon for second language learners to reach the ability of native speakers in experimental paradigms. Feedback during learning appears to be particularly important for speech category learning in a second language—although unsupervised learning can result in some gains in performance, significant improvement mostly comes from paradigms that involve some amount of supervision (McClelland, Fiez, & McCandliss, 2002; Goudbeek, Swingley, &

Smits, 2009).

Second-language speech sound learning is challenging, particularly when sounds are identified along a dimension or cue which is not part of the native language, or when the categories conflict with native language categories. These speakers rarely reach native-like ability, even when presented with extensive training paradigms.

1.1.3 The Speech Learning Model and Perceptual Assimilation Model

Two previous theories have considered why the speech representations that one has for a native language may interfere with the correct perception and learning of new perceptual categories—the Speech Learning Model (SLM) and the Perceptual Assimilation Model (PAM). The work in this dissertation and my proposal do not bear specifically on these theories, however, it may provide a more detailed mechanistic account of the larger theoretical claims. Each of these models deals with a slightly different class of speaker, where the SLM discusses second language learners (ie. someone who has learned a language other than their native language) and the PAM focuses on naive listeners (ie. someone listening to a non-native language that they do not know). However both theories make similar predictions about how listeners perceive sounds that are not in their native language.

The PAM predicts that incoming non-native speech sounds that are similar to sounds in the native language will be assimilated to these categories. This model specifically looks at contrasts between two sounds, where each sound will either assimilate to an existing L1 category, if it is close enough, or remain 'uncategorized' if it is not close to any L1 category. Discrimination between two sounds will be most challenging when both sounds are matched to the same

L1 category. Applied to the example of /r/ and /l/ discrimination for native Japanese speakers, this theory predicts that native Japanese speakers who do not know English will struggle to discriminate between these two categories as they are both assimilated to the same Japanese native category.

The SLM posits a similar reason that second language learners (distinguished from a naive listener as someone who has some knowledge of the second language) find discrimination of some second language speech sounds challenging. This theory discusses how second language learners generate new phonemic categories in their second language and posits that learners will only assign a new category to incoming input if it is perceptually far enough away from any native language sounds. Any incoming input that is close enough to a native category will be perceived as a member of that category. Again for native Japanese speakers who are learning English (a subject group common in the studies that are discussed above), discrimination between English /r/ and /l/ is challenging because neither of these sounds is distinct enough from the native representation that they will be mapped to a single category.

The SLM posits that L2 learners can "gradually discern" the differences between L2 sounds and only then are they able to form a stable phonetic category for them. This learning process is influenced by the complexity of the L2 phonetic categories, and how frequently they occur modulates the amount of exposure needed to discern them. The most recent version of this model—the SLM-r (Flege & Bohn, 2021)—makes more specific predictions about the link between L1 and L2 learning mechanisms. The SLM-r proposes that phonetic categories for word recognition are generated by statistical learning and that L2 and L1 learners across the lifespan use the same mechanisms to learn speech sounds, rather than learners losing the ability to learn speech sounds after the critical period. Any differences in the outcomes of learning are attributed to the specific

input that is received during learning and the initial state of the system. The same mechanisms are suggested to be used during native and second language learning, however for second language learners, the system starts from a place where a learner already knows and has strong representations of their native language. This interferes with second language sound processing in the ways outlined above, to the extent that speech sounds are difficult to learn. The fact that incoming L2 input is initially perceived according to L1 phonetic categories means that there are necessary differences in the observed outcomes of L1 and L2 learning and sometimes L1 categories can block L2 category learning entirely.

In this dissertation, I am not concerned with whether L1 and L2 learners have access to the same learning mechanisms across the lifespan. As discussed above, L2 learners seem to take advantage of specific learning strategies to form representations of L2 speech sounds that are not the same as those used during infancy. While on the face, the situations which give rise to learning appear to be different than the ones used by L1 learners, that does not mean that L2 learners do not have access to the mechanisms that L1 learners have—the strategies used to learn one's native language may simply not be beneficial for second language learning due to the issues laid out above. For this reason in this dissertation, I focus on the paradigms where L2 learners *can* learn speech categorization and why these techniques are particularly useful.

I do not discuss the claim that second language input is initially categorized as native categories until evidence suggests otherwise; however, the general claim that new input is filtered through the lens and processes of a native language does fit with the proposal I present. Instead of new input being categorized as L1 sounds, I instead propose that there is a perceptual space that L2 sounds are represented in which leads to categorization difficulties. In order to propose this, I integrate ideas from recent literature, particularly that the learning of a perceptual space

could be relevant during different kinds of adult phonetic learning

1.1.4 Infant Phonetic Learning

The ability to distinguish sounds in one's native language is a vital skill that infants learn at a very young age. At birth, infants are able to distinguish speech sounds in any language, before becoming showing specialization (Werker & Tees, 1984; Kuhl et al., 1992) in the ability to discriminate sounds in their native language as young as six months old (Kuhl et al., 1992, 2006). At this stage, infants lose the ability to discriminate between sounds that are not in their native language. This is usually taken as evidence that infants as young as six months old have knowledge of native phonetic categories.

One prominent proposal is that infants use statistical learning (Saffran, Aslin, & Newport, 1996; Saffran & Kirkham, 2018) to form representations of the speech sounds of their native language (Best, 1994). By accumulating distributional information about their environment and uncovering structure in the input, they are able to form categories around clusters of sounds that they hear regularly. Infants are able to use this information in laboratory studies where they are either presented with sounds that form a unimodal or bimodal distribution across an acoustic continuum. Infants in the latter condition are much better at discriminating sounds at either end of the continuum than those in the former condition, showing that infants are sensitive to the distributions of speech sounds along the spectrum and suggesting that they could leverage distributional information to cluster sounds into categories (Maye, Werker, & Gerken, 2002). This effect has been demonstrated in infants as young as 3-4 months of age (Wanrooij, Boersma, & Van Zuijen, 2014), although the mechanism may be less effective at around 10 months old,

once infants have passed the point where they have become specialized at their own language (Yoshida, Pons, Maye, & Werker, 2010). While statistical learning of phonetic categories is an appealing proposition and in-lab studies appear to show that infants are able to learn from this information, this learning mechanism may break down in some situations as real-world speech is much more variable than that presented in the lab (Bion, Miyazawa, Kikuchi, & Mazuka, 2013; Antetomaso et al., 2017). Speech can be affected by speech rate, gender, dialect and prosody, meaning that distributions that are often represented as discrete bimodal distributions in lab experiments, rarely surface as such in real speech data.

One recent theory of how infants' phonetic learning occurs is that incoming acoustic information causes the warping of a perceptual space towards the center of a category. For example, one argument for internal category structure comes from the perceptual magnet effect (Kuhl, 1991), which shows that speech sounds are more challenging to discriminate closer to the center of a category - dubbed a 'prototype' of the category. This effect is seen in both adults and infants and is taken as evidence that even infants have a structured representation of the acoustic space. While initial theories of perceptual space warping proposed that this space was formed around prototypes—a 'typical' example of a particular category—it is in fact possible to capture this effect without positing discrete categories (Guenther & Gjaja, 1996). By simply accounting for acoustic input and modeling a perceptual space according to the distribution received as input, it is possible to demonstrate this effect.

Computational modeling can be particularly helpful here as it allows the comparison of various learning mechanisms and test the effects captured by different algorithms (Matusevych, Schatz, Kamper, Feldman, & Goldwater, 2020; Schatz & Feldman, 2018; Schatz, Feldman, Goldwater, Cao, & Dupoux, 2019). This work demonstrates that it is possible to observe effects more

conventionally attributed to category knowledge only using information about the acoustic environment. Schatz, Feldman, Goldwater, Cao, and Dupoux (2021) illustrate that a model which describes acoustic input using only Gaussian distributions can exhibit learning effects seen in infants. Other effects that are commonly attributed to discrete phonetic categories such as the Language Familiarity Effect (Johnson, Westrek, Nazzi, & Cutler, 2011) can also be captured with this kind of model (Thorburn, Feldman, & Schatz, 2019).

More recently, Feldman, Goldwater, Dupoux, and Schatz (2021) propose that infants may not learn phonetic categories in the first year of life but instead learn a perceptual space that is not yet divided into discrete categories. Infant studies typically provide evidence for one particular 'categorical' effect—that they can discriminate between sounds along some dimensions but not others. They do not, however, show a peak at the category boundary—which requires knowledge of distinct categories. The processes in the first year of life may result in the infant learning the particular dimensions in their perceptual space that are necessary for distinguishing sounds in their input. Crucially, this proposal that behavior observed in infant experiments may not arise due to category knowledge, does not mean that humans in general do not possess discrete phonetic categories. This hypothesis simply states that the kinds of effects seen in infants do not arise due to phonetic categories. The idea of perceptual space learning as a separate process from category learning is important for the ideas I present in Chapter 4, where I propose that differences between learning paradigms and the systems engaged in adults arise because of the existence of perceptual space learning.

Regardless of whether listeners learn discrete categories in infancy or childhood, it is clear that knowledge learned during the early years of life is challenging to update. While the contents of this dissertation do not directly bear on the question of infant language acquisition, it is ben-

eficial to acknowledge the mechanisms that could contribute to native learning when discussing L2 learning, as these may be similar or different to the ones used later in life. It also helps to understand the form of the representations that could be learned in infancy.

1.1.5 Perceptual Space Learning

In order to distinguish perceptual space learning from the knowledge of discrete categories more clearly, it is helpful to first define what behavioral effects are typically used as evidence of category knowledge. I consider the perspective of Feldman et al. (2021) as a good summary of the literature on category and perceptual knowledge, which is typically illustrated as three different effects (Figure 1.1). The first is a sharp identification curve along the category boundary. This means that the transition from identifying a sound consistently as a member of one category and another is very steep–there is very little space for ambiguous sounds at this crossover point. This can usually be seen as 'cue-weighting' where the steepness of a boundary between categories along a particular dimension is equivalent to how strongly one weights that dimension in categorization. It is possible to see the steepness of a categorization curve change as a listener learns a specific category. If the curve gets shallower on one dimension and steeper on a corresponding dimension, then we can say that one has reweighted their cues (Idemaru & Holt, 2011, 2014; Francis & Nusbaum, 2002; Toscano & McMurray, 2010).

The second effect is a peak in discrimination ability at the category boundary. Listeners will typically show decreased sensitivity to differences between two sounds that are within a category and will be better at discriminating sounds that are members of different categories. This could arise either from a large distance between two groups of sounds within a perceptual space, or

due to the formation of discrete category knowledge. While the former is possible without the existence of explicit categories, it would indicate that the sounds are mostly separable in the perceptual space, and so this effect is usually attributed to category knowledge, which is what I will assume in this dissertation. Related to a peak at the category boundary, it is also possible that decreased discrimination ability is observed within a category, particularly around where most examples of a particular category are within the acoustic space. This effect once again requires either distinct category knowledge of a 'prototypical' example of that category (Kuhl, 1991), or well-separated peaks in the input that are able to warp the perceptual space. For this reason, I presume that both of these effects observed in experiments demonstrate category knowledge.

The third effect is the fact that listeners show different sensitivities to changes in stimuli across different dimensions. After being trained in categorization between two sounds, listeners can still discriminate between sounds along an irrelevant dimension (Lehet & Holt, 2020), suggesting that this ability may not be inherently linked to category knowledge. The third effect is distinct from the second effect in that it does not require a peak in discrimination across a category boundary. While this may not indicate category knowledge, it does involve some knowledge of a perceptual space more generally.

For this dissertation I take the position that effects one and two demonstrate knowledge of categories, whereas effect three indicates that a listener has some knowledge about the nature of the acoustic input (ie. they have learned a perceptual space that accounts for the distribution of the acoustic information), but does not indicate that the listener has fully formed knowledge of abstract categories.

So how might this effect appear without the presence of phonetic categories? It could be due to the learning of a perceptual space— a space that encodes perceptual information and is

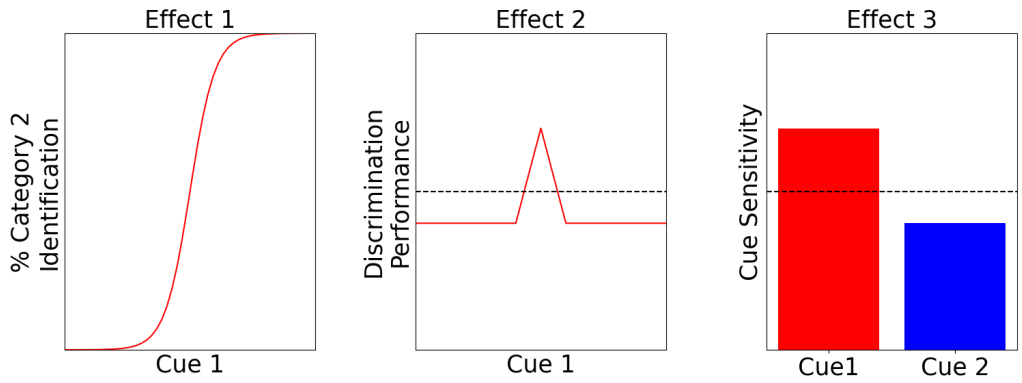


Figure 1.1: Diagrams of the three effects discussed in this dissertation. In this example, two categories exist along Cue 1 shown in red, and Cue 2 shown in blue is irrelevant distinguishing any categories. **Effect 1:** A sharp categorization boundary is observed for Cue 1. **Effect 2:** A peak in discrimination is observed at the category boundary. **Effect 3:** There will be more sensitivity to Cue 1 than Cue 2 - ie. general discrimination ability across Cue 1 will be better than discrimination for the irrelevant Cue 2. In this dissertation I presume that Effects 1 and 2 require discrete category knowledge and Effect 3 can be seen by only knowing information about the acoustic environment.

learned to best represent a specific language. This means that even without the knowledge of perceptual categories, it is still possible to see changes in general discrimination ability across different cues as one learns dimensions through the acoustic space. By accumulating information about the distribution of the input across many dimensions, one is able to learn the acoustic cues or combinations of acoustic cues that are required to best account for the incoming input. Importantly this knowledge does not include abstract categories.

The learning of a perceptual space is one way that it is possible to see effect three (that listeners can learn to discriminate sounds generally across dimensions that exist in their native language), without requiring phonetic categories (Feldman et al., 2021). This effect is separate from the increased peak at the category boundary and decreased discrimination around typical examples of the category (Figure 1.1). While it has been shown that the latter effects can be captured with the 'warping' of a perceptual space without category knowledge (Guenther & Gjaja,

1996), the algorithms that can capture this require a clear bimodal distribution with two separated peaks. Since naturalistic speech input does not typically have these properties, it is unlikely that listeners could learn categories from this information in a typical natural environment (Schatz et al., 2021). Alternatively, warping around typical category centers could be driven by category information, where a listener learns what typical examples of a category sound like and representations within the category are drawn closer to this prototype (Kuhl, 1991). For this work, I assume that the peak at the discrimination boundary—whether the changes are driven by category information or bimodal acoustic input—reflects underlying category knowledge rather than perceptual space learning. However the general sensitivity to a cue may arise from a learned perceptual space, and updating this space could be crucial to successful second language speech category learning.

This idea has already been proposed in slightly different terminology. Iverson et al. (2003) suggest that the issue of second language category learning for Japanese adults is that their perceptual space could be 'mistuned' for learning English. This space, which is formed during infancy interferes with second language sound learning as it brings representations along the F3 dimension closer together, making it challenging to form categories along this dimension. However, representations along the F2 dimension are more spread apart, which is what results in L2 speakers trying to rely more heavily on this dimension.

Similarly, work from Roark, Plaut, and Holt (2022) demonstrates through a neural network model that long-term priors are important to place restrictions on what can be learned later in life. I interpret this idea similarly to perceptual space learning—the perceptual space places restrictions on the types of categories can be learned. Early language experience alters the relatively low-level perceptual processing, which one builds categories across later in life and when learn-

ing a second language. The proposal in Chapter 4 puts forward a specific role for perceptual space learning in second-language speech sound learning.

1.2 Implicit Learning and The Video Game Paradigm

While the work that I have presented so far illustrates a challenging landscape for second language learners, there is a particularly effective paradigm for speech sound learning, which comes in the form of a video game paradigm. First proposed by Wade and Holt (2005), in this paradigm participants control a spaceship at the center of a screen with aliens entering from various locations. Participants must shoot or capture these aliens and increase their score every time they do so. Auditory tokens are presented before each alien enters and are sampled from a category that corresponds to the alien's location and color, providing an early cue for identifying which direction the participant should face. As the experiment progresses, aliens begin to move faster and it becomes increasingly difficult to turn and shoot or capture without attending to the auditory information.

After playing the video game for 30 minutes, participants are able to reliably discriminate between the different auditory categories that corresponded to each side of the screen from which aliens entered. This learning occurs despite the participants never being explicitly told that there is a correlation between the sounds and the video game. There is also a correlation between the performance in the video game across several measures and the ability for categorization post-game.

When participants play this video game where speech categories are the sounds that are functionally relevant to good performance, participants show significant improvement in identi-

fication of these sounds. This is shown by presenting Japanese native-speaking participants with a version of the game where the /r/ and /l/ sounds correspond to the directions that aliens enter from (Lim & Holt, 2011). The authors claim that for Japanese speakers learning these sounds, 5 hours playing the video game is equivalent to approximately 2-4 weeks of explicit training, and importantly, speakers show a reweighting from relying on F2 to categorize /r/ and /l/ to relying more on F3.

To simplify the paradigm to allow more systematic comparisons to other situations while maintaining the important aspects associated with the video game environment, Gabay et al. (2015) developed the Systematic Multimodal Associations Reaction Time (SMART) paradigm. In this paradigm, participants hear a collection of auditory tokens before seeing a visual stimulus appear in one of four regions of the screen. Participants must react as fast as they can by pressing a button associated with the area of the screen in which the visual stimulus appears. In the same way that aliens can enter from a side of the screen corresponding to the auditory tokens heard before the alien enters in the video game paradigm, the auditory tokens in this paradigm can correspond with the region in which the visual stimulus appears, allowing participants to have a shorter reaction time if they are able to implicitly match categories with screen regions. Performance on this task can be measured in two ways—measures of categorization (ie. discrimination or identification performance) can be presented after training within the SMART paradigm. Alternatively, a block can be presented during training, where the relationship between sound and visual region is broken, by presenting randomized sounds before each visual stimulus. If participants have generated associations between auditory tokens and the location of the visual stimulus, then reaction times should increase in this block, since stimuli are not appearing where participants would expect them to. This is indeed the behavior observed (Gabay et al., 2015)

Participants are unlikely to use an unsupervised learning mechanism (ie statistical learning)—where discrimination increases by just passively listening to the stimuli—within these paradigms. Lim et al. (2019) show that when tokens are randomized and do not correspond to a specific action set, participants do not improve in identification. Furthermore, a study by Roark, Lehet, Dick, and Holt (2022) using the SMART paradigm demonstrates that task-aligned actions are specifically required in order to see learning effects in a similar paradigm. This means that listeners cannot be simply mapping visual and auditory cues by seeing a specific alien appear after a specific sound.

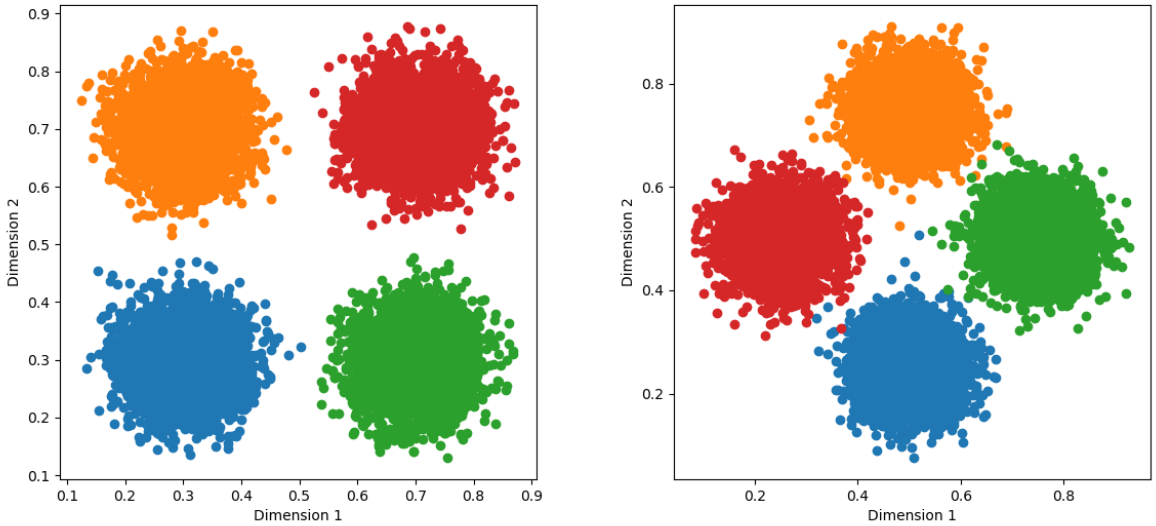
1.3 The Dual Systems Model

One recent model that may be important in explaining the effectiveness of the video game paradigm is the Dual Learning Systems (DLS) Model of auditory category learning. Taking influence from other domains of category learning—particularly vision—this theory posits that humans have access to two distinct learning systems which perform better at learning different category types. The *reflective* system is most similar to explicit learning and is where a category boundary is clearly defined over one dimension, can be verbalized and is easily testable. These are considered to be 'rule-based' categories and this system engages primarily the anterior regions of the striatum and prefrontal cortex. Alternatively, categories that require the integration of cues from various categories are proposed to use the *reflexive* system. These categories are classified as 'information integration' categories and are proposed to engage the caudate head, which is active during implicit learning tasks. (Chandrasekaran, Yi, & Maddox, 2014; Chandrasekaran, Koslov, & Maddox, 2014)

Theories that propose two learning systems in humans are very common (Sloman, 1996) and the DLS model is founded on the Competition between Verbal and Implicit Systems (COVIS) model first proposed by Ashby, Alfonso-Reese, Turken, and Waldron (1998). This is a domain-general perspective on two competing learning systems in humans, although it has mostly focused on visual category learning. Under this proposal, a procedural learning system is engaged by the prefrontal cortex which specializes in information integration categories and a declarative system engaged by the anterior cingulate which specializes in learning categories that are 'verbalizable'.

In the original COVIS theory, 'verbalizable' is described to highly correlate with unidimensional categories, where the decision boundary is exactly orthogonal to the dimension which is being categorized, however rules that can be described easily that may rely on more than one dimension are also discussed as falling into this category. In the Dual Systems Model this notion has evolved slightly to solely refer to unidimensional categories. Regardless of the definitions set out previously, in this dissertation, I will presume that the categories that are preferably learned through the rule-based system are unidimensional, and reference to only one dimension is required to describe the category boundary. Usually, a task that is proposed to involve this learning system consists of two perceptual dimensions, where one dimension is irrelevant for categorization of the stimuli, and in the other dimension, there is a clear boundary that will maximize categorization performance. An example of categories that would be considered under this system is shown in Figure 1.2a.

The other learning system is a procedural learning system that engages the tail of the caudate in the striatum and is primarily mediated by dopaminergic circuits (Ashby, Queller, & Berretty, 1999). The procedural system is proposed to be more effective at learning information integration categories. These are categories where categorization performance is maximized



(a) Rule-Based

(b) Information Integration

Figure 1.2: Example of rule-based and information integration category types. Under the COVIS and DLS models of learning, rule-based categories are best learned through a declarative system and information integration categories are best learned through a procedural learning system

through a boundary that requires the integration of information from both dimensions (Figure 1.2b). Motor actions are vital for this type of learning as these appear to induce the grouping of perceptual input in this system. Ashby et al. (1998) propose that in the caudate tail, groups of neurons associate particular actions with specific regions of the perceptual space, leading to categorization of the sounds in this region. One particularly important factor for procedural learning appears to be an association with actions, leading to a focus on the striatum as well as the putamen (Ashby & Maddox, 2011; Ashby, Ell, & Waldron, 2003).

Both of these learning types are considered in the COVIS model to rely on corticostriatal loops where information flows from cortical sensory areas to the striatum, the globus pallidus, the thalamus and back to cortex. In the initial COVIS model, learning is first dominated by the verbal system, and—with training—the implicit system can be engaged to give rise to better category

learning, depending on the type of category being presented (Ashby et al., 1998). This system associates responses generated in the striatum with groups of cells that represent higher-level (in this case, visual) representations of the input in inferotemporal cortex.

In the DLS model, a very similar framework is considered for auditory category learning, with different regions of the striatum which correspond to each of these learning systems performing different functions. During declarative learning (in the DLS model, reflective learning), the connection between the caudate head and the prefrontal cortex is strengthened whenever a new rule is generated and the rule is maintained through this connection. Alternatively in procedural learning (in the DLS model, reflexive learning), groups of cells that represent regions of the perceptual space are associated with a specific cortical-motor response through a unit (or small group of units) in the tail of the caudate. Once again, action is important in the procedural learning system as groups of sensory neurons are associated with specific actions during learning. The DLS model is proposed as a model of auditory category learning generally, however in this dissertation I will focus on its implication for speech categories and specifically those of a second language.

During experiments, it is possible to push people toward one of either of these systems by making subtle manipulations to a category learning task (Chandrasekaran, Yi, & Maddox, 2014; Maddox & Ashby, 2004). For example, the reflective system requires explicit feedback about category membership, so by providing minimal feedback (correct or incorrect) instead of full feedback (specific category membership), learning should be impacted in the reflective system, but not the reflexive system. Similarly, the reflexive system depends more on immediate feedback than the reflective system and therefore by delaying feedback, learning in the reflexive system should be impeded. Experimental evidence shows that when making these manipulations

that change the amount or timing of feedback and impede the reflective system, rule-based category learning decreases, where information-integration category learning remains the same, with the opposite true of manipulations that impede the reflexive system, confirming that each category type is better learned by a different system. These results are consistent with findings that feedback during training is more beneficial for unidimensional categories than multidimensional categories (Goudbeek et al., 2009).

Additionally, there is evidence that people who have impairments to specific aspects of these learning systems have difficulty with speech categories consistent with the dual system split in speech category learning. For example, people with elevated depressive symptoms would be thought to have an impaired reflective system since that system relies on frontal regions which are impacted by these symptoms. Since speech learning is considered reflexive-optimal and these participants are being pushed to use the reflexive system, since their reflective learning is impaired, participants who have these symptoms show better learning of second language category structure than those without the impairment to the reflective system (Chandrasekaran, Koslov, & Maddox, 2014) .

The setup of the video game paradigm such that actions are important during learning, explicit category information is not given, and the dorsal striatal circuits are engaged (Lim et al., 2019) means it clearly engages the reflexive system under this theory. The learning of speech categories is *reflexive optimal* given that the video game paradigm gives rise to particularly effective category learning, as well as evidence from dual systems experiments, where experimental manipulations pushed participants towards one learning system or the other. Because speech categories tend to require the integration of a variety of different acoustic cues, learning phonetic categories as an adult potentially benefits from the involvement of the reflexive learning system.

The evidence and proposal from the DLS model of category learning, show that the caudate head is particularly important for second language sound category learning.

The results from DLS experiments are consistent with previous findings that variability is helpful in speech sound acquisition. Chandrasekaran, Yi, and Maddox (2014) show that presenting stimuli that include multiple talkers increases accuracy on speech category learning tasks and propose that this is because participants are pushed away from a rule-based system as cues for a specific talker can no longer be used in categorization. If the existence of multiple talkers forces participants away from a reflective system and towards a reflexive system which is optimal for speech, this could explain why presenting a variety of input during speech category learning tasks is particularly effective. It appears that people switch off between the two systems during learning to switch to the more effective strategy depending on the category types being learned (rule-based or information integration) and performance for an individual between different category types is consistent (Roark & Chandrasekaran, 2023).

A comparison between the two learning systems in the DLS model comes in the form of the SMART paradigm (Gabay et al., 2015), which allows for a controlled comparison between implicit learning and explicit or overt learning paradigms. The order between the response from the participants and the presentation of the visual stimuli can be flipped to keep the experiment as closely aligned as possible. A comparison between the SMART paradigm and explicit learning for auditory categories surprisingly did not show a significant advantage for implicit learning over explicit learning (Roark & Holt, 2018a), however, no other direct comparisons have been made. These results may also arise due to the specific setup of the SMART paradigm. One of the primary measures in this paradigm is the increase in reaction time during a block where auditory input and visual stimulus are randomized, with the assumption that the extent to which

a participant has learned the correlation between stimulus and action will be reflected in how much slower they are when this correlation does not exist. The authors suggest though, that this randomized block could impact the categorization results presented at the end of the experiment in that the randomized block undoes some of the learning that has happened in previous blocks. Since there is no equivalent block in the explicit paradigm, the direct comparison may not be entirely equivalent. Although differences between the two learning paradigms did not arise here, the effectiveness seen in some papers using the video game paradigm, (Lim & Holt, 2011; Lim, Fiez, & Holt, 2014; Lim et al., 2019) alongside the evidence posited by the DLS model outlined above, suggest that there are some differences between implicit and explicit learning.

The split in systems in the Dual Learning Systems model may have interesting commonalities with the Dual Stream Model of speech processing proposed by Hickok and Poeppel (2007). This model suggests two different systems for speech processing — the dorsal stream connects to the motor system and is used during tasks that require 'phonological awareness' and enables phonological short-term memory. The ventral stream does computations for lexical access and interacts with higher-level language networks. While the DLS proposal does not specifically reference the Dual Stream Model, the task differences that appear in the DLS model, indicate that the two may use similar ideas. In the case of DLS, implicit and explicit tasks may need access to different information and performance is increased when using a system that may provide this information. Explicit learning requires knowing discrete category knowledge in order to assign identity to acoustic input, whereas implicit learning needs only to associate a certain region of the perceptual space with a given action. In the Dual Stream Model, tasks engage different processing streams despite acting over the same information, and each stream does not have the same access to update perceptual representations. Similarly, the implicit and explicit systems may not

be able to manipulate representations of phonetic information in the same way.

I aim to shed light on the differences between the implicit and explicit learning systems in Chapters 3 and 4 of this dissertation with modeling and experimental work. However, before presenting a computational model of the video game paradigm, I must provide some background on the algorithm that I will use.

Chapter 2: Reinforcement Learning

Given the engagement of the striatal system and setup of the video game paradigm, I explore the role of reinforcement learning in implicit learning. I propose a specific model after addressing what reinforcement learning looks like, the neural circuitry involved, and the other uses of reward-based learning in second language speech sound acquisition.

2.1 Algorithms

Reinforcement learning as an algorithm was first proposed by Sutton and Barto (1998) where the main learning signal for the model comes in the form of a reward. An agent takes actions within an environment and receives a reward upon reaching a favorable outcome. At each timestep, the agent receives information about the state of the environment and predicts the value of each action available to it. The agent then takes an action, observes the reward received, and updates its parameters according to the mismatch between the predicted and observed reward value. This process is iterative, enabling the agent to take actions that lead to reward in the future. Even if the agent does not receive a reward immediately, it will still take actions that put it in a better position to receive a reward at a future timestep.

A classic example of reinforcement learning from the domain of robotics comes in the form of a robot that is programmed to pick up trash scattered around a large building. The robot has

only limited battery power and must return to its docking station to charge at regular intervals. At every point in time, the robot has numerous actions available to it—the different directions it can move. Some directions may take it to find additional trash, whereas some may take it back to the charging station. When the robot picks up a piece of trash it may get 1 point of reward, however if it runs out of battery when not at the charging station it may receive -100 reward. The job of the agent (in this case the robot) is, therefore, to take in information about the state of the environment (ie. what it can see around it, the location in the building, its current battery level) and choose an action that will give it the most reward in the long-term. With a high battery it should be incentivized to explore and find more trash to receive a greater reward. With a low battery, it should return to the charging station to not receive a negative reward. While each individual action will not necessarily result in a reward, the robot will try to maximize its total reward over time, taking actions that lead to states which will eventually lead to higher reward.

Reinforcement learning algorithms are usually framed as Markov Decision Processes (Bellman, 1957), where the agent makes a decision at each timestep based solely on the current states. There are two main categories of reinforcement learning—model-free reinforcement learning and model-based reinforcement learning. In model-based reinforcement learning the agent has a model of the structure of the environment and information about the possible states that will follow each action. It can leverage this information to take actions that put it in the position to receive the highest reward. In other words, the agent knows transition patterns or probabilities between different states and makes decisions based on which state will come next. A model-free reinforcement learning algorithm does not have this model and instead makes decisions based purely on the learned value of each action, given the current state. It does not use specific information about what the proceeding state will be. In this work, I discuss a specific implementation

of model-free reinforcement learning.

The construction of the video game paradigm has components that lend itself clearly to an implementation of reinforcement learning. A participant receives information about the state of the environment (visual and auditory cues within the video game), takes an action within that environment (turning and shooting aliens), and receives a reward for correct actions (increases in score).

To illustrate the specific properties that reinforcement learning has and as a baseline and point of comparison throughout this dissertation, I will contrast it with a much more commonly used algorithm of supervised learning. In supervised learning a function is created which maps between input and a corresponding output—in this setting a function which maps auditory input directly to abstract categories. For any input, the model generates a corresponding output and receives the ground truth value of the output. The model can directly compare the predicted output from the model with the ground truth output and update its parameters so that its predicted output will be closer to the ground truth in the future. I make a clear distinction between these two algorithms, where reinforcement learning is given a reward signal, and the supervised algorithm is given specific category information. These are different learning signals — I consider that a reinforcement learning agent is only told *when* it performs correctly, but is not told the correct action when it does not. In contrast, the supervised network receives information about the correct decision on each trial.

I will focus on a Deep Q Learning model of reinforcement learning (Mnih et al., 2015), although several other formulations of reinforcement learning exist (ie. TD learning Sutton & Barto, 1998 Actor-Critic Learning Konda & Tsitsiklis, 1999).

2.2 The Striatum and Basal Ganglia

The basal ganglia are a collection of deep subcortical nuclei which are responsible for a variety of functions including executive function, motor learning, reward encoding and implicit learning (Lanciego, Luquin, & Obeso, 2012). These nuclei have wide connections to all cortical lobes and the entire region is collectively heavily implicated in motor processing. Of particular interest in this work, is the largest substructure of the basal ganglia in the mammalian brains—the striatum. The striatum contains the input nuclei of the basal ganglia and consists of the dorsal striatum containing the caudate nucleus and putamen and the ventral striatum which consists of the nucleus accumbens. Inputs received in the striatum project signal to the output nuclei consisting of the substantial nigra and the globus pallidus. Signals are processed through the thalamus before being sent back to cortex. The connections in this overall pathway are dubbed corticostriatal loops and are particularly important for the learning signals discussed here (Ashby et al., 1998)

For a long time, the basal ganglia have been known to encode reward and reward prediction through dopamine neurons in the midbrain which project to the striatum—neurons from the substantia nigra project to the dorsal striatum, and neurons from the ventral tegmental area to the ventral striatum (Swanson, 1982). The firing of these neurons encodes reward information (Hikosaka, Kim, Yasuda, & Yamamoto, 2014) and the caudate nucleus is also demonstrated to be involved in 'habit learning' where one takes actions based on past rewarding experiences. Work with rats and monkeys shows that caudate lesions impair tasks that require repeated action across a specific input type (Lanciego et al., 2012).

There are many models which show that activity within the striatum correlates to models

of reinforcement learning (Kawagoe, Takikawa, & Hikosaka, 1998; Joel, Niv, & Ruppin, 2002; Cohen & Frank, 2009; Dabney et al., 2020), and a full summary of the contributions of the striatum to learning can be seen in Lim et al. (2019). One formulation of reinforcement learning is an *actor-critic* model (Konda & Tsitsiklis, 1999), where the 'actor' is the agent that must take actions, and the 'critic' predicts how rewarding each action will be. Through fMRI imaging, it is possible to see responses in the dorsal striatum which correlate to the 'actor' which makes action decisions, and activity in the ventral striatum is modulated according to the 'critic' which predicts whether a specific action will be rewarding or not (O'Doherty et al., 2004).

Another formulation of reinforcement learning is temporal difference (TD) learning (Sutton & Barto, 1998). In this framework, the agent assigns value to different states based on a reward and then at each timestep compares the actual reward received in that state to the estimated value of that state. The difference between these values is then used to update the expected reward for this state in the future. Dopamine neurons in the midbrain which project to the striatum have been shown to modulate activity based on the expected value of the upcoming action (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006). It is proposed that these neurons have an effect on long-term representations in the basal ganglia that encode how rewarding a specific state will be.

In a similar vein to the COVIS and DLS models, other proposals suggest that learning in the brain can be split into different 'modules', where circuits in the cerebellum deal with supervised learning, the cerebral cortex deals with unsupervised learning, and reinforcement learning is represented in the basal ganglia where upcoming states and actions can be evaluated, once again matching the idea that the striatum could have circuits which specifically encode reinforcement learning mechanisms (Doya, 1999).

I will not cover the entire literature on the use of reinforcement learning and the basal

ganglia and striatum in this dissertation, as I focus specifically on the use of these algorithms in the domain of speech sound learning, and the case of a second language. In summary, much work has shown that the circuits that are active during implicit training paradigms and appear to correspond to the reflexive learning system for speech are strongly implicated in these reward-based and more specifically reinforcement learning algorithms.

2.3 Usage in Speech

2.3.1 Implicit Learning and Dual Systems Theories

There are several reasons to believe that reward-based learning and reinforcement learning specifically could play a role in speech category learning in humans. While the video game paradigm has facets that lend itself to the use of reinforcement learning including actions and rewards, there is other evidence that suggests the use of a reward-based algorithm in this paradigm. In Lim et al. (2019), participants play a video game where non-speech auditory categories are presented while activity within subcortical areas is measured using fMRI. When these categories are correlated with rewarding actions in the game, participants show increased activity in the dorsal striatum and specifically the caudate head. Not only is activity observed in areas that have been determined to deal with reward processing, but the amount of activity in this region correlates directly with performance in the video game and more importantly on a post-test categorization task (Lim et al., 2014, 2019), implying that the more that striatal circuits are engaged, the better outcomes for learning. This neural region is responsible for action encoding as well as being involved in reward prediction as discussed above, however these correlations between activity and performance disappear when categories are randomized and the speech sounds do

not provide a cue that is relevant for playing the game, suggesting that the activity observed is related in particular to the category learning in the experiment.

The striatum, however, has regions that perform different functions and the Dual Learning Systems (DLS) Model proposes that each learning system as discussed above, engages different neural circuits within this region (Chandrasekaran, Yi, & Maddox, 2014; Chandrasekaran, Koslov, & Maddox, 2014). The *reflective* system which is most similar to supervised learning and is proposed to learn rule-based categories better is thought to primarily engage the anterior regions of the striatum, whereas the *reflexive* system which preferably learns information integration categories is proposed to engage the caudate head and dorsal striatum—the area in which Lim et al. (2019) observe increased activity during video game learning

The results from the SMART paradigm also indicate a role for a reinforcement-based learning for speech categories. One of the primary results of Roark, Lehet, et al. (2022) is that task-aligned behavior is necessary for the learning of speech sound categories in this paradigm. This means that listeners cannot be simply mapping visual and auditory cues by seeing a specific visual stimulus appear after a certain sound, but that they need an action to be associated with this pairing in order for learning to occur. Given that reinforcement learning and particularly the algorithm I use requires matching actions with rewards, the algorithm appears to be a good account of this paradigm.

Lim et al. (2019) describe how reward prediction could work to update speech category representations within this reinforcement paradigm and an explicit paradigm. In an explicit paradigm, the goal of a learner is to map a specific input to a specific category. After making a prediction, the category is updated based on the difference between the prediction and the actual category. The learner's attention is drawn explicitly to the auditory input that they are trying

to categorize.

However in the implicit condition, a participant is drawing connections between multiple domains—in the video game paradigm visual and auditory. Here the learner is not mapping to categories but to actions and the perceptual input must be grouped together in order to take the correct actions, but they are not assigned to a specific category. The reward prediction error now encodes the difference between the expected reward of a specific action and groups both the auditory and visual input. While the reinforcement model that I use in this dissertation is formulated slightly differently in that I have layers in the model which correspond to category information explicitly, before predicting an action, the model resembles this framework, where a reinforcement learning algorithm is learning over the difference in reward for the action and must integrate this multimodal information.

Both the neural evidence and the nature of the video game and SMART paradigms therefore implicate the usage of some reward-based algorithms in these environments. In these paradigms it is important to note that participants are not told explicitly that categories are present, nor do they receive explicit feedback on what the correct categories are after taking actions. They take actions within the environment and the auditory categories help them to take actions that lead to rewarding outcomes - ie. successfully shooting the alien in the video game paradigm or responding with a short reaction time in the SMART paradigm). The importance of reward within these paradigms suggests that algorithms where reward is emphasized could be the reason for the observed results. The involvement of reward-based learning and specifically reinforcement learning forms the motivation for the simulations shown in chapter 3 of this dissertation.

2.3.2 Other Evidence For Reinforcement Learning Use

This constrained paradigm and set of theories are not the only investigations of the possible role of reward-based learning in speech category learning. Harmon, Idemaru, and Kapatsinski (2019) demonstrate that there are particular experimental effects in speech sound learning that can be accounted for with reinforcement learning model and not with supervised learning models. For example, when presented with a cue that has previously been predictive of a certain phonetic category and is no longer informative, humans will only stop using this cue for categorization if there is another more informative cue present. When no other informative cue has been presented, one will still use the uninformative cue to attempt categorization, even though it is clear that it is no longer useful. Reinforcement learning predicts human behavior in this paradigm, whereas supervised learning does not. Relatedly, experimental effects favoring the cue-outcome order of presentation can be described with reinforcement learning, but not with supervised learning (Nixon, 2020).

Similarly statistical learning, as described above, cannot predict other phenomena in speech sound learning. Another cue-outcome relation observed in human behavior is blocking—an informative cue that has already been associated with a particular outcome will block the learning of any other cues that correspond to the same outcome. While reinforcement learning can account for these effects, supervised learning does not show the same results.

Recent work has even indicated that some forms of statistical learning itself—a learning algorithm that has long been implicated in infant phonetic learning—may engage striatal circuits (Karuza et al., 2013). Orpella, Mas-Herrero, Ripollés, Marco-Pallarés, and Diego-Balaguer (2021) show that during statistical learning tasks for speech, activity can be observed in the stri-

atal system is consistent with a Temporal Difference (TD) model of reinforcement learning. This reinforcement learning model is a model-free account of a Markov decision process where the policy that assigns values to states is updated at each time step based on the difference between the expected reward and the actual reward received. Trial-by-trial predictions using the TD model can model online learning and predict activity in the ventral and dorsal striatum, providing evidence that reinforcement learning principles could underlie speech sound learning more generally. Evidence from fMRI also demonstrates a significant increase in activity in the caudate head for non-native speech contrasts after extensive learning on this contrast (Golestani & Zatorre, 2004). While this is proposed to arise due to increased engagement of the motor system in discriminating these sounds, I suggest that this could relate to other results of caudate involvement in speech category learning.

It is important to note that in proposing the use of reinforcement learning mechanisms for second language learning, I am not proposing a return to behaviorist notions that have previously been rejected in the field of linguistics (Chomsky, 1959). I am instead proposing precise ways in which neural circuits responsible for reward encoding and prediction in the brain may play an important role in speech sound learning. The work and proposals in this dissertation are only intended to apply to phonetic category learning specifically and how humans form the mapping from continuous acoustic information to abstract categories, where there is a clear role for these kinds of learning algorithms. The concept of reward in this dissertation specifically refers to constrained situations where a participant is rewarded for taking a particular action within a paradigm. Although being able to successfully parse and categorize the incoming input is required for taking the correct action, the reward is for the action itself, not the ability to comprehend the input. I discuss how this could possibly relate more generally to language learning and

speculate how this could be generalized later in this dissertation. I do not take a claim on whether these apply to higher levels of linguistic processing or not, however, I believe this is an important area for future work

Another attractive reason to investigate the use of reinforcement in speech is the fact that it does not require explicit supervision. As discussed in Chapter 1, explicit learning paradigms are not always effective in second-language speech sound learning, and during initial native-language phonetic learning, infants are not thought to benefit from explicit information. Considering a mechanism that can induce learning without explicit supervision could shed light on how these processes could function without this kind of information. Reward information as part of reinforcement learning could come from higher levels of cognition—an idea further discussed in Chapter 6.

2.4 Summary

There is clearly a role for reinforcement learning in speech sound learning, as evidenced by the work discussed here and the connections of reward-based learning to the Dual Learning Systems model of speech sound learning. However, there are few explicit models of what computations learners could be performing in these paradigms and why these paradigms are so effective in learning speech categories of a second language. The next two chapters of this dissertation focus on elaborating on these ideas.

I first formulate a computational model of the video game paradigm and test this on two implicit learning experiments. I demonstrate that a deep Q learning algorithm can successfully capture results in these two experiments and compare performance to a supervised learning model.

Although reinforcement learning provides a better model of learning, the differences between this and a supervised model are minimal, suggesting that the differences between the implicit and explicit learning systems do not arise due to algorithmic differences. In Chapter 4 I take this idea further and propose a specific split in learning between these two systems. I extend the model to a toy scenario and provide an experiment that could shed light on this proposal.

Chapter 3: The Video Game Paradigm: A Computational Model

3.1 Introduction

I have now established that reinforcement learning plays a role in implicit learning in the video game paradigm—but why is this learning paradigm particularly effective for speech sound learning? In this chapter, I formulate a computational model of the video game paradigm and compare it to a supervised learning model. I address why a paradigm where category information is presented implicitly results in at least as good performance as one where categories are explicitly given to the participants. The inherent computational properties of a reward-based algorithm that is implemented by the dorsal striatal implicit learning system could distinguish it from the anterior dorsal and prefrontal explicit learning system, resulting in differences in the rate and quality of learning different categories when each system is activated. Alternatively, the results of the algorithm deployed by each system may be similar, and any differences in learning may arise due to the neural properties of each system, such as differential connectivity between each learning regions or that each system is able to update different parts of the acoustic-phonetic mapping.

I provide new computational evidence that bears on this question. I show explicitly that a reinforcement implementation of reward-based learning provides a good model of human behavior in the video game paradigm across two experiments - learning noise categories (Lim et

al., 2019) and speech sound categories (Lim & Holt, 2011). I compare these models with a simulation of supervised learning and show that in the noise category experiment, reinforcement learning specifically reflects human behavior better than a supervised learning model. In the speech sound learning experiment, supervised learning and reinforcement learning provide a similar match to human behavior.

This work provides the first computational account of video game training for category learning and provides clear evidence that a reinforcement mechanism is used during video game training. In addition to providing a baseline for my modeling, the supervised algorithm provides a strong resemblance to explicit category learning paradigms. While the reinforcement learning algorithm provides better improvement in speech sound discrimination and a better fit to human data, the benefit of the algorithm is minimal, suggesting that any differences in effectiveness between the two learning paradigms may not lie in the algorithm but in the neural circuitry involved. I argue that the fact that the learning signal in the video game paradigm comes in the form of a reward rather than simple category information causes reinforcement learning circuits in the dorsal striatum to be active. Using these circuits versus more prefrontal and anterior striatal circuits better enables humans to update category representations. This raises new questions about the role of reinforcement learning in phonetic learning more generally. Importantly these models also provide one of the first explicit frameworks for further studying any computational properties of implicit learning that could lead to particular learning outcomes in this paradigm.

In each of these simulations, I compare reinforcement learning to supervised learning. I use supervised learning as a comparison, not because I believe it is a particularly strong model of a specific task but because it provides a baseline to consider the computational properties of reinforcement learning. However, since the supervised model receives direct feedback about correct

categorizations, it most closely resembles an explicit learning paradigm, where participants are asked to categorize sounds and receive feedback on the correct category after every trial.

The remainder of the chapter is organized as follows. I first provide background on reinforcement learning and its applications to this paradigm, and describe the models I use in this work. In the first simulation, I implement the reinforcement learning algorithm in a video game paradigm where synthesized noise sounds are presented and demonstrate that it reflects a specific pattern in human data better than a supervised algorithm. The second simulation shows that the algorithm can overcome native language knowledge and improves in discrimination of non-native speech sounds. I discuss the implications of these results and what may give rise to the effectiveness of the video game paradigm, suggesting that the specific neural circuitry engaged during the task may contribute to the particular strength of this form of learning.

3.2 Background

3.2.1 Computational Models

3.2.1.1 Reinforcement Learning

I define the reinforcement framework formally as a Markov Decision Process where the environment is given by a set of states $S = \{s_1, s_2, s_3, ...\}$, one of which is presented to the agent at each time step t. The agent then has a set of actions $A = \{a_1, a_2, a_3, ...\}$ that it can take. At each timestep, the agent receives information about the state, chooses an action, and receives a reward r_t before the environment transitions to the next state. The transition between states can depend on both the actions of the agent and external factors of the environment. The goal of the

agent is to find a policy mapping between states and actions (s, a) that will maximize the total reward for this and future steps, by updating its parameters depending on the reward it receives.

In this chapter, I use model-free reinforcement learning, where the policy of the agent is determined by the Q-function, defined as follows.

$$Q(s_t, a_t) = r_t + \gamma * \text{max}Q(s_{t+1}, a_{t+1}) \tag{3.1}$$

This function outputs the maximum possible reward for a specific action, by summing the reward at the current timestep r_t plus the maximum possible reward from the next timestep max $Q(s_{t+1}, a_{t+1})$ multiplied by a discount parameter γ. This function is iterative since the maximum value at the next timestep will apply the same function, allowing the model to account for future reward. This mapping is not directly observable for the agent, so it must estimate the value of each action at each state by approximating this function. While there are various methods for performing this approximation, I use a specific implementation of deep Q learning (Mnih et al., 2015) to build a deep neural network that represents the Q-function (a DQN). This network takes state information as input and outputs a distribution of values over all actions the agent can take. During training, transitions between states, the action taken, and reward received are stored in memory. At each timestep, I sample a batch of transitions, calculate the error across this batch using a Smooth L1 loss function to calculate the difference between expected and actual reward, and update parameters using stochastic gradient descent.

The agent's decisions in this model follow an ϵ-greedy algorithm to allow for a trade-off between exploring the environment to discover patterns and choosing the most optimal action to receive a reward. Throughout training, the agent will choose a random action with probability ϵ

and choose the optimal outcome with probability $1 - \epsilon$, where the optimal outcome is the action with the highest value, as determined by the network. The parameter decays during training to ensure that the agent has enough time to explore the full spectrum of state-action pairings at the start of the simulation, before starting to behave more optimally.

In the experimental paradigm (Lim & Holt, 2011; Lim et al., 2019), the player receives both auditory and visual cues as they can see the location/color of the alien and hear the corresponding sound. Over time, the game speed increases and in order to have a chance at shooting the alien, the participant must rely increasingly on the auditory cues as these are presented before the alien enters the screen. In these simulations, I present the agent with one repetition of an auditory token on each timestep, as well as the current direction the spaceship is facing. Visual information identifying where the alien is situated is not presented to the model, meaning that the agent does not have grounding for the correspondence between alien locations and the auditory category. This would be like a human playing a version of the game where they can see the direction that the spaceship is facing, but cannot see the location of any aliens. In theory, this should make the task more difficult as the agent in this model cannot associate alien locations with visual information - only the auditory input. This ensures that any success from this model is due to the exploration of the reinforcement algorithm and the actions the agent takes, rather than a simple correlation between visual and auditory signals. This is important to ensure given the important role that actions play in engaging reward circuits in prior experiments

I discretize the continuous nature of the game, where the participant can take actions at any time, into timesteps where at each step the participant receives information about their location and environment and selects one action out of those available to them. Actions consist of turning left or right, or shooting an alien. The reward is derived from the increase in score by correctly

performing these actions.

The environment for the video game in these simulations is constructed with 8 different directions in which the agent can face (N, NE, E, SE, S, SW, W, NW), where a left or right action turns the agent by 45°. Aliens can enter from 4 directions (N, E, S, and W), and each of these is associated with a different category. Each timestep consists of the presentation of one auditory token and one action by the agent. If the agent is facing the opposite direction to the location of a newly presented alien, it will need a minimum of 5 actions to turn toward and successfully shoot it. To allow enough time for this to happen, each stimulus is presented 8 times. If the agent does not shoot correctly within this time, then the alien disappears without the agent receiving any reward and the next stimulus is presented. During training, there are 4 tokens within each category presented during training and the agent has 3 actions available to it: TURN LEFT, TURN RIGHT or SHOOT. The agent receives one point of reward for every alien it shoots (unlike the original study, these simulations do not have a 'capture' mechanic). Location information is presented to the model as a one-hot vector of length 8 — ie., the value for the corresponding direction where the agent is facing is equal to 1 with all other values set to 0. This is not visual information and simply instills knowledge of the direction that the agent is facing into the model. The model does not receive any visual information about the location of the alien.

The particular architecture for each model differs between the two simulations and is discussed in the corresponding sections. An algorithm for Deep-Q learning as used in this dissertation is shown below.

Algorithm 1: Deep Q Algorithm

Initialize replay memory D with capacity N

Initialize policy network Q with random weights θ

Initialize target network $\hat{Q}$ with weights $\theta^{-} = \theta$

for *episode* $= 1, M$ **do**

Gather starting location l_1 and auditory token x_1 as first state $s_1 = \{l_1, x_1\}$

for *t=1,T* **do**

Select a random action a_t with probability ϵ

otherwise select $a_t = \text{argmax}_a Q(s_t, a|\theta)$

Perform action a_t and observe reward r_t and next state $s_t = \{l_t, x_t\}$

Store transition (s_t, a_t, r_t, s_{t+1}) in D

Sample random minibatch of B transitions $(s_j, a_j, r_j, s_{j+1}$ from D

Set $y_j = \{\ r_j$, if episode ends at step j+1

$r_j + \gamma \cdot \text{max}_{a'} \hat{Q}(s_{j+1}, a'|\theta^{-})$, otherwise

Perform a gradient descent step on $(y_j - Q(s_j, a_j|\theta))^2$ with respect to network

parameters θ

end

Every C episodes clone $\hat{Q} = Q$

end

3.2.1.2 Supervised Learning

For this algorithm I use an identical structure for initial network layers, however location information is not included and instead of outputting an action, the network outputs a category directly. Throughout training, the network is presented with a batch of inputs and predicts the

corresponding category for each. These predicted categories are compared to the ground truth categories for each input, and the model updates its parameters using this difference. I use a Smooth L1 loss function to calculate the difference in prediction and true label and backpropagate this signal, updating parameters using stochastic gradient descent.

For each of these comparisons, I aim to make the amount of data presented to the reinforcement and supervised learning models as close as possible. I keep the batch size and number of presentations of each stimulus constant between the two simulations, although the exact number of presentations for each token in the reinforcement learning model will vary as it will take the agent different numbers of timesteps to face and shoot the alien. As the reinforcement algorithm outputs actions and not categories, the architecture of the two models cannot be identical, however the architecture of the initial layers of each model is kept constant.

3.3 Simulation 1 - Synthesized Sounds

The first simulation models the results of Lim et al. (2019) and demonstrates that a reinforcement algorithm is able to reflect human behavior in discrimination of auditory noise categories with simple boundaries. I simulate human experiments by training a reinforcement learning network to map between states and actions within a video game environment and a supervised model which is directly given category labels. I evaluate the models by observing how much reward the reinforcement agent receives for various categories of stimuli and how many trials the supervised network can correctly identify. A comparison between the outcomes of the behavioral results across various token types shows that human behavior is better modeled by reinforcement learning than supervised learning.

3.3.1 Methods

I use stimuli constructed following the parameters from Lim et al. (2019), with offset and onset categories created across an experimental and control condition (Figure 3.1). Offset categories are consistent across the two conditions and can be identified by just one dimension — the trajectory of change after 150ms — where onset categories are constructed differently between the two conditions. In the original experiment, onset and offset categories also differ in the type of wave—a noise carrier versus a sawtooth carrier. In the experimental condition, one must attend to both the trajectory of change during the first 150ms of the stimulus and the starting frequency in order to identify the correct token. In the control condition, onset tokens are randomized and have no identifiable category boundary. The construction of stimuli in this way yields three types of category boundaries, which I refer to as SIMPLE, COMPLEX, and INCOHERENT. Offset categories in both conditions have a SIMPLE boundary, as these can be distinguished along only one dimension. Onset categories have a COMPLEX boundary in the experimental condition due to the two-dimensional nature of the category boundary, whereas onset categories in the control condition are INCOHERENT as there is no category boundary available.

I simulate the synthesized sounds in this experiment using a Gaussian distribution centered around each frequency peak at each timestep. This means that the simulation does not differentiate the offset and onset stimuli by wave type, requiring it to discover specific patterns in the pseudo-acoustic signal. Each Gaussian distribution has a peak of 10 and a variance of 150Hz. For each of the 4 categories, 11 tokens are synthesized with 4 presented during training and the remaining 7 withheld for testing. Results for both learning types are averaged over 6 model runs. To simplify the presentation of tokens for the INCOHERENT boundary condition in this simula-

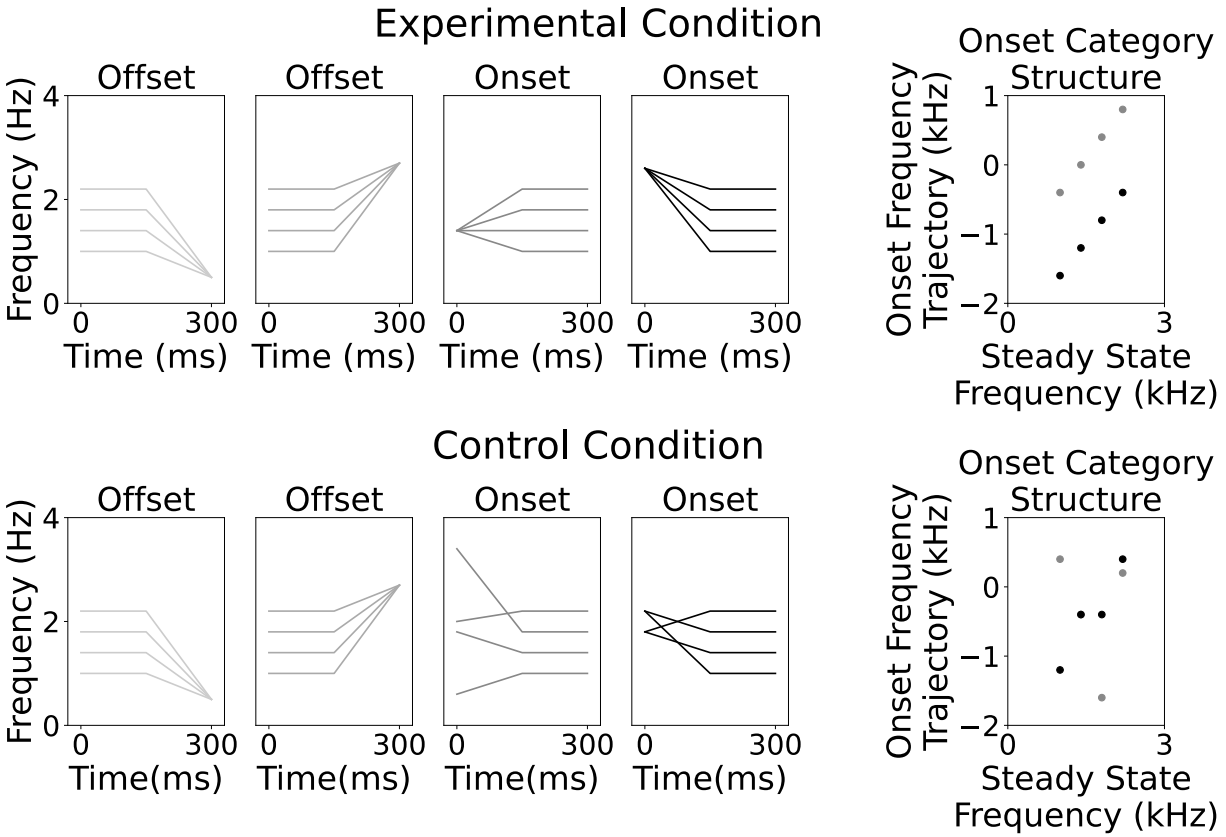


Figure 3.1: Frequency profiles (left) and onset category structure (right) for simulation 1 noise categories in the experimental (top) and control (bottom) condition. Profiles from Lim et al. (2019).

tion, 4 specific tokens are presented rather than a range of randomized tokens as in the original experiment.

The reinforcement network (Figure 3.2) combines the auditory representation with the location information and outputs a value assigned to each action using a deep Q network trained with the algorithm outlined above. The network is trained over 1000 episodes, each consisting of 128 tokens and tested on an additional 500 episodes of 176 tokens with weights frozen. Performance is defined by looking at the proportion of total possible reward the agent receives for each category type during this test phase.

The supervised network (Figure 3.3) uses solely the auditory representation and is trained to

predict the category associated with each input, where each node in the output vector corresponds to one category. It is trained similarly over 128,000 tokens (equaling the 128 tokens per 1000 episodes for reinforcement learning) and evaluated by taking the node with the highest value as the predicted category for a specific stimulus. This is taken for each of the 11 tokens and averaged over different model runs.

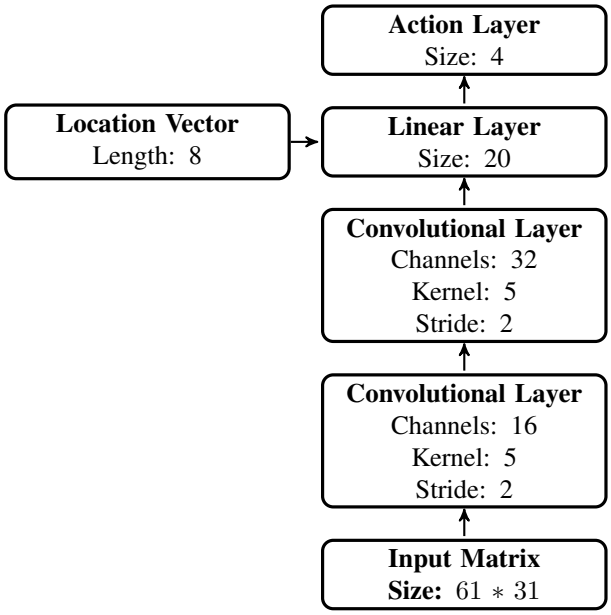


Figure 3.2: Reinforcement Learning Network Diagram for Simulation 1.

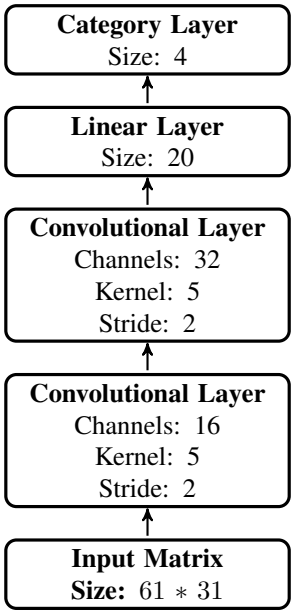


Figure 3.3: Supervised Leaning Network diagram for simulation 1.

3.3.2 Results

Results during the testing period align closely with the post-test categorization responses from the original experiment (Figure 3.4). As expected, both models perform successfully in the paradigm and exhibit better performance for tokens presented during training than novel tokens. Neural networks perform well at memorizing specific tokens during training, so novel tokens provide a better test of generalization and comparison point to the experimental results. Generally, higher performance is observed for offset tokens than onset tokens for both models, replicating

human behavior. This is because the boundary between categories has lower dimensionality in the offset condition than in the onset condition: onset tokens have a more complex boundary defined by both the frequency trajectory and the steady-state frequency (Figure 3.1). As I present 4 specific tokens for the INCOHERENT boundary, rather than a range of randomized tokens as in the original experiment, this model has a higher success rate for these tokens than the original experiments. As there is no boundary to generalize, however, I see that performance for novel tokens is low across both models - mirroring the human results.

For each model, the average values for each token type across models is treated as the mean of a distribution reflecting performance and measure how close this is to the mean of the performance distribution of human experiments using cross entropy[1]. A high number (i.e. a negative number closer to zero) indicates that the models' distribution is closer to that of the original results. These values (Table 3.1) demonstrate that a reinforcement learning model captures human data better in novel offset tokens across both conditions, whereas the supervised model captures novel token data better in the onset condition. Summing over these values, an overall cross-entropy of -9.76 for the reinforcement learning model and -10.53 for the supervised model shows that the reinforcement learning reflects human results better than supervised learning.

Where supervised and reinforcement learning models diverge is the difference in general-

[1]We use cross entropy to calculate model fit here — log likelihoods cannot be calculated as I do not have access to the original experimental data and the exact number of data points for each token. Furthermore, the parameters in these models are not fit to the experimental data, but rather are learned during a simulated learning process. However, if there were equal number of trials between the experimental data and these models, then the cross entropy values shown here would be proportional to the log likelihoods. Cross entropy is calculated with the formula: $C = \sum_i x_i * log(\hat{x_i})$, where x_i is a probability distribution representing the experimental data (the mean value presented in the original paper), $\hat{x}_i$ is the estimate of this value produced by this model, and $\sum_i$ indicates the sum over both successes and failures. I subtract 0.01 from the performance of each model to avoid undefined values by calculating $log(0)$.

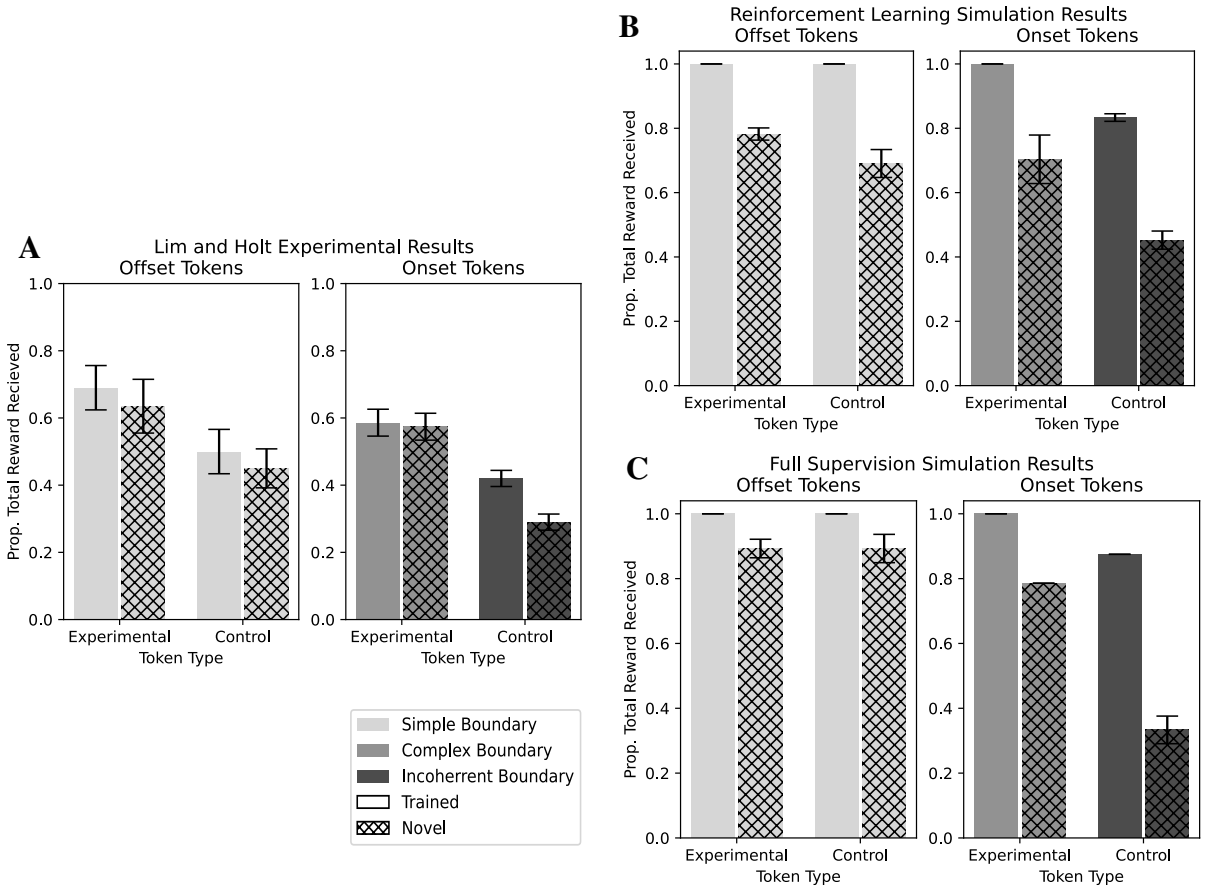


Figure 3.4: Simulation 1 Results: A) Experimental results from (Lim et al., 2019). B) Reinforcement Learning Results and C) Supervised Learning Results with 95% confidence intervals.

Table 3.1: Cross Entropy values comparing the two models to experimental data for each type of token. Reinforcement Learning captures the results better for novel off set tokens in both conditions and trained onset tokens in the control condition.

	Offset Tokens				Onset Tokens			
	Experimental		Control		Experimental		Control	
	Trained	Novel	Trained	Novel	Trained	Novel	Trained	Novel
Reinforcement Learning	-1.47	**-0.71**	-2.35	**-0.81**	-1.95	**-0.72**	**-1.10**	-0.65
Supervised Learning	-1.47	-0.88	-2.35	-1.25	-1.95	-0.79	-1.24	**-0.60**

ization for SIMPLE stimuli between the two conditions for novel tokens. Despite the fact that these tokens are the same, humans are worse at discriminating between SIMPLE stimuli in the control condition than they are at discriminating between them in the experimental condition.

This suggests that the noise provided by the incoherent tokens is enough to decrease generalization across the board and not just for the onset tokens themselves. The reinforcement learning model exhibits the same pattern of lower performance for these tokens in the control condition. This specific pattern of results is not exhibited by the supervised learning network, suggesting that a reinforcement learning algorithm may better capture human data than supervised learning. The above results are consistent across several learning rates, game types and ϵ-deacy values.

This simulation demonstrates that it is possible to gather implicit knowledge about auditory categories through a reinforcement signal, giving rise to human-like behavior and that specific patterns in the data are mirrored by a reinforcement model but not a supervised model. The reinforcement learning algorithm successfully discovers the structure of input present through the reward function and actions within the environment. Having shown that a reinforcement learning algorithm is able to use perceptual information within this task, the next step is to show that the same algorithm can overcome native language knowledge and will simulate human performance in its improvement in discrimination of non-native speech categories.

3.4 Simulation 2 - Speech Sounds

The second simulation models a learning experiment where Japanese and English speakers participate in the video game with English speech tokens presented throughout play (Lim & Holt, 2011). Participants hear tokens of [ra], [la], [da], and [ga] before each alien enters and Japanese speakers show improvement on [r]/[l] discrimination after 5 days of training without reaching native-level performance, as measured on an identification task presented before and after the training period. I once again compare a supervised learning algorithm to a reinforcement learning

algorithm in this paradigm to investigate how each algorithm overcomes native knowledge to improve on discrimination between the auditory categories presented. This simulation consists of two parts: native language training, where the knowledge of a native Japanese speaker is learned, and non-native training, where this native model is exposed to either the video game environment or supervised learning.

3.4.1 Methods

I model native language knowledge by training a supervised phoneme classifier on Japanese speech. This is a simple and efficient way to instill native language knowledge into the model. The network takes acoustic parameters as input and outputs a distribution over phonemes, indicating the identity of the phone at the center of the input window. Training data consist of auditory tokens sampled from labeled corpora as described below. The model is presented with a window of speech and a corresponding vector indicating the phone at the center of the window and is trained using stochastic gradient descent over the network's parameters. I do not mean this supervised training to be an accurate model of first language acquisition, but rather, a way to quickly create networks that have knowledge of a "native" language. I similarly train an English model to use as a native language control. In this experiment, real speech tokens are sampled from corpora - stimuli which are inherently noisier than controlled lab-prepared stimuli. Using this more naturalistic data for both training and testing demonstrates that the model can deal with variability and is not sensitive only to stimuli created specifically for this experiment.

The native network (shaded in Figure 3.5) consists of 3 convolutional and batch normalization layers followed by 1 linear layer. Data for native language training is derived from the Wall

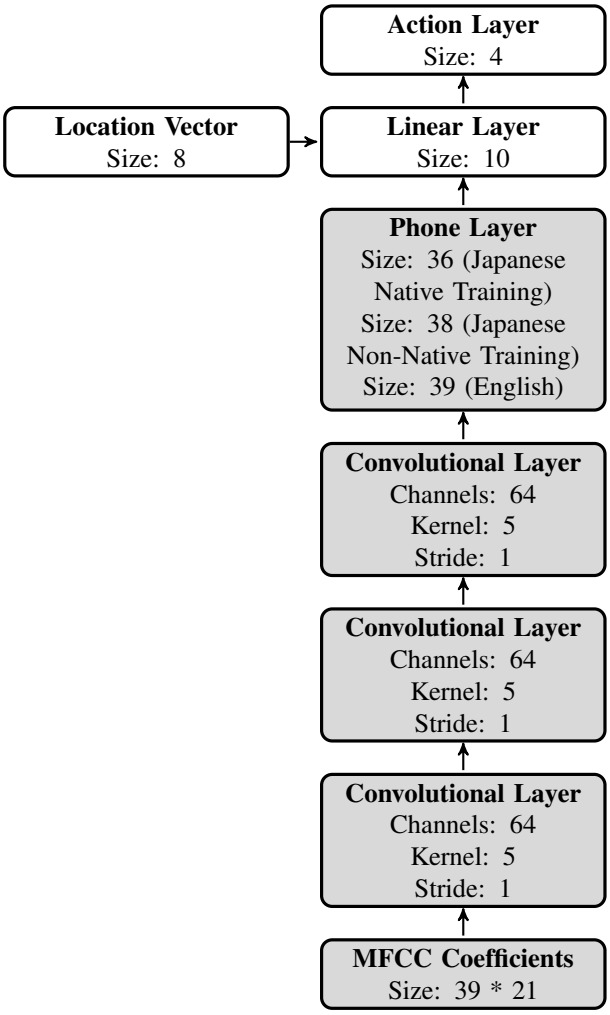


Figure 3.5: Reinforcement Learning Network diagram for simulation 2. Phoneme layer is different size for English and Japanese due to a different number of native language phonemes.

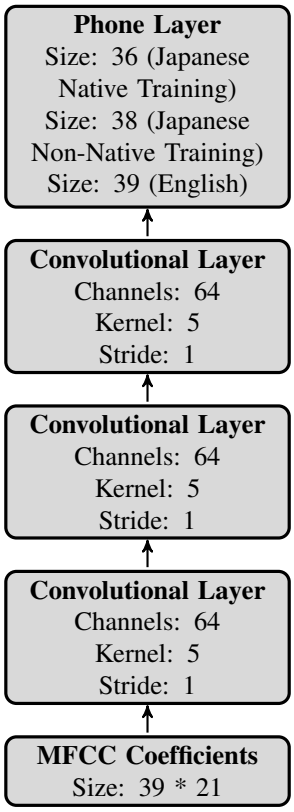


Figure 3.6: Supervised Learning Network diagram for simulation 2

Street Journal Corpus for English (Paul & Baker, 1992) and the GlobalPhone Japanese Corpus (Schultz, Vu, & Schlippe, 2013) for Japanese. The model is presented with 19.5 hours of speech and training occurs for 25 epochs.

For subsequent video game training, I sample four sounds in each category ([r], [l], [d], and [g]) from the Wall Street Journal Corpus, with each category reflecting a different location of the alien. These stimuli are those that the English native language training successfully categorizes

and are kept constant throughout the experiment. I expose the native language model to training in the video game paradigm, where phonetic information is combined with the location vector into a linear layer as in simulation 1 (unshaded in Figure 3.5) and the model once again outputs its expected value of each action. I add two new nodes to the phoneme layer to mirror the supervised model and allow the model to map new information at this layer. The training signal is allowed to backpropagate through all layers of the network and can change the initial parameters relating to the phone classification native language training. The environment and presentation of stimuli in the video game occur as described in experiment 1, over 1500 episodes consisting of 128 tokens (32 for each category). As the primary goal in this experiment is seeing the outcome of video game training when there is some native knowledge already in the system, the reinforcement learning architecture is simply attached to the output of this model at the phoneme layer.

The supervised learning models are presented with only examples of /r/ and /l/ sounds - the same tokens used in the video game training. Since the native model does not include nodes representing /r/ and /l/, I add two nodes to the phoneme layer corresponding to these two phonemes. I then continue training the model as a phoneme classifier with the same /r/ and /l/ input as in the reinforcement learning model. The model updates its parameters using the difference between its prediction of the phoneme category and the ground truth category (either /r/ or /l/). This is presented as a one-hot vector, except its length is two larger than in the native training, where the new /r/ or /l/ value is set to 1 and the other values, including those corresponding to categories learned during training, is set to 0. I aim to make the amount of training data equivalent between the reinforcement and supervised learning models. The reinforcement learning model makes a parameter update over a batch of 32 tokens every timestep. I present the supervised model with 1500 epochs of 64 datapoints - 32 from each of the /r/ and /l/ categories. I do not present /d/ and

/g/ tokens as part of the supervised model as, these phonemes are not my main focus in this study. It is true that these specific phonemes may be acoustically different in Japanese and English and mapping these different tokens to the native categories for /d/ and /g/ may not be trivial, however this is not a learning problem I aim to address, and nodes exist in the supervised model for /d/ and /g/ after the native language training.

Neural networks often require large sets of data to train, so an additional version of each model is trained, where I sample a large number of tokens during non-native training. Instead of sampling only four tokens of each category, I instead present a new token from the training portion of the Wall Street Journal corpus at each training step.

Acoustic input during all training is given as Mel Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976), which are designed to describe the overall shape of the acoustic information represented by the cochlea at any one point in time. I take 13 coefficients and first- and second-order derivatives, giving 39 total dimensions. A 200ms speech window is segmented into frames that are 25ms wide, advanced every 10ms and zero-padded, yielding a total of 21 frames.

Neural networks are known to suffer from an issue known as catastrophic forgetting, where performance on one task is greatly reduced after training on a second task. This issue is relevant for this paradigm as I first train on a native language classification task before presenting the network with the video game training—two tasks that output quite different outputs. In this work I am specifically focused on the discrimination of L2 contrasts in this work and use the native language model only as a starting point for second language learning. While I am not concerned with the continued perception of the native language, allowing the network to update all parameters in an unconstrained way would make the native language knowledge that is instilled irrelevant. For this reason, a regularization term is added during training which aims to find a solution to

the second task (the non-native sound learning) which also performs well on the native language training (Kirkpatrick et al., 2017; Aljundi, Babiloni, Elhoseiny, Rohrbach, & Tuytelaars, 2018). This term penalizes parameters as they move away from the parameters learned during native language training and is implemented for both supervised and reinforcement networks.

I simulate the [r]/[l] pre- and post-test discrimination tasks from the original experiment as well as measure performance on the native language using a machine ABX test, which is a parameter-free method of measuring the distance between model representations (Schatz et al., 2013; Schatz, 2016). I take a vector of activations for a presented token at the 'phone' layer of the model. Two tokens are taken from one category - A and X - and a third token from a different category - B. The Euclidean distance between vectors A and X, and B and X is taken to determine which distance is shorter. If X is determined to be close to A, then the trial is a success, if X is closer to B, then the trial has failed. The ABX success rate is defined as the probability of success for two tokens from these categories selected at random from the corpus. An ABX success of 1 indicates perfect discrimination, with 0.5 being chance performance. The ABX task is run over 9 million samples of [r] and [l] drawn from a portion of the WSJ and GPJ corpora withheld for testing, including 143 different speakers. Training and evaluation sets are completely independent and no token or speaker used during training is present during evaluation. For the ABX task used in these experiments, A, B and X are all sampled from the same speaker, where X is sampled from a different 'context' (ie. surrounding phonemes), than A and B. To see approximately how discriminable the different sounds are through the convolutional layers of the model, I also run an ABX discrimination task over each of these layers for both learning mechanisms.

3.4.2 Results

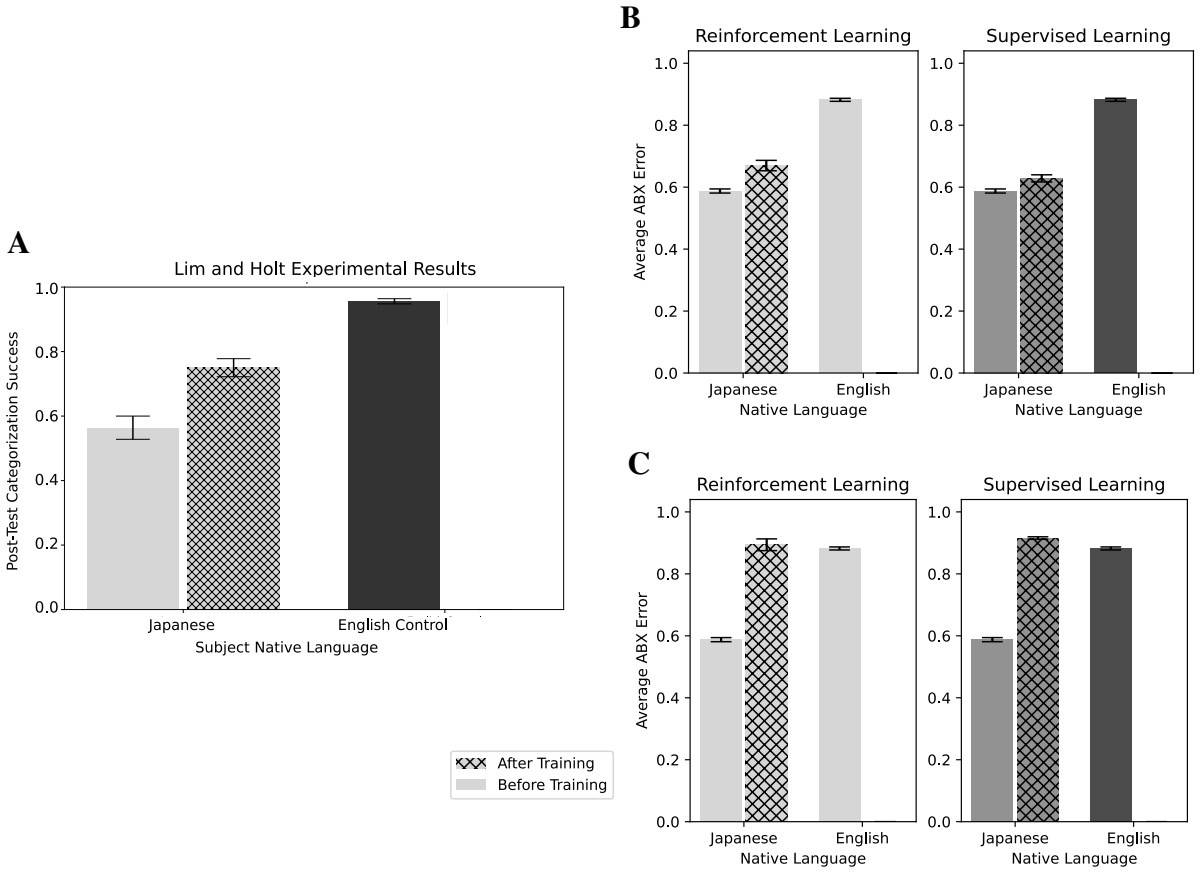


Figure 3.7: Simulation 2 Results: A) Experimental results from (Lim & Holt, 2011). B) Simulation results from run with only 4 tokens of each category presented during training. Reinforcement Learning Results (left) and Supervised Learning Results (right) with 95% confidence intervals. C) Simulation results from run with large variety of tokens from each category presented during training.

Pre- and post-training ABX success rates are shown in Figure 3.7. The Japanese native model improves in [r]/[l] discrimination ability after non-native training across both supervised ($t = 5.37$, $p = 0.002$) and reinforcement learning ($t = 8.72$, $p < 0.001$) models when presented

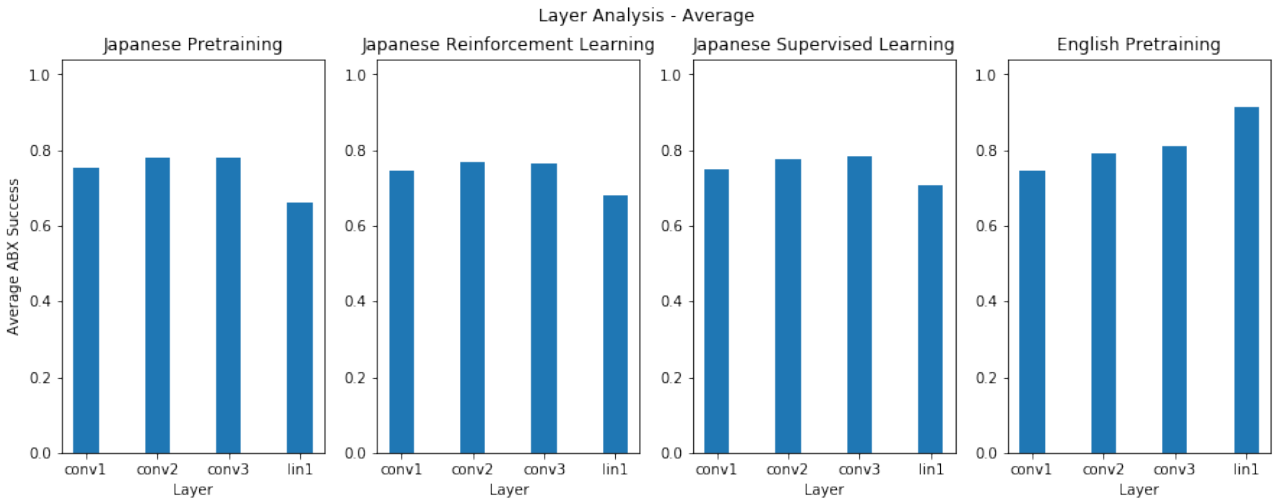


Figure 3.8: Simulation 2 Results at Each Network Layer with for model with only 4 tokens of each category averaged over all runs. Convolutional Layers denoted by 'conv' and Linear Layers denoted by 'lin'

Table 3.2: Cross Entropy values comparing the two models to experimental data.

	Native Language Training		Limited Token Video Game Training	All Tokens Video Game Training
	English	Japanese	Japanese	Japanese
Reinforcement Learning	-0.23	-0.68	**-0.58**	**-0.64**
Supervised Learning			-0.68	-0.68

with only 4 tokens. This improvement is small but significant for both models and neither performance reaches that of the native English model. This is consistent with results from Japanese native speakers who participate in the experiment and show increased [r]/[l] discrimination performance but do not reach native-level performance. I note a statistically significant ($t = 3.07$, $p = 0.014$), but very small difference between the performance of the supervised and reinforcement learning models after training (two tailed t-test), with reinforcement learning exhibiting slightly higher performance than supervised learning.

For models presented with a large number of tokens sampled from the English corpus, the size of the improvement before and after video game training is much larger for both supervised

($t = 24.01$, $p < 0.001$), and reinforcement training ($t = 70.83$, $p < 0.001$). This model does not have an experimental counterpart, but the similar performance between supervised and reinforcement learning models is consistent over this different input. There is not a significant difference between the supervised and reinforcement-trained models for this input ($t = 1.93$, $p = 0.111$). I calculate cross entropy to compare these results to experimental data (Table 3.2). Once again the results observed are consistent over various parameter settings of the network including learning rate and decay values.

Figure 3.8 shows the results of an ABX task at each of the three convolutional layers as well as the linear 'phone' layer. The reinforcement learning and supervised learning models both start from identical native training, meaning that the representations and weights before training are the same for both simulations. The ABX discrimination performance for the Japanese native model is higher for the convolutional layers than it is for the phone layer, and the opposite is true for the English native model suggesting that abstraction to phonemic categories is happening specifically at the last layer of the network. This appears to be somewhat true for the reinforcement learning and supervised training also, as the largest difference in discrimination before and after training, is also at this final layer. Despite having access to update all layers of the network, it appears that the second language training mostly changes the linear 'phone' layer.

In the reinforcement learning model, the 'phoneme' layer may not be encoding discrete phonemes after the video game training since all layers are allowed to update, and training is no longer taking place at this layer. The elastic weight consolidation may keep this layer somewhat constant throughout the video game training as this is designed to find a solution to the native language and non-native training at the same time. While in this model I am looking specifically

at how these two algorithms can learn a second language, a future consideration may be to enforce sparseness at this layer. Since the target during native language training is a one-hot vector with only one value that signals the specific phoneme, a spareness requirement for this layer may keep representations similar during training and force the new nodes to specifically encode the two new categories presented to the model.

The fact that the increase in improvement is significantly higher when the model is presented with a large variety of tokens sampled from the corpus, is consistent with evidence that High Variability training (Logan et al., 1991) can lead to more successful discrimination in second language learners.

These results show that in principle, a reinforcement signal is enough to alter speech representations and enable an agent to improve its discrimination of 'non-native' speech sounds, providing additional evidence that the video game paradigm results arise due to the use of reinforcement learning. There is a statistically significant but small difference in discrimination after training between the two models, where reinforcement learning shows better performance than supervised learning. This demonstrates that a reinforcement learning agent set within a video game paradigm can overcome native language knowledge and performs slightly better than a supervised learning algorithm. The reinforcement algorithm can use the category information implicitly provided through action and reward pairings to uncover structure in the stimuli to the same extent as an algorithm that takes explicit category information. This provides a specific account that can explain how humans are able to learn sound categories in the video game environment.

3.5 Discussion

3.5.1 Implications For Second Language Phonetic Learning

Simulations in this chapter show that a reinforcement learning network can capture human behavior in the video game implicit speech learning paradigm better than a supervised learning algorithm. This proposed model of the video game environment consists of a deep Q network—a reinforcement learning algorithm that maps an input of locations and auditory tokens to actions and updates parameters at each timestep through a reward signal. In the first simulation, I demonstrate that reinforcement learning provides a better model of specific human data than supervised learning, although both are able to learn synthesized sound categories. I provide evidence that a reward-based learning algorithm can produce results in line with human behavior and could be what is driving learning in the video game paradigm. In the second simulation I build a model of a native Japanese and English speaker before exposing the Japanese-trained model to the video game paradigm—again comparing this with supervised learning. I show that both the reinforcement and supervised networks perform similarly and reflect the improvement that humans show when learning /r/ and /l/ sounds. That the reinforcement learning model successfully captures behavioral results in both paradigms provides additional evidence towards the theory that reinforcement learning plays a role in adults' speech category acquisition. During learning, activation is shown in reward-learning systems in the striatum with a high degree of functional connectivity observed between these regions and the speech perceptual system in superior temporal gyrus (Lim et al., 2014, 2019). This model demonstrates that reinforcement learning—a specific implementation of reward-based learning—captures human data better than a supervised algorithm,

supports the idea that usage of these circuits is important in speech category learning and that corticostriatal loops may be particularly effective in updating perceptual representations.

While the model proposed in this chapter shows a statistically significant difference compared to the performance of a supervised algorithm, the effect size is minimal. Previous work has given conflicting evidence on whether implicit learning is better than explicit learning. Behaviorally, the supervised learning simulation could be considered to model a situation where an L2 learner recognizes there are two additional categories to learn and immediately matches any new information to one of those categories before receiving explicit feedback at the end of each trial. According to Lim and Holt (2011), 5 days of training in the video game paradigm appears to be more effective than 2-4 weeks of explicit training of this sort (Lively et al., 1993), although Roark, Lehet, et al. (2022) show minimal difference between the two learning paradigms with non-speech categories. The implicit learning paradigm is clearly effective however and there appear to be differences in the kind of information learned through this training and through a paradigm that engages frontal and other striatal circuits, as proposed by the Dual Learning Systems model. What causes this difference? If one considers that supervised learning could model explicit paradigms, then why do these models show very similar results and what could give rise to the differences seen in prior literature?

I suggest that differences between these paradigms must arise from differences in the neural circuitry recruited in each task. Neural imaging of the video game paradigm not only shows activation of the striatum during category learning, but participants who exhibit learning show strong functional connectivity between the striatum and the superior temporal sulcus (STS)—the primary cortical center of speech processing (Liebenthal, Binder, Spitzer, Possing, & Medler, 2005). This suggests that the basal ganglia are using reward information to update auditory

perceptual networks via corticostriatal loops. When the noise stimuli are randomized in the experimental counterpart to my first simulation (such that the aliens do not correspond to distinct auditory categories), activity in the striatum is no longer predictive of game and categorization performance and less functional connectivity is observed overall (Lim et al., 2019). This means that the connectivity is directly related to learning information about category membership of the stimuli. If this functional connectivity plays an important role in category learning, then the effectiveness of reinforcement paradigms could lie in the engagement of these specific reward circuits. When the basal ganglia are active, connections to the STS allow for effective learning; when the basal ganglia are less active, or not engaged at all, updates to the perceptual system in STS are reduced.

The neural signatures observed in initial video game experiments are also seen in other category learning paradigms (Golestani & Zatorre, 2004). These show that when learning is effective in explicit learning paradigms, greater activation is also seen in the basal ganglia. Basal ganglia activation may be particularly important for effective speech category learning in a wide range of situations. While explicit learning paradigms are able to activate these neural circuits required for effective learning, the video game by its nature and setup as a reward-based learning paradigm, may engage basal ganglial circuits more specifically and effectively. The fact that feedback in the video game paradigm comes as a reward signal, rather than simple category information, activates the reward-based circuits in this region.

However, this still leaves open the question of why updates induced by the basal ganglia are particularly effective. What is special about activating these circuits? I suggest the possibility that a frontal system may perform separate updates from the striatal system activated through reward-based learning. These circuits may have particular properties that modulate their effectiveness

for speech sound learning.

One difference between the systems could arise from more rapid plasticity induced by basal ganglial feedback circuits. Theories in visual category learning propose that different feedback loops may result in plasticity on different timescales. As one example, Seger and Miller (2010) suggest that the striatum forms rapid representations of reward prediction that can gate updates to cortex. Under this theory, striatal areas are responsible for forming the connection between specific cues and actions and frontal areas are responsible for generalization and identifying the properties that are consistent across categories. Within this framework, the effectiveness of the implicit learning paradigm would arise because—even though both striatal and frontal neural circuitry update the perceptual system in the same way—the striatal circuitry causes faster plasticity in perceptual regions.

Alternatively, this work conforms with the dual learning systems (DLS) model for speech category learning (Chandrasekaran, Koslov, & Maddox, 2014). This theory builds on two system models in vision literature and posits that a reflective system generates category boundaries by forming explicit rule-based updates and a reflexive system generates implicit associations between tokens. A neurobiological split between the systems is similar to that proposed in vision research is suggested (Ashby et al., 1998; Ashby & Maddox, 2011), where the reflective system uses frontal circuits and the reflexive system uses circuitry in the dorsal striatum. The framework is supported by both behavioral (Chandrasekaran, Yi, & Maddox, 2014), and neural studies (Feng et al., 2021).

As part of the DLS model, Chandrasekaran, Yi, and Maddox (2014) propose that speech category learning is *reflexive-optimal*, due to the fact that paradigms which engage this system are particularly effective for learning speech sound categories and that these categories tend to

require integration across different acoustic dimensions—the type of category type which this system is specialized for Interestingly, this optimality could coincide with recent work in infant category learning which indicates that learning dimensions of an auditory perceptual space could be a separate process from the formation of specific categories during infancy (Feldman et al., 2021). I speculate one possibility that the reason the reflexive system is particularly effective in humans, could be because it targets an earlier stage of the mapping to phonetic categories — ie. it targets the warping of perceptual space, from which phonetic categories can be built, as opposed to the stage which forms phonetic categories.

3.5.2 Conclusion And Future Directions

Next steps require identifying the specific range of tasks that appear to activate the corticostriatal pathways recruited in this paradigm. Harmon et al. (2019) demonstrate that reinforcement learning reflects human behavior over other algorithms in a paradigm where the down-weighting of a phonetic cue will occur only if there is a more informative cue present. This task may be another which selectively activates basal ganglia circuits. There are, however, few additional examples in which we know that the basal ganglia are recruited or that specific predictions differ between reinforcement and supervised learning.

As discussed above, I focus on learning to perceive L2 contrasts in this work, and have assumed that learning starts from native language speech perception. Since I do not model the continued perception of speech in the first language, I am agnostic with regard to the extent to which the first and second language speech systems are shared or separate. For this reason, I do not discuss the ability of the system to discriminate native language contrasts after [r]/[l] train-

ing. The regularization term ensures that the network uses native knowledge during non-native training, however both models that I study in this chapter would show catastrophic forgetting for the native language tokens. Thus, I assume I am modeling an L2 perceptual system that is simply initialized to be identical to the L1.

This framework opens avenues to investigate the flexibility of representations of speech and future work can manipulate which layers of the network can be influenced by the reinforcement learning signal, and how those weight updates occur. This model could provide a basic framework to test whether there are particular parts of the processing system that are less flexibly updated, or not updated at all, during second language learning. Further exploration of the regularization term as a way to constrain the plasticity of the system is also an interesting direction for future modeling work.

In the future, this model could also be augmented to include changing levels of motivation as part of the simulation. Varying motivation or engagement could come from varying the amount of reward that the agent receives for taking correct actions, or the precise actions for which the agent receives reward. In the potential extension to infant phonetic learning, intrinsic motivation can be incorporated into the model, where the reward function accounts for actions that explore the entire action/reward space.

One lingering question is why some types of reward seem more effective than others in updating speech sound representations and how this relates to real-world learning scenarios. If reward is particularly important for improvement in categorization performance then one might expect that motivation to learn a language and everyday reward inside a language classroom may result in the more effective learning we see in these paradigms. Many people also use mobile game-like applications to learn foreign languages, which often include a level system where

learners receive rewards such as items, or points on a leaderboard. This work raises questions about the exact role that these rewards play in second language learning. Everyday reward may not be enough to see this boost in performance and it is likely that some engagement of the basal ganglia is required in order to see effectiveness from reinforcement learning.

In experimental settings, it appears that the specific presentation of reward is important to engage the reinforcement learning system. Delaying reward or changing the exact feedback information can change the learning system that is engaged (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014), suggesting that specific conditions may have to be met in order to engage the most effective learning mechanisms. A specific correlation between reward and actions also appears to be important (Roark, Lehet, et al., 2022) for learning in implicit paradigms. While there are still unanswered questions here, I speculate that for the striatum to be used during learning, there must be a clear connection between the actions that one takes given a specific input and the reward received. This connection may not be possible outside of specific carefully curated paradigms. Motivation to learn a language could provide reward to a learner, but it is not associated with specific actions that a learner can take when hearing incoming acoustic information. Similarly in a language classroom, the correlation between action and reward is challenging to define specifically for speech category learning, and the fact that a specific action cannot be mapped to an particular input and reward, may mean that the reinforcement system cannot be engaged.

Although there is still much room for future work, the simulations in this chapter have implications for second language teaching pedagogy. Learning the sounds of a second language has always been challenging, particularly when those sounds do not match up with one's native language phonology. My work reinforces the idea that implicit learning can provide an effective

strategy for learning non-native speech sounds. Understanding why this paradigm is so effective will allow us to consider other tasks and activities that could lead to successful learning. Reinforcement learning—where improvement is motivated through rewards—is a particularly relevant topic and further simulations on the impact of rewards could potentially benefit second language education.

Chapter 4: Implications For The Dual System Model

In the previous chapter I established that reward-based neural circuits, and specifically reinforcement learning, play a role in speech sound learning in an implicit video game paradigm. Although reinforcement learning provides a good model of learning in this paradigm, there do not appear to be strong algorithmic advantages to the use of these computations over a supervised learning algorithm where explicit feedback is given. Supervised learning resembles explicit learning paradigms in many ways, however, these paradigms do not appear to be as effective as implicit paradigms in updating speech sound representations in second-language learners.

So why does this apparent difference in effectiveness between these two paradigms arise and what causes the differences in performance between implicit and explicit learning? Many studies have investigated the different learning paradigms, processes, and stimulus types that each of the learning systems performs best, but there is little work that attempts to address why such differences exist. The Dual Learning Systems model suggests that speech is reflexive optimal partly due to the fact that different category boundaries are better acquired by different learning systems and that speech categories are such that the striatal learning system is required. However there is little explanation of *why* each system preferably learns a different category type. In the previous chapter, I propose that any differences must be due to differences in the neural circuitry that is engaged during each of these tasks. This appears to be what is suggested by the DLS

model, and in this chapter, I explore this idea further and what the differences between these two systems may look like. I propose a novel split between the reflexive and reflective auditory learning systems, suggesting that each learning system may update different parts of the mapping from acoustics to phonetics. I suggest one specific way in which these systems may be split and how this idea is consistent with previous literature in this area. I then introduce a brief simulation that explores this idea further before presenting an experiment aimed at demonstrating this effect in humans. While the results of this experiment are inconclusive, it still provides an idea of how these effects could be studied in the future. The results presented in this chapter still require much development and should be considered 'vignettes' of how this proposal could play out. They provide the basis for further investigation and demonstrate how the modeling work in the previous chapter can be expanded to provide a more in-depth account of human speech category learning.

4.1 Background

In order to break down any differences between implicit and explicit learning, we must return to the Dual Learning Systems model for auditory category learning. Chandrasekaran, Yi, and Maddox (2014) propose a split learning system where a reward-based system (reflexive) that uses dorsal striatal circuits is separate from a more explicit rule-based system (reflective) which uses anterior dorsal circuits and is supported by prefrontal cortex. The reflexive system is considered to use reward-based learning given the neural circuits active during these learning paradigms, as well as the nature of the specific paradigms that engage these circuits and particular manipulations (ie. delaying the presentation of feedback), that push people towards the reflective

system.

The reflective system has been demonstrated to better learn input when categories have a 'rule-based' relationship—meaning that simple boundaries across one dimension exist between the two categories and individual tokens can be classified along one dimension. This is in line with broader two-system theories of category learning that apply to visual category learning such as the COVIS model (Ashby et al., 1998). This other system seems to perform better at learning information integration categories, where the boundary involves integrating two separate perceptual dimensions.

Chandrasekaran, Yi, and Maddox (2014) suggest that the speech is *reflexive-optimal*, meaning that speech sound categories are learned best by the reflexive system. Research with the video game paradigm shows that implicit learning can be a very effective method of learning second language speech sounds and may lead to more improvement in categorization of sounds than classic supervised learning.

It is also proposed that speech sounds are preferably learned by the reflexive system because they are complex perceptual categories that often require the integration of information across various dimensions. For example, there are many cues that can indicate whether a sound is /p/ or /b/ in English, including voice onset time, fundamental frequency, and others. If learning speech sounds requires combining information across these dimensions then these tokens should be learned better through the reflexive learning system, although it is unclear under this theory what dimensions are considered 'fundamental' that humans must integrate across to process speech sounds.

While a controlled comparison between implicit learning through the video game paradigm and explicit learning for speech has not been conducted, it is clear that there are differences

between each of these learning systems and paradigms when it comes to auditory categories. The fact that speech sound learning may be more effective when the reflexive system is engaged because it performs best at learning complex categories, does not explain *why* this system is so good at learning complex categories. The findings in the previous chapter indicate that the specific computations that could be implemented by each system are not what underlies this difference, so what explains the difference? I propose that the actual knowledge that is changed by each of these systems is fundamentally different, leading to a reflexive-optimal system for speech and a preference for rule-based category acquisition in the reflective system.

4.2 Proposal: Dual Systems Change Different Parts of the Acoustic-Phonetic Mapping

I propose that any differences between learning systems in speech-sound learning arise because they update substantially different parts of the acoustic-phonetic mapping. This explains why speech sound learning is reflexive-optimal and implicit learning paradigms are so effective for learning the sounds of a second language. The knowledge that is required to perform this mapping successfully can only be accessed by the reflexive system.

The purpose of the speech categorization system is to map incoming continuous acoustic information to a discrete abstract representation of a phoneme that can be used by higher levels of linguistic processing. This mapping is complex and could be considered to involve two separate processes which are consistent with prior results under the dual systems hypothesis. To explain this I return to the question of infant phonetic learning and work in speech sound acquisition, specifically to the idea of perceptual space learning.

Recent work in infant phonetic learning has proposed that effects that are often attributed to phonetic category knowledge in infants could arise from the learning of a perceptual space caused by statistical learning and the distribution of the input that infants receive, as opposed to the formation of discrete explicit categories. This theory suggests that infants learn the dimensions that they must attend to during the first years of life, which they then can form categories over at a later point. A discrimination peak at a category boundary and the relative weighting of perceptual cues for categorization are hallmarks of category knowledge, but discrimination ability across a dimension more generally may not require this (Feldman et al., 2021) (see Figure 1.1). This last effect could arise from learning that particular dimensions are important in the language and can be acquired in infants by accumulating statistical information about the environment without discrete category knowledge. Many second language perception studies— in particular studies comparing the effectiveness of different training paradigms—look at the outcomes of training generally, typically with categorization or identification tests. Given that it is possible for people to retain sensitivity to an irrelevant cue for categorization (Lehet & Holt, 2020), it is clear that not all perpetual knowledge is inherently linked to category knowledge. If it were, then we would expect that when learning a cue is irrelevant for categorization, one would lose sensitivity to this category. There is a need to tease apart the effects from different training paradigms that require explicit category knowledge and what could be influenced by mere perceptual space and dimension learning.

Under the ideas presented in Feldman et al. (2021), the three effects that are commonly attributed to categorization ability would fit into my proposal in the following ways. The updating of a perceptual space would give rise to the ability to discriminate between sounds across a specific dimension generally, without any differences between pairs of sounds that are within

a category or in different categories (Effect 3). This is because the perceptual space learning module would help one to learn that representing a specific dimension is important. This results in the dimension being 'expanded' in the perceptual space where every pair of sounds along the specific dimension becomes further away from each other, improving perceptual acuity and allowing for better discrimination.

In the case of second language learners, perceptual space learning may involve attending to a dimension that is not required for the discrimination of sounds in the native language. Under my proposal, learning to represent this dimension in the perceptual space when it has been collapsed previously is the same process as combining two cues along one dimension as is proposed for information integration categories. Both of these situations require the manipulation of perceptual cues and shifting attention that can happen with purely continuous input, and does not require top-down categories.

The other two effects commonly observed for category learning require updates from the category system. A peak in discrimination (Effect 1), where there is increased discrimination performance for pairs of sounds across the category boundary compared to equally space pairs of sounds within a category, either requires explicit knowledge of category membership or a clearly separated bimodal input. While this bimodal distribution does not necessarily require labels and it is possible to see warping effects with input that is clearly separated around two peaks (ie. a Self Organizing Map Model—Guenther & Gjaja, 1996) , this is only possible with input that is clearly separated. I assume this is effectively discrete category input and falls under the category system. Speech in a natural environment is typically noisy and affected by dialect, individual differences between speakers, and speaking rate and rarely follows this clear distribution pattern. In most instances, this input will not give rise to this effect unless explicit labels are given. I

also consider the case of decreased discrimination ability for within-category pairs of sounds as part the same effect, falling under the explicit category system. Similarly, changes in the relevant weighting for different cues in categorization (Effect 1) require updates from the category system. By definition, cue weights are linked to category identity and without categories to group sounds into, how much a cue contributes to category identity cannot be defined.

Since the perceptual space learned to represent a native language can make it difficult to categorize sounds in a second language, being able to update this space and learn new dimensions is incredibly important for successful categorization. This is specifically posited as a reason that non-native listeners struggle to accommodate new phonetic categories (Iverson et al., 2003). I propose that the reflexive system accesses an earlier stage of the acoustic-phonetic mapping — ie. the perceptual space, which the reflective system that draws boundaries upon the existing space cannot. This is illustrated in Figure 4.1. The implicit system is able to learn the dimension along which the acoustic space must be represented by associating groups of cells in sensory cortex with specific actions, such as turning or shooting in the video game paradigm. The reflective system updates later parts of this mapping by drawing boundaries along the dimensions which are represented in the perceptual space, forming explicit categories over these dimensions.

While this hypothesis is perhaps most easily expressed as a dichotomy between the two systems where each is only able to update one discrete aspect of the acoustic-phonetic mapping, this is most likely overly simplistic and too strict in reality. Perceptual space learning and category learning are inherently very closely linked and updating one representation likely has some effect on the other. Improving the dimensions that one can access when categorizing sounds will inherently lead to better performance on categorization and it would be very unlikely that during implicit learning one only learns to attend to a certain dimension and does not learn categories.

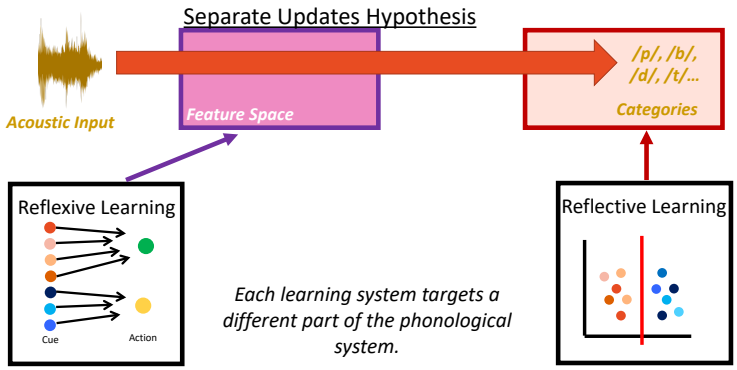


Figure 4.1: Separate Update Proposal Version 1: Different Updates - the reflexive learning system only targets the perceptual space, whereas the reflective system can only target category learning

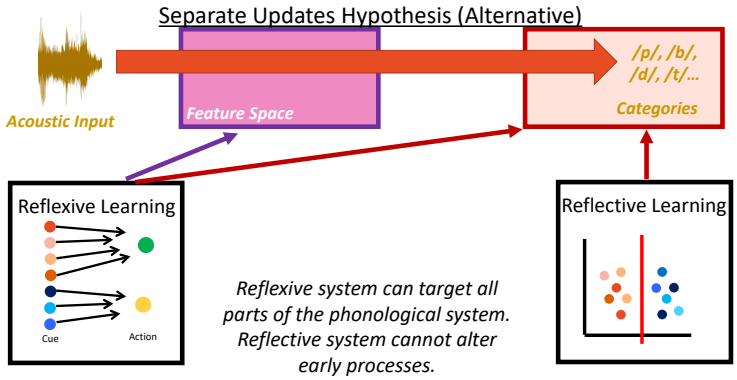


Figure 4.2: Separate Update Proposal Version 2: Different Overlapping Updates - the reflexive learning system can target all stages of the mapping, whereas the reflective system can only target later stages of the mapping

In fact, given that participants in the video game paradigm do show effective categorization after the experiment, a strict proposal is not fully consistent with experimental results. Similarly one cannot assign items to categories without creating some boundary and while Japanese native speakers learning English do struggle in discrimination of /r/ and /l/, they are not at chance after explicit learning, even if it does not involve an implicit component.

A weaker version of this hypothesis which is perhaps more likely is that access may be overlapping and both systems could be used simultaneously during all tasks, simply with more activation for the system which is preferably used given the paradigm that is presented. One can imagine that the mapping from acoustic to phonetic representations lies on a spectrum, where earlier processes must extract important features and dimensions and later processes form categories over these dimensions. An example of what this proposal could look like is given in Figure 4.2, where the implicit system could update representations at all stages of the mapping, where the explicit system can only access later 'category' stages. Given the fact that implicit paradigms clearly do allow for the learning of categories, this hypothesis would appear more likely than a binary split between the two systems.

For example, consider the neural network model used in simulations in Chapter 3. In this network, there are several convolutional layers– feature extractors that learn the patterns in the signal which are necessary to categorize phonemes—followed by a linear layer that combines these features to create predictions about what phoneme has been heard. A strict version of this hypothesis would suggest that one system could access purely the convolutional layers, whereas another could only access later layers which draw the rule-based boundary. The more lenient version of the proposal indicates that the reflexive system could make updates all the way through the network and the reflective system preferentially makes updates at the decision layer but also

has some access to make small updates to earlier layers

How this proposal integrates with previous literature on the Dual Systems model is discussed further below, but it is important to consider what the alternative hypothesis would be to a split in knowledge updated by each system. One possibility is that each learning system performs the same updates to the acoustic-phonetic mapping, but connections between the dorsal striatum and the auditory system mean that updates made through this learning system are more effective in some way. An illustration of what this would look like in comparison to my proposal is shown in Figure 4.3. Perhaps some properties of this neural circuitry mean that any changes made when the striatum is active are more effective, but ultimately each system has access to the same information in the perceptual system. However, this would not account for the fact that different types of categories appear to be preferentially learned by each system. One would expect to see much more similar outcomes to implicit and explicit learning paradigms which only vary on the level of improvement, rather than distinct learning patterns which differ between the two systems. It could be proposed that since information integration categories tend to be more challenging to learn, one requires a more effective system or more engagement to learn them, but this does not fit with the experimental data that demonstrates clear differences between the systems.

Another possibility is that the effects seen in previous studies are due to motivational or engagement effects. Implicit learning paradigms are inherently more engaging due to their 'gamified' nature, in particular the video game paradigm which first opened up this work in second language learners. While this should be considered when comparing these paradigms and could play a significant role in overall performance, this once again does not account for differences in preferential category types between the two systems.

My proposal is one of the first attempts to explain *why* there are differences in category

learning between an implicit and explicit learning system. It is a start towards a larger study of differences between these two systems and requires more work to show that this is in fact a difference between explicit and implicit paradigms. In the sections that follow, I show that the proposal is consistent with most prior literature, before providing a simulation that illustrates what this theory would look like. I then present an experiment that attempts to tease apart these differences. The following sections should be viewed as an initial foray into this theory and its implications as well as demonstrating the role that modeling—particularly the framework developed in the previous chapter—can play in the investigation of this effect.

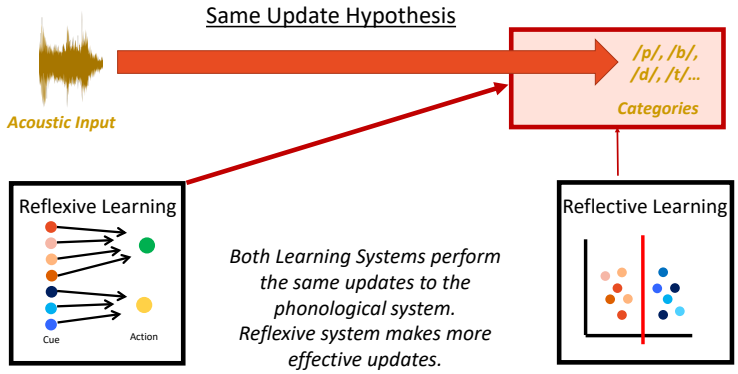


Figure 4.3: Alternative Hypothesis: Same Update - each system performs the exact same updates to the mapping

4.3 Application to /r/-/l/ Learning

As a specific example, we can imagine a native Japanese speaker, whose perceptual space is specialized for Japanese from birth. The dimension of F3 does not need to be learned or represented for consonants as it is not involved in any categorization and the distribution of sounds along this dimension is not important for distinguishing sounds in this language. This means that once a listener has learned their native language, they do not need to represent F3 for liq-

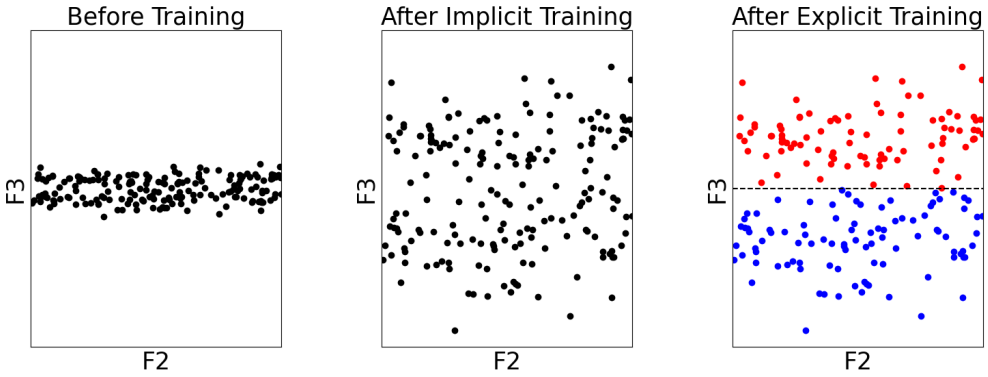
uid sounds and it becomes collapsed, leading to decreased ability in discrimination. When this speaker tries to discriminate between /r/ and /l/ in English this will be challenging as there is not a perceptual dimension along which these sounds differ. (Figure 4.4a, left)

While top-down information could potentially help to improve the categorization by drawing 'rule-based' boundaries across the space, without learning that F3 is a dimension that must be attended to, discrimination will always be difficult as the sounds along this dimension will remain close together within the perceptual space. The listener is attempting to draw boundaries over a space that does not represent the dimension necessary to sort sounds into these categories (Figure 4.4a, right).

However, if this Japanese speaker is able to reshape their perceptual space and learn that F3 is a dimension to which they must attend, the task of drawing boundaries across this space becomes much easier. My proposal suggests this happens through the reflexive system, engaging corticostriatal loops. This dimension becomes expanded allowing for better discrimination between pairs of sounds that vary only by F3. However, they do not yet have discrete categories for these sounds. At this stage, we would expect to see increased discrimination generally over the F3 dimension (effect 3), but would not yet observe a peak in discrimination at the category boundary (effect 2), nor a steep identification curve which represents a high relative F3 cue weight (Effect 1, Figure 4.4b, left $\rightarrow$ center).

It may be argued that the participant does not need to learn to represent F3 to discriminate between liquid sounds if they already use this cue to identify other sounds (ie. vowels), which this proposal cannot account for. However, given that listeners can use F3 for discrimination of vowels but not /r/-/l/ discrimination, there are clear differences between the representation of F3 for different sounds and this is an issue for other theories of /r/-/l/ discrimination and not

(a) F3 and F2 representation for native Japanese speakers before training and after explicit training without implicit training which alters perceptual space.



(b) F3 and F2 representation for native Japanese speakers before training and after explicit training when implicit training occurs first.

Figure 4.4: Before training, Japanese native speakers have collapsed the F3 dimension for liquids as it is not required for discrimination. If explicit training operates over this space (top) then categorization is challenging. Ideal categorization is shown with blue and red corresponding to different categories, however due to the compressed nature of this dimension it is difficult to achieve this category split and listeners may attempt to categorize using F2 as a more expanded dimension. If implicit training occurs first (below) it can expand this dimension by learning the perceptual space, allowing for better discrimination and then easier categorization when explicit training occurs following.

just this proposal. Under my proposal, the cue which represents F3 for liquids may also encode other acoustic information that distinguishes liquids from other sounds. For example, the cue that encodes F3 for vowels may jointly encode acoustic properties of vowels and therefore cannot be used to distinguish between /r/ and /l/. The new cue learned through an implicit paradigm will allow this since F3 will be encoded alongside acoustic properties of liquid sounds.

The learned perceptual space then makes it easier to form categories, and the reflective system that is able to express clear rules over these dimensions will be able to draw a boundary over this learned F3 dimension. This allows for better explicit categorization, and we would expect to see the discrimination peak at the category boundary (effect 2), as well as a steep identification curve—both effects that are associated with category knowledge (Figure 4.4b, right)

This example illustrates an ordering which becomes a necessary part of this proposal, where the reflective system is more effective when it follows learning by the reflexive system. This is the case because the outcome of perceptual space learning is an earlier, less abstract stage of the acoustic-phonetic mapping processes and is required in order to form category information. This ordering makes a specific prediction that if participants are presented with both implicit and explicit training paradigms, learning should be better when the implicit paradigm precedes explicit training.

4.4 Integration with Prior Literature

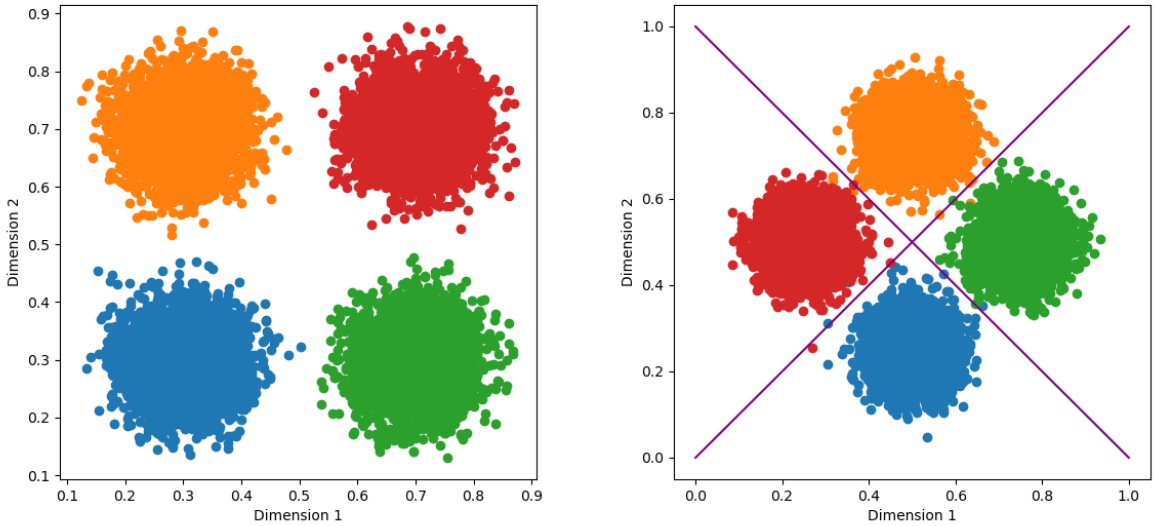
The theory presented here is consistent in several ways with prior neural, behavioral and modeling work that has been discussed so far in this dissertation. There are three specific ideas in the literature which are substantiated by this proposal—the fact that each system responds better

to a particular category type, the fact that speech is reflexive optimal, and recent work in infant phonetic learning.

First, this proposal accounts for the differences in category types that are better learned by each of the dual systems. The rule-based categories that are better learned by the reflective system only require drawing a boundary across already existing dimensions. If the reflective system can only access later-stage categorization processes, then this is not problematic as the formation of an explicit boundary along these dimensions only requires this simple declarative rule. Information-integration categories, however require the combination of multiple dimensions in order to create category boundaries. The reflexive system under this proposal which can update the dimensions that are required for categorization is able to learn a perceptual space that the reflective system can then draw explicit boundaries across. This dimension learning could consist of picking out the dimensions which are necessary for categorization.

To demonstrate this I illustrate an acoustic space (Figure 4.5), where the x and y axes are two fundamental dimensions. In Figure 4.5a it is possible to draw simple boundaries along each of these dimensions without any integration. In the second example where the categories would be considered to require integration of dimensions, this is no longer possible. However if one was able to learn dimensions represented by the purple lines displayed in Figure 4.5b, then the space would be transformed and this would become the same classification problem as the rule-based categories. The proposal I make is that perceptual space-learning processes of the reflexive system are able to induce learning of new dimensions, which allows the reflective processes to then form the explicit categories.

Secondly, this theory is consistent with the fact that speech sound learning appears to be *reflexive-optimal* in adults and provides an account that does not rely solely on the fact that



(a) Rule-Based

(b) Information Integration

Figure 4.5: Rule-based and information integration category types, where learning particular dimensions for information integration category types (purple lines, right), allow for information integration categories to become rule-based categories.

different categories are learned better through each paradigm. This theory specifically explains why these different categories may be learning more effectively through different paradigms. While supervised learning in adults can help with learning to some extent, it appears that it may be more effective to learn speech sounds through an implicit paradigm such as the video game paradigm.

As discussed previously, when someone learns a second language, they start with a perceptual space (ie. learned dimensions) that may not be optimal for learning categories in the new language. If the perceptual space starts from the point of their first language, then they need to update their space to accommodate the new input, as well as form categories over these perceptual representations. If explicit learning focuses on forming rules and specific boundaries between categories, then this will not be able to perform successful learning over bad dimensions. Per-

haps any effectiveness of implicit learning is because it helps to update these dimensions. This explains why the implicit learning observed in the video game paradigm appears to be so effective.

There is a more subtle prediction that this theory makes about second language learning using these different approaches. If a second language listener is learning to discriminate between sounds that exist across a dimension that is in their native language, then implicit learning may not have as much of an advantage. An example may be if there are 3 distinct categories that exist along a VOT continuum in the second language compared to only two in the native language. One may expect that if a space that accommodates these new sounds does not need to be learned, then there may be little difference between implicit and explicit methods. While this idea needs to be considered in much more detail, this could be why the advantage of implicit learning is not always observed (Roark & Holt, 2018b).

Thirdly, this theory bridges the gap between some recent developments in infant phonetic learning and second-language speech sound acquisition. Perceptual space learning has more recently been considered a separate process that occurs in infant phonetic learning, but the idea of this as a distinct learning task has not been fully incorporated into second language learning theories. It is not unreasonable to expect that the learning of a perceptual space and the learning of important dimensions may be a factor in second-language phonetic learning as well.

Finally, this theory makes more specific some of the predictions made by the Speech Learning Model (SLM) and the Perceptual Assimilation Model (PAM), discussed in Chapter 1. These models suggest that until there is evidence otherwise, second language phonetic input is perceived as belonging to the closest native category. The proposal I make is not consistent with this exact argument, however, it instead provides a similar explanation that second language speech sounds

are perceived through a perceptual space that is suited to the native language. Once this perceptual space is updated, categories can be more easily formed for the second language. These ideas are closely related and likely have similar outcomes, and future work can look more carefully at the relation between this proposal and these earlier theories.

Having described how this proposal fits with prior literature, I now present a simulation and an experiment that aim to provide some evidence for this hypothesis.

4.5 Simulation 3 - Information Integration vs Rule-Based Categories

The simulations in Chapter 3 demonstrate that there are few algorithmic differences between a reinforcement learning and a supervised learning algorithm for learning speech sound categories. However I do not specifically tease apart rule-based and information integration categories in these simulations. In line with the proposal I put forward, I demonstrate that the algorithms are not inherently better at learning one kind of stimuli or another. I provide a toy model which is presented with either information integration or rule-based categories showing that—at least in this framework—there are no differences in learning between the two algorithms.

I take inspiration from Roark, Plaut, and Holt (2022) in generating a network that uses population coding to take input and either predict categories or actions within a video game paradigm as described above. Input is sampled across two dimensions, where categories are either arranged with category boundaries that rely on only one dimension or require the integration of the two dimensions as shown in Figure 4.5. There are 10 input nodes to the model for each of the two dimensions, which encode the dimension through a Gaussian distribution with a mean at the sampled value. For example, a value of x is sampled from one of the 4 categories, and

then a Gaussian distribution is generated with a mean of the chosen value of x and evaluated at 10 points across the continuum. The same is repeated for y. This results in a population code with each input node corresponding to a fixed value across each dimension, mirroring population codes in human sensory cortex.

Each model has two linear layers, the second of length 4—the number of categories in the simulation. Since this is an easy task, I run two versions of the model with the first layer of size 10 and size 6 to see if reducing the number of nodes the network has access to changes the findings. For the supervised model each of these nodes represents a category and the model is trained to output a value of 1 for the category present and a value of 0 for all other categories. The reinforcement learning model is set up as in the previous chapter and represents a video game paradigm where an agent turns and needs to 'shoot' an alien that enters from a direction that corresponds to the category being presented. The 'category' layer is combined with a location vector of length 8 as in the, before outputting one of three actions that the agent can take. In this model, the 'category' layer in reinforcement learning is not enforced or pushed towards discrete category representations since there is no pretraining.

Each model is trained on 2000 tokens of each category type sampled at random and results are averaged over 20 runs of the model. For testing the weights of each model are frozen and an additional 200 categories are sampled at random from each of the categories. The proportion of tokens that the model successfully identifies (for the supervised learning model) or 'shoots' within the paradigm (for the reinforcement learning model) is taken as a measure of performance

Results are shown in Figure 4.6. We can see that there is no difference between the learning of each type of category across the two models. *Both* models better learn the rule-based categories where the boundary is defined over only one dimension than the information integration

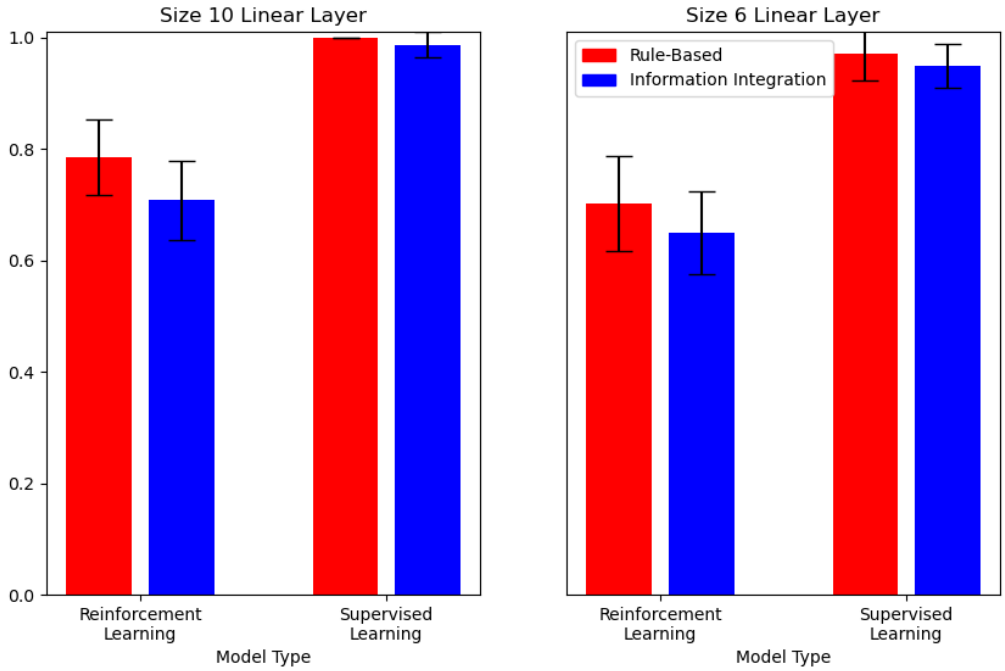


Figure 4.6: Proportion of correct categories during test for a reinforcement learning and supervised learning model using rule-based and information integration categories as input

categories which require two dimensions to categorize. This demonstrates that there is not a split in which category type is preferentially learned by each model. Under the Dual Learning Systems theory, rule-based categories are better learned by the reflective system (supervised learning here) and information integration categories by the reflexive system (reinforcement learning here). If algorithmic differences were to give rise to these effects, then we would expect that reinforcement learning would perform better on information integration categories than rule-based categories. Instead, rule-based categories are learned better across both models. The reinforcement learning model has lower performance generally, likely because this is an easy task for supervised learning to acquire.

This is simply a toy model and it is possible that other models of reward-based and explicit learning could show differences in how well these categories are learned. This does not dis-

pute the Dual System Model, but it illustrates that the computational properties of the algorithm implemented by each system are not behind the effects observed in behavior. According to the theory I propose where each learning system has access to update different parts of the perceptual mapping, it is consistent that each learning system does not inherently learn one category type better than another. Instead the parts of the mapping that each system can access lead to different results for each in humans

4.6 Experiment 1 - Comparison Between Implicit and Explicit Learning

This proposal makes specific predictions about what happens when only one part of the acoustic-phonetic mapping is updated. What I lay out in this dissertation indicates that there are two separate processes that follow each other when mapping from acoustic input to abstract phonetic categories. Early stages of this mapping encode an acoustic space that performs a transformation on the incoming acoustic input before later stages deal with the mapping to discrete phonetic categories.

If early stages of the mapping are not allowed to update, then we would expect to see decreased performance in the learning of information integration categories, as well as lower performance on contrasts which require the updating of a perceptual space, such as speech categories which differ across a dimension that has not previously been learned. Alternatively, if later stages of the mapping are not allowed to update, then we would expect decreased learning on rule-based category types, as well as discrete identification tasks.

The work in this dissertation has so far focused on using computational models to simulate learning in different environments. It would be possible to explore this proposal further with

these models, specifically hindering the model's ability to update early transformations that could mirror perceptual space learning or later transformations which could mirror discrete category learning. For example, in the models presented above, one could consider convolutional layers which implement feature extraction to simulate perceptual space learning and linear layers at the end of the model to implement discrete category mapping. By freezing the weights in each of these layers one can simulate what different learning paradigms under this proposal might look like and is one example of how the computational models presented in the previous chapter can be used to extend our knowledge of the Dual Systems Model.

At this point in this dissertation, however, I move away from computational modelling work to instead look for experimental data which could provide evidence for my proposed theory. In this section, I will present the results from a preliminary behavioral experiment aimed at teasing apart perceptual update differences in implicit and explicit learning. Although the results will turn out not to be conclusive, they provide useful insights for future work.

Previous experiments that have compared explicit and implicit learning have primarily focused on the tokens that are best learned through each paradigm. These results align with the theory I propose as outlined above, however prior experimental work is not enough to show that the procedural and declarative learning systems map to different stages of the acoustic-phonetic mapping. For that reason, I conduct an experiment that aims to tease apart any differences in perceptual updates between the two learning systems. I propose a specific split between the information that each system can update, and while this experiment is designed with that theory in mind, it is also has the goal of looking more generally at any differences between the knowledge altered in different paradigms. I explore whether the systems differ purely in the effectiveness of their updates to the perceptual system, or whether the nature of the knowledge that they update

is different. In other words, are the striatal/reflexive and frontal/reflective learning systems proposed by the dual systems model of speech category learning performing the same updates to the acoustic-phonetic mapping of speech sound categories?

To study this question, I look at the /b/-/p/ contrast in English, which has VOT as a primary perceptual cue and F0 as a secondary perceptual cue, and examine a situation where listeners must learn to use F0 as the primary cue to that existing contrast in their language. I measure how within-experiment discrimination and identification using this new cue differ as a function of whether participants learned the cue in an implicit or explicit paradigm. I present a controlled comparison between implicit learning and explicit learning, using a version of the SMART paradigm as a basis for implicit learning and an explicit paradigm that is as closely aligned with the implicit paradigm as possible.

For results consistent with my proposal, I would expect each of these paradigms to give rise to different behavioral patterns depending on whether a particular effect depends on category knowledge or general dimension learning. I predict that the perceptual effects which correspond to category learning—a change in cue weighting and a discrimination peak at the category boundary—will be exhibited more strongly by participants who learn in an explicit paradigm. Moreover, participants that learn the F0 cue for /b/-/p/ discrimination in an implicit paradigm are more likely to show the behavioral effect that could be attributed to perceptual space learning—namely an overall increased sensitivity to sounds along the F0 continuum. These results would indicate that each system is performing different updates to the acoustic-phonetic mapping

Furthermore, this experiment provides an explicit comparison of implicit and explicit learning for speech sounds. While the video game paradigm is clearly very effective in second-language speech sound learning, a controlled comparison between implicit and explicit learning

for speech has not been conducted. To my knowledge, the only controlled comparison was restricted solely to non-speech sounds (Roark & Holt, 2018b), and did not show any differences between the two learning paradigms. However these results may be due to the specific experimental design, where participants were presented with a learning block where the correlation between sounds and categories was randomized before the final identification task which could obscure any learning differences. It is clear that video game learning is particularly effective, and a controlled comparison with speech sounds is helpful to confirm these findings.

4.6.1 Methods

4.6.1.1 Experimental Design

A total of 38 native English speakers were recruited for the experiment consisting mostly of students from the University of Maryland and were randomly assigned to either an implicit or explicit training condition. All participants gave informed written consent and all experimental procedures were approved by the University of Maryland Institutional Review Board. Participants were asked to list their native language(s) and any other languages they knew and their experience with them. This information was used to exclude any participants who knew languages that included the cue used in the training stimuli. A total of 3 participants were excluded for not completing the experiment and 2 were excluded for incorrect responses in the identification task, leaving 17 participants in the implicit condition and 16 participants in the explicit condition. All participants participated in two training phases, the first consisting of 4 blocks and the second consisting of 2 blocks with a discrimination and identification task at the start of the experiment and after each training block.

Both training paradigms consisted of four rectangles presented on the screen, each of which corresponded to a different key on the keyboard (‘u’, ‘i’, ‘o’, and ‘p’). Participants in the explicit training group were presented with a 5 auditory tokens sampled from one category and asked to determine which rectangle the sound corresponded to by pressing the corresponding key. After pressing the key, the participants saw and ‘X’ presented in the center of the rectangle that corresponded to the correct answer. The implicit paradigm closely resembles the SMART paradigm (Gabay et al., 2015), where participants similarly heard 5 tokens, before a small image of an ‘alien’ appeared in one of the rectangles on screen. Participants were told that this was a reaction time task and they had to press the button corresponding to the box in which the alien appeared as quickly as possible after seeing it. After each training block in the implicit condition, participants were told their average reaction time during the block and how many times they pressed the correct key.

While the first four training blocks in each condition corresponded to the respective training condition (implicit training versus explicit training), the final two blocks in each condition both consisted of explicit training. The reason to present explicit training after implicit training in that condition was to investigate the effects of one training type followed by the other. In the proposal laid out above, implicit training alters an earlier stage of acoustic-phonetic mapping than explicit learning, suggesting that following implicit training with explicit training would result in the highest overall improvement in categorization performance on perceptual tasks. The exact block design of the experiment is demonstrated in Figure 4.7 and an example of each trial type is given in Figure 4.8.

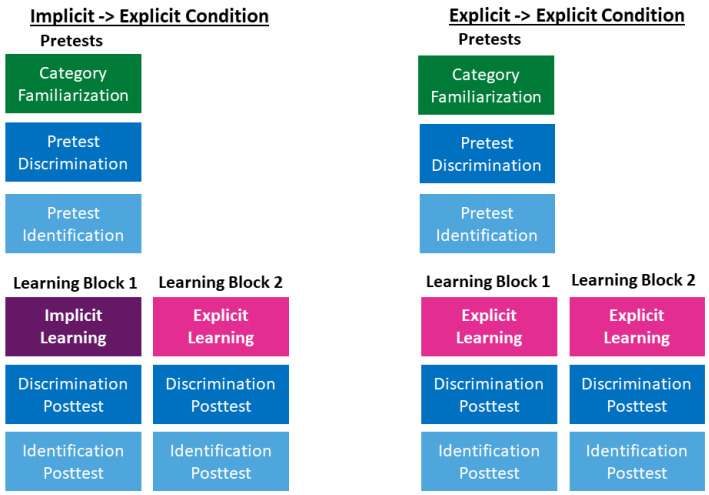


Figure 4.7: Overview of the experiment for participants in each condition

4.6.1.2 Stimuli

The critical stimuli presented during training consisted of /b/ and /p/ stimuli across both a VOT and F0 varied across a spectrum, using tokens from a previous experiment (Wu & Holt, 2022), illustrated with the words 'beer' and 'pier'. F0 is a secondary cue for identifying the voicing of stop consonants in English, however it is not commonly used for categorization except in cases where there is noise in the acoustic input.

This contrast was chosen as I expected that it would give a baseline discrimination performance in native English speakers that is not as chance but is also not at ceiling, allowing me to see if discrimination improvement increases or decreases with each of the training paradigms. In addition, an /s/-/ʃ continuum demonstrated over the words 'Sue' and shoe' was used as a filler to give 4 options during training and identification testing.

During discrimination testing, pairs of sounds with a difference of 20 Hz in F0 were presented with a neutral VOT in the center of the VOT spectrum of 15 ms. This allowed for the

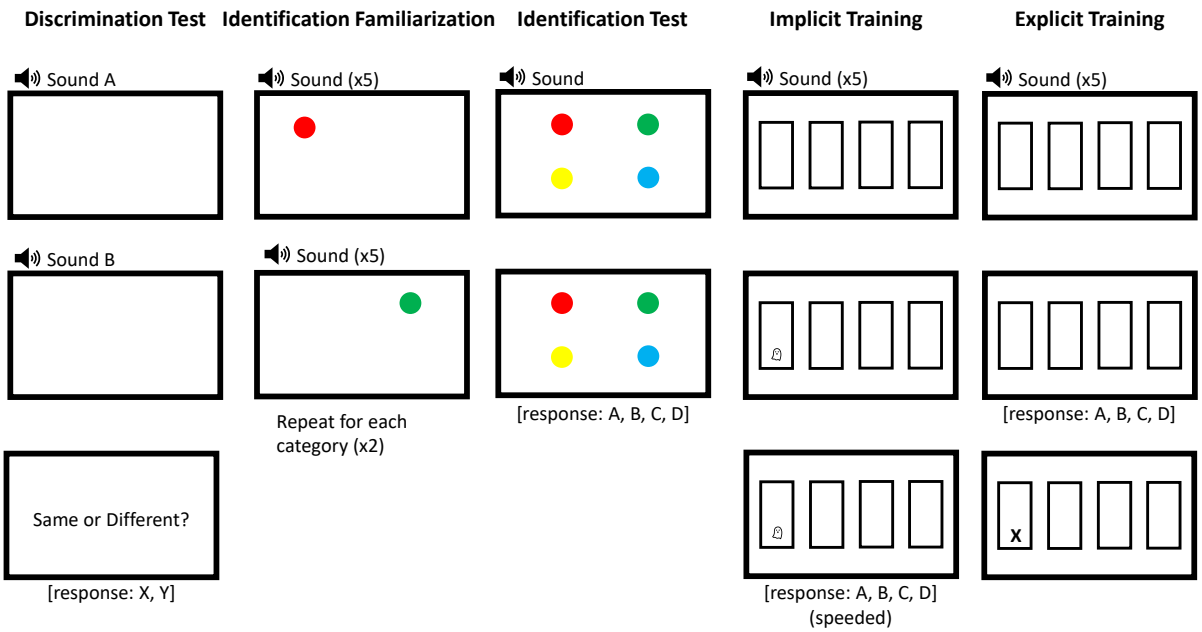


Figure 4.8: Layout of the different components of the experimental paradigm.
Discrimination Test: Sound A and Sound B are presented followed by a prompt asking the participant to select whether the sounds were the same or different.
Identification Familiarization: Participants hear 5 examples of each category presented alongside a colored shape. All stimuli are looped through twice to familiarize participants with the identity of each category
Identification Test: All four colored circles are presented on the screen and the participant hears a sound and must choose the category that it belongs to.
Implicit Training: Four rectangles illustrating each region (each with a corresponding key) are displayed on screen and participants hear 5 examples of a specific category. Following this, a stimulus is displayed in one of the rectangles. Participants must react as fast as they can with a button press corresponding to the location of the stimulus
Explicit Training: Four rectangles are displayed as in implicit learning, however after the sound presentation, the participant must make a decision of the category that the sound corresponds to. After this, the participant sees the location that does correspond to the sound category (and will infer if they were correct or incorrect.)

evaluation of F0 discrimination with an ambiguous value of VOT. For identification testing, a spectrum of sounds were presented with 4 possible values of F0 and VOT (for a total of 16 tokens). This gave a spectrum of responses across both continua to measure the relative cue weighting of participants. (see Figure 4.9)

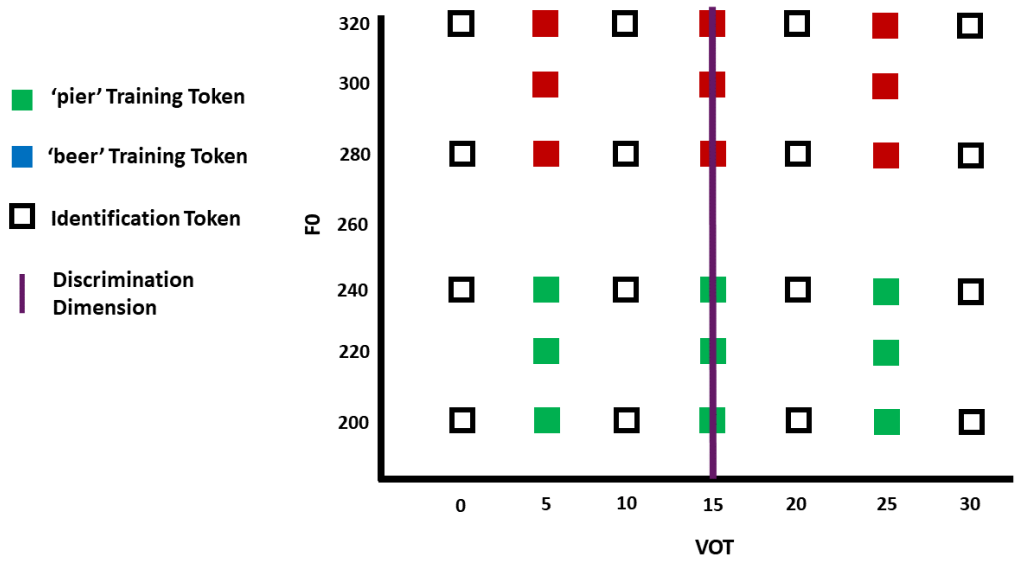


Figure 4.9: Stimuli used in the experiment. During training, tokens in each category are sampled across a range of VOT for either high or low F0. For identification tasks, in order to determine cue weighting, tokens with black boxes are presented, comprising of the full spectrum of VOT and F0. For discrimination, equal spaces tokens are presented along the specified dimension.

4.6.1.3 Perceptual Measures

Two tasks were used to measure the learning of the cues used in the experiment. A discrimination task presented two sounds along the F0 continuum with a neutral VOT—either two sounds with a 20 Hz difference in frequency or the same sound twice. Participants were asked to answer whether they heard the same sound or different sounds.

For the identification task participants heard one token at a time and were asked to sort

tokens across a variety of F0 and VOT into categories. In order to not bias the participants towards specific representations by giving labels of /pier/ and /beer/, each token was associated with a particular colored circle. Before each identification task, participants were familiarized with samples of sounds that corresponded to each category (the mapping remained consistent throughout the experiment) and during testing the participants were asked to press to assign the token to that category. One colored circle was associated with low F0 and one with high F0. with the other two associated with endpoints of the /s/-/ʃ/ continuum. Tokens presented during familiarization were sampled from the limits of the F0 spectrum. This task was used to calculate categorization curves as well as measure cue weighting.

4.6.1.4 Predictions

In general, this experiment has been designed such that discrimination results would change in different directions depending on whether a learning system is performing perceptual space or dimension learning versus category learning. I expect that if participants in the implicit condition are performing dimension or space learning generally, as compared to category learning, then the sensitivity to changes within categories may not decrease to the extent that it does when learning categories. This is because learning a perceptual space under this definition means learning to represent a particular dimension, but not warping to change this distance between different tokens along that dimension. However, if a participant is learning categories, then differentiation between sounds that are on the same side of the category boundary is not required for the task and discrimination of these sounds will decrease. In other words, I would look for a discrimination peak forming in the explicit condition, but more general attention to the F0 dimension with no

discrimination peak in the implicit condition.

However for measures of identification, and particularly cue re-weighting, I would expect to see the performance move in the opposite direction. Increasing discrimination across the F0 dimension and learning that this dimension is important for stimuli in this experiment, would inherently lead to better categorization performance generally, however, cue-weighting specifically is a factor of category learning (Lim & Holt, 2011; Lehet & Holt, 2020). If implicit learning is only updating just the perceptual space then participants in this condition should not improve in explicit identification as much as those in the explicit paradigm who update category representations.

The measures in this experiment are very closely linked and it is possible that because these are so closely linked even if the proposal laid out in this chapter is true, this experiment could show no difference between the two conditions. This still is a first step in attempting to tease apart differences in each system in a way that has not yet been attempted.

4.6.2 Results

4.6.2.1 Within versus Across Category Discrimination

For discrimination measures, I calculate a d' as a measure of sensitivity to differences in stimuli at each point along the F0 continuum. For each pair of 'different' F0 trials that differ by 20 Hz, I calculate d' by combining the 'same' trials for each of the two F0 values. These are plotted in figures at the midpoint between the two tokens. This calculation is performed across the F0 continuum after each training block for each condition. In order to see any potential changes caused by learning, I calculate the difference in d' between the pre- and mid-test and the pre- and

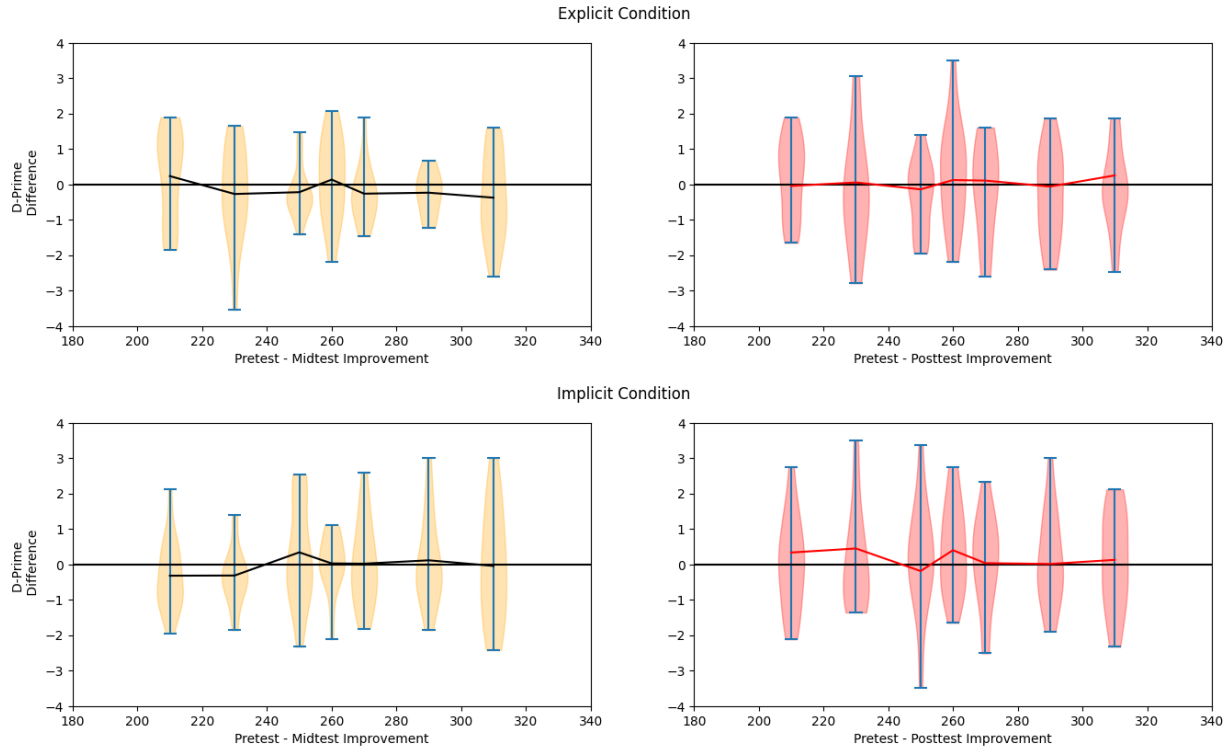


Figure 4.10: Difference in sensitivity across the F0 continuum between the pre-test and after training block 1 (Yellow-) and difference after training block 2 (Red—Explicit learning for both conditions). Violin plot shows overall profile across participants with average of the solid line

post-test across the F0 continuum. This is plotted in Figure 4.10.

T-tests are calculated for three points in the spectrum to see whether there are any significant differences in discrimination performance before and after training. While the figures show that there appears to be an improvement in discrimination, particularly at the category boundary, it is clear that this data is noisy and there are no significant differences in d' across the F0 continuum after each training block.

4.6.2.2 Cue Weighting

For the identification task, all responses to stimuli on the 'beer'/'pier' continuum where the participant pressed one of the corresponding circles were included—ie. any responses where the

participant sorted these sounds into one of the ‘sue’/‘shoe’ categories were excluded (in cases where participants responded with a ‘sue’/‘shoe’ category to more than three sounds on this continuum, the participant was omitted entirely from the analysis). From the remaining data, cue weights are calculated for each participant by performing a logistic regression where the F0 and VOT value of a category are used to predict whether a participant will respond that the category. The weighting of each dimension in the logistic regression is then normalized to equal 1, showing the relative weight of each cue for a specific participant.

Raw responses averaged across all participants for each of the 16 tokens presented during this task are shown in Figure 4.11 We can see that across learning, participants start out with a slightly more diagonal categorization boundary, before relying more on F0 after each training block.

Cue weights for each participant are shown in Figure 4.12. Again we can see a slight reweighting of perceptual cues towards F0 after each training block. Once again improvement is not significant after each training block, although there may be signs of subtle differences between the two conditions. It appears that cue weighting increases more after the first block for the explicit condition than implicit condition (Table 4.1).

4.6.2.3 Identification Performance

Finally, we can plot simple identification curves at each point during training. For each value of F0 across the spectrum, I average across all VOT presentations to plot the identification curve across this dimension for all participants. The results of this are shown in Figure 4.13. I also report average identification performance across participants (Table 4.1)

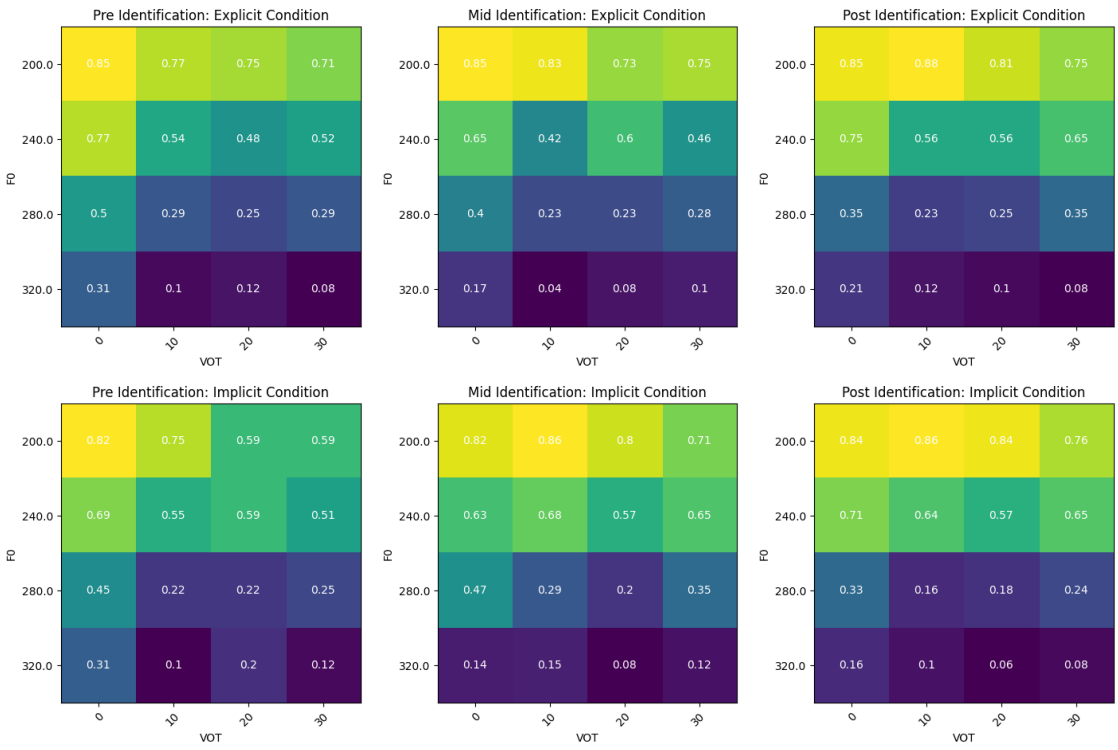


Figure 4.11: Proportion of /b/ responses across the VOT and F0 Continuum at each point in training for explicit (top) and implicit (bottom) conditions.

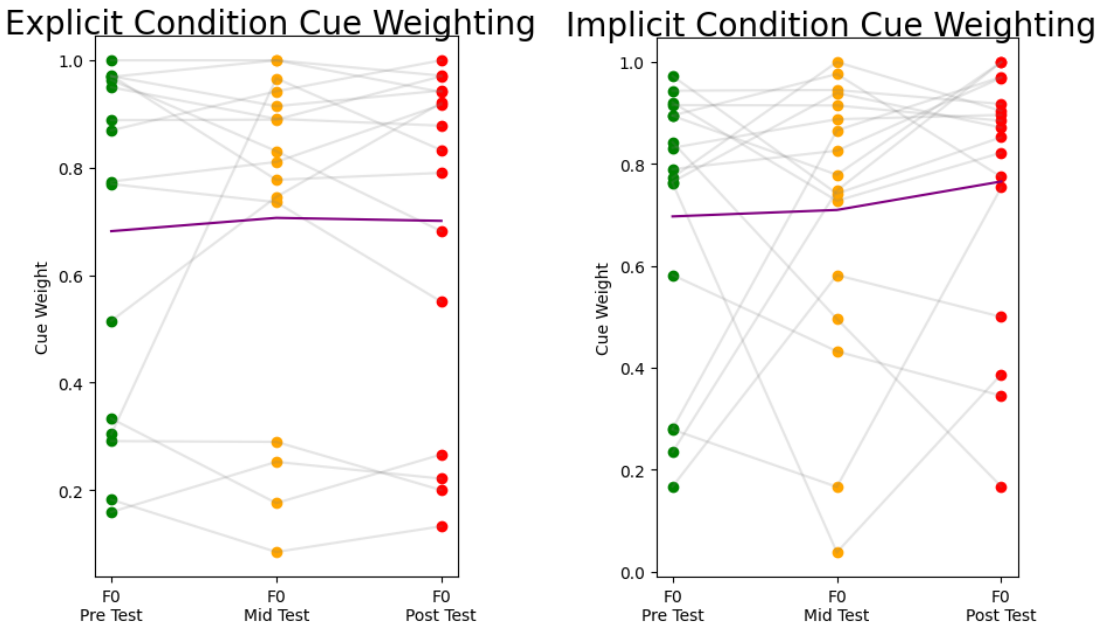


Figure 4.12: F0 Cue weights at each point in training—average shown by solid purple line with each gray line representing one participant. Higher cue weight means participant relies more heavily on F0 for /p/-/b/ identification during testing.

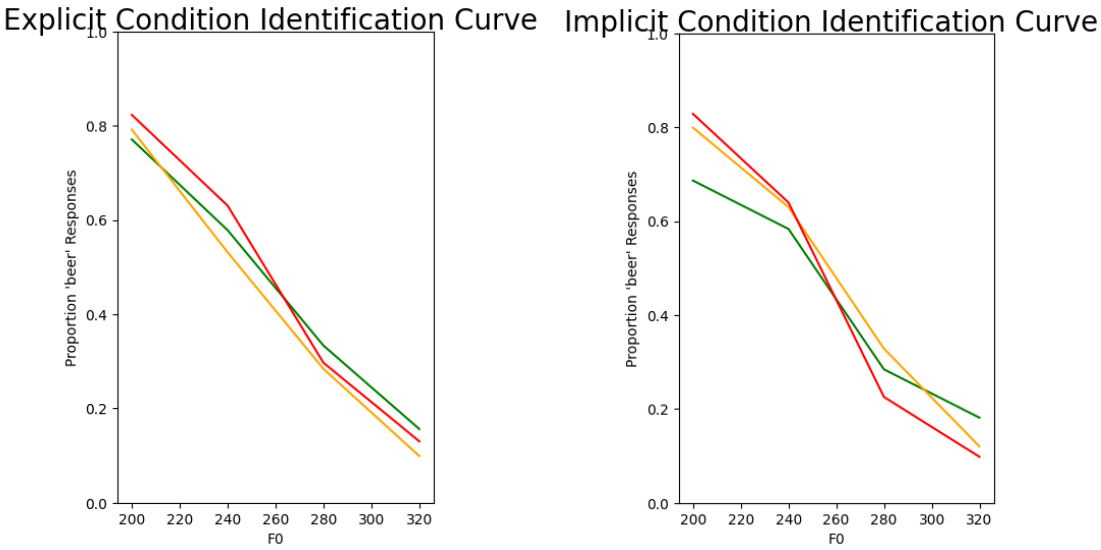


Figure 4.13: Identification curves for before training (Green) after block 1 (Yellow) and after block 2 (Red). Y axis represent proportion of /b/ responses averaged across participants.

	Implicit Condition	
	Average F0 Cue Weight	Categorization Performance
Pre Test	0.70	0.70
Mid Test	0.71 [p(df=16) = 0.87)	0.75 [p(df=16) = 0.18]
Post Test	0.77 [p(df=16) = 0.31)	0.79 [p(df=16) = 0.12]

	Explicit Condition	
	Average F0 Cue Weight	Categorization Perfomance
Pre Test	0.68	0.71
Mid Test	0.71 [p(df=15) = 0.63]	0.74 [p(df=15) = 0.37]
Post Test	0.70 [p(df=15) = 0.82]	0.76 *[p(df=15) = 0.04]

Table 4.1: Identification performance and Cue Weights after each condition. Significance values are presented as calculated for a paired t-test between the specific identification task and the previous identification task

4.7 Discussion

The results of this experiment are inconclusive. While it is possible to see small signatures of learning generally, particularly when looking at identification performance and cue-reweighting, the discrimination results are very noisy and it is not possible to make strong conclusions from these data. The hypothesis behind this experiment is that different learning systems update different parts of the perceptual system. This predicts that learning profiles will differ between each of the conditions, however none of these measures differ significantly between the implicit and explicit conditions. There are two small differences that could hint that these effects may come out with additional participants and a more developed experimental design. First, there appears to be more average improvement for discrimination performance at the category boundary for the first block of the explicit condition compared to the first block of the implicit condition. According to my proposal, this would indicate that participants in the implicit condition could be learning that they should attend to this dimension, but are not developing the acquired dissimilarity which comes with category knowledge. Secondly, it appears that any improvement in cue weighting for the implicit condition comes in the second block of training, where they are in the explicit paradigm. Again, since cue weights are inherently associated with categories, this could indicate that category knowledge is forming more in this block than the implicit block.

Despite these subtle differences, given the low improvement generally and non-significant discrimination improvement, few conclusions can be drawn on the proposal specifically. The number of participants and power are simply too small to show effects. The lack of effects particularly in discrimination could be because the task is not at a good level for these participants. Discrimination even before training is high for some participants and could be considered at ceil-

ing, since they show a d-prime even before training greater than 2 (a d-prime of 2 would indicate that the distributions are 2 standard deviations away from each other). Varying the discrimination trials such that some tokens are closer together could provide a better measure of discrimination performance across this dimension.

In speech perception studies it can be challenging to gather clean data, and in the real world learning of phonetic cues is messy. This experiment was particularly difficult given that it required an extensive training regime, meaning that participants may lose engagement through testing and training. The participants heard the same categories repeatedly for an hour and fatigue effects could strongly influence the data, particularly in the second block. I likely need many more participants in order to get enough data to see any clear differences or similarities between groups. Recent work has had much success running perception studies online (Obasih, Luthra, Dick, & Holt, 2022), which may be an avenue to get more data for this experiment, in order to increase the power of the results.

Even with additional data there are other reasons that this experiment may not give the expected results given my proposal. In looking at the average sensitivity to discrimination, we see that on average, even before training participants do show reasonable sensitivity along the F0 dimension. It may be the case that this task was not challenging enough and some participants were at ceiling before training, meaning that they could not improve after training. Similar results are seen when looking at cue weighting, where some participants already are relying almost entirely on F0 for categorization before training. Other non-speech knowledge could also be playing a role here and someone with musical training for example may show more sensitivity to F0 from the beginning.

It is also possible that there is no difference between implicit and explicit learning here

because the F0 cue already exists as a secondary cue in English. While the proposal I make in this chapter would suggest that discrimination performance would increase more with implicit learning due to its effect on the perceptual space, it is possible that due to its nature as a secondary cue for voicing in English, and role in other aspects of speech and audition more generally, there is not space for more attention to this cue and the dimension is already learned as good as it can be. In this situation, the reflexive system would not have any additional effect than the reflective system given that the perceptual space is already fully formed in a way that allows learning of this cue.

This fact itself does make any future studies of this proposal rather challenging. In order to study any differences between these two learning systems one would be best placed to look for an effect where performance moves in different directions depending on the system which is active—the precise reason for choosing the F0 contrast for voicing in this experiment as I expected that participants would start above chance on this contrast, but not with entirely accurate performance. Discrimination for some participants in this experiment may be at ceiling and they do not have room to show improvement, particularly since the cue used here is a secondary cue for an existing contrast. This more generally is a useful consideration in other studies which do not show a difference between implicit and explicit learning. If these two learning paradigms do indeed update different knowledge in the perceptual system, then it may be the case that when no difference between the paradigms is seen, the knowledge that the implicit learning helps with has already been learned,

This idea would actually be consistent with previous findings showing that the video game is particularly effective for improving Japanese native speakers' discrimination, but that implicit learning does not always show a significant benefit, even for information integration categories

generally. For native Japanese speakers, F3 is not a relevant cue, and therefore implicit learning can play a role in helping to shape this perceptual space that explicit learning cannot. However when the cue that is learned during training involves dimension(s) that the participant already knows and attends to (ie. F0 in the case of this experiment), then I would predict no difference between the two training techniques. Future experiments that contrast these two techniques could account for this and consider what participants' previous knowledge is of the cues that are being trained.

Overall while the results of this experiment are inconclusive and do not provide direct evidence for the theory proposed in this chapter, this experiment provides a good framework for the type of experimental evidence which would shed light on this proposal and a step towards further work which can investigate the specific updates that the different learning systems make to the perceptual system.

4.8 Summary

This chapter has integrated the previous literature on a dual systems approach to category learning with simulations in the previous chapter to posit a theory where each system can preferentially update different parts of the perceptual system. While this is only an initial theory, I lay down a simulation that illustrates how this could play out and how modeling could help elucidate how these systems differ, followed by an experiment which, while the results are not concrete, is the start to teasing these differences apart experimentally. Distinguishing why these systems differ in the categories that they best learn is an important step in developing dual learning systems theories of auditory category learning.

These are just the beginnings of research trajectories that can develop our knowledge of how these learning systems function and work presented in this chapter is only a start. The simulation which compares different stimulus types in this chapter only deals with simple dimensions and should be expanded to more realistic auditory categories with real audio input as the basis for the model. This could confirm whether there are any differences between the categories that a model such as this could deal with.

Additionally, the neural network model with convolutional layers as presented here is only one type of model that could deal with incoming auditory input. Other models such as Gaussian Mixture Models (Schatz et al., 2021) have had success in modeling infant phonetic learning and perhaps could be combined with reinforcement learning components to test how the perceptual space operates across a variety of computational representations.

In addition to the experimental and modeling work presented here, there are other avenues for considering this paradigm. One place to look is to the specific neural connectivity between each of the regions that deal with category learning and cortex. For example, it is known that there are relatively few connections directly between the striatum and auditory cortex (Yeterian & Pandya, 1998), instead, the densest connections are between the caudate and putamen and auditory belt region. One way to investigate the specific updates that each learning system is able to make may be to look at the specific connections that each region of the striatum and prefrontal cortex has to the auditory system and whether there could be different connections to earlier or later stages of the auditory stream.

Alternatively, if these results do not work out and there is not a split in the knowledge that implicit and explicit systems can update, then alternative explanations are clearly necessary. An explanation for the effectiveness of implicit paradigms for second language speech category

learning that relies on a distinction between category types is helpful but does not fully explain why such learning circuits are more effective for these category types. It is clear that differences exist and it is important that an explanation of these differences is grounded in cognitive neuroscience.

One aspect of these paradigms that should continue to be considered is participant engagement and motivation. Anecdotal evidence from this experiment showed that participants had better overall experiences when performing implicit learning over explicit learning, most likely because this kind of learning be in a more 'gamified' format was more entertaining. While there are many clear effects in the literature which can be attributed to implicit versus explicit learning of speech sounds and more than just participant motivation, this effect should not be discounted when comparing explicit and implicit paradigms.

This hypothesis has strong implications for second language speech sound learning specifically. One of the main examples of auditory category learning as an adult is learning speech sounds of another language, and this proposal ties in the recent results which are more general to the auditory system with data from second language acquisition literature and the video game paradigm specifically. Understanding the systems and paradigms for auditory category learning more generally, allows us to build stronger theories of how humans could learn the speech sounds of a second language.

Chapter 5: Prediction During Online Processing

So far I have demonstrated that listeners struggle to form phonetic representations of a second language and discussed what effect learning techniques may exist to acquire these categories. In this chapter, I will address how second language learners leverage prediction during online speech processing by studying neural activity observed while listeners attend to an audiobook in a continuous listening environment and discuss how this might be possible with the phonetic representations they have.

While listening to speech it is well established that listeners are constantly predicting upcoming concepts, words and sounds (Ferreira & Chantavarin, 2018). Recent work has shown that in native speakers, expectations of upcoming input are generated based on various context levels that listeners can integrate while listening to speech (Brodbeck et al., 2022). These predictions are formed using context at the phoneme level (information about what phonemes the listener has heard), word level (information about the current word that this listener is hearing) and sentence level (what previous words the listener has heard).

Given that second language learners do not have the same phonetic representations as native speakers and it is challenging for these listeners to ever reach native-like ability, one may expect that they do not have the same processes or do not employ the same information in predicting upcoming phonemic information. Alternatively, one might expect that they rely more on these

predictions since there exists a less robust representation of incoming input. I first explore the prior literature in prediction in a first and second language before introducing a new MEG corpus of second-language naturalistic listening data. I analyze these data for second language learners and compare them with the responses of native speakers, showing that there are fewer differences between the responses from these two groups than one might expect. I then present some initial ideas about how this could tie in with the phonemic representations of second-language listeners and what future work could be done to bring together ideas of prediction and usage of phonetic information during online speech processing with theories of prediction in second-language learners.

5.1 Background

5.1.1 Prediction

When hearing speech, listeners constantly anticipate upcoming concepts, words and sounds, and these predictions are proposed to be an essential part of linguistic processing (Ferreira & Chantavarin, 2018). However, there are various pieces of information in the linguistic hierarchy that humans could integrate over when they make these predictions——this includes using transition probabilities between phonemes to predict what phoneme will come next, using part of the current word to restrict the cohort of phonemes that could come next, or even using previous words and sentence information to predict the next word. Two primary theories have emerged that describe how the brain may integrate these different levels of contextual information.

A local model of processing proposes that listeners have access to information at a particular level of information to predict information at that context level. For example, speakers can

use information about transition probabilities of phonemes to predict upcoming phonemes but cannot integrate higher-level linguistic knowledge with these predictions, such as lexical or sentence information. Evidence from a visual world paradigm and cross-modal priming (Gaston & Marantz, 2018; Gagnepain, Henson, & Davis, 2012) provide evidence that listeners will activate predictions of upcoming words, even if they are not valid in the current syntactic or semantic context. If humans can activate predictions of words that do not fit with these contexts, then they may not be able to integrate higher-level information when forming predictions—consistent with a local context model.

Alternatively, listeners may be able to form one unified global representation of context and use this to predict upcoming material. In the case of phoneme prediction, one would integrate lexical- and sentence-level information to inform their prediction of upcoming phonemic information. While this model of prediction conflicts with some of the experimental results above, it is consistent with a Bayesian account of speech processing (Norris & McQueen, 2008; Jurafsky, 1996) where one should use as much information as is available to form a prior on incoming input to increase the accuracy of identifying the incoming signal.

Recent work has shown that both of these models are potentially at play during speech processing and that local and global information is integrated simultaneously in different neural regions. Using continuous neural data while participants listen to an audiobook, Brodbeck et al. (2022) demonstrate that predictions using sublexical, lexical, and sentence-level context can all be decoded from neural data, with each prediction model exhibiting a distinct neural signature and uniquely describing a portion of the signal. This suggests that the brain constructs predictions from its statistical knowledge at various levels of representation. Illustrations of the three context levels and how they fit into local and global prediction models are shown in Figure 5.1.

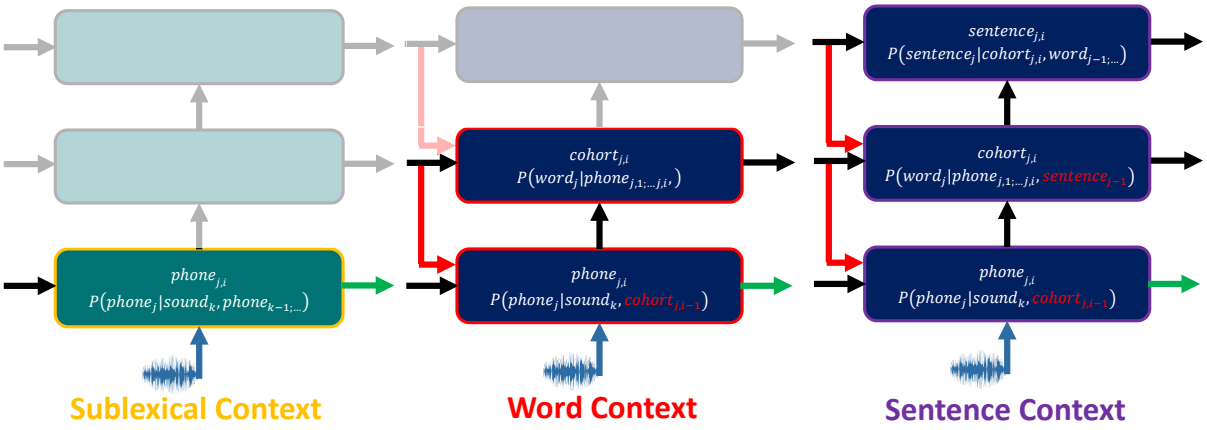


Figure 5.1: Integration of different levels of linguistic context during online prediction: The sublexical context model corresponds to a local model of prediction where the only information accessible when predicting upcoming phonemes is the string of phonemes that have just been heard. The word and sentence context models correspond to a global model of prediction, where higher-level linguistic information—statistical knowledge about the current word and sentence—can be used to restrict the prediction of upcoming phonemes. For the models in this chapter, surprisal and phoneme entropy are calculated over the distributions in the bottom box in each column (information theoretic values for the next phoneme, shown by the green arrow). Cohort entropy is only valid for word and sentence context models and is the entropy for the current word—a calculation over the probability in the middle box for each column.

5.1.2 Non-Native Speech Prediction

While prediction is an established mechanism in native listeners and this evidence suggests integration of different context levels, the mechanisms of prediction in second-language learners are much less clear and on first glance second-language listeners appear to exhibit less prediction than native speakers. Various theories have been posited as to why second language learners do not appear to give behavioral and neural responses which indicate predictive processes in the same experiments that native speakers do. It has previously been suggested that non-native speakers may not represent syntactic information in the same way as native speakers (ie. the Shallow Structure Hypothesis - Clahsen and Felser (2006a, 2006b, 2006c)). More recently however, a strong case has been made that second-language learners do not predict upcoming material to

the same extent as native speakers, rather than exhibiting entirely different syntactic representations. There are three specific areas that have demonstrated differences between native and second language speakers: grammatical gender processing, event-related neural measures and gap-filler contexts.

One of the strongest pieces of evidence against the use of prediction in second-language learners is the use of grammatical gender of a determiner in restricting predictions of upcoming nouns. Generally, second-language learners do not appear to use gender information to guide predictions, in contrast to native speakers (Lew-Williams & Fernald, 2010). The ability to use grammatical gender in prediction appears to be related directly to the knowledge of gender, demonstrated as participants rely less on this as a cue when they are presented with consistent incorrect gender in an experiment (Hopp, 2016). This applies to second language learners in similar ways, with more proficient learners showing the most similar effects to native speakers (Dussias, Kroff, Tamargo, & Gerfen, 2013).

Neural event-related responses (McLaughlin, Osterhout, & Kim, 2004; McLaughlin et al., 2010) also demonstrate a difference in prediction and processing between native and non-native speakers. Various studies have demonstrated differences in ERP responses to syntactic and semantic violations between native and non-native speakers. Hahne and Friederici (2001) show that the typical early anterior negativity and P600 responses to syntactically incongruent sentences exhibited by native speakers do not appear in second-language learners. Additionally, the typical response to a syntactically correct sentence appears to be higher, which could exhibit more effort of information integration in second language learners. Other studies show that while the response to a semantic violation in L2 speakers may be similar to native speakers, the response may be smaller and slightly delayed to native speakers (Hahne, 2001). This effect is also mod-

ulated by proficiency — after intense instruction in a second language, L2 speakers can show a P600 response for grammatical features that are not in their native language and the amplitude of this response is modulated by proficiency scores.

A final effect that demonstrates differences in processing for L2 speakers is filler-gap contexts where long-distance dependencies occur due to a non-canonical order of syntactic constituents. When this situation arises, there is a gap in the sentence that needs to be filled. L1 speakers show rapid reflexes when encountering these filler-gap contexts and immediately attempt to fill the gap and revise this prediction with any new information. L2 speakers may not fill the gap as quickly as L1 speakers, and may not use the same abstract syntactic information in restricting what could fill this gap during online processing. (Omaki & Schulz, 2011)

However why do second language learners not show the same markers of prediction to native speakers? Is this because there are mechanistic differences in the processing of second-language speech that do not allow the formation of predictions, or is this due to the nature of statistical representations and knowledge in a second language? Recent work has argued that the overall body of evidence points to the latter. Kaan (2014) argues that these differences may arise due to the presence of different statistical distributions compared to native speakers. Second language learners receive very different input from native speakers learning a first language, and the differences in the input during this time may cause speakers to accumulate different statistical knowledge about the environment. While a native language is learned as an infant by listening to speech in the environment, a second language is typically learned in a class and through explicit instruction from a teacher. The statistical knowledge one might gain during first-language learning may not be gathered until later in the second-language learning process. The existence of proficiency differences in these results is also consistent with this argument, as more profi-

cient speakers tend to show more hallmarks of prediction. In context of the equations shown in Figure 5.1, this would mean a different probability distribution over which surprisal and entropy values are calculated. The current phoneme and context could be identical to the representations used in calculations in native speakers, but a different value for surprisal is generated because the distributions themselves are different. Therefore responses in specific prediction studies would be different even though second language learners are using the same processes. This theory fits with results that proficiency modulates responses in second language learners. Proficiency clearly modulates neural responses when listening to a second language (Ihara et al., 2021; Di Liberto et al., 2021) and in high proficiency second language listeners it is possible to see responses such as left anterior negativity. As a learner becomes more proficient, they build a more 'native-like' representation of the statistics of the language.

Kaan (2014) also argues that the lack of prediction observed in L2 experiments may arise because of task-related effects. Small changes to experimental paradigms can have larger effects on outcomes including that visual world paradigm outcomes change depending on how many things are being displayed (Ferreira, Foucart, & Engelhardt, 2013), gender mismatches can cause participants to rely less on gender as a cue throughout an experiment (Van Heugten, Dahan, Johnson, & Christophe, 2012), and in predictable contexts, participants may not wait until the critical disambiguation point to predict upcoming material. That these changes can alter predictive processes in controlled experiments means that we may want to look more closely at responses to naturalistic speech to conclude whether second-language prediction responses are so different to those in native speakers or whether second-language learners are simply more sensitive to task-related effects.

To address these issues and observe whether prediction can be decoded from L2 neural data

in a more ecological environment, I turn in this chapter to naturalistic learning studies, which have seen a surge in recent years. This experimental style allows us to see what second language learners could encode as a whole without designing specific conditions. Studies of this type involved collecting MEG or EEG data from participants while they listen to a recorded audiobook or other continuous stimulus. The data is then analyzed by regressing various predictors to the data to discover structure within the neural signal. Previous L1 studies to use this technique include Brodbeck et al. (2022); Brodbeck, Hong, and Simon (2018); Di Liberto et al. (2021, 2022); Llanos, German, Gnanateja, and Chandrasekaran (2021).

This technique has been used in a few previous studies to discover the underlying information neurally encoded by L2 speakers. For second language learners, recent work from Di Liberto et al. (2021) has used this style of experiment and analysis to discover that L2 learners show encoding of phonemes that is closer to native speakers as proficiency increases. However prior work has not specifically looked at naturalistic responses in the context of prediction. In this experiment, I turn specifically to the question of prediction at various levels. I use the experimental design and stimuli from (Brodbeck et al., 2022) to test the encoding of prediction at three different levels of context — sublexical, lexical, and sentence. Using the same data and setup that has been used for native English speakers, we have a direct comparison between native and non-native participants to investigate differences between these two groups.

Given the spatial resolution of MEG, this work also allows me to compare the lateralization of responses as it compares to native speakers. The difference between responses in the right and left hemispheres and how this might compare with native speakers is not something that has been looked at in-depth for second-language listeners. While this is still an open question more generally in neurolinguistic research (McGettigan & Scott, 2012) and any differences here will

provide only initial data on what could be different between the two language groups, an effect of naturalistic listening studies such as this is that this information can be gathered and speculated about without additional effort in data collection.

This project also creates a new corpus of neural responses to naturalistic listening of speech in L2 listeners more generally. While several fMRI and EEG datasets exist, this work provides a novel corpus of naturalistic neural responses in L2 listeners recoreded with MEG—providing good spatial and temporal resolution.

5.2 Experiment 2: A Naturalistic Listening MEG Corpus of Second Language Learners

This experiment mirrors the paradigm performed by (Brodbeck et al., 2022) with the current study looking at second-language English learners. Participants listen to an audiobook in the MEG scanner while neural responses are recorded during continuous listening. In order to keep variability across participants to a minimum, while still allowing for comparison between different languages, two specific languages were chosen for this study—Mandarin and Sinhala. These languages were chosen primarily for ease of recruitment, while also being a member of significantly different language families to English.

5.2.1 Participants

Participants were recruited primarily from the University of Maryland with criteria that subjects be a native speaker of Sinhala or Mandarin, had not lived in a primarily English-speaking country for more than 6 months before age 12, and had not learned English before age 5. Addi-

tionally, all subjects self-reported no neurological or hearing impairments and were right-handed according to the Edinburgh handedness inventory (Oldfield, 1971). All language background questions were self-reported. A total of 16 native Sinhala speakers and 10 native Mandarin speakers were recorded with 4 Sinhala speakers excluded — two were exposed to English before 5 years of age, one lived in an English-speaking country for 1 year before age 12 and did not disclose this on the pre-experiment screening and one was excluded due to excessive movement in the scanner. All participants gave informed written consent and procedures were approved by the University of Maryland Institutional Review Board. Each subject received $\$15$/Hour for their participation in the 2-2.5 hour experiment.

5.2.2 Procedure

Before scanning commenced, English proficiency was assessed with two tasks and an English exposure questionnaire. The questionnaire was a short survey that asked about participants' exposure to the English language, length of residency in the United States, and comfort with English. Questions on this survey and responses are reported in Table 5.1.

A lexical decision task was presented to participants using the LexTALE framework (Lemhöfer & Broersma, 2012) and online application (www.lextale.com). Participants were shown 60 strings of letters and had to decide whether each was an English word or not. The test includes 40 strings that are valid words of English and 20 that are not and a score is calculated by averaging the percentage of correct and incorrect strings as follows.

$$Score = (\%CorrectValidWords + \%CorrectNonWords)/2 \tag{5.1}$$

LexTALE is a standard task to demonstrate proficiency of English and has been shown to correlate with vocabulary knowledge as well as general language proficiency. The average score reported from a large group of advanced Dutch and Korean learners of English is 70.7%.

Participants also completed a cloze task, where they were asked to complete words from an introductory passage from 'The Botany of Desire' (Pollan, 2002) (not a passage presented during scanning) that had the latter half of the word omitted. After the first sentence, half of every second word was removed from the passage, with participants tasked with completing each work. A passage from the same book was chosen to assess speakers' familiarity with the writing style that they would hear while in the scanner. Cloze tasks are also considered to be a better measure of language proficiency than simple language exposure questions (Park et al., 2022), and may reflect distributional knowledge required for prediction in this study.

I recognize that these two methods are not a perfect measure of language proficiency and will not allow as close a comparison of proficiency as prior studies of this kind (ie. Di Liberto et al. (2021)). They were selected primarily to balance an approximate measure of language knowledge with time and ease of presentation before starting the experiment.

5.2.3 Stimuli

In the scanner, participants listened to approximately 50 minutes of the audiobook 'The Botany of Desire' by Michael Pollan (Pollan, 2002). The book was split into 11 segments between 205 and 331 seconds in length (identical to those in Brodbeck et al. (2022)). Between each segment, participants were asked 3-4 comprehension questions to ensure that they were actively attending to the content. To ensure that participants processed the excerpts for meaning, they

were presented in chronological order.

5.2.4 Data acquisition and Preprocessing

Neural activity was recorded using a whole-head MEG system (KIT, Kanazawa, Japan) with 157 axial gradiometers. The MEG system is inside a magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany) at the University of Maryland, College Park. Data was recorded with a sampling rate of 1 kHz and was filtered with a 200 Hz low-pass filter and a 60 Hz notch filter. Participants lay down in the scanner and were asked to either keep their eyes open or closed for the duration of the scan and avoid excessive blinking. Foam earbuds inserted in the ear canal presented audio during the experiment.

Before imaging, participants' head shape was digitized (Polhemus 3SPACE FASTRAK) and head position in the scanner was measured using 5 marker coils— 3 placed on the forehead and two placed 1 cm forward from the tragus on each ear. Head position was measured at the beginning of the recording, after pre-experiment tests, and at the end of the experiment. For most subjects, the second and third marker measurements were averaged for the purposes of calculation of the inverse function for mapping to the cortical surface. One Sinhala participant was deemed to have moved their head too much during recording, meaning that the position in the scanner could not be determined and they were excluded from the analysis. Pre-experiment in-scanner tasks consisted of 60 seconds of resting state recording and a tone task where participants heard 100 jittered tones and were asked to count how many were played.

Data were processed and analyzed using MNEPython (Gramfort et al., 2014) and Eelbrain (Brodbeck, 2023a) with additional TRF-Tools helper scripts (Brodbeck, 2023b). Spatiotemporal

signal space separation (Taulu & Simola, 2006) was applied to remove artifacts and all frequencies below 1 Hz and above 40 Hz were filtered. Independent component analysis (ICA) (Bell & Sejnowski, 1995) was then used to remove additional artifacts of eye blinks, heart rate and scanner noise from the MEG signal and I decided which components represented artifacts and should be excluded from further analysis. Channels with no data were automatically excluded and a bandpass filter was applied to further restrict the data to frequencies between 1 and 10 Hz. The head shape of each participant was fit to the FreeSurfer 'fsaverage' average brain (Fischl, 2012) using an average scaling technique and this coregistration was used to create an inverse function such that activity could be mapped to virtual source dipoles across a common cortical surface. For any results that were masked by brain region, regions were generated according to the '*aparc*' freesurfer parcellation (Desikan et al., 2006).

5.2.5 Proficiency Profile

While I do not go into a detailed analysis of proficiency measures and how they related to neural activity, statistics for participant proficiency are shown in Table 5.1 including answers to questions on the English exposure survey, LexTALE scores and Cloze task scores. One participant did not complete the proficiency tasks and is not reported for these analyses. As expected, LexTALE and Cloze task scores were more correlated with each other than they were correlated with any other measure ($r = 0.38$), although these are not highly correlated (Table 5.2). Length of Residence in the US did not correlate highly with either of these measures ($r_{lextale} = 0.15$, $r_{cloze} = -0.02$), which is unsurprising given that this is not often a good measure of proficiency. The speakers tested in this chapter would most likely be described as 'advanced' English learners

	Sinhala	Mandarin	All
Age			
Language Exposure Questions			
At What age did you first learn English?	6.0 (2.1)	7.5 (3.3)	6.7 (3.2)
At what age did you feel proficient in English?	18.5 (5.0)	18.4 (3.5)	18.4 (4.4)
How long have you lived in the US?	2.4 (1.2)	3.7 (2.7)	3.0 (2.1)
At what age did you spend more than 3 hours a day in an English speaking environment?	14.4 (5.6)	15.8 (3.9)	15.1 (4.9)
Proficiency Tests			
Cloze Score	68.3 (13.3)	68.6 (10.1)	68.4 (11.9)
LexTALE Score	78.9 (11.1)	73.3 (11.1)	76.2 (11.4)

Table 5.1: Profiency Profile [Mean (Standard Deviation)] for all included participants

for two reasons. Firstly, almost all of our subjects were recruited from the University of Maryland, and require a level of fluency as students of an English-speaker university in addition to language requirements required for admission[1]. Secondly, LexTALE scores of the participants are above the averages reported by advanced learners of English (Lemhöfer & Broersma, 2012)—the participants presented here have an aggregate average score of 75.8 compared to an average of 70.7 previously reported for advanced learners of English.

5.2.6 mTRF Analysis

The continuous neural data was analyzed using an mTRF analysis technique (Brodbeck, Presacco, & Simon, 2018; Lalor, Power, Reilly, & Foxe, 2009) where the signal is deconvolved with various predictors to uncover any neural responses to specific features in the audiobook. Under this method, it is assumed that some of the neural signal can be generated by convolving a particular response function with the value of each predictor at each point in time. These predictors can either be single pulses—such as at the onset of a phoneme—or continuous predictors

[1]The University of Maryland requires a TOEFL score of 96 for full enrollment

	Age Learned	Three Hours Exposure	Age Felt Proficient	Length Of Residence	LexTALE	Cloze
Age Learned						
Three Hours Exposure	**0.35**					
Age Felt Proficient	**0.48**	**0.81**				
Length Of Residence	0.22	-0.02	-0.26			
LexTALE	-0.19	-0.13	-0.13	0.25		
Cloze	-0.02	-0.33	-0.30	0.05	**0.40**	

Table 5.2: Correlation between each proficiency measure. Bold illustrates the highest correlations over all comparisons.

such as the acoustic profile of the stimulus.

The deconvolution allows the generation of a specific temporal response function (TRF) that models the neural response to the presence of each feature. The deconvolution happens individually for each virtual source dipole, allowing the calculation of how significantly each predictor is able to explain the neural signal across the cortical surface.

The models and predictors used follow the methodology from Brodbeck et al. (2022) except where noted below. A model of the continuous data is generated with combinations of the following predictors as explained below.

5.2.6.1 Phonetic Transcription

A phonetic transcription of the audiobook was obtained using a combination of the CMU pronunciation dictionary and the Montreal forced aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), which is also used to align the phonetic transcription with the acoustic

stimulus. Additional manual adjustment was performed to align acoustic data with transcriptions.

5.2.6.2 Acoustic Predictors

To absorb any activity in the neural signal that can be described by low-level acoustic properties of the stimulus, gammatone frequencies and gammatone onsets are included in all models with 8 log spaced frequency bands between 10 and 5000 Hz (Fishbach, Nelken, & Yeshurun, 2001). In addition, to account for neural signals which correspond to the segmentation of phonemes and to account for any variance that can be represented purely with phoneme boundaries, two predictors are included to mark phoneme and word onset. The word onset predictor includes a pulse for every word onset and phoneme onset includes a pulse at each phoneme onset that is not also a word onset to avoid colinearity issues. This means that any neural responses described by each of the context model predictors can be attributed to the specific value that it encodes and not simply because there is a word or phoneme onset.

5.2.6.3 Phoneme Feature Predictors

To investigate the encoding of phonetic features in the neural signal and the representations of particular phonemes, I follow the methods of Di Liberto et al. (2021) in building a phoneme encoding space from the neural data. To do this a phoneme feature predictor with phonetic features (Chomsky & Halle, 1968) is included with a pulse for the onset of each phoneme which includes a specific phonetic feature. There are 19 of these features: voiced, unvoiced, sonorant, obstruent, syllabic, consonantal, plosive, fricative, strident, nasal, approximate, labial, coronal, anterior, dorsal, front, back, high and low. In addition, a separate predictor set is included which

denotes whether a particular phoneme onset is a vowel or consonant onset[2].

5.2.6.4 Context Models

Models of surprisal and entropy of phonemes at three different context levels are used to study the integration of local and unified global context models as described in previous work with native English speakers, in addition to uncovering the predictive processes of second language learners more generally.

Each of these models influences the prediction at the phoneme level with entropy and surprisal encoded at the onset of each phoneme. Under the local context model, predictions at a particular level of linguistic representation can only integrate information at the same level or below. For example, the prediction of an upcoming phoneme can only use the prior string of phonemes as information to make this prediction and any statistical knowledge about the current word or previous words cannot be used. This form of prediction corresponds to the sublexical context model, which integrates only information at the phoneme level.

The global context model, however, predicts that listeners can predict using information at higher levels of the linguistic hierarchy, meaning that the prediction of an upcoming phoneme can be influenced by information about the current word and sentence context (the red arrows in Figure 5.1). In my analysis, this theory corresponds to the word and sentence context models, as these include a prediction of phoneme accounting for this information. In the word context model, the current word influences the surprisal and entropy of the upcoming phoneme, and in the sentence context model, a model of the sentence context influences the predictability of a

[2]The vowel/consonant feature will be colinear with some of the phonetic features included in this analysis and was included to be consistent with the prior work. Given that the phonetic feature analysis does not yield significant results generally, it is unlikely any colinearity changed the outcome of the analysis, although the inclusion of such correlated features should be considered more deeply in future work.

word which then influences the predictability of a phoneme.

Sublexical Context Model

The sublexical predictors are calculated over the probability of encountering a particular phoneme, given the previous 4 phonemes heard, yielding a 5-gram phoneme prediction model. These values are generated by training a 5-gram model on the SUBTLEX-US corpus (Brysbaert & New, 2009). Two values of this distribution are used as predictors in the TRF model: phoneme surprisal, and phoneme entropy. Surprisal is calculated using the formula below, where $phoneme_k$ is the current phoneme and $context$ include the previous 4 phonemes $phoneme_{k-4;...k-1}$.

$$I(phoneme_k) = -log_2(p(phoneme_k|context)) \tag{5.2}$$

Phoneme entropy is defined as the average expected surprisal of the following phoneme and calculated using the equation below. Here the context is the previous 3 phonemes and current phoneme $phoneme_{k-3;...k}$, and I represents the entire inventory of phonemes in the language.

$$H(phoneme_k) = \sum_{i}^{I} p(phoneme = i|context) * ln(p(phoneme = i|context)) \tag{5.3}$$

Lexical Context Model

For the lexical context model, the same calculations above were made, however with a different context. Here at the onset of a word, the probability of a given word is the initial probability of the word calculated from word frequency extract from the Corpus of Contemporary American English (Davies, 2017). At each following phonemes in a word, the probability of any word

that is not consistent with the phonemes already heard is set to zero and the distributions are normalized. The probability of a phoneme is calculated given the probability of every possible word completion.

In addition to the measures described above, an additional statistical measure can be calculated as the entropy over the distribution of words. This is calculated the same way as phoneme entropy above, except with word probability instead of phoneme probability. Here $phoneme_{j,k}$ represents the k-th phoneme in the j-th word and W represents all words in the corpus.

$$H(phoneme_{j,k}) = \sum_{w}^{W} p(word_j = w|context) * ln(p(word_j = w|context)) \tag{5.4}$$

Sentence Context Model

For the sentence context model a 5-gram model is trained on the Corpus of America English (Davies, 2017) with KenLM (Heafield, 2011), where word probability is calculated given the previous 5 words seen. For sentence context models, the same measures are above are calculated, however instead of the probability of a word proportionate to its frequency, it is equal to its probability given the 5-gram model.

5.2.6.5 Models

The full mTRF model includes acoustic and phoneme onset predictors as well as surprisal, cohort entropy and phoneme entropy calculated using the lexical and sentence context levels and phoneme entropy and surprisal calculated using the sublexical context level. To calculate the predictive power of a predictor or set of predictors, a reduced model is run including all predic-

Predictor	Type	Dims	Full Model	Context Reduced	Extended	Phonetic Feature
Gammatone	Continuous	8	✓	✓	✓	✓
Gammatone Onsets	Continuous	8	✓	✓	✓	
Word Onset	Pulse	1	✓	✓	✓	✓
Phoneme Onset	Pulse	1	✓	✓	✓	✓
Vowel/Consonant	Pulse	2			✓	✓
Phonetic Features	Pulse	19			✓	✓
Sublexical Surpisal	Pulse	1	✓		✓	
Sublexical Entropy	Pulse	1	✓		✓	
Lexical Surprisal	Pulse	1	✓		✓	
Lexical Cohort Entropy	Pulse	1	✓		✓	
Lexical Phoneme Entropy	Pulse	1	✓		✓	
Sentence Surprisal	Pulse	1	✓		✓	
Sentence Cohort Entropy	Pulse	1	✓		✓	
Sentence Phoneme Entropy	Pulse	1	✓		✓	

Table 5.3: Predictors included in each TRF model: To calculate the predictive power of features at a specific context levels, additional analyses are run with the full model omitting all sublexical, lexical and sentence level predictors respectively. As Gammatone onsets encode similar information to phonetic features, this predictor is not included in the phonetic feature model. (See Section 5.5)

tors in the full model, except the predictor under investigation. The difference in the proportion of variation explained[3] by the reduced model can be subtracted from the proportion of variation explained by the full model to give a conservative estimate of the improvement that this specific predictor or predictors contribute to the model fit. Some analyses in this chapter are conducted over the entire cortical surface, where some analyses are restricted to a constructed ROI that covers bilateral superior temporal areas which primarily deal with speech processing—this includes the *superiortemporal* and *transversetemporal* labeled '*aparc*' parcellations.

A Phoneme Feature model is calculated using the full model as well as the phonetic feature predictors in order to decode phonetic features from the signal as in (Di Liberto et al., 2021).

5.3 Second Language Prediction in Continuous Naturalistic Speech

5.3.1 Results

5.3.1.1 Model Fit

Model prediction improvement is calculated as described above and results are shown in Figure 5.2 separated by each language group. Improvement in prediction is measured as a proportion of the maximum explained variability across all virtual source dipoles for the full model and averaged across dipoles within the superior temporal region.

The difference in predictive power between the reduced and full models across each individual virtual source dipole is shown in Figures 5.3 smoothed using a Gaussian window (SD = 5mm). A t-test was calculated between each of the reduced models and the full model by taking

[3] Variation explained by a model is calculated as $1 - \big(l1(\text{residuals})/l1(y)\big)$, where y is the continuous neural signal.

the smoothed map and a mass univariate one-sample t-test with threshold-free cluster enhancement (TFCE) (Smith & Nichols, 2009) was applied. Statistical summaries of the t-test are shown in Table 5.4 and cortical maps of significance are shown in Figure 5.4

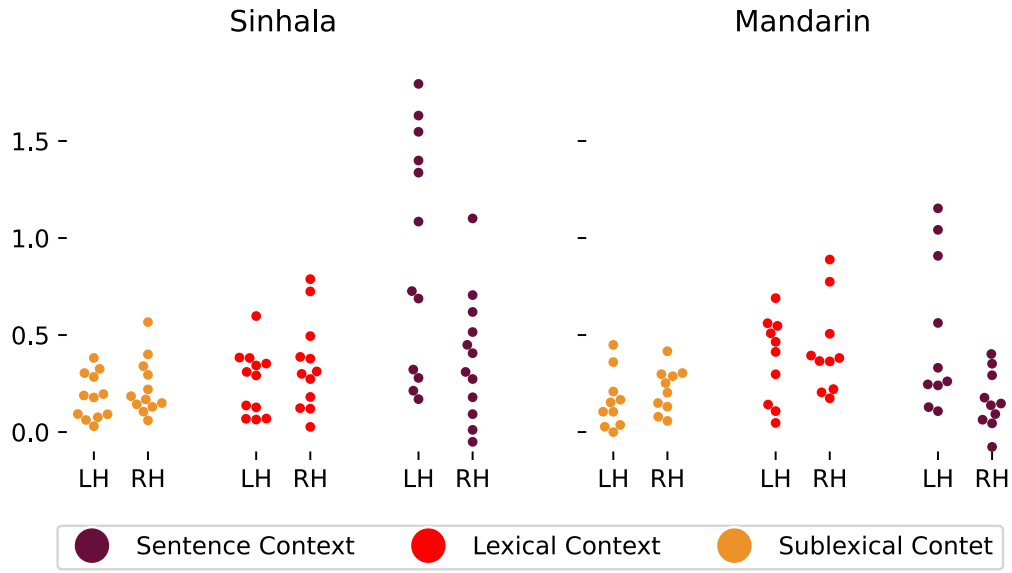


Figure 5.2: Model Prediction Improvement. Each participant is an individual point and this plot shows the difference in predictive power between a reduced model without predictors at each level compared to a full model. Yellow = Sublexical Context, Red = Lexical Context, Purple = Sentence Context. Scale of y-axes is improvement expressed as a proportion of the maximum explained variability across all virtual source dipoles in the ROI for the full model.

	Sinhala					
	Left Hemisphere			Right Hemisphere		
	Δ %	t(11)	p	Δ %	t(11)	p
Sublexical	0.18	5.45***	<0.001	0.23	5.48***	<0.001
Lexical	0.26	5.40***	<0.001	0.34	5.07***	<0.001
Sentence	0.93	5.35***	<0.001	0.38	4.09**	0.002
Linguistic	2.4	5.51***	<0.001	1.8	5.87***	<0.001

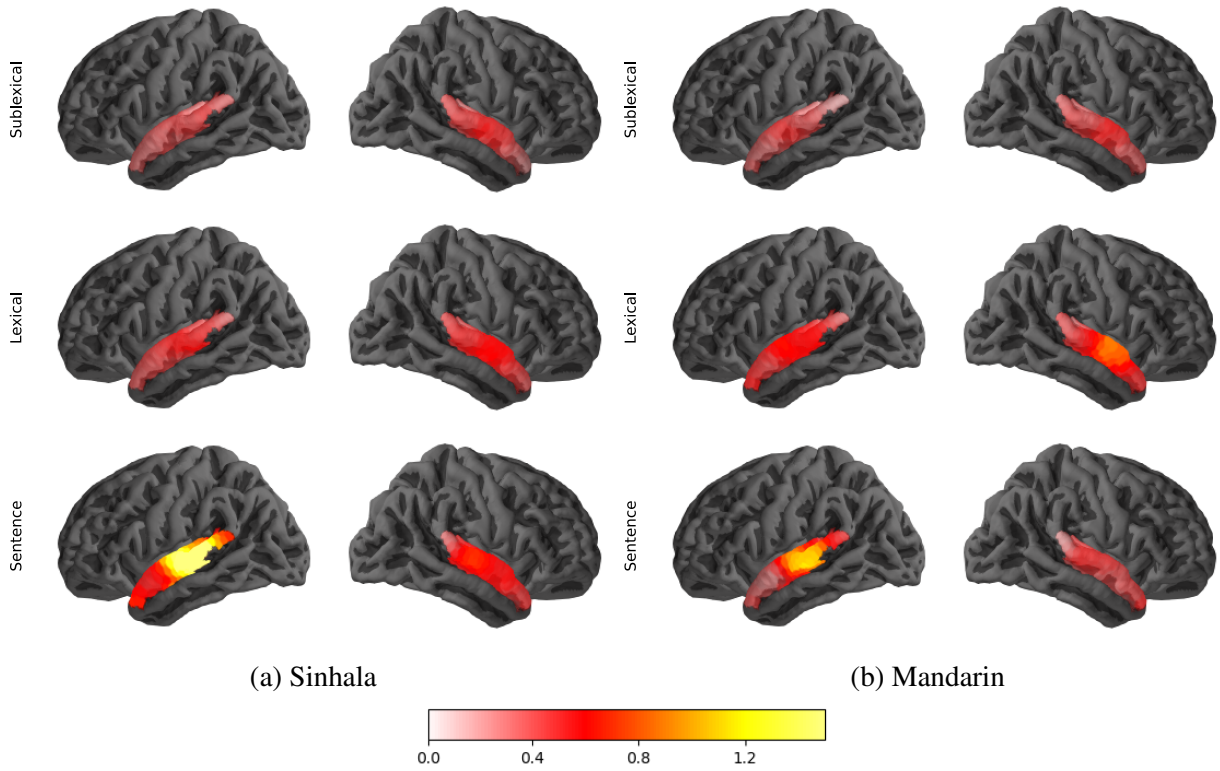
	Mandarin					
	Left Hemisphere			Right Hemisphere		
	Δ %	t(9)	p	Δ %	t(9)	p
Sublexical	0.16	3.5**	0.007	0.22	6.04***	<0.001
Lexical	0.37	5.47***	<0.001	0.43	5.70***	<0.001
Sentence	0.50	3.99**	0.003	0.16	3.49**	0.007
Linguistic	1.88	5.25***	<0.001	1.56	6.28***	<0.001

Table 5.4: Statistical tests for model prediction improvement between full model and a reduced model excluding predictors at a particular context level. Test run is a mass univariate one-sample t-test with threshold free cluster enhancement. Linguistic test compares difference between full model and a reduced model excluding context predictors at all levels.

5.3.1.2 TRFs

In order to plot the temporal response functions generated by the mTRF analysis in the region of interest the TRF for each predictor is extracted from the full model. For each predictor, TRFs are then averaged for each participant across all virtual source dipoles in the region for each hemisphere, before averaging across all participants within a language group. While this may include averaging over dipoles that do not contribute to the predictive power of the model, most dipoles in this region significantly contribute to the power of one of the context models, so there are very few for which this is the case.

Due to the nature of MEG, each source dipole may show either a positive or negative response. For this reason, two different averages are conducted — one where the absolute value



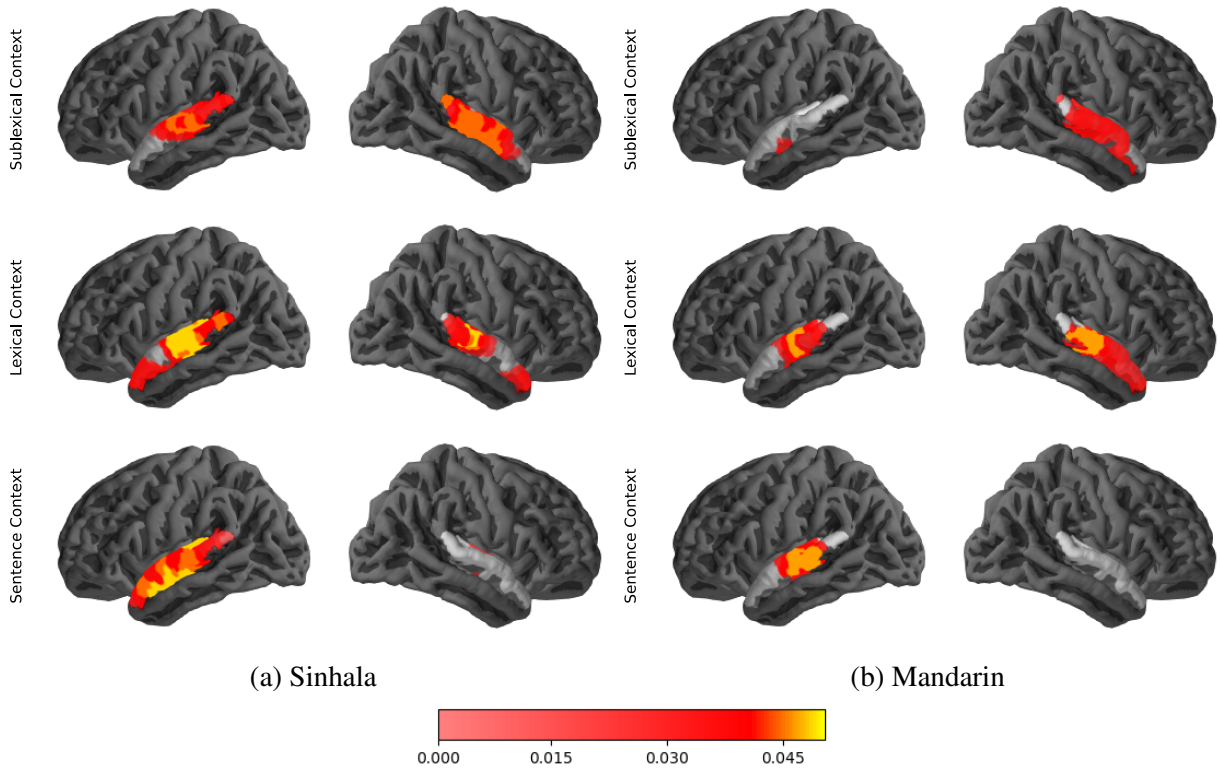
(a) Sinhala (b) Mandarin

Figure 5.3: Cortical Difference Maps For Sublexical, Lexical and Sentence Context Levels: Difference in predictive power between full model and reduced model for each virtual source dipole across the region of interest. The averaged data from these plots is plotted in 5.2.

of each TRF is taken before averaging and one where it isn't. This allows for the analysis of the overall response of the TRF as well as the polarity of the response in this region for each predictor. This is especially useful for responses where there are two peaks, to see whether both have the same polarity or different polarity and could represent different underlying responses. Absolute value TRFs for each language group are shown in Figure 5.5 and raw averaged TRFs are shown in Figure 5.6.

5.3.1.3 Effects of Proficiency

Three different measures were used to investigate the effects of proficiency on the neural signal - the first peak of the decoded mTRF for each predictor, the time latency of that peak, and

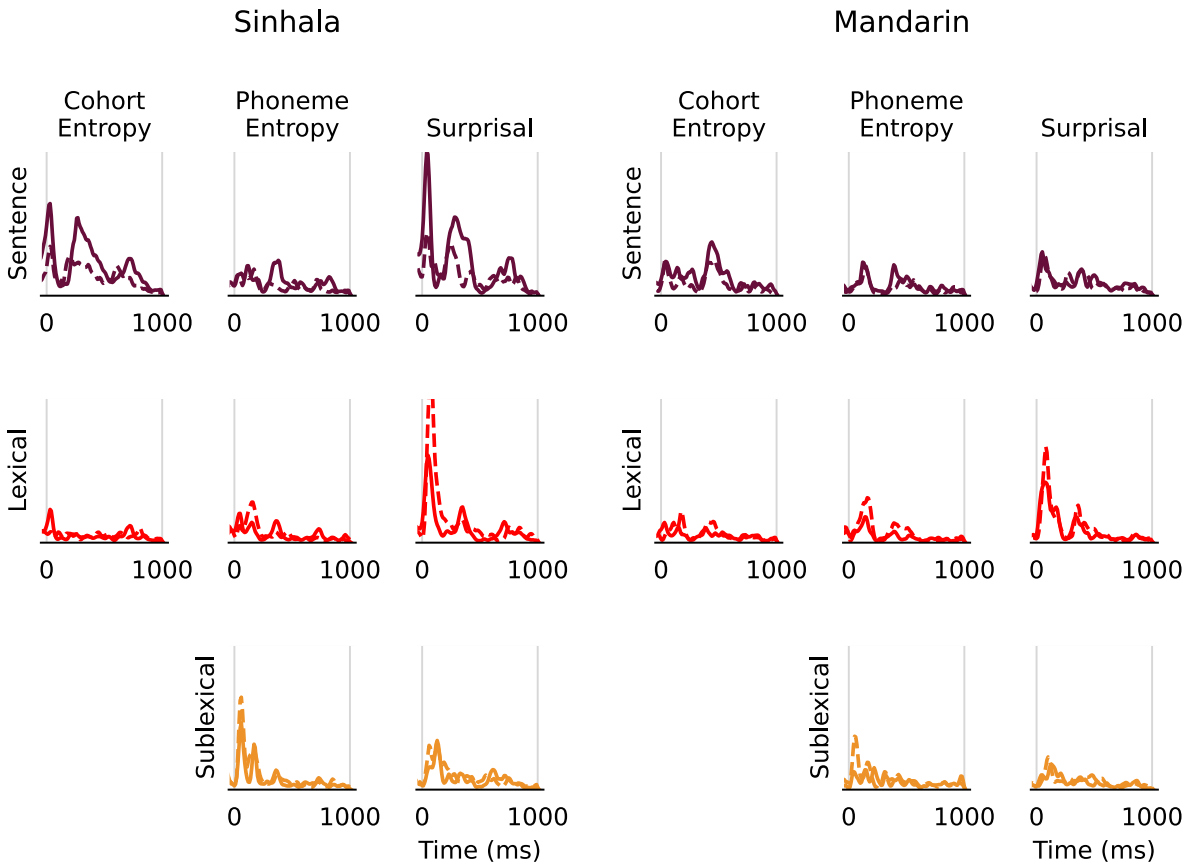


(a) Sinhala (b) Mandarin

Figure 5.4: Significance Maps of Model Improvement For Sublexical, Lexical and Sentence Context Levels: p-value for the model improvement for each virtual source dipole over the region of interest. Any significance values above 0.05 are not plotted.

the increase of model productivity for each of the three context levels. Each of these measures was compared to participant performance in the lexTALE task and cloze score. Neither of these measures had any significant correlation with any of the three measures used—results of this analysis can be seen in Appendix C.2.

Although none of these measures correlate with the proficiency measures used in the experiment, this does not mean that proficiency does not have an effect on the neural response. As discussed above, the participants in this experiment would be considered advanced learners of English and so they may have reached a level of competence where proficiency is not reflected as strongly within the neural signal (ie. we have reached a ceiling effect). Additionally, the measures of proficiency used in this study are only approximate and may not provide a clean enough



(a) Sinhala

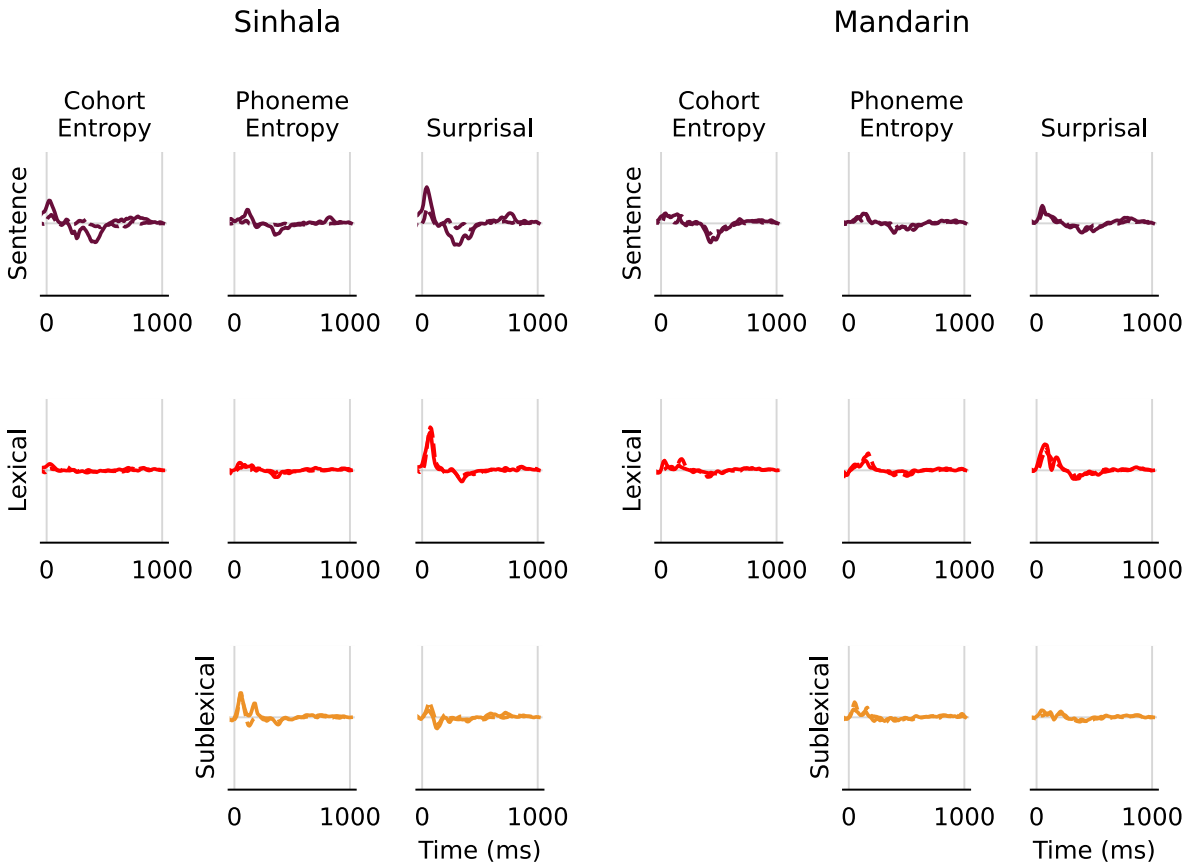
(b) Mandarin

Figure 5.5: TRF function for each predictor in the full model. The absolute value for the TRF function in each virtual source dipole is averaged first for each participant and then across participants. Dotted line is averaged over right hemisphere ROI dipoles and solid line is averaged of left hemisphere ROI dipoles.

measure of proficiency to see effects at this high level.

5.3.2 Discussion

The goal of this chapter is to investigate how second language learners use prediction during online sentence processing and integrate different levels of linguistic information in contextual-



(a) Sinhala (b) Mandarin

Figure 5.6: TRF function for each predictor as in Figure 5.5, except the absolute value is not taken. This shows the polarity of the response to each predictor across the ROI.

izing upcoming input. The results here show clearly that these speakers are using information at sublexical, lexical and sentence context levels to anticipate upcoming phonemes, providing additional evidence that second language learners can exhibit similar prediction processes to native speakers, particularly in naturalistic environments.

In the context of the results from Brodbeck et al. (2022), these data show that non-native speakers are integrating both global information at the scale of lexical items and sentences, as

well as predicting using local information from just the sublexical level. Each of the context models significantly improves the prediction of the neural data across both hemispheres, with the sentence level context accounting for the most data, followed by lexical and sublexical level information. Despite the fact that measures predicted by each of these models are correlated and each model is predicting the same unit of speech (the upcoming phoneme) with different information, each context model uniquely contributes to the explanation of the neural signal. These models go further than exhibiting prediction in second language learners but also unify theories of local and global context representations previously only illustrated in native speakers.

Work on second language prediction has typically studied neural or behavioral responses from paradigms that present carefully constructed sentences to study a specific effect or phenomenon; or used constrained measures such as eye-tracking within a visual word paradigm, to see whether these subjects are indeed predicting upcoming material. In this study, I specifically show that strong predictive processes are active in naturalistic speech listening where there are no particular task requirements. This work provides further evidence that the lack of prediction supposedly reflected in previous studies may indeed be due to task constraints as proposed by (Kaan, 2014), and not because second language learners do not have the ability to predict in their second language.

As one of the context predictors used here is prior sublexical information, it is interesting to also consider these findings in light of work that focuses on the representations that these learners have of the sounds of the second language. It is commonly established that second language learners can struggle to discriminate between sounds that are not present in their native language, particularly when two sounds differ along a dimension that is not present in the native language (Goto, 1971; Yamada & Tohkura, 1992). While these studies are often presented in

very constrained settings where participants must differentiate between two words or sounds in isolation, one may expect that due to the instability of these representations, second language learners may rely more heavily on top-down contextual information when making predictions about upcoming input. These results show that this is not the case, and non-native speakers do rely on contextual information at the sublexical level as well as these higher-level representations. While the sentence model does appear to explain more data than the lexical and sublexical context models in this analysis, the same is true of native-speakers (Brodbeck et al., 2022). A more detailed study of these differences would be helpful to distinguish any differences, but these current results do not appear to show any different reliance on different context models to that of native speakers.

Cortical maps of predictive power and significance show that context predictor features can be decoded from regions of the brain typically associated with speech and language processing—namely the superior temporal gyrus and surrounding areas. Furthermore, the evidence presented here corroborates findings from native speakers that responses to different context predictors are spatially separated across cortex. Similar lateralization effects as native speakers are observed, with sublexical- and lexical-level context measures exhibiting bilateral activation and sentence-level context predictors showing most activation in the right hemisphere superior temporal regions. When looking at the significance of t-tests between a reduced model without sentence-level context predictors and a full model, almost all significance attributed to these predictors is shown in the right hemisphere. A more detailed comparison between native and non-native speakers is necessary to conclude whether this activity differs significantly between the two groups.

It is important to note the role that proficiency may play in the results stated here. While we measure proficiency approximately through two tasks, these measures do not appear to correlate

with any of our neural measures. In looking at phoneme space analysis, Di Liberto et al. (2021) show an effect of proficiency in the neural encoding, along with countless other studies which show that evoked responses are modulated by proficiency. The main goal of the study here was to gather a corpus of proficient second language learners split across two languages, and I did not set out to find a participant group that varied across multiple proficiencies. The results stated here do not discount that proficiency modulates the responses of second language learners, though. While I do use more than simply years of residence (which is known to be a poor measure of second language proficiency), the proficiency measures I use are quick and approximate measures that are not as detailed as those present in prior studies (Di Liberto et al., 2021). This likely leads to more noise in these measures and may result in null effects in the analysis. Additionally as discussed above, the participants in this study are all likely to have high English proficiency given that they are attending a university where English is the language of instruction. For that reason, I would place these results amongst results of 'advanced' second language learners. These results however are consistent with those studies, given that it is often seen that as speakers become more proficient they show responses that are more similar to native speakers.

This work also provides a corpus of naturalistic data, which is becoming increasingly important within the study of online sentence and speech processing in humans. While at least one EEG dataset of second language naturalistic listening exists (Di Liberto et al., 2021), in addition to several continuous fMRI datasets (Brennan & Hale, 2019; Bhattasali et al., 2019), this is one of the first datasets of MEG naturalistic listening data for second language learners. MEG allows not only good spatial resolution across the cortical surface when creating an inverse solution with head shape as I have done, but it also allows for fine temporal resolution to see rapid responses to predictors. The corpus here closely parallels that of Brodbeck et al. (2022), with the same

protocols and stimuli as native speakers, allowing for detailed comparisons between native and second-language speakers.

Overall in this chapter, I have demonstrated the nature of online predictions in second language learners through analysis of a newly created corpus of MEG naturalistic listening data. While there are many more analyses that can be conducted on this dataset, I present initial findings which illustrate that second language learners show predictive processes that reflect native speakers in integrating information from both local and global context models.

5.4 Comparison with Native Speakers

In order to compare the results collected in this experiment with native speakers, I reanalyze the raw data from Brodbeck et al. (2022)—where participants listen to the same audiobook—through the same pipeline as the results presented above, including filtering and ICA analysis to remove artifacts from the data. Refiltering and reapplying these analyses ensures that I have a clear and controlled comparison between native and non-speakers. This analysis results in small differences between the results shown here and those in the original Brodbeck et al. (2022) chapter, however any changes are not large and overall the results here align with the data and findings of that paper.

5.4.1 Results

5.4.1.1 Model Fits

Shown in Figure 5.7 is the improvement in predictive power for predictors at each context level for both L1 English speakers (Brodbeck et al., 2022) and L2 English speakers (Data col-

lected as above). This is once again defined as the difference between the predictive power of a full and reduced model. Statistical tests over this data for L1 speakers and L2 speakers combined language group are included in Table 5.5.

Overall we see that each level of context predictor provides significant improvement in the explanatory power of the model and there is no significant difference between the improvement provided for native speakers and second language learners.

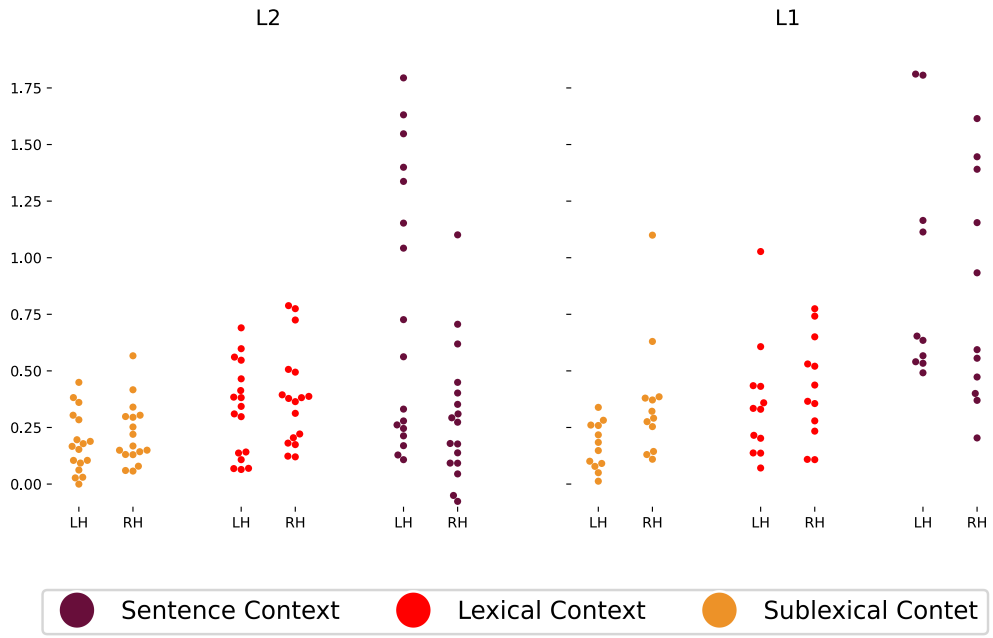


Figure 5.7: Model Prediction Improvement for L1 and L2 Speakers

5.4.1.2 Lateralization

To investigate any differences in lateralization between native speakers and second language learners, I construct a 'lateralization index' for each level of context predictor. This is calculated by taking the increase in model predictive power as calculated above for each hemisphere and normalizing to see where activity is more predicted for each participant. For each

	L1 (English)					
	Left Hemisphere			Right Hemisphere		
	Δ %	t(10)	p	Δ %	t(10)	p
Sublexical	0.17	5.62***	<0.001	0.37	4.68***	<0.001
Lexical	0.36	4.75***	<0.001	0.43	6.53***	<0.001
Sentence	1.13	5.97***	<0.001	1.03	4.25**	0.001
	L2 (Mandarin/Sinhala)					
	Left Hemisphere			Right Hemisphere		
	Δ %	t(21)	p	Δ %	t(21)	p
Sublexical	0.17	6.36***	<0.001	0.22	8.15***	<0.001
Lexical	0.31	7.49***	<0.001	0.38	7.64***	<0.001
Sentence	0.73	6.22***	<0.001	0.28	4.78***	<0.001

Table 5.5: Statistical Tests for Each Language Group

participant, the index is calculated using the following equation:

$$LI = RHImprovement/(LHImprovement + RHImprovement) \tag{5.5}$$

A value above 0.5 indicates that this context level increases predictive power more in the right hemisphere and a value less than 0.5 indicates that this context level increases predictive power more in the left hemisphere. Plots of lateralization index are shown in Figure 5.8.

5.4.1.3 Latency

A final analysis of differences between native and second language learners of English is the latency of the neural responses to each predictor. To see the neural responses to each context level over time, the response of all predictors at each context level is summed for each virtual source dipole across a particular time window and then averaged across all participants in each language group. The time windows for responses are 200 ms: -50—150 ms, 150-350 ms, and

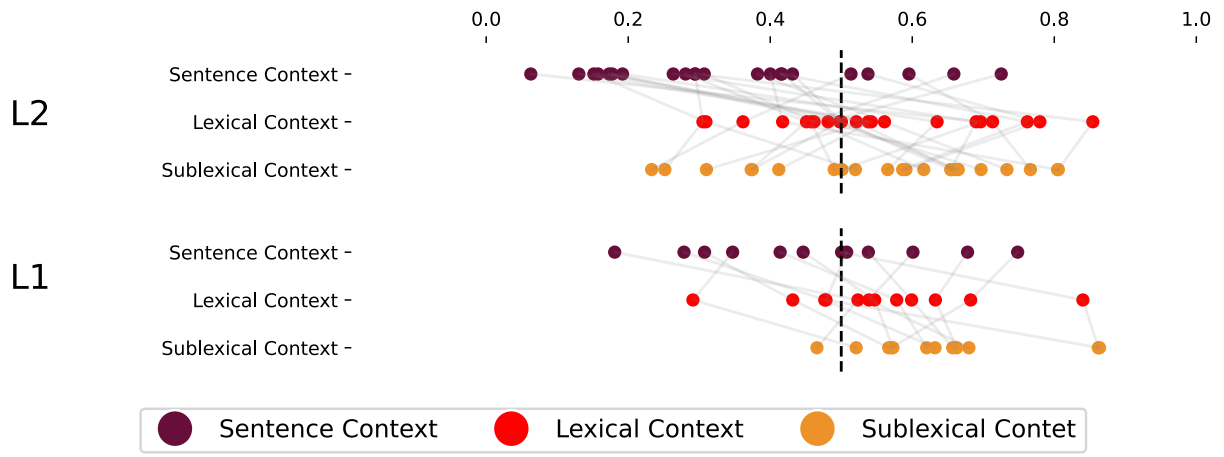


Figure 5.8: Lateralization Index For L1 and L2 Speakers: Calculated using predictive power for each context level

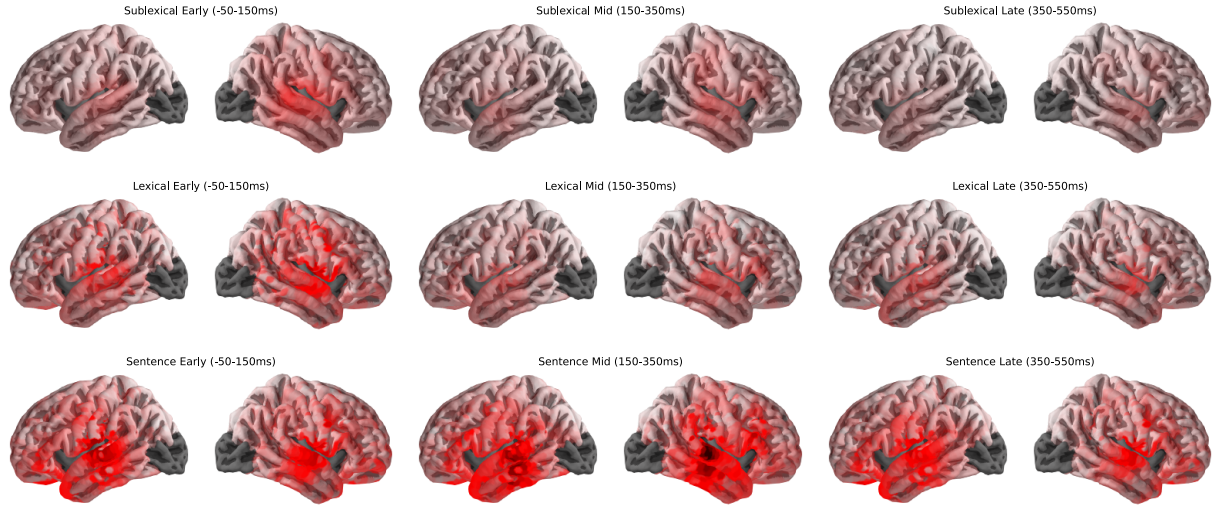


Figure 5.9: Map of summed responses of TRFs at each context level split into early (-50—150ms), mid (150-350ms) and late (250-550ms) time scales for native speakers

250-550 ms. For plotting purposes, responses are smoothed using a Gaussian window. We see that the sentence context level predicts activity across all timescales whereas the sublexical and lexical context levels predict data only at the earliest timescale, however these findings appear to be consistent across native and second language learners (Figures 5.9 and 5.10)

All TRFs generated for the full model were averaged across the STG ROI for each language group, both using the absolute value of the response in each virtual source dipole within the ROI

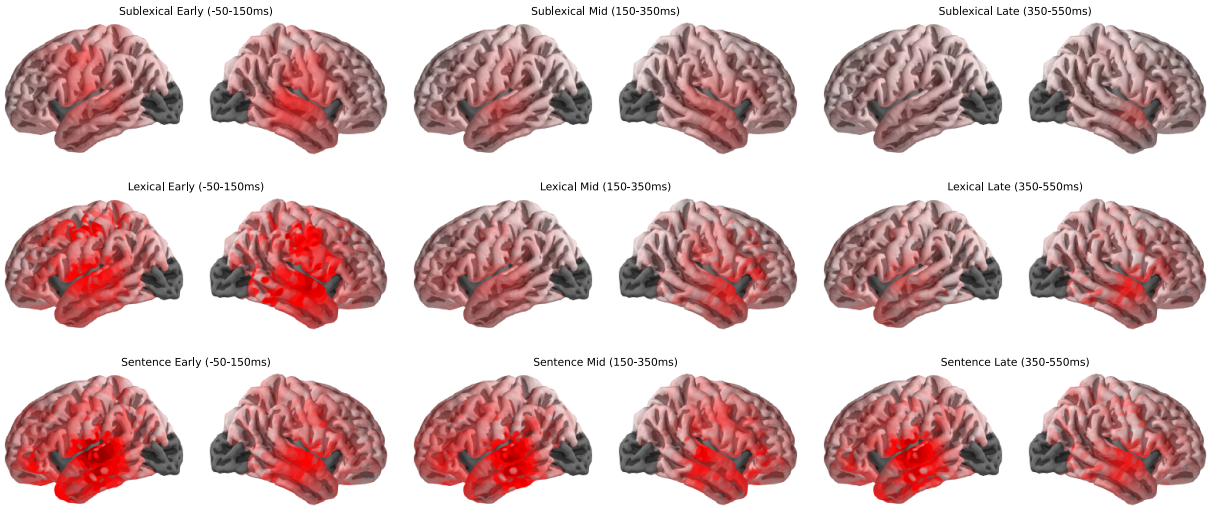


Figure 5.10: Map of summed responses of TRFs at each context level split into early (-50—150ms), mid (150-350ms) and late (250-550ms) time scales for second language learners

as well as an average of the overall value. Once again this allows a clarification of the polarity of the TRF responses to each predictor. Finally, a peak is extracted from the TRF for each predictor by finding the maximum value of the TRF in the first 250ms of the response. The time point of this peak is extracted and plotted in Figure 5.11 for each language group.

5.4.2 Discussion

In this section, I perform a controlled comparison between responses from native English speakers and non-native English speakers within the continuous natural listening paradigm. While these participants come from two different studies, the participants all listened to identical stimuli [4] and were analyzed using a consistent mTRF pipeline. The overall results indicate that the second language learners in the current study exhibit similar neural responses to predictive measures to native English speakers across various context levels. One small difference may arise

[4]Two participants in the native speaker condition form Brodbeck et al. (2022) listened to a slightly different version of 1 of the 11 total presented segments

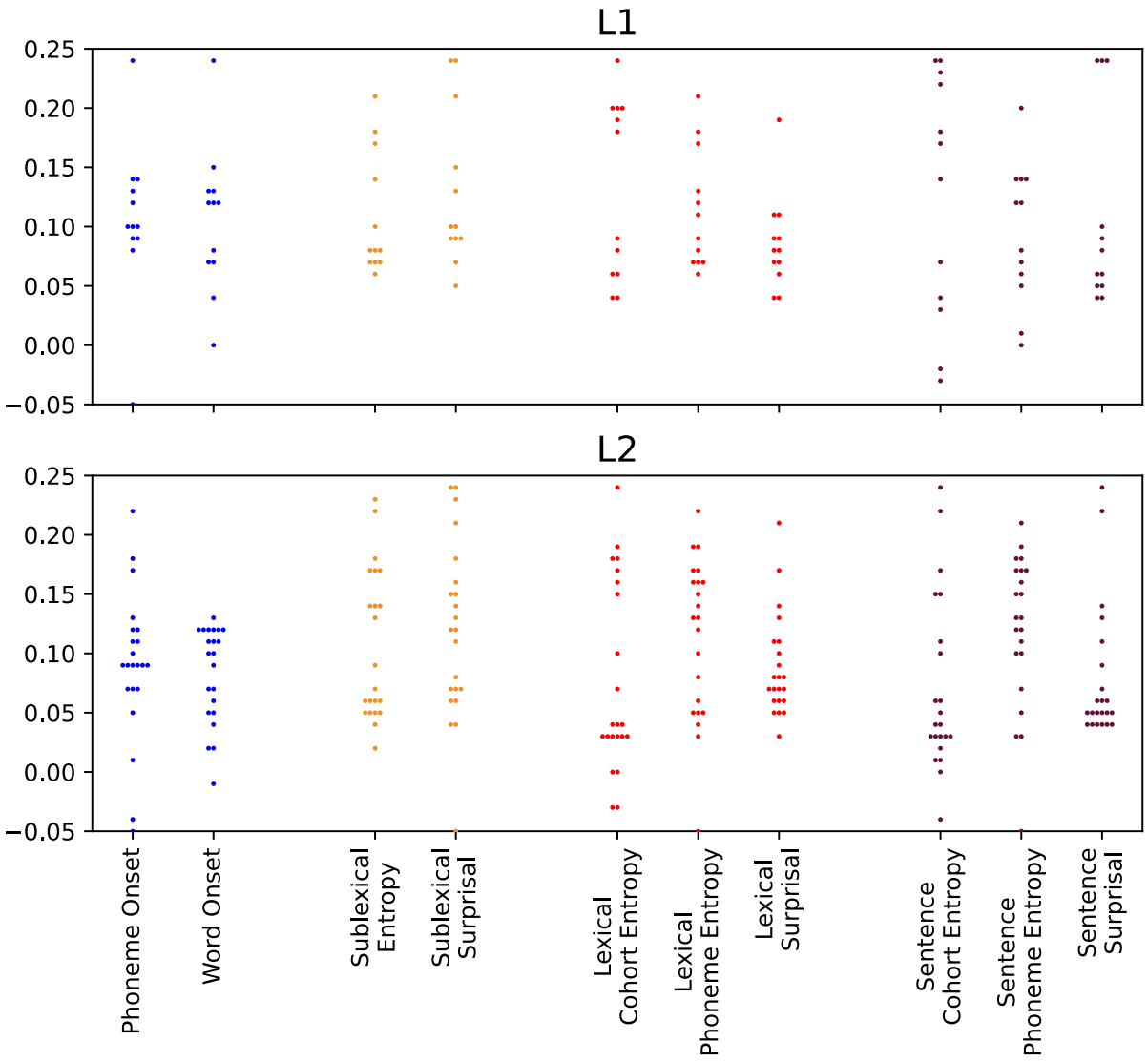
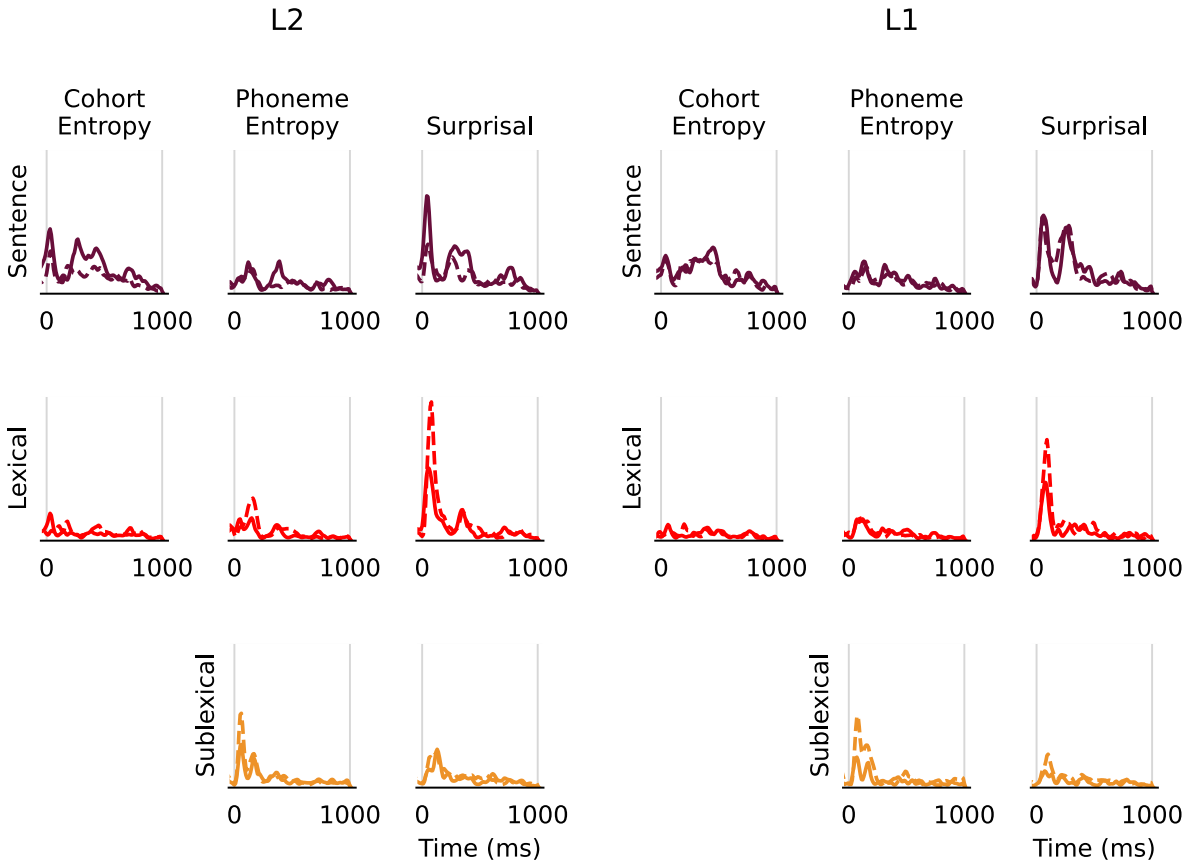


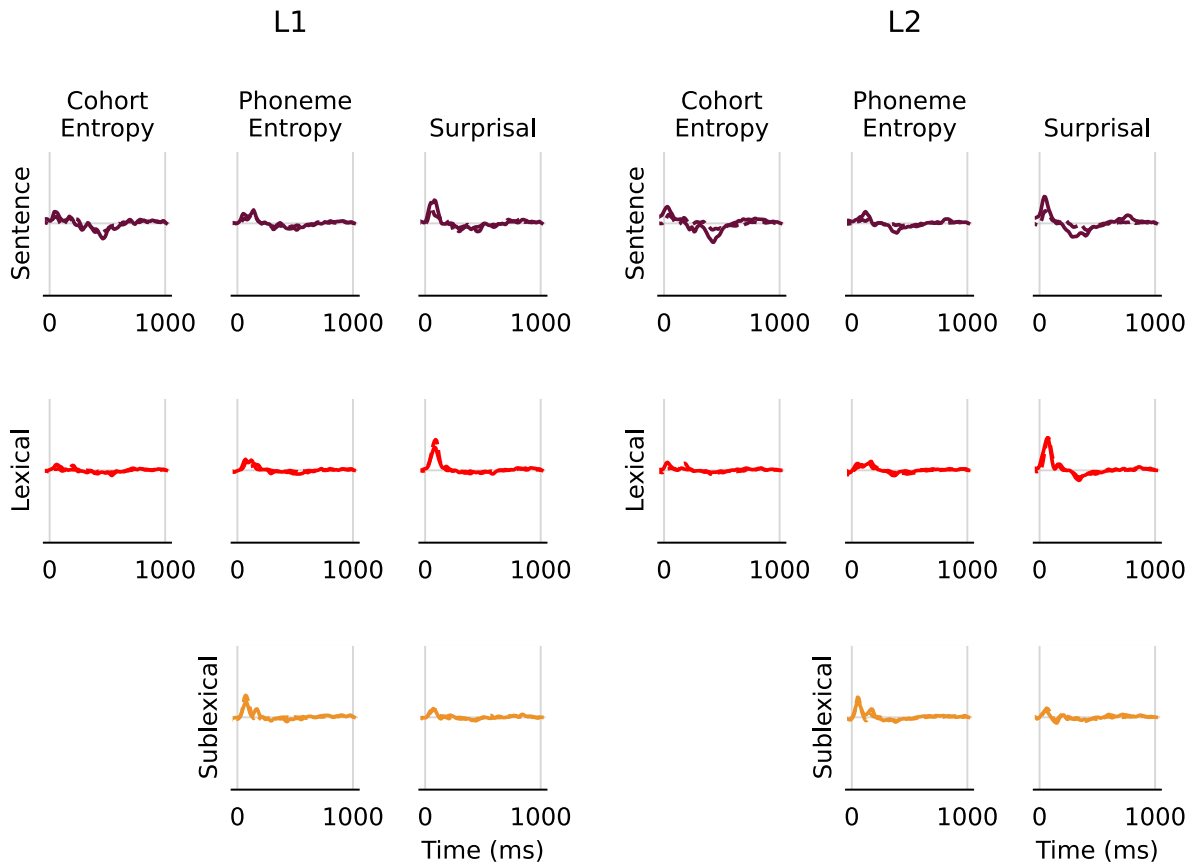
Figure 5.11: Time of TRF peak across predictors within first 250 ms of responses plotted for native English speakers and second language learners.



(a) L2 (b) L1

Figure 5.12: TRF function for each predictor in the full model. The absolute value for the TRF function in each virtual source dipole is averaged first for each participant and then across participants. Dotted line is averaged over right hemisphere ROI dipoles and solid line is averaged of left hemisphere ROI dipoles.

in the lateralization of responses with second language learners showing more activity that correlates with these measures in the left hemisphere than right. To measure response to each context level, I use the amount of additional explanatory power provided by a specific set of predictors in the mTRF model. Using this measure, we see that in each of the three context level predictors—sublexical, lexical and sentence—the responses decoded for the native English speakers is not



(a) L1 (b) L2

Figure 5.13: TRF function for each predictor in the full model. The absolute value for the TRF function in each virtual source dipole is averaged first for each participant and then across participants. Dotted line is averaged over right hemisphere ROI dipoles and solid line is averaged of left hemisphere ROI dipoles.

significantly different than the responses decoded for second language learners.

I also study the difference between the two groups by looking at the time of the peak of the TRF responses decoded for each predictor (Figure 5.11. In the previous section, I used this as a measure to compare the responses to the proficiency of each second language learner. Here I do a larger comparison between the two groups of native and second language learners. This again

shows no significant difference between the two groups of language backgrounds, suggesting that second-language learners' responses are aligned with those of native speakers.

These results together indicate that not only do second language learners predict upcoming input at various levels of linguistic context, but that these responses align closely with those of native speakers. This complements previous evidence which suggests that second language learners are able to build toward strong representations of prediction. While the participants in this study have reached a high level of proficiency in English as discussed previously in this chapter, none of the participants analyzed learned English before age 5, and on average were not exposed to more than 3 hours of English a day until age 14. Therefore despite having advanced proficiency at the time of scanning, they are clearly considered to be second-language learners of English.

These results also do not indicate any latency in the neural response to prediction in second language learners. Given previous findings that second language learners do not predict to the same extent as native speakers, one may expect that any neural responses reflecting prediction may be delayed if more processing is required to integrate incoming information. However there is no evidence of any differences in timing in these data, suggesting that if there are differences in the online integration of phonetic information with predictions, it is reflected in a different response at the same time, rather than showing a different temporal profile.

One small difference can be seen when looking at the individual TRF responses to each predictor in the full model. For the surprisal response at the lexical context level, there is a clear second peak in second language learners compared to native speakers, and in looking at the raw averages it is clear that this is a negative bilateral response (Figure 5.13). A similar, albeit smaller peak is also present at the lexical level. While this could be insignificant, it may signal additional

processing in the second language learners or a second stage of integration at the lexical level. These responses are likely reflected in the slight increase in activity that can be seen in Figure 5.10 at the late lexical level at 350—550 ms. The current analyses do not provide the power to tease apart these small differences, but this could indicate that while the results here show very similar responses in second language learners and native speakers, there are still small differences that exist when predicting phonetic information in a second language.

Since so many previous studies have shown different patterns of prediction in second language learners, I return to the question proposed at the beginning of this chapter. Is the lack of prediction typically seen in second language learners due to mechanistic differences in the processing of a second language or other differences in statistical or distributional knowledge that a second language learner has? In the previous section, I show that second language learners are able to integrate information at various context levels when predicting upcoming input, and the results in this chapter go one step further in demonstrating that it is possible for second language learners to show the same neural responses to predictive measures as native speakers. This is consistent with the latter hypothesis that differences in markers of prediction in second language learners arise due to differences in knowledge and not mechanistic differences with the processing between ones first and second language. These results are inconsistent with the Shallow Structure (Clahsen & Felser, 2006a) or similar theories, which suggest that the online representations of speech in second language learners fundamentally differ from that of native speakers. In the current analysis, the model used to generate predictions at the sentence level is a 5-gram model that does not contain syntactic structure, however, it is expected to provide a baseline of sentence-level information that can be used to predict upcoming phonemes. These kinds of models are commonly used as an approximation of information that is available at the

sentence level (Brodbeck et al., 2022). The fact that this model can describe neural data as well in second-language learners as it can in English speakers, suggests a similar processing of sentence information. Future analyses of this data could look specifically at a model which contains syntactic and a greater depth of structural information in order to ensure that this is true, and check whether there are differences in the encoding of more abstract syntactic structure in native speakers and second language learners.

Given that the results here show no differences in prediction between second language learners and native speakers, one may wonder how this could be consistent with the numerous findings that demonstrate that second language learners do not predict in the same way as native speakers and do not use information such as gender markers to predict upcoming input. Even previous neural results have shown that second language learners do not exhibit the same evoked responses as native speakers in experiments which appear closely linked to what is being measured in this experiment. I return to two main points posited by Kaan (2014)—namely that second language learners may not have the same distributional information about a language as native speakers, and that task effects could play a big role in influencing experimental results.

On the first point, Kaan (2014) suggests that second language learners may show different levels of prediction to native speakers, not because they cannot form predictions, but because they have a different representation of the statistics of that language compared to native speakers. For example, Kaan considers that the input an infant receives during native language learning is very different from the kind of input that one receives when learning a second language. Infants may learn by listening to those around them talking, where a second language learner may acquire a language in a formal classroom with instruction from a teacher. The model of the environment that one may build, including what words are commonly heard together, could be different be-

tween these two settings. Second language learners may therefore be predicting using different contextual information to native speakers. For example, in the equations proposed in this analysis (Figure 5.1), this would correspond to performing the same mathematical calculation but with a different probability distribution given a specific phoneme and context. In the current study the participants are highly proficient in English and this may have allowed them to already reach a point where this distributional information is the same as a native speaker of English. Given that these participants are attending an university where English is the language of instruction and may have been in English speaking environments for several years, this would not be surprising.

Furthermore, it is known that speakers can be quick to adapt to new contexts and this could influence how participants form predictions. The participants in this study reside in the US and therefore are exposed to English daily, allowing them to build up this distributional knowledge which would mirror native English speakers. If speakers with the same proficiency were tested in an environment where they do not listen to English as frequently, they may show different responses if the statistical knowledge is less robust in the moment they are tested.

Second, in regard to the task-related effects of prediction studies, the work here once again highlights the importance of naturalistic continuous listening studies to uncover neural representations of speech and language. Kaan (2014) suggests that the lack of prediction shown in some studies could arise from the constraints of specific experiments. For example, in experiments where expectations are contradicted, participants throughout the course of the experiment begin to rely less on prior information to predict upcoming input, which could lead to results that do not truly reflect underlying processes in natural speech processing environments. If second language learners are more sensitive to these effects than native listeners, then results would show differences between the two groups despite the same underlying mechanisms. These results are

presented with the understanding that similar biases could be present in naturalistic listening studies also, where second language listeners could become fatigued throughout by additional attentional effort required to process the incoming content and future analyses breaking down sessions into sections could be useful to see responses over the course of the experiment. However the results presented here are less likely to be influenced by task-related effects and provide an insight into prediction processes without some of these effects of experimental conditions or stimuli.

These are two potential reasons that these speakers show such similar signatures of prediction to native speakers despite prior results, although the current results cannot tease apart these outcomes. It could be that there are responses to prediction during early language learning, but typical experiments to test this cannot pick up on the responses, perhaps due to the experimental paradigm. Alternatively, these may indeed be responses that are only observed in the highest proficiency second language learners.

One additional explanation is that markers of predictions that are typically not present in second language N400 or grammatical gender prediction studies use different mechanisms than what gives rise to the results observed here. The continuous listening paradigm captures effects during online speech comprehension that are fully subconscious in participants. These responses could be separate to those that are observed during previous studies which may involve more conscious processes. Grammatical gender may be a different type of cue that could be treated differently from phoneme prediction and these kinds of effects not be captured by the current analysis (grammatical gender agreement for nouns is not relevant for English and so would not be present in the current dataset regardless, but similar cues discussed above, such as other agreement effects or filler gap contexts may be). The results presented here could be driven

by situations where second language listeners are predicting phonemes, and any situations where prediction breaks down are averaged out over the entire audiobook.

Future work should look more closely at proficiency level in relation to these responses, perhaps along the lines of Di Liberto et al. (2021), who present a more elaborate test of English proficiency before scanning participants. The author's test proficiency through a 20-minute test before using EEG to scan participants' responses as they listen to an audiobook, although the analysis in that study mainly tests the phonetic space encoded in the EEG signal. The same analysis that is presented in this chapter could be conducted on that dataset to see whether the context responses differ between proficiency groups. Otherwise, a similar set of participants could be recruited and tested on both specifically constructed experimental paradigms such as prediction using markers such as grammatical gender, as well as recorded listening to naturalistic speech. This would allow us to see whether other—perhaps more conscious—markers of prediction correlate with online neural responses. This data could illuminate much more clearly the reason that these results differ from prior work

One potential difference between native and second language processing that does arise from this analysis is the lateralization of the neural response to these predictors. It appears from the lateralization index as well as looking at the TRF responses to each predictor that responses in the second language learners are more shifted towards the left hemisphere than in native speakers. Lateralization of neural responses is not something that is usually studied in the neural responses of second language learners, however I propose that one potential reason for this could be that second language learners are not integrating information over as long timescales as native speakers. A notable theory of differences in hemispheres in language comprehension is the 'asymmetric sampling in time' hypothesis proposed by (Poeppel, 2003), who suggests that the

left hemisphere integrates information over a shorter timescale and the right hemisphere integrates information over longer timescales. The results in this chapter do not appear to follow this model since the highest level linguistic context—the sentence level—which integrates information over the longest timescale shows more bilateral activations where the smaller timescale integration shows mostly activation in the right hemisphere. The results of the native speakers also do not align with this hypothesis and there is clearly more work to be done to integrate these results with such a theory.

The Dual Stream Model (Hickok & Poeppel, 2007) and later work may align with this proposal which suggests that a combinatorial network for language processing as well as the lexical interface are lateralized to the left hemisphere. These networks are likely more necessary for the responses shown in the results presented in this analysis. I propose that greater left lateralization may arise due to greater processing costs for integrating information in second language learners. Although overall the effects are similar between the native and non-native speakers it is possible that second language learners require more effort to integrate surprising information online during speech processing. Signatures of these processes would arise more in left hemisphere than right hemisphere given that, particularly the lexical and sentence context predictors, require integration via these combinatorial and lexical networks. Linguistic processing is generally left lateralized and so this result could reflect a larger general engagement of the linguistic system as a whole.

There are other slight differences in the neural responses, including what we may see in this data as a second component to some responses in second language learners—particularly for lexical and sublexical surprisal—which could reflect an additional process of integrating information. I do not have specific hypotheses about what this response could reflect, and deeper

analysis of this data and specifically these predictors could indicate the mechanisms behind this response.

Overall, this section has demonstrated that the responses to predicting upcoming phonemes in native speakers and second language learners are incredibly similar, both in strength and timescale of the response. I have discussed how these results are still consistent with many previous findings and indicate that prediction in second language learners may ultimately depend on the knowledge of the statistical distributions of a language that second language learners may take time to acquire. I propose several new lines of research which will allow further investigation of these effects, including additional analyses on the neural data presented here. An advantage of naturalistic continuous listening studies is that this data can be analyzed in many ways and with many different algorithms, meaning that the experiment here can provide a resource for study into second language speech sound processing.

5.5 Phoneme Representations

Previous chapters in this dissertation have discussed the representation of speech sounds that non-native speakers have when learning a second language. It is clear that second language learners have difficulty in discriminating between phonetic categories that are not present in their native language (Goto, 1971; Broersma & Cutler, 2011) and rarely reach the level of native speakers. One primary issue appears to be that these cues cannot always be used efficiently during online sentence comprehension.

In this experiment, I have shown that when predicting upcoming phonemes, second-language learners show similar responses to native speakers in integrating information at various context

levels. One may consider that in order to predict phonemes and show the same neural signatures of native speakers, a second language learner must have a robust representation of these phonemes that are being predicted. Di Liberto et al. (2021) demonstrate that the neural encoding of phonemes as measured by EEG responses is scaled with proficiency with more proficient speakers showing a more similar phonetic feature space to native speakers than less proficient speakers.

Given that I see such similar prediction responses in the previous analysis, I try to uncover this phonetic feature space in this MEG data, in order to see whether there is a robust phonetic feature space for the second language learners in this study that mirrors native speakers, or whether there are clear differences. If the spaces are similar, then it may suggest that the participants in this study are proficient enough that they may be able to show similar prediction signatures to native speakers, because they have similar phonetic representations to native speakers. However if there are clear differences between the phonetic feature space then it would suggest that second language learners are able to build prediction representations and show incredibly similar responses to native speakers, even building off different representations.

I analyze phoneme representations through an mTRF model that includes phoneme features as predictors. Since the Gammatone Onset predictors are closely related to phonetic categories, I compare the effectiveness for phonetic features in the mTRF analysis by calculating the improvement in model fits that results from the inclusion of each set of features. Results are shown in Figure 5.14.

It is clear that phonetic features in this analysis do not account for any additional neural data than simply including gammatone onsets. Despite the fact that gammatone predictors explain more of the data than phonetic predictors, I follow a similar analysis to Di Liberto et al. (2021) in

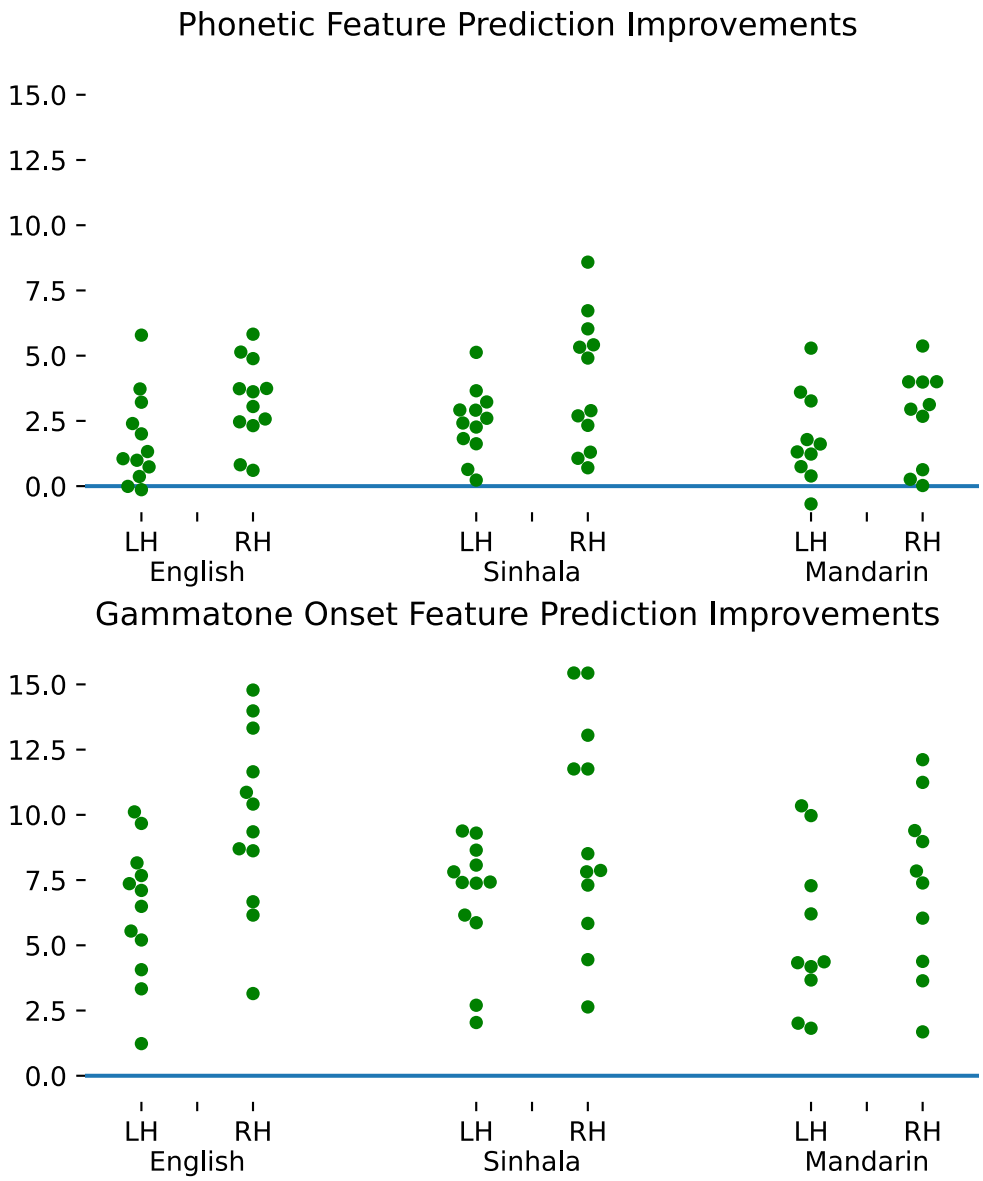


Figure 5.14: Total improvement in model fit when gammatone onsets and phonetic features are included in the TRF analysis.

constructing TRFs that correspond to each phoneme, by summing across TRFs for the features present in that phoneme. It is then possible to compare cortical responses to different phonemes across participants and language groups. This analysis does not produce any significant results and I was not able to replicate results from the previous work. In addition, there are no significant differences when looking at individual phonemic contrasts that exist in English, but not in the native language of the listeners.

This finding is not consistent with the results of Di Liberto et al. (2021) who see that the phonetic space of second language learners becomes closer to that of native speakers as proficiency increases. Despite the fact that the previous analysis was conducted on EEG data rather than MEG data, with similar temporal resolution and increased spatial resolution, it may be expected that the same analysis could be performed on this data. It is possible that there are no significant results from this analysis because we do not have varying levels of proficiency. The lack of differences may be because these listeners in this study already have well-formed representations of all of the phonetic categories of English—ie. unlike the previous study, all of these speakers had perhaps reached the endpoint of their learning. However one would still expect that there would be differences between each language group and between contrasts that exist or do not exist in the native language of participants. This may be expected given the persistence of native language effects in other more controlled discrimination paradigms, such as the /r/-/l/ contrast for native Japanese speakers, where effects are observed even when one has reached the end point of English learning. This clearly requires much future work to investigate this data and see if the lack of differences here is purely methodological or down to the properties of this participant group. I discuss some of these possibilities below and in the general discussion.

5.6 Summary

This chapter looks at how the phonetic categories in second language learners can be used during online speech processing, particularly if second language learners can use these representations to form predictions at various context levels—sublexical, lexical and sentence. One may think that given the previous results in this dissertation showing the difficulties of updating the perceptual system to accommodate a second language, second language learners may not be able to use the phonemic representations to predict upcoming material to the extent of native speakers. This is in fact borne out in previous research on prediction, showing that second language learners do not appear to use context to anticipate upcoming information to the extent of native speakers.

By looking at neural responses during a naturalistic listening paradigm, we see very similar responses in second language learners as we see in native speakers, illustrating that they may be using the same underlying predictive processing during online speech processing. While this seemingly conflicts with previous results that show that native and second language learners have different neural signatures of prediction, I follow the hypothesis that these results could be due to task-related effects or differences in knowledge of lexical and phoneme distributions in the second language. It is clear that there is much work to reconcile these findings.

These results have interesting implications with the work discussed in previous chapters of this dissertation. The participants in the current dataset were highly proficient of English, but were not exposed to English early in life and were not native speakers. However they were still able to exhibit neural responses to prediction which closely map to those of native speakers. This is in contrast to the results discussed at the very beginning of this dissertation and what Chapters

3 and 4 deal with that second language learners struggle to reach native-level performance on the discrimination of speech sounds in their second language. While these results are not necessarily inconsistent it is worth considering what implications this conflict has for theories of speech processing.

Chapter 6: General Discussion

In this dissertation, I have discussed phonetic category learning for second language learners, the effective training techniques, and a proposal that different training techniques update different information in the mapping from acoustic to phonetic information. I then examined the usage of these categories and demonstrated that in prediction in a naturalistic setting, second language learners exhibit responses similar to native speakers.

I have provided a clear computational model for how humans learn speech sounds and auditory tokens more generally within an implicit video game paradigm, suggesting that little algorithmic advantage exists for using reinforcement learning over other mechanisms in this situation. I use these algorithms to motivate a proposal where there exists a split in the knowledge of the acoustic-phonetic mapping of speech sounds where updates are made by a striatal reflexive learning system and a frontal reflective learning system as proposed by dual systems models of auditory category learning. My proposal states that the striatal system enables the formation of a perceptual space for acoustic input, learning the auditory dimensions and cues necessary to classify the stimulus. The frontal system is able to form discrete categories over this space. I explain how this proposal fits with multiple previous research areas and provide a simulation and an experiment to demonstrate how these ideas play out. This is one of few proposals which specifically addresses *why* different category boundary types are best learned by each of these

systems.

I conclude by looking at how second language learners are able to predict upcoming phonetic input using different levels of context and show that neural responses to prediction are remarkably similar to those of native speakers, despite evidence that indicates fewer markers of prediction in second language learners in controlled experimental settings. I discuss that these results can be complementary under a theory where listeners have different representations of the statistical environment of a second language. I demonstrate the merits of naturalistic continuous listening studies and provide a new corpus of second language learners listening to an audiobook while MEG responses are recorded.

I see this dissertation as being at the representational/algorithmic level of Marr's proposed levels of analysis (Marr, 1982). Through this work, I take influence from theories at the computational level, such as linguistic theory, as well as neural implementation. However, this dissertation itself makes proposals specifically about the representations of speech sounds in second language learners and the nature of the algorithms that can apply.

This work offers some important observations into the use of modeling in speech perception research. Modeling studies have been a large part of this field for many years, and the simulations here provide two specific examples of how advances in computational research can provide insight into recent theories.

First, the choice of algorithm to model the video game paradigm was founded on specific experimental research into the effectiveness of the paradigm, including the neural circuits engaged and the DLS and COVIS models. The simulations provided by this model then in turn influenced experimental work and the proposal of additions to the current theory. While modeling work on its own is useful in discovering cognitive processes, it can be particularly powerful in

a feedback loop with experimental research. While the particular experiment presented here did not yield significant results, I hope that future work in this area can continue to use computational and experimental work hand-in-hand to explore this area of human learning.

Secondly, the analysis presented in Chapter 5 of this dissertation demonstrates a separate use for advances in computational work—the ability to extract structure from a continuous neural signal. This type of analysis has shown increasing promise over recent years and this dissertation shows once again that it can yield useful results that may not be possible through other methods. While carefully constructed experimental paradigms have a place in research, these paradigms are important to test theories about speech processing in a more ecological environment and investigate whether ideas are still consistent during online processing.

6.1 Future Directions

6.1.1 Phonetic Representations and Continuous Neural Data

It is clear that listeners have difficulty learning phonetic categories in a second language. Although much work has been dedicated to studying what these representations could look like and what the most effective training techniques are, little work has aligned this with the usage of these categories during online processing. Given that the representations of sounds in a second language do not support some aspects of perceptual knowledge, such as discrimination effects, it might be surprising that it is possible to build such neural responses to prediction from these representations.

For example, take the surprisal and entropy calculations used to decode neural responses in Chapter 5. If a second language learner only has one category for English /r/ and /l/, then their

probability distributions that are used to calculate the probability of a given upcoming phoneme will not be the same as that of a native speaker. Even if they do have two different phonetic categories for these sounds, if the boundary between them is not the same as native speakers or they are assigning sounds to each category at chance, then a different probability distribution will arise.

It could be the case that these differences make little difference ultimately in online perception, and the effects may not make a large difference overall when listening to speech. For example, over the course of my experiment, any differences in neural responses to sounds that are not present in the native language of participants could average out over the 45 minutes of analysis. However, it is important to remember that even sounds that are discriminable for second language learners could have slightly different representations to native speakers. While these contrasts may be easier to learn for a second language than contrasts where the critical dimension is not important in the native language, the small differences could still impact perception and should be studied in the context of online speech processing. I propose some strands of future research that could investigate the intersection of these different processes.

First, more analysis should be conducted on the continuous data gathered in this dissertation. While the languages recorded here do not have well-documented contrasts in English that are not in Sinhala and Mandarin (ie. there is no such developed research program such as the /r/-/l/ distinction), one could look into generating a model of English where the phonemes used as building blocks are those that are well represented in the native language. For example, for the phoneme context model used here, one could map each phoneme in the English corpus to the closest phoneme in Sinhala and use this to calculate surprisal and entropy. This would show whether the representations from the native language are still being used during processing and

whether the representations used to build predictions are any different from those of a native English speaker.

This, however is likely not nuanced enough to uncover differences in processing between native and non-native speakers fully, and I believe this raises a larger issue in decoding phonetic and phoneme representations in continuous naturalistic listening studies—what are the phonetic representations that we should look for in the brain? In the analysis presented here, I cannot uncover significant responses to phonetic features; at least this does not explain any more of the data than using simple gammatone onset acoustic features. Although this has been successful in the past with EEG recordings (Di Liberto et al., 2021), these may not be the best features to encode phonetic processing from the acoustic signal. While the literature in phonetic learning has considered the mapping from features to phonemes at great length and phonetic features are a foundational building block in linguistic theory, it is not clear that the naturalistic literature has yet found the best way to decode this information from continuous data. More consideration should be put into decoding acoustic-perceptual mapping from this neural data.

One example that could provide some insight would be to look at models that are built on data from a specific language and are designed to capture the acoustic environment of a language while remaining agnostic to the specific discrete phonetic categories that are present. One example would be a Gaussian Mixture Model, which has been shown to capture specific effects of infant phonetic learning (Schatz et al., 2019, 2021). This model presumes that each frame of data in the acoustic signal is generated from one of a number of Gaussian distributions and, after training, can provide a posterior over the distributions that an individual frame is likely to be generated. The model is not supervised by phonetic categories or features and does not depend on precise and potentially subjective definitions of these features. After training a model like this

on Sinhala and Mandarin, it could be seen whether second language learners may still be representing some information through the lens of their native language when forming representations and calculating predictions over the incoming signal.

This also could provide a perspective that may align with my proposal of the Dual Learning Systems model of speech category learning. I suggest that the two different learning systems may update different parts of the acoustic-phonetic mapping with a reflexive system updating the perceptual space and learning the dimensions over which the reflective system can categorize. This could tie into the different representations that can be decoded from the continuous neural data. The representations of prediction that can be extracted and built upon are the discrete phonetic categories that would be learned by the reflective system. The phonetic space that models such as the Gaussian Mixture model could decode are those that are updated by the reflexive system. There is still work required to combine these ideas explicitly, but I think there are important considerations regarding the exact representations used during online processing.

Similarly, examining how second language listeners use information during processing is important in uncovering the exact representations that they have. Even if these listeners are able to categorize and discriminate sounds in the second language successfully, it does not mean that they have the same representations as native speakers. Looking further at what predictions are made when listening to speech and whether these differ from native speakers will show whether the second language learners have the same representations as native speakers or whether they are able to successfully perform identification and discrimination successfully, but by using other methods.

6.1.2 Motivation and Reward

One facet of implicit learning paradigms that requires further development is the exact role of reward and motivation. As implicit learning and reinforcement learning are generally discussed, reward is a specific feature of performing a correct action rather than a motivating factor to perform the task well. This is reasonable when considering models of the paradigm since reward is specifically an attribute of reinforcement learning and the circuits which perform this type of learning. However, given that dopamine could play a role in motivation more generally, there are other effects of these paradigms that could be missing from our current thinking. For example, anecdotal evidence from running participants in the experiment from Chapter 4 suggests that participants have a better overall experience when in the implicit condition versus the explicit condition. These tasks, even when pared down to a simple version that closely aligns with explicit learning paradigms, may engage participants to an extent that explicit paradigms do not, leading to better overall performance in the experiment. This is not the sole reason for the results in the literature and the DLS model as a whole since there are many attributes to the different learning systems. However, this is not something that should be neglected when building theories of second language speech sound learning. Given the rise of language learning apps which provide many avenues to motivate someone to continue learning, these rewards should not be discounted and could play a role in second language learning.

Another consideration with reward and motivation is to extend these models to cover paradigms where a distinct reward is not defined. In the video game paradigm, it is clear that a learning signal could come from successfully shooting an alien, however there may be situa-

tions where this system is engaged with a less obvious external reward. This is where the idea of intrinsic reward could play a role. This is a recent idea in reinforcement learning where reward is internal to an agent where the reward can come from the result of the action itself rather than the environment. One example is the 'Diversity Is All You Need' (DIAYN) framework (Eysenbach, Gupta, Ibarz, & Levine, 2018), where an agent is incentivized to explore a solution space to the widest extent possible and reaching new states is inherently rewarding.

In speech-sound learning, these internal rewards could come from higher levels of the linguistic hierarchy. For example, perhaps a reward for successfully categorizing phonemes is that this helps to be able to recognize common strings of phonemes in the input and eventually words. Although the work I show here is restricted to these domains with a clear reward, the algorithms could be expanded in some way to account for this.

In order to expand on this, however, it is also important to consider why more general rewards may not show the same learning outcomes. As discussed in Chapter 3, 'everyday rewards' do not seem to activate the reinforcement system in the same way as implicit paradigms. This means that rewards stemming from knowing a language better, or improving understanding don't seem to have the same effect as rewards gathered from a controlled paradigm. Given the importance of task-related actions in engaging striatal circuits in experimental settings, it appears that this circuitry may require actions that are specifically correlated with reward—more nebulous rewards such as 'better understanding' that could be relevant in a more naturalistic setting do not have this property. The ideas have not been explored thoroughly though, and to fully recognize the role of these mechanisms, future work must address these questions.

6.1.3 Infant Phonetic Learning

In the future, I hope to expand the algorithms presented in this dissertation to model native language learning in infants, as they appear to acquire speech sound categories in a passive environment without direct feedback. Studies on auditory category learning show that adult listeners can learn one-dimensional stimuli through passive exposure. However, they usually require explicit feedback to learn auditory stimuli that vary on multiple dimensions (Goudbeek et al., 2009; Yi & Chandrasekaran, 2016). When infants learn the sounds of their native language, however, they are not told what specific sounds belong in what phoneme categories. While they receive little to no direct feedback from caregivers, they can still learn to discriminate speech sounds at a young age. Many theories posit that they may accumulate statistical information about their acoustic environment (Vallabha & McClelland, 2007), although there are some effects that purely statistical learning cannot account for (Nixon & Tomaschek, 2021). The idea of 'intrinsic reward' within reinforcement learning literature posits algorithms where rewards come from within the system rather than from the external environment, (Singh, Lewis, Barto, & Sorg, 2010). For example, an algorithm that explores a wide area of the solution space may receive a reward for doing just that, even if it does not take rewarding actions (Eysenbach et al., 2018). These ideas could perhaps be applied to infant phonetic learning, where rewards are provided from higher levels of cognition, such as being able to form categories that allow for successful word and pattern recognition. This would provide some explanation of how they are able to learn speech sound categories effectively at a young age without explicit feedback.

Reinforcement learning may also provide a model of experimental paradigms typically used with infants. A common infant testing procedure is the conditioned headturn (CHT) paradigm,

where an infant on a caregiver's lap is trained to turn their head to a toy whenever they notice a change in a stimulus. When they do this correctly, they are rewarded with the toy lighting up and making a sound. This very easily fits as a reinforcement learning algorithm where the learning signal during the experiment that the infant receives comes in the form of a reward for good actions. While these paradigms are typically thought to reflect pre-existing knowledge that the infant has, the effectiveness of a reward-based paradigm in adults may lead us to consider that infants learn during the experiment as well.

6.1.4 Algorithmic Differences

In this dissertation, I use one specific implementation of reinforcement learning—Deep Q Learning. This type of learning is best suited to the video game paradigm as the algorithm has clear counterparts of state and action within the experiment. Many other reinforcement learning algorithms exist, however, including TD-Learning (Sutton & Barto, 1998) and Actor-Critic Models (Konda & Tsitsiklis, 1999). Until now, most work looking at reinforcement learning algorithms in speech has used different algorithms, with little coherent theory of how these models should be used in simulating cognitive processes.

There are now several studies that point to the importance of reinforcement learning algorithms in speech sound learning, including the work presented in this dissertation, work showing that statistical learning processes could be supported by reinforcement learning principles (Orpella et al., 2021) and that reinforcement learning captures specific cue learning effects (Nixon, 2020; Harmon et al., 2019). Given the body of work that points to the use of these algorithms, it is now time for the field to work towards a combined theory of the role that these algorithms

play.

One specific goal that future work should aim to address is the specific definition of reinforcement learning in speech category learning. For example, in this dissertation, I compare reinforcement learning to supervised learning throughout since supervised learning would somewhat mirror an explicit learning paradigm. However, at its limit deep Q learning could be very similar to supervised learning. If we consider a situation where there are four actions - each corresponding to a specific category, then a reinforcement learning model that outputs actions aligns very closely with the architecture of a supervised learning model.

So what defines a reinforcement learning model in speech sound learning? This is an open area for investigation, and there may be terminology differences between cognitive science and recent uses of these algorithms in computer science. In speech category learning, previous literature seems to indicate that reinforcement learning is classified by learning that engages dorsal striatal circuits neural and for paradigms where actions are particularly important for successful learning. Further than this, however, many questions still remain to combine these disparate studies and generate a clear theory of the role that this learning type plays in speech category learning.

6.1.5 Extension to Other Domains

The Dual Learning Systems model originated with the COVIS model of competing learning systems, mostly studied in vision. Although the extension of this theory to auditory categories is a recent development, the findings from auditory categories and speech specifically could have implications for the theory as a whole. Speech category learning is a unique domain where there

are perceptual categories that are learned quickly in infancy and are robust to change later in life. The fact that it is so difficult to learn the speech categories in a second language provides a particular case that these learning systems must tackle, and through the application of competing learning systems to this situation, we may learn more about these theories generally that we may not otherwise.

For example, the split between the learning systems that I propose could apply more widely to other domains. Perhaps there are similar examples in vision processing, where the implicit system is able to learn earlier stages of visual mapping, and an explicit system generates discrete categories. It is possible that some of these findings cannot be found in vision alone since visual categories do not come with the same biases that language does. Once the DLS theories and the extension I propose here are investigated further, these findings should be considered in the context of the original COVIS model.

Appendix A: Reinforcement Learning

A.1 Elastic Weight Consolidation

Elastic weight consolidation Kirkpatrick et al. (2017) is designed to overcome the problem of 'catastrophic forgettting' when training a network on more than one task. When a network is trained on multiple tasks consecutively, it is possible that a new task will cause the network to no longer be able to perform a previous task—essentially 'overwriting' the weights learned in the original task. In this paper, the problem manifests in simulation 2, where video game training can update the weights of the network in such a way that it loses its pretraining ability and is no longer able to discriminate phones. I implement the EWC during video game training to counteract this problem. This involves adding a regularization term to the loss function as follows.

During video game training we start with the loss function $L_{vg}(\theta)$ which represents the smooth L1 loss over the current parameters θ. We add to this a function P(θ) that penalizes parameters further from those learned during pretraining.

$$L(\theta) = L_G(\theta) + P(\theta) \tag{A.1}$$

The penalty term is calculated by taking the distance between the current parameters θ and the parameters after phone classification pretraining θ^*_C. Each individual weight is multiplied by

the Fischer information coefficient F, representing the importance of the given weight to the classification task. The penalty is multiplied by a weight factor β which determines how closely the model should stay to the original task. This results in a final loss function of

$$L(\theta) = L_G(\theta) + \beta \sum_i F_i(\theta_i - \theta^*_{C,i})^2 \tag{A.2}$$

Appendix B: Experiment D-Prime Tables

B.1 Implicit Condition

		F3 Frequency (Hz)						
		210	**230**	**250**	**260**	**270**	**290**	**310**
Minimum	**Pre Test**	-0.23	-1.18	-1.18	-0.88	-0.75	0.00	-1.65
	Mid Test	-1.18	-1.18	-0.43	-1.18	-0.97	-0.65	-1.06
	Post Test	-1.94	-1.86	-2.33	-2.12	-1.82	-1.86	-2.42
Maximum	**Pre Test**	4.65	3.71	3.00	3.29	3.00	3.48	2.76
	Mid Test	4.65	4.65	4.65	3.00	4.65	4.65	4.65
	Post Test	2.12	1.38	2.53	1.12	2.60	3.01	3.00
Average	**Pre Test**	1.41	1.30	0.77	0.78	1.21	1.09	0.99
	Mid Test	1.09	0.98	1.11	0.81	1.23	1.21	0.94
	Post Test	-0.32	-0.32	0.34	0.03	0.02	0.12	-0.04
Standard Deviation	**Pre Test**	1.42	1.39	1.36	1.26	1.23	1.13	1.33
	Mid Test	1.43	1.49	1.45	1.09	1.67	1.71	1.68
	Post Test	1.00	0.86	1.48	0.84	1.23	1.28	1.62

Table B.1: Implicit Condition D-Prime Results: Minimum, Maximum, Mean and Standard Deviation for participants in the implicit condition. Mid-Test follows implicit learning and Post-Test follows explicit learning.

B.2 Explicit Condition

		F3 Frequency (Hz)						
		210	**230**	**250**	**260**	**270**	**290**	**310**
Minimum	**Pre Test**	-0.43	-0.29	0.00	-0.71	0.11	0.00	-0.36
	Mid Test	-0.29	-1.65	-0.67	-0.29	-1.06	-0.53	-1.18
	Post Test	-1.85	-3.55	-1.42	-2.19	-1.47	-1.23	-2.60
Maximum	**Pre Test**	2.76	4.65	3.71	1.90	3.71	3.71	3.29
	Mid Test	4.65	3.71	3.29	2.76	4.65	3.71	4.65
	Post Test	1.90	1.65	1.47	2.08	1.90	0.67	1.59
Average	**Pre Test**	1.13	1.42	1.13	0.69	1.23	1.28	1.26
	Mid Test	1.36	1.15	0.91	0.83	0.96	1.05	0.88
	Post Test	0.23	-0.27	-0.22	0.14	-0.26	-0.24	-0.38
Standard Deviation	**Pre Test**	0.99	1.34	1.19	0.69	0.98	1.16	1.03
	Mid Test	1.34	1.45	1.06	1.04	1.42	1.20	1.51
	Post Test	1.32	1.39	0.75	1.21	0.84	0.61	1.26

Table B.2: Explicit Condition D-Prime Results: Minimum, Maximum, Mean and Standard Deviation for participants in the explicit condition. Mid-Test and Post-Test both follow blocks of explicit training.

Appendix C: Additional Neural Analyses

C.1 Acoustic Predictor TRFs

Responses to acoustic predictors show similar effects to those of the linguistic predictors discussed in Chapter 5. Slightly higher TRFs in left hemisphere are recorded for second-language learners of English over native speakers, exhibiting the same lateralization responses as linguistic predictors.

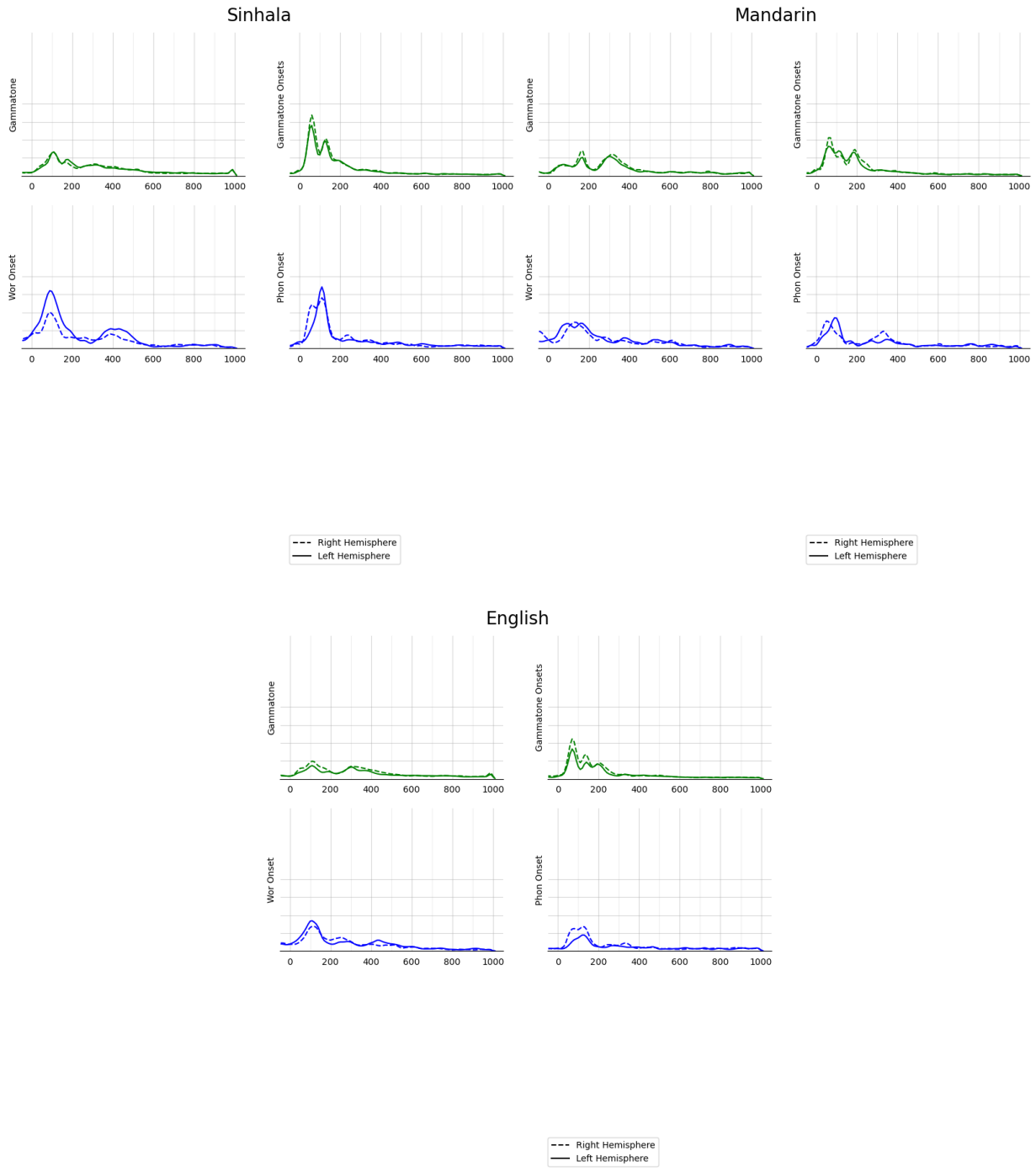


Figure C.1: Average of all subject acoustic/baseline TRFs combined over all frontotemporalparietal voxels separated by native language

C.2 Effects of Proficiency

Both LexTale score and Cloze task performance are plotted for each second language listener against the height of the peak (Figure C.2) and the timepoint of this peak (Figure C.3) for each TRF in the full model. Peaks are calculated over the first 250ms. There is no clear correlation between either of these proficiency measures and the profile of the TRF.

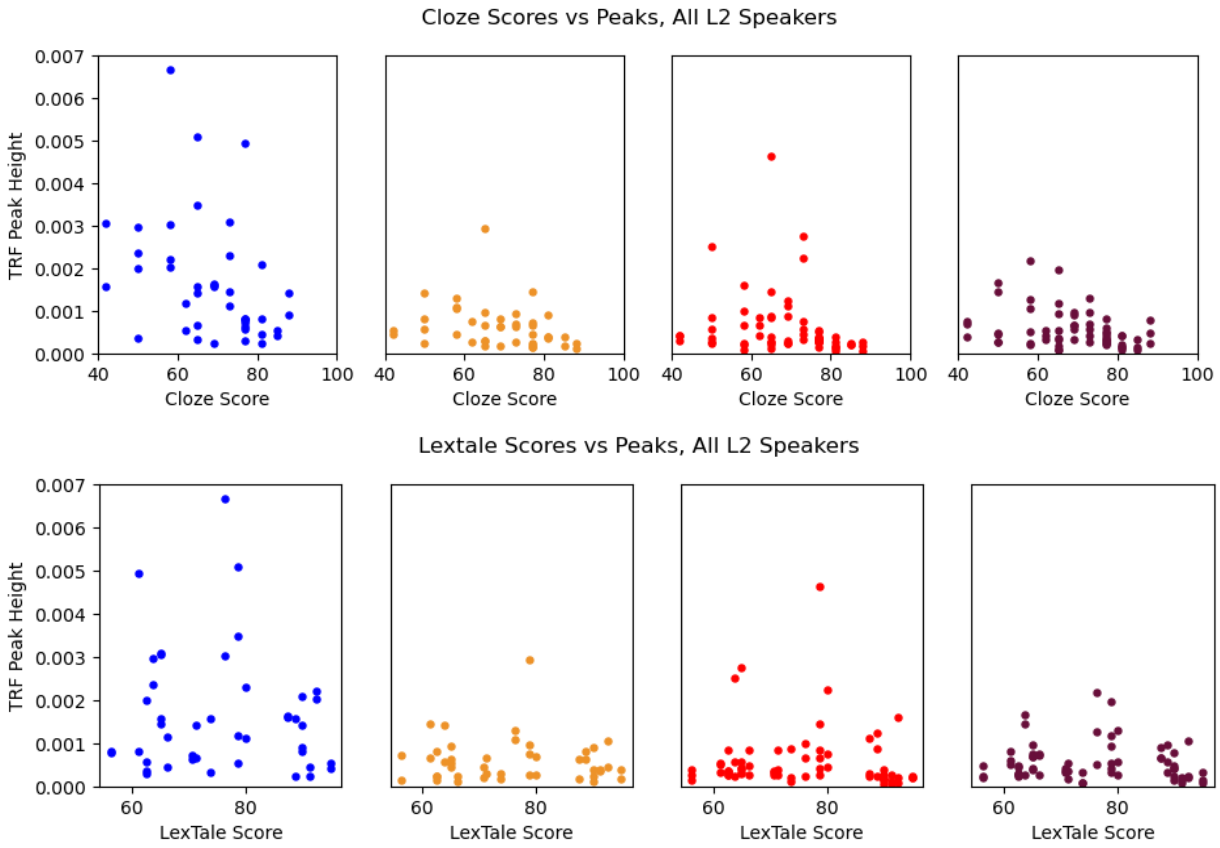


Figure C.2: Height of peak in TRF signal plotted against cloze and lextale proficiency measures for each feature in the full model: Surprisal (circle), Cohort Entropy (diamond), Phoneme Entropy (plus), Phoneme Onset (plus) and Word Onset (x)

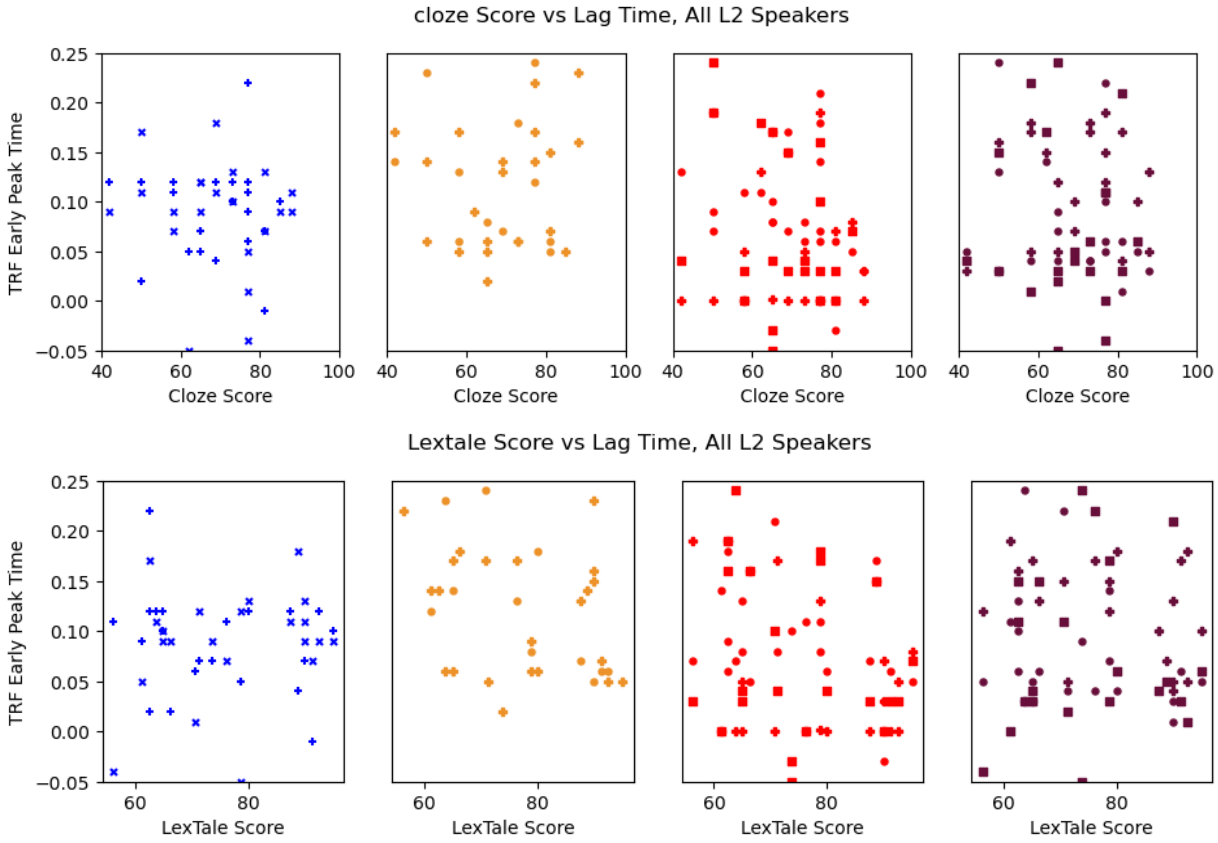


Figure C.3: Time lag of peak in TRF signal plotted against cloze and lextale proficiency measures for each feature in the full model: Surprisal (circle), Cohort Entropy (diamond), Phoneme Entropy (plus), Phoneme Onset (plus) and Word Onset (x)

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). *Memory Aware Synapses: Learning what (not) to forget.* arXiv. doi: 10.48550/arXiv.1711.09601

Antetomaso, S., Miyazawa, K., Feldman, N., Elsner, M., Hitczenko, K., & Mazuka, R. (2017). Modeling Phonetic Category Learning from Natural Acoustic Data.

Aoyama, K., & Flege, J. E. (2011). Effects of L2 Experience on Perception of English /r/ and /l/ by Native Japanese Speakers. *Journal of the Phonetic Society of Japan*, *15*(3), 5–13. doi: 10.24467/onseikenkyu.15.3_5

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481. doi: 10.1037/0033-295x.105.3.442

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125. doi: 10.3758/BF03196132

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*(1), 147–161. doi: 10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199. doi: 10.3758/BF03207622

Bell, A. J., & Sejnowski, T. J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, *7*(6), 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Bellman, R. (1957). A Markovian Decision Process. *Journal of Mathematics and Mechanics*, *6*(5), 679–684.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA, US: The MIT Press.

Bhattasali, S., Fabre, M., Luh, W.-M., Al Saied, H., Constant, M., Pallier, C., … Hale, J. (2019). Localising memory retrieval and syntactic composition: an fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, *34*(4), 491–510. doi: 10.1080/23273798.2018.1518533

Bion, R. A. H., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning Phonemic Vowel Length from Naturalistic Recordings of Japanese Infant-Directed Speech. *PLOS ONE*, *8*(2), e51594. doi: 10.1371/journal.pone.0051594

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

Brennan, J. R., & Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions

during naturalistic listening. *PLOS ONE*, *14*(1), e0207741. doi: 10.1371/journal.pone.0207741

Brodbeck, C. (2023a). *Eelbrian.*

Brodbeck, C. (2023b). *TRF-Tools.*

Brodbeck, C., Bhattasali, S., Cruz Heredia, A. A., Resnik, P., Simon, J. Z., & Lau, E. (2022). Parallel processing in speech perception with local and global representations of linguistic context. *eLife*, *11*, e72056. doi: 10.7554/eLife.72056

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976–3983.e5. doi: 10.1016/j.cub.2018.10.042

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, *172*, 162–174. doi: 10.1016/j.neuroimage.2018.01.042

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *The Journal of the Acoustical Society of America*, *117*(6), 3890–3901. doi: 10.1121/1.1906060

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *Quarterly Journal of Experimental Psychology*, *64*(1), 74–95. doi: 10.1080/17470218.2010.499174

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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. doi: 10.3758/BRM.41.4.977

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, *5*.

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495. doi: 10.3758/s13423-013-0501-5

Chomsky, N. (1959). Review of B.F. Skinner, Verbal Behaviour. *Language*, *35*(1), 26–28.

Chomsky, N., & Halle, M. (1968). *THE SOUND PATTERN OF ENGLISH* (Tech. Rep.).

Clahsen, H., & Felser, C. (2006a). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, *27*(1), 107–126. doi: 10.1017/S0142716406060206

Clahsen, H., & Felser, C. (2006b). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42. doi: 10.1017/S0142716406060024

Clahsen, H., & Felser, C. (2006c). How native-like is non-native language processing? *Trends in Cognitive Sciences*, *10*(12), 564–570. doi: 10.1016/j.tics.2006.10.002

Cohen, M. X., & Frank, M. J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research*, *199*(1), 141–156. doi: 10.1016/j.bbr.2008.09.029

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. doi: 10.1038/s41586-019-1924-6

Davies, M. (2017). *Corpus of Contemporary American English (COCA).* Harvard Dataverse. doi: 10.7910/DVN/AMUDUW

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., ... Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex

on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021

Di Liberto, G. M., Attaheri, A., Cantisani, G., Reilly, R. B., Choisdealbha, N., Rocha, S., … Goswami, U. (2022). *Emergence of the cortical encoding of phonetic features in the first year of life.* bioRxiv. doi: 10.1101/2022.10.11.511716

Di Liberto, G. M., Nie, J., Yeaton, J., Khalighinejad, B., Shamma, S. A., & Mesgarani, N. (2021). Neural representation of linguistic feature hierarchy reflects second-language proficiency. *NeuroImage*, *227*, 117586. doi: 10.1016/j.neuroimage.2020.117586

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7), 961–974. doi: 10.1016/S0893-6080(99)00046-5

Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., & Gerfen, C. (2013). WHEN GENDER AND LOOKING GO HAND IN HAND: Grammatical Gender Processing In L2 Spanish. *Studies in Second Language Acquisition*, *35*(2), 353–387. doi: 10.1017/S0272263112000915

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is All You Need: Learning Skills without a Reward Function. *arXiv:1802.06070 [cs]*.

Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do Infants Really Learn Phonetic Categories? *Open Mind*, *5*, 113–131. doi: 10.1162/opmi_a_00046

Feng, G., Gan, Z., Yi, H. G., Ell, S. W., Roark, C. L., Wang, S., … Chandrasekaran, B. (2021). Neural dynamics underlying the acquisition of distinct auditory category structures. *NeuroImage*, *244*, 118565. doi: 10.1016/j.neuroimage.2021.118565

Ferreira, F., & Chantavarin, S. (2018). Integration and Prediction in Language Processing: A Synthesis of Old and New. *Current Directions in Psychological Science*, *27*(6), 443–448. doi: 10.1177/0963721418794491

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, *69*, 165–182. doi: 10.1016/j.jml.2013.06.001

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Fishbach, A., Nelken, I., & Yeshurun, Y. (2001). Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. *Journal of Neurophysiology*, *85*(6), 2303–2323. doi: 10.1152/jn.2001.85.6.2303

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3–83). Cambridge: Cambridge University Press. doi: 10.1017/9781108886901.002

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366. doi: 10.1037/0096-1523.28.2.349

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental Auditory Category Learning. *Journal of experimental psychology. Human perception and performance*, *41*(4), 1124–1138. doi: 10.1037/xhp0000073

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal Predictive Codes for Spoken Words in Auditory Cortex. *Current Biology*, *22*(7), 615–621. doi: 10.1016/j.cub.2012.02.015

Gaston, P., & Marantz, A. (2018). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single "clash" as noun or

verb. *Language, Cognition and Neuroscience*, *33*(4), 402–423. doi: 10.1080/23273798.2017.1395466

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506. doi: 10.1016/j.neuroimage.2003.09.071

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "l" and "r.". *Neuropsychologia*, *9*(3), 317–323. doi: 10.1016/0028-3932(71)90027-3

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of experimental psychology. Human perception and performance*, *35*(6), 1913–1933. doi: 10.1037/a0015781

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. doi: 10.1016/j.neuroimage.2013.10.027

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2, Pt 1), 1111–1121. doi: 10.1121/1.416296

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: the case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*, *107*(5 Pt 1), 2711–2724. doi: 10.1121/1.428657

Hahne, A. (2001). What's Different in Second-Language Processing? Evidence from Event-Related Brain Potentials. *Journal of Psycholinguistic Research*, *30*(3), 251–266. doi: 10.1023/A:1010490917575

Hahne, A., & Friederici, A. D. (2001). Processing a second language: late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, *4*(2), 123–141. doi: 10.1017/S1366728901000232

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88. doi: 10.1016/j.cognition.2019.03.011

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 187–197). Edinburgh, Scotland: Association for Computational Linguistics.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. doi: 10.1038/nrn2113

Hikosaka, O., Kim, H. F., Yasuda, M., & Yamamoto, S. (2014). Basal ganglia circuits for reward value-guided behavior. *Annual review of neuroscience*, *37*, 289–306. doi: 10.1146/annurev-neuro-071013-013924

Holliday, J. J. (2015). A longitudinal study of the second language acquisition of a three-way stop contrast. *Journal of Phonetics*, *50*, 1–14. doi: 10.1016/j.wocn.2015.01.004

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307. doi: 10.1177/0267658315624960

Idemaru, K., & Holt, L. L. (2011). Word Recognition Reflects Dimension-based Statistical Learning. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(6), 1939–1956. doi: 10.1037/a0025641

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*,

40, 1009–1021. doi: 10.1037/a0035269

Ihara, A. S., Matsumoto, A., Ojima, S., Katayama, J., Nakamura, K., Yokota, Y., ... Naruse, Y. (2021). Prediction of Second Language Proficiency Based on Electroencephalographic Signals Measured While Listening to Natural Speech. *Frontiers in Human Neuroscience*, *15*, 665809. doi: 10.3389/fnhum.2021.665809

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: a comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. doi: 10.1121/1.2062307

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. doi: 10.1016/S0010-0277(02)00198-1

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks: The Official Journal of the International Neural Network Society*, *15*(4-6), 535–547. doi: 10.1016/s0893-6080(02)00047-3

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011. doi: 10.1111/j.1467-7687.2011.01052.x

Jurafsky, D. (1996). A Probabilistic Model of Lexical and Syntactic Access and Disambiguation. *Cognitive Science*, *20*(2), 137–194. doi: 10.1207/s15516709cog2002_1

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?* :. *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. doi: 10.1075/lab.4.2.05kaa

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*(1), 46–54. doi: 10.1016/j.bandl.2012.11.007

Kawagoe, R., Takikawa, Y., & Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience*, *1*(5), 411–416. doi: 10.1038/1625

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526. doi: 10.1073/pnas.1611835114

Konda, V., & Tsitsiklis, J. (1999). Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems* (Vol. 12). MIT Press.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*, 93–107. doi: 10.3758/BF03212211

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21. doi: 10.1111/j.1467-7687.2006.00468.x

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science (New York, N.Y.)*, *255*(5044), 606–608. doi: 10.1126/science.1736364

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. doi: 10.1152/jn.90896.2008

Lanciego, J. L., Luquin, N., & Obeso, J. A. (2012). Functional Neuroanatomy of the Basal Ganglia. *Cold Spring Harbor Perspectives in Medicine*, *2*(12), a009621. doi: 10.1101/

cshperspect.a009621

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*. doi: 10.1016/j.cognition.2020.104328

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. doi: 10.3758/s13428-011-0146-0

Levy, E. S. (2009). On the assimilation-discrimination relationship in American English adults' French vowel learning. *The Journal of the Acoustical Society of America*, *126*(5), 2670–2682. doi: 10.1121/1.3224715

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of memory and language*, *63*(4), 447–464. doi: 10.1016/j.jml.2010.07.003

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex (New York, N.Y.: 1991)*, *15*(10), 1621–1631. doi: 10.1093/cercor/bhi040

Lim, S. J., Fiez, J., & Holt, L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences*, *116*, 201811992. doi: 10.1073/pnas.1811992116

Lim, S. J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience*, *8*. doi: 10.3389/fnins.2014.00230

Lim, S. J., & Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science*, *35*(7), 1390–1405. doi: https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242–1255. doi: 10.1121/1.408177

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076–2087. doi: 10.1121/1.410149

Llanos, F., German, J. S., Gnanateja, G. N., & Chandrasekaran, B. (2021). The neural processing of pitch accents in continuous speech. *Neuropsychologia*, *158*, 107883. doi: 10.1016/j.neuropsychologia.2021.107883

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

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*, 309–332. doi: 10.1016/j.beproc.2004.03.011

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

Matusevych, Y., Schatz, T., Kamper, H., Feldman, N. H., & Goldwater, S. (2020). Evaluating computational models of infant phonetic learning across languages. *arXiv preprint*

arXiv:2008.02888.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–111. doi: 10.1016/s0010-0277(01)00157-3

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech 2017* (pp. 498–502). ISCA. doi: 10.21437/Interspeech.2017-1386

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/-/l/ discrimination to Japanese adults: behavioral and neural aspects. *Physiology & Behavior*, *77*(4-5), 657–662. doi: 10.1016/s0031-9384(02)00916-2

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: what's wrong, what's right, and what's left? *Trends in cognitive sciences*, *16*(5), 269–276. doi: 10.1016/j.tics.2012.04.006

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience*, *7*(7), 703–704. doi: 10.1038/nn1264

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain Potentials Reveal Discrete Stages of L2 Grammatical Learning. *Language Learning*, *60*(s2), 123–150. doi: 10.1111/j.1467-9922.2010.00604.x

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 374–388.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331–340. doi: 10.3758/BF03211209

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, *9*(8), 1057–1063. doi: 10.1038/nn1743

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081. doi: 10.1016/j.cognition.2019.104081

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition*, *212*, 104697. doi: 10.1016/j.cognition.2021.104697

Noguchi, M., & Kam, C. L. H. (2018). The Emergence of the Allophonic Perception of Unfamiliar Speech Sounds: The Effects of Contextual Distribution and Phonetic Naturalness. *Language Learning*, *68*(1), 147–176. doi: 10.1111/lang.12267

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395. doi: 10.1037/0033-295X.115.2.357

Obasih, C. O., Luthra, S., Dick, F., & Holt, L. L. (2022). *Auditory category learning is robust across training regimes.* PsyArXiv. doi: 10.31234/osf.io/ygwhd

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. *Science*, *304*(5669),

452–454. doi: 10.1126/science.1094285

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. doi: 10.1016/0028-3932(71)90067-4

Omaki, A., & Schulz, B. (2011). Filler-Gap Dependencies and Island Constraints in Second-Language Sentence Processing. *Studies in Second Language Acquisition*, *33*(4), 563–588.

Orpella, J., Mas-Herrero, E., Ripollés, P., Marco-Pallarés, J., & Diego-Balaguer, R. d. (2021). Language statistical learning responds to reinforcement learning principles rooted in the striatum. *PLOS Biology*, *19*(9), e3001119. doi: 10.1371/journal.pbio.3001119

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). Proficiency Reporting Practices in Research on Second Language Acquisition: Have We Made any Progress? *Language Learning*, *72*(1), 198–236. doi: 10.1111/lang.12475

Paul, D. B., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, *41*(1), 245–255. doi: 10.1016/S0167-6393(02)00107-3

Pollan, M. (2002). *The Botany of Desire: A Plant's-Eye View of the World*. Random House Publishing Group.

Roark, C. L., & Chandrasekaran, B. (2023). Stable, flexible, common, and distinct behaviors support rule-based and information-integration category learning. *npj Science of Learning*, *8*(1), 1–19. doi: 10.1038/s41539-023-00163-0

Roark, C. L., & Holt, L. L. (2018a). *Perceptual dimensions influence auditory category learning*. PsyArXiv. doi: 10.31234/osf.io/r673v

Roark, C. L., & Holt, L. L. (2018b). Task and distribution sampling affect auditory category learning. *Attention, Perception, & Psychophysics*, *80*(7), 1804–1822. doi: 10.3758/s13414-018-1552-5

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The representational glue for incidental category learning is alignment with task-relevant behavior. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(6), 769–784. doi: 10.1037/xlm0001078

Roark, C. L., Plaut, D. C., & Holt, L. L. (2022). A neural network model of the effect of prior experience with regularities on subsequent category learning. *Cognition*, *222*, 104997. doi: 10.1016/j.cognition.2021.104997

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928. doi: 10.1126/science.274.5294.1926

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181–203. doi: 10.1146/annurev-psych-122216-011805

Schatz, T. (2016). *ABX-disciminability measures and applications* (Doctoral Dissertation). Universite Paris 6 (UPMC).

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., & Dupoux, E. (2019). *Early phonetic learning without phonetic categories – Insights from machine learning* (preprint). PsyArXiv. doi: 10.31234/osf.io/fc4wh

Schatz, T., & Feldman, N. H. (2018). Neural network vs . HMM speech recognition systems as models of human cross-linguistic phonetic perception..

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, *118*(7), e2001844118. doi: 10 .1073/pnas.2001844118

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 : 14th Annual Conference of the International Speech Communication Association* (pp. 1–5). Lyon, France.

Schultz, T., Vu, T., & Schlippe, T. (2013). GlobalPhone: A Multilingual Text & Speech Database in 20 Languages.. doi: 10.1109/ICASSP.2013.6639248

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219. doi: 10.1146/annurev.neuro.051508.135546

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. doi: 10.1016/j.wocn.2017.11.002

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82. doi: 10.1109/TAMD.2010.2051031

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22. doi: 10.1037/0033-2909.119.1.3

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. doi: 10.1016/j.neuroimage.2008.03.061

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. doi: 10.3758/BF03202673

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Swanson, L. W. (1982). The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Research Bulletin*, *9*(1-6), 321–353. doi: 10.1016/0361-9230(82)90145-9

Taulu, S., & Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine & Biology*, *51*(7), 1759. doi: 10.1088/0031-9155/51/7/008

Thorburn, C., Feldman, N., & Schatz, T. (2019). A quantitative model of the language familiarity effect in infancy. In *2019 Conference on Cognitive Computational Neuroscience*. Berlin, Germany: Cognitive Computational Neuroscience. doi: 10.32470/CCN.2019.1353-0

Toscano, J. C., & McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, *34*(3), 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(1), 53–73. doi: 10.3758/CABN.7.1.53

Van Heugten, M., Dahan, D., Johnson, E. K., & Christophe, A. (2012). Accommodating syntactic violations during online speech perception. In *Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY*. Citeseer.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950. doi: 10 .3758/BF03193621

Wanrooij, K., Boersma, P., & Van Zuijen, T. L. (2014). Fast phonetic learning occurs already in 2-to-3-month old infants: an ERP study. *Frontiers in psychology*, *5*, 77.

Weissler, R. E., Drake, S., Kampf, K., Diantoro, C., Foster, K., Kirkpatrick, A., ... Baese-Berk, M. M. (2023). Examining linguistic and experimenter biases through "non-native" versus "native" speech. *Applied Psycholinguistics*, 1–15. doi: 10.1017/S0142716423000115

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development*, *52*, 349–355. doi: 10.2307/1129249

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, *7*(1), 49–63. doi: 10.1016/S0163-6383(84)80022-3

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology. Human Perception and Performance*, *48*(9), 913–925. doi: 10.1037/xhp0001037

Yamada, R. A., Strange, W., Magnuson, J. S., Pruitt, J. S., & Clarke, W. D. (1994). The Intelligibility of Japanese Speakers' Production of American English /r/, /i/, and /w/, as Evaluated by Native Speakers of American English. In *Third International Conference on Spoken Language Processing (ICSLP 94)*.

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, *52*(4), 376–392. doi: 10.3758/BF03206698

Yeterian, E. H., & Pandya, D. N. (1998). Corticostriatal connections of the superior temporal region in rhesus monkeys. *The Journal of Comparative Neurology*, *399*(3), 384–402.

Yi, H. G., & Chandrasekaran, B. (2016). Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. *The Journal of the Acoustical Society of America*, *140*(2), 1332–1335. doi: 10.1121/1.4961163

Yoshida, K., Pons, F., Maye, J., & Werker, J. (2010). Distributional Phonetic Learning at 10Months of Age. *Infancy*, *15*, 420–433. doi: 10.1111/j.1532-7078.2009.00024.x