ABSTRACT

Title of dissertation: HOW TO USE CONTEXT FOR PHONETIC LEARNING AND PERCEPTION FROM NATURALISTIC SPEECH

Kasia Hitczenko
Doctor of Philosophy, 2019

Dissertation directed by: Professor Naomi Feldman
Department of Linguistics & UMIACS

Infants learn about the sounds of their language and adults process the sounds they hear, even though sound categories often overlap in their acoustics. This dissertation is about how contextual information (e.g. who spoke the sound and what the neighboring sounds were) can help in phonetic learning and speech perception. The role of contextual information in these tasks is well-studied, but almost exclusively using simplified, controlled lab speech data. In this dissertation, we study naturalistic speech of the type that listeners primarily hear.

The dissertation centers around two main theories about how context could be used: top-down information accounts, which argue that listeners use context to predict which sound will be produced, and normalization accounts, which argue that listeners compensate for the fact that the same sound is produced differently in different contexts by factoring out this systematic context-dependent variability from the acoustics. These ideas have been somewhat conflated in past research, and have rarely been tested on naturalistic speech. We start by implementing top-down and

normalization accounts separately and evaluating their relative efficacy on spontaneous speech, using the test case of Japanese vowel length. We find that top-down information strategies are effective even on spontaneous speech. Surprisingly, we find that normalization is ineffective on spontaneous speech, in contrast to what has been found on lab speech. We, then, provide analyses showing that when there are systematic regularities in which contexts different sounds occur in - which are common in naturalistic speech, but generally controlled for in lab speech - normalization can actually increase category overlap rather than decrease it. Finally, we present a new proposal for how infants might learn which dimensions of their language are contrastive that takes advantage of these systematic regularities in which contexts different sounds occur in. We propose that infants might learn that a particular dimension of their language is contrastive, by tracking the acoustic distribution of speech sounds across contexts, and learning that a dimension is contrastive when the shape changes substantially across contexts. We show that this learning account makes critical predictions that hold true in naturalistic speech, and is one of the first accounts that can qualitatively explain why infants learn what they do.

The results in this dissertation teach us about how listeners might use context to overcome variability in their input. More generally, they reveal that results from lab speech do not necessarily generalize to spontaneous speech, and that using realistic data matters. Turning to spontaneous speech not only gives us a more realistic view of language learning and processing, but can actually help us decide between different theories that all have support from lab speech and, therefore, can complement work on lab data well.

HOW TO USE CONTEXT FOR PHONETIC LEARNING AND PERCEPTION FROM NATURALISTIC SPEECH

by

Kasia Hitczenko

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
2019

Advisory Committee:
Professor Naomi Feldman, Chair/Advisor
Professor Micha Elsner
Professor William Idsardi
Professor Jeff Lidz
Professor Rochelle Newman

© Copyright by
Kasia Hitczenko
2019

Foreword

Chapter 3 and 4 discuss joint work with Reiko Mazuka, Micha Elsner and Naomi Feldman. Portions of Chapter 3 have been published in Hitczenko et al. (2018). Chapter 5 discusses joint work with Naomi Feldman. This work was supported by National Science Foundation grants #IIS-1421695, #IIS-1422987, #DGE-1449815, as well as NSF/JSPS EAPSI grant #1713974.

Acknowledgments

Thank you to my advisor, Naomi Feldman, for being the greatest in every way. I'll never be able to do justice to how much you've done for me and this dissertation these past five years, so I'll just say – I wish that everybody could have an advisor like you. I admire you so much as a researcher: you're brilliant, you're very open-minded to new ideas, you effortlessly glide in and out of different academic circles, leaving a giant mark wherever you go, and you push the field in interesting directions. At the same time, you're the most supportive advisor. Thank you for being a nonstop advocate of mine; for caring about my happiness; for giving insightful and detailed feedback on everything I showed you or talked to you about; for pushing me outside of my comfort zone (in good ways!); for having confidence in me; and for so ardently fighting against me not having confidence in me. I don't know how you do it all, but I can't thank you enough.

A huge thank you to the rest of my committee. Thank you to Micha Elsner, who gave so much to this work, and was a constant source of on-the-nose remarks and suggestions these past five years. It is truly a pleasure collaborating with you. Thank you to Bill Idsardi, who not only has an insane wealth of knowledge, but is so generous in sharing it with others. Thank you to Jeff Lidz, who never failed to asked the question "But what about acquisition?" and in so doing, inspired Chapter 5 because I finally wanted to have an answer to it. Thank you to Rochelle Newman who was always willing to meet with me to discuss my work and provide a different perspective on it. Those meetings guided this dissertation in more ways than she knows.

There are so many other people who contributed to this dissertation. Thank you to Reiko Mazuka and her whole lab for so generously welcoming me to RIKEN and Japan, for sharing your expertise, for letting me use your incredible corpus, and for creating the corpus in the first place. This dissertation literally could not have existed without you. Thank you to everyone who so generously shared data with me: Mirjam Ernestus, Reiko Mazuka, Janet Werker, Dan Swingley, Laurel Fais, and Paul Hikkert.

Thank you to each and every one of the faculty in the department for prioritizing us students, for creating a lovely environment to go into every day, and for being so available to us. Thank you especially to Norbert Hornstein for being so supportive - checking in with me often, giving insightful feedback on practice interviews and talks, and helping me keep everything in perspective. Thank you to Peggy Antonisse and Tonia Bleam for pouring your hearts into your students. Teaching was a lot less intimidating knowing that you had my back.

Thank you to Kim Kwok for everything you've done for me and for keeping me afloat. Thank you to Roger Levy for so generously hosting me in his lab I'm so grateful that I had the opportunity to spend some time with the MIT BCS and Linguistics communities.

Thank you, thank you, thank you to my cohort, Laurel Perkins, Nick Huang, and Gesoel Mendes - you are all absolute stars and I feel very lucky to have gotten

to experience these years by your side. It's been a wild ride, and I couldn't imagine smarter, nicer, humbler, or more fun people to experience it with. I can't wait to see you take over the world!

Thank you to Alayo Tripp for helping me make the right decision to come to Maryland and for giving me so much to look up to. Thank you to Thomas Schatz for being such a fun person to talk to about research and more. Thank you to Cassidy Henry for all of the lovely meals – I wish we could have overlapped longer.

Thank you to the Hyattsville crew for all of the adventures, which made the last 5 years so much fun. Thank you to Phoebe Gaston for being the most amazing friend ever, for taking all of my anxiety in stride, somehow always knowing exactly what to say to make me feel better and introducing me to mead and all the other cool Hyattsville spots. I owe you big time next year. Thank you to Allyson Ettinger, for being our organizer extraordinaire, for all of the bus rides and chats about anything and everything, for spending a lot of time eating Thai food and in supermarkets with me, and for being my Chicago buddy. I can't wait for next year. Thank you to Christian Brodbeck for being a pizza whiz, and always being so incredibly generous (except when you ruined our Christmas cookies). Thank you to Lara Ehrenhofer for always checking in and reminding us to think about how far we've come, not how far we have left to go. Those reminders always came at the right time. Thank you to Paulina Lyskawa for being such a lovely presence, and one of the most thoughtful people I know. Thank you for sending positive vibes my way at every occasion possible – I appreciated every single one of them. Thank you to Mina Hirzel for being such a light in the department. You're so easy to talk to, and very refreshing to be around. It makes me so happy that you were there to celebrate the moment I finished my draft. Thank you, Anton Malko, for helping keep things light-hearted and fun. Thank you to Aura Cruz Heredia. I wish we had become friends earlier, but you helped me so much in this final dissertation/job push. Thanks for being such a great cheerleader. Thank you to Rachel Dudley for all the wonderful advice that I wish I'd followed more, and for being so funny and calming to be around. Thank you to Jeff Green for setting a wonderful example of what it means to be a well-balanced individual. Thank you to Max Papillon for being so generous with your time, and helping to shape my thoughts about phonology and teaching. Thank you to Hanna Muller for making the department a better place. Thank you to Chris Baron for checking in often, always speaking your mind, and pushing people to be people. Thank you to all of the grad students, Baggetts, postdocs, etc. who have made my time at UMD so memorable.

Thank you to Kevin Howard and Jeff Adler, my EAPSI buds. I was at a point in my research where I felt like nothing would ever work out, but you filled my personal life with so much joy, and I'm so grateful for that. Thank you to my roommates, Laura and Ryan, Niha and Nami, Courtney (and Puff of course), for making and being a part of my home these past 5 years. They were grounding and comfortable places and I appreciate each and every one of you.

Thank you to my non-grad school friends for always reminding me that this dissertation is a tiny piece of who I am, and for always keeping me on the track of my values. Thank you, Gianna Fote, for being the best friend ever and for always

being a facetime away. You are the silliest, funniest, smartest, most passionate person I know, who can put a smile on my face no matter what. Thank you for always reminding me to not take things too seriously. Thank you to Michael and Amanda Sherman! I can't express how happy I am that we ended up in the same city these last years. Thank you for providing me a place to escape my dissertation with sleepovers, sushi, John Mulaney, ski trips, spike ball, concerts, surprise engagements, game nights, and video games. Also, thank you for teaching me to be more of a burden – making sure I ask and receive help.

And finally, to Blake and my family – I hope you know how much you all mean to me. We did it!

Table of Contents

Foreword ii

Acknowledgements iii

Table of Contents vi

List of Tables x

List of Figures xi

1 Introduction 1

2 Background 8
2.1 Phonetic Learning . . . 8
2.1.1 General Phonetic Learning . . . 8
2.1.2 The Test Case of Japanese Vowel Length . . . 11
2.2 Categorization - Using Unnormalized Acoustic Cues . . . 15
2.3 How Context Could Be Used . . . 17
2.3.1 Top-down information accounts . . . 18
2.3.2 Normalization accounts . . . 21
2.3.3 Adaptation accounts . . . 26
2.4 Top-Down Information . . . 28
2.4.1 Evidence for Top-Down Information Accounts . . . 28
2.4.2 Evidence that there is top-down information in Japanese . . . 32
2.5 Normalization . . . 33
2.5.1 Evidence for Normalization . . . 34
2.5.2 Evidence that factoring out systematic variability might be useful in Japanese . . . 40

3 Top-down information vs. Normalization on Spontaneous Speech 43
3.1 Top-down information is effective for the Japanese vowel length contrast 43
3.1.1 Data . . . 44
3.1.1.1 Acoustic cues . . . 45
3.1.1.2 Contextual Factors . . . 46
3.1.2 Methods . . . 49
3.1.3 Results . . . 51
3.1.3.1 Baseline Model . . . 51

3.1.3.2 Acoustic and Higher-Level Contextual Information Model . . . 53
3.1.3.3 Higher-Level Contextual Information Model Without Acoustics . . . 53
3.1.4 Discussion . . . 54
3.2 Normalization is ineffective for the Japanese vowel length contrast . . 55
3.2.1 Data . . . 56
3.2.2 Methods . . . 56
3.2.2.1 Normalization Implementation . . . 56
3.2.2.2 Logistic Regressions . . . 59
3.2.3 Results . . . 60
3.2.3.1 Unnormalized Model . . . 62
3.2.3.2 Linear Regression Normalization Models . . . 62
3.2.3.3 Neural Network Normalization Models . . . 63
3.2.4 Discussion . . . 64
3.3 Summary . . . 66

4 The discrepancy in results arises from differences between controlled lab speech and naturalistic spontaneous speech 67
4.1 Controlled vs. Naturalistic Speech . . . 67
4.1.1 Data . . . 68
4.1.1.1 Acoustic cues . . . 70
4.1.1.2 Contextual Factors . . . 70
4.1.2 Methods . . . 71
4.1.3 Results . . . 72
4.1.3.1 Werker Read Speech Data . . . 72
4.1.3.2 Werker Spontaneous Speech Data . . . 73
4.1.4 Discussion . . . 75
4.2 Simulation: Imbalance between categories in the contexts they occur in can hurt normalization . . . 78
4.2.0.1 Methods and Data . . . 79
4.2.0.2 Results . . . 80
4.2.0.3 Discussion . . . 82
4.3 Mathematical Analysis: Contextual category imbalances can hurt normalization, but systematic variability does not hurt top-down information accounts . . . 83
4.3.0.1 Contextual category imbalances can impede normalization . . . 85
4.3.0.2 Systematic variability does not impede top-down information accounts . . . 91
4.3.0.3 Discussion . . . 92
4.4 Summary . . . 93

5 A Learning Story: Naturalistic Data Support Distributional Learning Across Contexts 95
5.1 Proposal: Distributional Learning Across Contexts . . . 97
5.2 Analysis 1: French vs. Japanese . . . 100
5.2.1 Data . . . 100
5.2.1.1 The Corpus of Spontaneous Japanese (CSJ) . . . 101
5.2.1.2 Nijmegen Corpus of Casual French (NCCFr) . . . 103
5.2.1.3 Information Extracted from Corpora . . . 105
5.2.2 Methods . . . 107
5.2.2.1 Analysis Description . . . 107
5.2.2.2 Earthmover's Distance . . . 108
5.2.3 Results . . . 110
5.2.4 Discussion . . . 114
5.3 Analyses 2 and 3: Dutch . . . 117
5.3.1 Data . . . 118
5.3.1.1 Ernestus Corpus of Spontaneous Dutch (ECSD; Dutch ADS) . . . 118
5.3.1.2 Swingley Dutch Infant-Directed Speech . . . 120
5.3.2 Methods . . . 121
5.3.2.1 Analysis 2: Dutch ADS . . . 122
5.3.2.2 Analysis 3: Dutch IDS . . . 122
5.3.3 Results . . . 123
5.3.3.1 Dutch ADS . . . 123
5.3.3.2 Dutch IDS . . . 125
5.3.4 Discussion . . . 127
5.4 Summary and Discussion . . . 128
5.4.1 Do these results generalize to infant-directed speech? . . . 129
5.4.2 Implications beyond language acquisition . . . 131
5.4.3 Corpus sizes . . . 131

6 General Discussion 134
6.1 Summary of this Dissertation . . . 134
6.2 How generalizable are these results? . . . 136
6.2.1 Top-down information is helpful for separating short and long vowels . . . 136
6.2.2 Normalization is ineffective for separating short and long vowels 137
6.2.3 Distribution shape changes can signal contrastive dimensions . 138
6.3 The Status of Normalization . . . 143
6.4 The Status of Distributional Learning Across Contexts . . . 149
6.4.1 Relationship with Other Phonetic Learning . . . 149
6.4.1.1 Learning how many categories exist along a contrastive dimension . . . 150
6.4.1.2 Learning the category parameters . . . 152
6.4.1.3 Learning to categorize individual sounds . . . 157
6.4.2 Can people do this? . . . 158

6.5 Adaptation as another top-down information account of acquisition . 159
6.6 Controlled Lab Speech vs. Naturalistic Speech 161
6.7 Conclusion 163

A Normalization and Top-Down Information Results on Japanese Adult-Directed Speech (R-JMICC) 165
A.1 Top-Down Information Results 165
A.1.0.1 Baseline Model 165
A.1.0.2 Acoustic and Higher-Level Contextual Information Model 167
A.1.0.3 Higher-Level Contextual Information Model Without Acoustics 168
A.2 Normalization Results 168
A.2.0.1 Unnormalized Model 170
A.2.1 Linear Regression Normalization Model Results 170
A.3 Summary of R-JMICC ADS vs. IDS Results 170

B Derivation for Equation 4.3 172

References 179

List of Tables

3.1 The full set of contextual factors available for the R-JMICC dataset . 46
3.2 Top-down information results on naturalistic Japanese IDS (R-JMICC) 52
3.3 Normalization results on naturalistic Japanese IDS (R-JMICC) 61

4.1 The full set of contextual factors available for the Werker Read and Werker Spontaneous datasets 70
4.2 Normalization results: Naturalistic vs. simplified speech 74
4.3 Normalization results from balanced simplified speech corpus 81

5.1 Description of contextual factors used to test learning account 106
5.2 Driving contexts in analysis comparing French and Japanese 111
5.3 Driving word frames in analysis comparing French and Japanese . . . 112

A.1 Top-down information results on naturalistic Japanese ADS (R-JMICC) 166
A.2 Normalization results on naturalistic Japanese ADS (R-JMICC) . . . 169

List of Figures

2.1 Distribution of Japanese vowels in naturalistic infant-directed speech 14
2.2 Toy example demonstrating how top-down information can be helpful 19
2.3 Toy example demonsrating how normalization can be helpful 22

4.1 Distribution of Werker Read IDS vowels 68
4.2 Distribution of Werker Spontaneous IDS vowels 68
4.3 Toy example demonstrating how category imbalances across contexts can hurt normalization 90

5.1 Toy example demonstrating distributional learning across contexts proposal 98
5.2 Distribution of vowels in Japanese ADS (CSJ corpus) 101
5.3 Distribution of vowels in French ADS (NCCFr corpus) 103
5.4 Distribution shape changes across contexts in French vs. Japanese (Context = Vowel Quality + Prosodic Position + Neighboring Sounds) 111
5.5 Distribution shape changes across contexts in French vs. Japanese (Context = Word Frame) 112
5.6 Distribution shape changes across contexts in Dutch ADS 124
5.7 Distribution shape changes across contexts in Dutch IDS 126

Chapter 1: Introduction

Listeners are exposed to highly variable, continuous speech and map it to discrete sound categories. To do so, they first learn as infants what the relevant sounds of their language are, and, subsequently, map incoming signal to learned categories. This is generally a robust process - infants learn about the sounds of their language as early as six months (Kuhl et al., 1992) and, for the most part, listeners process what they are hearing in an effortless manner. However, despite how seemingly easily listeners solve these tasks, they are computationally difficult problems. In fact, after decades of research in this area, researchers have not yet established a robust one-to-one mapping between signal and category that works to anywhere near the degree of success of human listeners.

The reason these tasks are so computationally difficult is because there is a large amount of variability in the speech signal, which can lead to acoustic overlap between different sound categories (Bion et al., 2013). One sound category can be acoustically realized in infinitely many ways, and two different sound categories can, in theory, have identical acoustic realizations. This makes establishing a one-to-one mapping between speech and category difficult. The basic problem is that absolute acoustic or other perceptual cues are insufficient to separate categories as well as

humans do.

This dissertation is about how contextual information (e.g. information about who spoke the sound, what the neighboring sounds were, where in an utterance the sound occurred, and so forth) can help in phonetic learning and speech perception. Researchers have long argued that listeners may be relying on context to help map from signal to categories. The role of context is widely-studied in cognitive science, and fundamental to many cognitive theories, with most researchers largely agreeing that it is crucial in speech perception and acquisition (e.g. Ainsworth, 1974; Auckland et al., 2007; Bar, 2003,0; Bar and Ullman, 1996; Biederman et al., 1982; Boyce et al., 1989; Cole et al., 2010; Davenport and Potter, 2004; Dillon et al., 2013; Feldman et al., 2013a,0; Fujisaki et al., 1975; Ganong, 1980; Green and Courtis, 1966; Kleinschmidt and Jaeger, 2015; Mann and Repp, 1980; McMurray and Jongman, 2011; Moreton and Amano, 1999; Nearey, 1978; Palmer, 1975; Port and Dalby, 1982; Post et al., 1998; Rose and Bressan, 2002; Todorović, 2010; Torralba, 2003; Warren, 1970). However, despite how much has been written about the role of context in these problems, the work has focused primarily on synthesized speech or controlled lab speech. In this dissertation, we study how context can be helpful in learning and perception on naturalistic, spontaneous speech of the type that listeners primarily encounter on a day-to-day basis.

This dissertation will center around two main, non-mutually exclusive ways that listeners could rely on context. Researchers have identified these theories based on two ways that context affects a speaker's production. The first is that context affects which sounds are likely to occur – e.g. /æ/ is much more likely than /ɛ/

to occur in the context th_t ('that' is a word, 'thet' is not), so listeners could be biased to perceive acoustics in that frame as /æ/ rather than /ɛ/. That is, top-down information could guide expectations about what category was likely to be heard. This information can supplement the acoustics, and we will refer to these as 'top-down information' accounts.

The second is that context affects how sounds are produced. For example, speech rate will significantly and systematically alter the acoustics of the signal. This leads to variability in how a particular sound is produced, and can lead to overlap between different sound categories (e.g. a /p/ uttered in fast speech could overlap with a /b/ uttered in slow speech). Listeners could, thus, factor out systematic variability stemming from contextual factors like speech rate from their input. Removing variability may lead to less overlap between categories, and make the mapping from acoustics to categories clearer. In other words, context could be used to pre-process the acoustics that are used for categorization decisions. These types of accounts have generally been termed 'normalization' accounts. The top-down information and normalization account examples provided above make use of two different contextual factors (i.e. neighboring sounds vs. speech rate), but many contextual factors can affect both stages of production (i.e. which category is produced and how it is produced).

These two ways of using context have both been studied extensively, as will be discussed in more detail in Chapter 2. There is a large body of experimental and computational work supporting the notions (i) that context does affect both which sound is produced and how it is produced, (ii) that listeners can make use of these

strategies, and (iii) that listeners do make use of these strategies to help overcome the overlapping categories problem. Both ways of using context are relatively well-accepted in the speech perception literature.

However, there are two main limitations with previous work that warrant further study. First, these two ways of using context - although different - have been somewhat conflated in previous work. In particular, experiments that have been used to argue for one or the other generally show that an acoustic signal is perceived as one category in a particular context, but when the same signal is placed in a different context, it is perceived differently. This type of finding has been used, in different literatures, to argue for both top-down information and normalization accounts, but depending on the specifics, merely shows that context is used, but not how. Therefore, it is not entirely clear whether listeners are using both of these strategies, and if not, which one they are using. Addressing this limitation requires separating these two accounts and testing them individually.

Second, these ideas have mostly been studied on synthetic or carefully controlled lab speech, which differs in important ways from the naturalistic and spontaneous speech that listeners actually learn from and process. It is not clear whether promising results from controlled lab speech generalize to more variable spontaneous speech; indeed, where tested, they have often not (e.g. Antetomaso et al., 2017). In addition, most of the debate so far has centered on whether listeners do or do not make use of these strategies, and has assumed that if listeners did use these strategies, doing so would help them process naturalistic speech. However, this assumption has not yet been supported, as there is actually little to no evi-

dence so far that these strategies are effective on naturalistic speech. Addressing this limitation requires applying these two strategies to naturalistic speech of the type that listeners are mostly exposed to, and testing whether they are effective in separating overlapping categories.

In this dissertation, we study how context can be effectively used in learning and processing, taking these two issues into account. In Chapter 3, we implement top-down information and normalization accounts separately and evaluate their relative contribution in the process of going from speech signal to categories - and we do so on spontaneous speech. We focus on speech perception, specifically on the test case of Japanese vowel length, a test case with particularly overlapping categories that current computational models fail to learn. We find that top-down information is helpful in separating sound categories, remaining robust even on spontaneously produced speech. However, contrary to expectations, we find that normalization is not helpful. In Chapter 4, we study why exactly the discrepancy between our results and previous findings occurs. We find that the discrepancy results from the difference between controlled lab speech and spontaneous speech, by showing that the exact same normalization process we use works if we apply it to controlled lab speech that is more similar to the speech used in previous work. Simulations and a mathematical analysis reveal that one property of spontaneous speech that seems to play a particularly important role is the fact that categories do not occur uniformly across contexts in spontaneous speech, as they do in controlled lab speech. Imbalances in where categories occur - precisely the type of signal that is helpful in top-down information accounts - can hurt normalization.

While Chapters 3 and 4 focus on how well sounds can be categorized when the category labels are known, in Chapter 5, we turn more directly to the question of learning. We propose a new account for how infants could learn which dimensions of their language are contrastive, which takes advantage of the signal that helps top-down information accounts. The idea is that imbalances in the relative proportion of different categories across contexts can cause the distribution shape to vary by context. Our proposal is that infants could learn that a particular dimension (like duration) is contrastive if there are large changes in distribution shape across contexts. I show that this account makes critical predictions that hold true in naturalistic speech, and that the signal needed for this account to be successful is present in all of the naturalistic speech corpora we test.

Past research on these cognitive theories has tended to focus on whether listeners do or do not use these strategies, making the key underlying assumption that using them would actually solve the overlapping categories problem present in speech. Our results validate this assumption for top-down information accounts, and even propose a new account for how contextual information could be helpful in a top-down fashion. However, our results show that in our case study, this assumption is wrong for normalization. Regardless of whether listeners use it, a normalization strategy would be ineffective for learning and processing from naturalistic speech, at least in its current proposed form. It is possible that our theory about how listeners normalize could be repaired, as we will discuss, and this warrants further study. However, these results highlight the importance of studying speech perception and phonetic learning using spontaneous speech, in addition to carefully controlled lab

speech, as results from one do not necessarily generalize to the other.

In what follows, we introduce the specifics of the test case we consider, as well as work that has been done in support of top-down information and normalization accounts. We then describe our simulations and analyses revealing that top-down information is helpful for overcoming overlapping categories on spontaneous speech, even in cases where normalization is not.

Chapter 2: Background

2.1 Phonetic Learning

2.1.1 General Phonetic Learning

Adults' speech perception depends on the languages they have learned (e.g. Goto, 1971; Liberman et al., 1957; Lisker and Abramson, 1970). In particular, adults have an easier time differentiating two sounds that are contrastive in their language (i.e. can change the meaning of a word) than two sounds that are not contrastive in their language. For example, American English adults can hear the difference between [r] and [l], which are contrastive in their language, but Japanese adults have a harder time differentiating these two sounds, which are non-contrastive in their language.

Young infants, on the other hand, can discriminate most of the phonetic contrasts that experimenters have tested them on, regardless of their language background (e.g. Lasky et al., 1975; Trehub, 1976; Werker and Tees, 1984; Werker et al., 1981), though there are some exceptions (see e.g. Aslin, 1980; Eilers and Minifie, 1975; Eimas, 1975; Polka et al., 2001). However, experiments have shown that this begins to change within the first year of life, when infants' phonetic perception seems

to become more attuned to the language(s) they hear. Just like adults, they seem to stop differentiating sounds that are not contrastive in their language. For example, in a classic study, Werker and Tees (1984) tested English and Salish infants on their ability to distinguish /k/ vs /q/, two sounds that are not contrastive in English, but are contrastive in Salish. They showed that English infants were initially able to discriminate these sounds, but lost the ability to do so between the ages of 6-12 months. Salish infants, meanwhile, were able to discriminate the sounds at every age they were tested at. In another classic study, Kuhl et al. (1992) showed that vowel perception seems to become affected by linguistic experience by 6 months of age. They tested infants learning Swedish and American English on their ability to discriminate different variants of /i/ (a sound in English, but not Swedish) and /y/ (a sound in Swedish, but not English). They showed that by 6 months of age, infants were worse at discriminating two within-category sounds if they came from their own native language. This same pattern of findings has been found for many different contrasts (e.g. Best and McRoberts, 2003; Best et al., 1995; Polka and Werker, 1994; Tsushima et al., 1994).

Taken together, there is a body of literature suggesting that infants learn something about the phonetic contrasts of their language by the first year of age - though some phonetic learning extends into the second year (e.g. Mugitani et al., 2009), and some extends even later (McMurray et al., 2018). These classic findings have traditionally been taken as evidence that infants have learn the phonetic categories of their language, though recent work calls this interpretation into question (Schatz et al., 2019). What can be agreed on is that infants seem to have learned something

about which acoustic dimensions are contrastive in their language. Whether at this point or later, they will eventually learn the phonetic categories of their language, and, as adults, they will be able to categorize the incoming speech they hear into these learned sound categories.

How do infants learn about the sounds of their language? Researchers have long thought that infants learn which dimensions of their language are contrastive by tracking the acoustic distributions of the speech sounds of their language. In particular, Maye et al. (2002) proposed that infants learn which dimensions of their language are contrastive based on the overall shape of the distribution. They proposed that an infant who observes a bimodal distribution along an acoustic dimension will learn that that dimension is contrastive (i.e. has multiple categories along it), whereas an infant who observes a unimodal distribution along an acoustic distribution will learn that that dimension is not contrastive in their language. This idea, known as distributional learning, has received experimental and computational support - though most of the support has come from studying simplified speech data. Experimentally, Maye et al. (2002) familiarized 6-8 month olds with a sounds from a /ta/-/da/ continuum that either formed a bimodal distribution or a unimodal distribution. At test, only infants who had heard the bimodal distribution learned the contrast (measured by their ability to discriminate the endpoints of the continuum). Computationally, distributional learning accounts have been shown to successfully learn English and Japanese phonetic categories (Vallabha et al., 2007), when the data used come from controlled lab speech which differs in important ways form naturalistic speech.

Although this story has received a lot of support, the support comes primarily from studying simplified speech, and this account is insufficient for learning from naturalistic speech. In particular, contrastive dimensions often have unimodal distributions in naturalistic speech (see Figure 2.1), and distributional learning fails to learn the correct categories (Bion et al., 2013). In naturalistic speech, there is substantially more variability and overlap between categories, and some categories have been overrepresented in simplified speech relative to their base rate of occurrence in naturalistic speech, which makes it easier to learn them.

The field has put forth a number of proposals for how learning might happen in spite of variability and overlap, and they generally involve using contextual information in a top-down fashion (e.g. Feldman et al., 2013b; Thiessen, 2007) or for normalization (e.g. Dillon et al., 2013). However, support for these various theories has also come primarily from simplified, controlled lab speech data: there is support that these accounts can reduce overlap on simplified speech, but it is unclear whether they can also reduce overlap on naturalistic speech. Here, we will test top-down information and normalization theories on their efficacy on the test case of Japanese vowel length, a particular phonetic learning problem that we now turn to.

2.1.2 The Test Case of Japanese Vowel Length

This dissertation uses the Japanese vowel length contrast as a test case to compare the relative efficacy of top-down information and normalization strategies. In Japanese, there are two sound categories along the duration dimension - referred

to as ‘short’ vowels and ‘long’ vowels (Han, 1962). Which category is used can change the meaning of a word. For example, /biru/ with a short vowel means ‘building,’ while /biːru/ with a long vowel means ‘beer’. Results from perception and production studies reveal that Japanese speakers differentiate short and long vowels: they produce short and long vowels differently and can identify which vowel length category a particular vowel belongs to (Chen et al., 2016; Hisagi et al., 2010; Mugitani et al., 2009; Werker et al., 2007). Based primarily on studies of controlled laboratory speech, vowel length is thought to be signaled primarily by the vowel duration cue, and to a lesser extent, formant values (e.g. Arai et al., 1999; Fujisaki et al., 1975; Kinoshita et al., 2002; Lehnert-LeHouillier, 2010). At this point, we wish to highlight an important terminological distinction between vowel length and vowel duration - and the corresponding two meanings that short/long can have in this context. Vowel length refers to the category status of a vowel - i.e. whether it is the vowel category that will result in /biru/ (‘building’) or /biːru/ (‘beer’). Vowel duration refers to the acoustic property of a vowel - i.e. how long it took the speaker to articulate the vowel - and is thought to be a cue to vowel length. Therefore, a vowel can be referred to as short (or long) if it belongs to the short (or long) category, but it can also be referred to short (or long) depending on its physical duration. In this dissertation, I will use ‘phonologically short/long,’ ‘phonemically short/long,’ or simply ‘short/long’ to refer to category status, and ‘acoustically short/long’ to refer to physical vowel duration.

This distinction is critical because a vowel’s duration and length do not always line up. Recent work has shown that although short vowels and long vowels

are different categories, the range of durations they can have overlap substantially (Bion et al., 2013). While long vowels are, on average, acoustically longer than short vowels, a particular production of a phonologically short vowel can be acoustically longer than a particular production of a phonologically long vowel. In fact, because only 9% of Japanese vowels are phonologically long, the combined distribution of all vowels is unimodal along the duration dimension (Figure 2.1). Therefore, while vowel duration is thought to be a cue to vowel length, it is insufficient to completely separate short and long vowels, and researchers are not yet sure how exactly Japanese listeners process or learn vowel length.

Note that vowel length is not the only way that Japanese listeners need to categorize incoming vowels. There are 10 total vowel categories in Japanese, a short and long version of five different vowel qualities (/a/, /e/, /i/, /o/, /u/) (Han, 1962), so Japanese listeners need to determine both the vowel length and the vowel quality of incoming vowels. However, the acquisition and processing of vowel quality and vowel length seem to be relatively independent processes. It is thought that Japanese infants learn the vowel length contrast at around 10 months of age, about 6 months after they have been argued to learn the vowel qualities (Sato et al., 2010) (but see Mugitani et al. (2009), which argues that infants have not learned about the vowel length contrast as a phonological property until 18 months of age). In this dissertation, I assume the problem of learning and processing vowel quality has been solved and simply consider how Japanese listeners may learn and process vowel length. This is the problem to which we apply the top-down information and normalization strategies to try to explain how adults categorize incoming vowels and

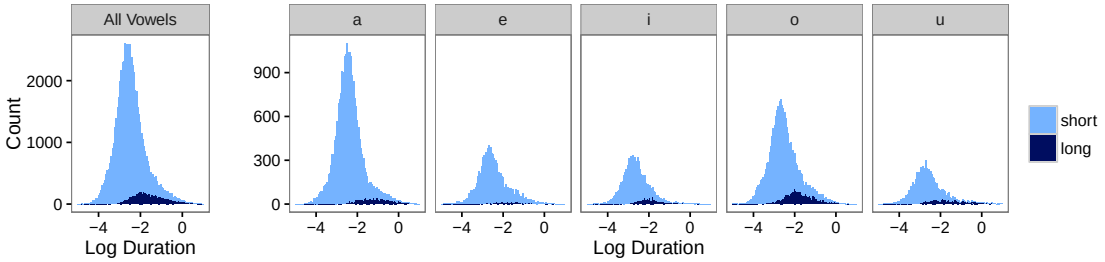


Figure 2.1: Distribution of R-JMICC dataset vowels (by log-duration): all are unimodal distributions. Values displayed are logs of the vowel durations in seconds. As a result, log-durations will be negative whenever the vowel is less than a second long.

how infants learn that vowel length is contrastive.

This problem is just one instance of the commonly observed overlapping categories problem, where the physical cues are insufficient on their own to explain human perception - and contextual information is thought to be used (e.g. Allen et al., 2003; Hillenbrand et al., 1995, 2001; Narayan, 2013; Narayan et al., 2017; Newman et al., 2001; Swingley, 2019; Swingley and Alarcon, 2018). The Japanese vowel length contrast is an ideal test case to use because there exists a hand-annotated dataset consisting of both child- and adult-directed Japanese spontaneous speech (Mazuka et al., 2006) that allows us to easily implement both hypotheses about the use of context and test their efficacy in improving categorization. In the following sections, I will discuss the various models we will apply to the problem of Japanese vowel length in Chapter 3.

2.2 Categorization - Using Unnormalized Acoustic Cues

Japanese listeners must first determine how many sounds there are along the duration dimension during acquisition and, once they have learned the language and its categories, they must decide which of the vowels they hear are short or long through a categorization process. We will start by testing the usefulness of top-down information and normalization strategies by implementing them computationally and seeing how well they perform in categorizing Japanese vowels as short or long. We will compare their performance against a baseline model that categorizes exclusively based on unaltered, unnormalized acoustic cues. Because we test categorization models, these results are directly applicable to adult speech perception, where the task is precisely to categorize vowels. That is, the better a strategy helps categorize vowels as short or long, the more helpful that strategy is in adult speech categorization. However, the results in Chapter 3 are also highly relevant to acquisition (as will be discussed in Chatper 5), where the task is to discover that there are two categories along the duration dimension. Studying how well particular strategies help categorize vowels reveals how well each strategy separates the vowels. This, in turn, shows us to what extent each strategy we study provides signal that infants could learn from. If a strategy does not separate vowels well, it would be hard for an infant to use it to learn, whereas one that does separate vowels is much more promising (although this does not necessarily mean that infants rely on it during learning). In what follows, we lay out what this base categorization model looks like, before turning to a discussion of how context could be used in the

process.

A categorization model can take many forms, but in this dissertation, we model categorization using logistic regression, following previous work (McMurray and Jongman, 2011). Our logistic regression models will take as input a set of cues and map them to vowel category (either short or long). The baseline categorization model - argued to be insufficient in Bion et al. (2013) as described in the previous section - will take as input a vowel's acoustic cues - duration and formant values - and will categorize the vowel as short or long depending on those cues. It will do so by weighting each of the cues (in terms of how much they contribute to whether the vowel should be short or long), summing the weighted acoustic cues, and then transforming this value into a probability that represents the probability that this vowel is short versus long. That is, if we consider the acoustic cues, d, f_1, f_2, f_3, logistic regression takes the following form:

$$P(\text{long}|d, f_1, f_2, f_3) = \frac{1}{1 + e^{\beta_0 + \beta_d d + \beta_{f_1} f_1 + \beta_{f_2} f_2 + \beta_{f_3} f_3}} \tag{2.1}$$

where the β terms are weights on the cues - duration, d, and formants, f_1-f_3. The probability that the vowel is short is $1 - P(\text{long}|d, f_1, f_2, f_3)$. The model categorizes the sound as belonging to the category (short or long) that has the higher probability.

Learning this function involves learning an intercept (β_0), as well as a weight for each cue ($\beta_d ... \beta_{f_3}$). The model is trained on data that consist of the unnormalized acoustic cues of a vowel, labelled with the category that vowel belongs to, and weights are learned so as to optimally separate the short vowels from the long

vowels. Once we learn this function, we can take any new vowel and calculate the probability that that vowel is long (or short). However, the only information that this model has access to is the acoustics of the vowel. As discussed previously, this model has been shown to be insufficient to categorize vowels well, because there is a large amount of overlap in acoustic cues between short and long vowels. We will test the efficacy of top-down information and normalization accounts by implementing them and seeing whether they result in improved categorization relative to this baseline model.

2.3 How Context Could Be Used

The fact that only using acoustics results in bad performance has led researchers to hypothesize that listeners make use of context when learning and processing sounds, with context being broadly defined to include neighboring sounds, but also position in a word/utterance, speaker, speech rate, part of speech of the word the sound was produced in, as well as other aspects of the sound itself (e.g. vowel quality).

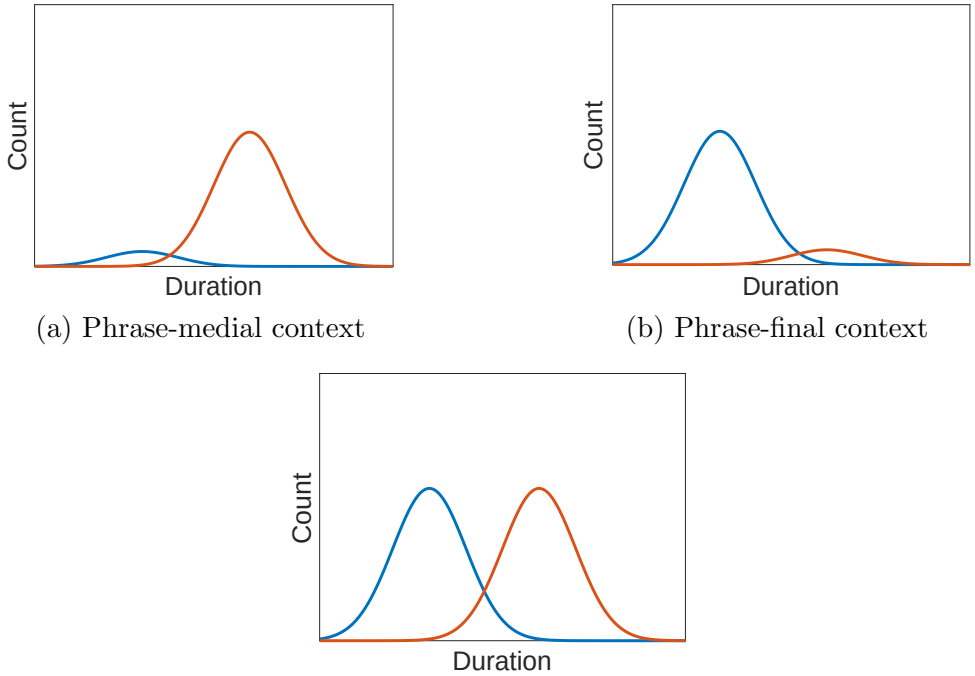
There are two ways that listeners could make use of context. To illustrate the two ways that context could be used to improve categorization, it is useful to consider the production process and the role that context plays in it, as well as how the baseline categorization model presented in the previous section fails to take this into account.

When a speaker produces a vowel, they first decide which category to produce

(short or long vowel) depending on what word they are producing, and then utter a particular acoustic value for that vowel based on the vowel category they are producing. Both of these components of the production process are affected by the context of the vowel, but this is ignored in the base categorization model. In the following two sections, we introduce the two ways context affects a speaker's sound production and, consequently, the two ways that context could in theory be used to improve categorization.

2.3.1 Top-down information accounts

First, the context of a sound directly relates to which vowel category is more or less likely to be produced. To provide an example from English, a speaker is much more likely to produce an /æ/ vowel (as in 'mat') than an /ɛ/ vowel (as in 'met') in the context th_t (the word 'that' exists, but 'thet' does not), but the opposite holds when the context is g_t instead (the word 'get' exists, but 'gat' does not). In Japanese, a speaker is relatively more likely to a produce a long vowel if they are saying an /o/ vowel than if they are saying an /a/ vowel, as can be seen in Figure 2.1. A listener could benefit from taking this type of prior knowledge into account, and indeed listeners are argued to make use of this type of information, as we discuss below. In particular, if they are making a categorization decision between the /æ/-/ɛ/, they could take advantage of the fact that a vowel in the context th_t is much more likely to be /æ/, while a vowel in the context g_t is much more likely to be /ɛ/. This type of strategy could in principle be used without access to any acoustic



(a) Phrase-medial context (b) Phrase-final context

(c) Overall

Figure 2.2: A toy example demonstrating how using contextual information as top-down information can be helpful. Although short vowels and long vowels are equally common overall, short vowels are much more common phrase-finally, and the opposite holds phrase-medially. Our baseline categorization model will not be able to take this into account and will miscategorize some vowels as long in phrase-final position and miscategorize some vowels as short in phrase-medial position.

information whatsoever. As a result, this type of strategy is often referred to a 'top down information' account, as it makes use of listeners' prior knowledge of which sounds are likely to occur in which contexts, in addition to the sounds' bottom-up acoustic cues.

The categorization model presented in the previous section does not, in its current form, take this type of information into account: it only takes into account whether one of the vowel categories is more likely to occur overall, not whether vowel categories are more likely to occur in particular contexts. To illustrate why this is problematic, consider the toy case shown in Figure 2.2, in which there are two

categories (short and long), and only two contexts (let's say phrase-medial vowels and phrase-final vowels). Overall, phonemically short and long vowels occur with identical frequency; however, the phonemically short category is much more likely to occur in phrase-final position and the phonemically long category is much more likely to occur in phrase-medial position. The base categorization model will simply place the category boundary halfway between the short and long vowel means in (c), when really this category boundary should be at a shorter duration for vowels that occur phrase-medially and at a longer duration for vowels that occur phrase-finally. This means that the base categorization model will overclassify phrase-medial short vowels (i.e. short vowels in contexts where long vowels are much more likely to occur) as long and will overclassify phrase-final long vowels as short. However, taking into account context as a top-down influence can help correct this problem. In particular, if the model or listener takes into account expectations about which vowel is more likely to occur in the context heard, then they will, all else being equal, be biased towards categorizing vowels as short in contexts where short vowels are more likely to occur, and biased towards categorizing vowels as long in contexts where long vowels are more likely to occur.

The base categorization model can be augmented to take this into account, in order to reflect what listeners are thought to do. In the logistic regression, this could be accomplished by adding the contexts as independent predictors. For example, in our Japanese example, if we added the vowel quality, q, of the vowel as an independent predictor, this could encode the fact that vowels that are /o/ are relatively more likely to be long than /a/ vowels:

$$P(\text{long}|d, f_1, f_2, f_3, q) = \frac{1}{1 + e^{\beta_0 + \beta_1 d + \beta_2 f_1 + \beta_3 f_2 + \beta_4 f_3 + \beta_5 q}} \tag{2.2}$$

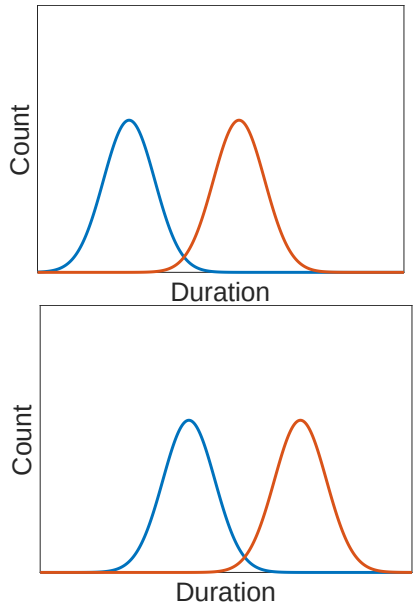
Essentially, this means that in addition to the acoustic cues affecting the relative probabilities of the vowel being short or long, the quality of vowel can also affect the categorization decision. Additional terms could be added depending on what other contextual factors are thought to predict category membership.

The effect is that instead of having one categorization boundary overall, the boundary between categorizing a vowel as short and categorizing it as long will shift depending on the context, and how likely short vs. long vowels are to occur in that context. Essentially, if phonemically long vowels are relatively more likely to occur in a particular context than short vowels, then the boundary between short and long vowels will shift towards vowels of shorter durations, such that more vowels are classified as long, and the opposite holds in contexts where phonemically short vowels are relatively more likely. Crucially, unlike our next model, this model assumes that once the vowel category type (short/long) is chosen, the acoustics do not depend on the context at all. That is, this model assumes that the acoustics of a sound will be the same regardless of the context it was produced in.

2.3.2 Normalization accounts

In the previous section, we saw that context can affect which category a speaker is likely to produce, and that listeners could use this type of contextual information in a top-down fashion to help categorize vowels. The next and final component

(a) Phrase-medial context



(b) Phrase-final context

Figure 2.3: A toy example demonstrating how using contextual information to normalize acoustics can be helpful. Phrase-final vowels are systematically acoustically lengthened, which introduces overlap in the overall distribution of short vowels and long vowels. However, a listener who knows that phrase-final vowels are systematically acoustically lengthened could normalize for this acoustic lengthening, and reduce the overall overlap between short vowels and long vowels in their input.

of the speaker's production process is to actually produce an acoustic value for the vowel category they have chosen. This portion of the production process is also affected by context, as context systematically and predictably affects how a particular sound category is acoustically realized. As an example, vowels uttered in fast speech are, all else being equal, acoustically shorter than vowels uttered in slow speech. Similarly, vowels uttered phrase-finally are, all else being equal, acoustically longer than vowels uttered phrase-medially.

This can introduce variability and overlap between short vowels and long vowels into the overall distribution - and is problematic for a categorization model

simply relying on absolute acoustic cues. Consider the toy case in Figure 2.3. Here again, there are two vowel categories (short vowels and long vowels) and there are two contexts (let's say phrase-medial and phrase-final). In phrase-medial position, short vowels are produced with an average acoustic duration of 150ms and long vowels are produced with an average acoustic duration of 300ms. In phrase-final position, vowels are systematically acoustically lengthened by 100ms. This scenario is problematic for the base categorization model because the overall distribution will reveal a lot of overlap between sound categories. In particular, long vowels in phrase-medial position will overlap with short vowels in phrase-final position. The baseline categorization model presented previously learns a categorization boundary between short vowels and long vowels, which is the same for all vowels, regardless of context. This will cause the model to overclassify vowels occurring in lengthening contexts as phonemically long, and overclassify vowels occurring in acoustically short contexts as phonemically short.

However, the shifts in acoustic cue values are systematic and predictable once the context is known, so using contextual information can help overcome these problems - and listeners have been argued to do so in listening situations. There are various ways this problem could be overcome, and corresponding ways the baseline logistic regression model could be augmented. One is that listeners might build a separate mapping between acoustics and category membership for each context they encounter, such that lengthening contexts will have a boundary between short/long vowels at a higher duration, and vice versa for shortening contexts. This idea is referred to as adaptation, and we will return to it in later discussion, but do not

directly study it in this dissertation. Instead, we focus on a second idea, referred to as normalization.

The idea behind normalization is that instead of creating a different acoustic boundary between short/long vowels for every context encountered, all acoustics are mapped to the same context-independent acoustic space and then one boundary is estimated in this context-independent space. This is done by estimating how much any particular context lengthens or shortens the vowels, and then undoing all lengthening or shortening processes. Returning to our example, normalization would work by estimating that the vowels in phrase-final context are on average 100ms longer than vowels in phrase-medial context, and then essentially shifting the distributions to compensate for this lengthening. Another way to think about it is that each vowel is represented relative to the mean duration of vowels that occurred in the same context. Acoustic cues that have been mapped to this context-independent space are referred to as normalized cues.

In the top-down information accounts, the logistic regression in Equation (2.1) was augmented by adding additional predictors based on the context of the sound in question (e.g. the vowel quality, q). In normalization accounts, the logistic regression in Equation (2.1) is changed by performing a preprocessing step (which will be described in detail below), and inputting normalized cues into the logistic regression, instead of unnormalized cues as before. That is, instead of adding predictors, the existing predictors in Equation (2.1) (d^{unnorm}, f_1^{unnorm}, f_2^{unnorm}, f_3^{unnorm}) are replaced with normalized predictors (d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}). This means that while information about the context a sound occurs in is a direct input to the logistic

regression categorization model in top-down information accounts, it is not in normalization accounts. Rather, in normalization accounts, the contextual information is used to obtain normalized acoustic cues, which are ultimately the only input to the categorization model.

The normalized cues of a sound are obtained by predicting its expected cue values based on the context it occurs in, and then comparing these expected cue values against its actual cue values.

$$cue^{norm} = cue^{unnorm} - cue^{expected} \tag{2.3}$$

The expected cues can be calculated from a sound's contextual information using various methods, and we make use of two such methods. First, we follow past work, and train a linear regression to predict a sound's acoustic cues from the context the sound occurs in (Cole et al., 2010; McMurray and Jongman, 2011). The second method involves training a neural network to predict a sound's acoustic cues from the context the sound occurs in. The benefit of this method is that it allows for more powerful normalization functions that cannot be implemented with a linear regression. Once the pre-processing step is complete and we have normalized all of the acoustic cues relative to context, we can then replace the unnormalized cues with normalized cues in the logistic regression categorization model.

$$P(\text{long}|d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}) = \frac{1}{1 + e^{\beta_0 + \beta_1 d^{norm} + \beta_2 f_1^{norm} + \beta_3 f_2^{norm} + \beta_4 f_3^{norm}}} \tag{2.4}$$

This will again have the effect of shifting where the boundary between short and long vowels falls. In particular, considering the example in Figure 2.3, the vowels in phrase-final position are perfectly shifted from those in phrase-medial position. Ignoring context will cause there to be a huge degree of variability and overlap between short and long vowel acoustics in the overall distributions. However, normalizing out this variability by shifting the two contexts so that they line up will help. In particular, vowels that are acoustically quite long will be readily classified as short because listeners could account for the fact that these vowels were lengthened and undo this effect. That is, a long acoustic duration presented in phrase-medial context may be perceived as long; however, when placed in phrase-final context, that same vowel with the same acoustics may now be perceived as phonemically short because it may be relatively short relative to other vowels that occur in that same lengthening context.

2.3.3 Adaptation accounts

Another idea that has been proposed is that of adaptation (Dahan et al., 2008; Kleinschmidt and Jaeger, 2015). Under 'adaptation' accounts, listeners build a separate model for each context they encounter, so they have a different mapping between acoustic space to categories for each context a sound occurs in. For example, a listener using an adaptation strategy would build a separate model for utterance-medial /o/ vowels, utterance-final /a/ vowels, etc. In doing so, adaptation allows listeners to take into account systematic acoustic variability that stems from the

context a sound occurs in, because they eliminate variability that results from the context a sound occurs in. That is, these models would encode the fact that a shorter absolute duration is required to classify a vowel as long in utterance-medial position than in utterance-final position, without transforming the vowels' acoustic cues as is done in normalization. At the same time, adaptation can also encode top-down information. Building a separate model for each encountered context necessarily encodes relative frequency of occurrence of different sound categories across different contexts, and this could bias perception. Therefore, adaptation can take advantage both of factoring out systematic variability and using top-down information. Because we wish to disentangle the relative contribution of these two ideas, we do not study the efficacy of adaptation strategies here, but we return to the idea of adaptation in the General Discussion in Chapter 6.

Crucially, we have seen that there are two different, non-mutually exclusive ways that context could affect sound production, and the two ways that listeners could make use of this contextual information are different - as top-down information or via normalization. Top-down information strategies involve taking advantage of how likely each category is to occur a priori and normalization takes advantage of how different contexts are known to affect acoustic realizations. They both involve making use of context, but do so at different stages of the process - one acting as a pre-processing step to categorization, and the other directly helping in the categorization. Nonetheless, they both produce similar changes in categorization, which has led researchers to sometimes conflate them in the literature. In what

follows, for each of the two strategies, we first review the evidence that has been used to argue in favor of listeners making use of these strategies, and then we review evidence that both ways of using context could potentially be helpful in Japanese.

2.4 Top-Down Information

2.4.1 Evidence for Top-Down Information Accounts

Experimental and computational work suggests that people can and do make use of higher-level linguistic information in a top-down fashion - at least on synthesized or carefully controlled laboratory speech. Researchers have long argued that adults make use of higher-level structure to bias perception of bottom-up acoustics. In various experiments, researchers have presented participants with stimuli that have portions of removed or degraded acoustic information, and showed that participants make use of contextual information to compensate. For example, Warren (1970) had participants listen to a recording of a sentence with a single phone (and its transitional cues) completely removed and replaced with a cough of the same duration. The listeners all reported hearing the sound, even though its acoustics were completely removed, and were unable to identify the position of the cough. The fact that participants perceived the phone despite it not being there shows that people are making use of linguistic context in speech perception.

Additional research has argued that top-down information is used even in cases where the full acoustic information is available. In a classic study, Ganong (1980) investigated how lexical information affected speech perception. He played

participants acoustic continuua that ranged between a non-word and a word (e.g. from dask to task, or from dash to tash), and had participants categorize the stop as voiced (in these cases /d/) or voiceless (in these cases /t/). He found that the adult participants were biased towards choosing the categorization that resulted in a word. He argued that this showed that listeners were not simply relying on the acoustic cue of the target sound (in this case, voice onset time), but were also using context in a top-down fashion to constrain the categorization decision they made.

In another study, Brown and Hildum (1956) played participants recordings of CCVC syllables, where the initial consonant cluster either violated phonotactic constraints of English (e.g. /tl/) or did not violate phonotactic constraints of English (e.g. /gl/). The researchers asked participants to identify what they had heard and found that participants made more identification errors when the stimuli violated phonotactic constraints than when they did not. The researchers argued that the participants were making use of top-down phonotactic information when processing the acoustics they heard. A similar finding was replicated by Massaro and Cohen (1983).

Particularly relevant to our test case, there is experimental evidence from Moreton and Amano (1999) that Japanese speakers may make use of higher-level contextual information to make decisions about vowel length. Japanese words can be divided into four "strata," depending on their historical origin. Moreton and Amano (1999) take advantage of two key differences between two of these strata - namely between Foreign words, which were recently borrowed into Japanese usually from European languages, and Sino-Japanese words, which were borrowed from Chinese

languages. First, long /a/ is found in Foreign words, but not Sino-Japanese words. Second, Sino-Japanese and Foreign sections of the Japanese lexicon have different frequency distributions over consonants. For example, /p/ occurs in Foreign words, but very rarely in Sino-Japanese words, while /hy/ occurs in Sino-Japanese words, but very rarely in Foreign words. Taken together, an /a/ vowel that occurs with /hy/ is almost certainly phonemically short (the presence of /hy/ suggests it occurs within a Sino-Japanese word, which do not tend to have long vowels), while an /a/ vowel that occurs with /p/ could also be long. In a series of experiments, the researchers found that Japanese speakers made use of these regularities when identifying vowels. The Japanese participants were less likely to identify an /a/ vowel as long if it occurred next to consonants that were more common in the Sino-Japanese strata (in which long /a/ does not occur), but more likely to identify an /a/ vowel as long if it occurred next to consonants that were more common in the Foreign strata. These findings suggest that participants are making use of top-down information.

The idea of using top-down linguistic information has been well-studied in the adult speech perception literature, and recent studies with children have also found evidence that infants are processing higher-level information (Bergelson and Swingley, 2012). Feldman et al. (2013b) showed that both adults and infants used lexical context while acquiring sound categories. In particular, they showed that both adults and infants were more likely to assign acoustically similar vowels (/ɔ/ vs. /ɑ/) to different sound categories when they were not exposed to minimal pairs between them (i.e. when they did not occur in the same phonetic contexts) than

when they were exposed to minimal pairs (i.e. when the vowels occurred in identical phonetic contexts). Feldman et al. (2013a) showed that a computational model that made use of information about the word frames that sounds occurred in resulted in an improvement in phonetic category learning over models that did not incorporate lexical information.

Overall, the idea that higher-level information is influencing speech perception and language acquisition has been replicated many times over, and is mostly accepted in the field (e.g. Feldman et al., 2013b; Ganong, 1980; Marslen-Wilson and Warren, 1994; McQueen et al., 1999; Miller et al., 1951; Rubin et al., 1976; Swingley, 2009; Swingley and Aslin, 2007; Thiessen, 2007). Most of the support for this idea, however, comes from work on simplified speech data. Furthermore, the model from Feldman et al. (2013a) was recently applied to the problem of Japanese vowel length we study here, and was found to be ineffective on spontaneous speech (Antetomaso et al., 2017). Therefore, there is some recent doubt that this strategy could be helpful on spontaneous speech. However, phonemically short vowels and phonemically long vowels have been shown to differ in the contexts that they are likely to occur in in Japanese, so there is potentially signal that would be helpful to a listener relying on such a strategy. We discuss this evidence in the following section.

2.4.2 Evidence that there is top-down information in Japanese

With the exception of Moreton and Amano (1999) and Antetomaso et al. (2017), there has not been much work on studying the role of top-down information in the acquisition and processing of Japanese vowel length. However, there is independent evidence that there are systematic differences between short and long vowels in the types of contexts/environments they occur in. That is, there are patterns that listeners could make use of to better process and learn the vowel length contrast.

First, different vowel qualities have different relative probabilities of short and long vowels, as seen in Figure 2.1, mainly due to the frequency of specific lexical items they occur in. In particular, as discussed previously, long vowels comprise a greater proportion of /o/ vowels than /a/ vowels. Therefore, the quality of the target vowel could bias listeners towards perceiving a vowel of a particular length.

Short and long vowels also differ in the types of sounds they co-occur with. As reviewed in the previous section, various strata of the Japanese lexicon differ in how commonly they make use of long vowels, and they also differ in the types of consonants they include. These facts result in distributional differences in the types of consonants that short and long vowels co-occur with, and participants could make use of these patterns to bias perception - and they do, as reviewed in the previous section (Moreton and Amano, 1999). Similarly, in some dialects of Japanese, long vowels do not occur before nasals, due to phonotactic constraints. Vowels also tend to be phonologically short when adjacent to long consonants. Therefore, the adjacent

sounds of a vowel could potentially provide useful, disambiguating information about the length status of a target vowel (Isei-Jaakkola, 2004).

Finally, prosodic position could also be helpful, as phonemically long vowels are less likely to occur phrase-finally. As a result, listeners could exploit the prosodic position of the vowel to help determine the length of a vowel: they could be biased towards classifying a phrase-final vowel as short.

Overall, there are various patterns due to phonological, historical, or lexical reasons that result in differences in how likely short versus long vowels are to occur in particular contexts. Listeners could exploit this information in a top-down fashion to categorize and learn the vowel length contrast. We test how effective this strategy could be by applying it to the Japanese vowel length contrast, and we propose a learning account that takes advantages of these systematic regularities in Chapter 5.

2.5 Normalization

The second strategy we consider is normalization. We first review literature that has argued in favor of listeners making use of normalization. We then review evidence suggesting that systematic differences in how vowels are produced across contexts introduce variability and overlap into the overall distribution of Japanese vowels.

2.5.1 Evidence for Normalization

A body of experimental work has argued that listeners can and do normalize when making categorization decisions - at least on the carefully controlled laboratory speech or synthetic speech that is typically studied (but see Johnson (1997), Johnson (2006), Pierrehumbert (2002), which argue against normalization). The work in support of normalization mostly relies on findings showing that listeners' perception of a particular sound can change by modifying the context it appears in. As we saw, modifying the context can also change listeners' perception if they are relying on a top-down information strategy. Therefore, this evidence is insufficient to argue uniquely for normalization as a useful strategy when the contextual factor being normalized out could also prove helpful in a top-down information account (e.g. neighboring sounds, prosodic position). However, for contextual factors that do not influence which category is more likely to be produced (e.g. speech rate and speaker), there is extensive evidence that that listeners factor out systematic variability from the acoustics of lab speech, though these studies do not necessarily pinpoint normalization as the involved mechanism (as opposed to adaptation, for example).

Nearey (1978) studied synthetic speech and showed that listeners were able to factor out systematic variability resulting from speaker differences. In his study, phonetically trained listeners identified a range of synthetic test vowels that varied in their first and second formants, that were presented either following a vowel synthesized with formants that corresponded to typical male speech (i.e. relatively

low formants) or following a vowel synthesized with formants that corresponded to typical child speech (i.e. relatively high formants). As the context changed from 'male speech' to 'child speech', listeners shifted all of their category boundaries upward in F1 and F2. This is strong evidence that listeners factor out systematic variability stemming from speaker, and this type of result has been repeatedly found (e.g Strand and Johnson, 1996).

Mann and Repp (1980) studied synthetic speech and argued that listeners also take into account coarticulatory influences. They played participants a fricative from the /ʃ/ to /s/ continuum, followed by either the rounded vowel /u/ or the unrounded vowel /a/, and asked participants to identify the fricative. They found that participants were more likely to identify the fricative as /s/ when it was followed by /u/ than when it was followed by /a/. Fujisaki and Kunisaki (1978) found a similar effect with Japanese speakers. Both studies interpreted this finding as evidence for participants taking into out systematic variability.

Various studies have also shown that listeners take into account the influence of speech rate. These findings are particularly relevant to the Japanese vowel length case, because they offer evidence that participants using durational cues also take into account systematic variability due to context. Using synthesized speech, Fujisaki et al. (1975) studied Japanese listeners' perception of the contrast between short and long consonants as a function of contextual speech rate. They played participants synthesized syllables ranging from /ise/ to /isse/ and found that the absolute duration at which participants' percept changed from a short consonant to a long consonant was affected by the speech rate of the utterance.

In the realm of English vowels, which exhibit durational differences that correlate with quality differences (e.g. the difference between /i/ and /ɪ/), Ainsworth (1974) studied synthesized speech and found that the speech rate of a precursor phrase affected whether listeners reported hearing a short vowel or a long vowel as the target. Specifically, when a vowel was preceded by three vowels of a given duration, participants were more likely to identify the test vowel as a longer vowel (e.g. /i/) if preceded by acoustically short vowels, and more likely to identify the test vowel as a short vowel (e.g. /ɪ/) if preceded by acoustically long vowels. This was especially true for test values with ambiguous formant values and intermediate durations.

Verbrugge et al. (1976) studied controlled laboratory speech and asked participants to identify the vowel of a rapidly articulated test /pVp/ syllable in the context it was originally uttered, in isolation, or in a different precursor phrase that was produced by the same speaker. They found that participants correctly identified the vowel in its original context, but when the syllable was presented in isolation or in the misleading context, participants perceived a reduced/shortened version of the vowel (i.e. /ɪ/ instead of /i/). They argued that participants were using the context to factor out the effect of speech rate and correctly perceive an acoustically short long vowel as long. When the context was removed or made misleading, participants were unable to take into account speech rate and incorrectly perceived the acoustically short long vowel as short.

These findings have been extended to voicing distinctions, as well as /b/-/w/ distinctions, and recent work has even suggested that changing the speech rate

of neighboring consonants can cause listeners to not hear or insert entire function words (e.g. Ainsworth, 1973; Dilley and Pitt, 2010; Miller and Liberman, 1979; Minifie et al., 1977; Summerfield, 1981a). Overall, the general finding that listeners' perceptions of a sound (or even a word) change as a function of the context it occurs in has been replicated many times (e.g. Crystal and House, 1990; Miller, 1981; Miller et al., 1984,9; Newman and Sawusch, 1996; Pickett and Decker, 1960; Sawusch and Newman, 2000; Wayland et al., 1992,9).

However, as mentioned above, recent work has also suggested that some of the experimental findings that have been taken as evidence for factoring out systematic variability may actually be support for participants making use of top-down information. In a classic study, Port and Dalby (1982) argued that listeners use durations of neighboring sounds, in addition to utterance speech rate, to calibrate (or normalize) the durational cues of the target sound. They ran several experiments studying English listeners' voicing judgments in synthesized minimal pairs like *rapid* versus *rabid*. They showed that the duration of a vowel neighboring a stop could affect listener's perception of whether that stop was voiced or voiceless (Port and Dalby, 1982), and similar findings have been reported in other research as well (e.g. Boucher, 2002; Summerfield, 1981b). These findings have classically been interpreted as evidence that listeners factor out the effect of speech rate, and use the relative duration of the stop's closure duration to the neighboring vowel to do so. However, Toscano and McMurray (2012) argued that these same findings were consistent with the alternative idea that listeners are using neighboring vowel duration as a direct cue to the voicing of the target stop (parallel to closure duration or

VOT), rather than normalizing for it. Although this reinterpretation has been discussed with reference to a particular set of studies (Boucher, 2002; Port and Dalby, 1982; Summerfield, 1981b), it raises the interesting possibility that many other of the studies arguing for normalization could also be used as evidence for a top-down information account, rather than for normalization. In particular, this holds true for all studies where the contextual factor that is normalized out could also prove helpful in a top-down information account - for example, neighboring sounds.

Experimental findings in support of normalization have been supplemented by recent computational work, which has generally found that models that normalize for systematic variability achieve better sound category identification results, and better match human performance than models that do not.

McMurray and Jongman (2011) showed that a model that normalized for multiple contextual factors better matched human behavior than a model that did not. They took lab recordings of the 8 English fricatives /f, v, θ, ð, s, z, ʃ, ʒ/ produced in the initial position of a CVC syllable, where the vowel was one of six vowels, and the final consonant was always /p/. They had measurements of 24 cues from these tokens (Jongman et al., 2000). They presented a subset of these recordings to listeners and asked them to identify the syllable-initial fricative. They then used a method from Nearey (1990) and Cole et al. (2010), that we also make use of in Chapters 3-4, to compare whether normalized or unnormalized cues led to more human-like identification in their model. They used logistic regression as their categorization model. The model was trained to identify the fricative, from a set of either normalized or unnormalized acoustic input cues. In the unnormalized

version, the input to this categorization model was simply the absolute acoustic information. In the normalized version, they used linear regression to factor out the influence of speaker and neighboring vowel on these acoustic cues, and used the normalized cues as input into the categorization model. They found that the version that used normalized cues yielded a better match to human categorization than the version that used unnormalized cues, which they use as support that people are normalizing. Many other researchers have found a better match with human performance when using acoustic representations that take into account predictable variability (e.g. Apfelbaum and McMurray, 2015; Cole et al., 2010; Richter et al., 2017); however, these models have, for the most part, only been applied to controlled and well-enunciated lab speech.

While most of the work on taking into account systematic variability has been done in the context of adult speech perception, there has been some work looking at how factoring out systematic variability interacts with the acquisition problem, again through normalization. Dillon et al. (2013) considered the problem of learning the phonological system of Inuktitut, using elicited speech. Inuktitut has three vowels (/i/, /u/, /a/), but these vowels are lowered when followed by uvular consonants. The researchers found that a computational model that learned from the unnormalized vowel formants failed to learn the correct sound categories of Inuktitut (learning six categories instead), but when they subtracted out the influence of the neighboring uvular and used these normalized vowel formants as input to the model, it was able to learn the correct three categories of Inuktitut, just as infants do, suggesting that normalization is a possible strategy that infants

could be using in learning the sounds of their language.

Because most cognitive research has focused on carefully controlled laboratory research or synthesized speech, and because many of the empirical studies supporting normalization could also be in support of top-down information accounts, it is hard to draw strong conclusions about the efficacy of normalization in naturalistic listening environments. Chapter 3 further tests its efficacy in naturalistic listening situations.

2.5.2 Evidence that factoring out systematic variability might be useful in Japanese

In this section, we present evidence that there are factors other than phonological length that influence the duration of Japanese vowels that could cause the overlap between short and long vowels. This is variability that normalization could, in principle, help reduce.

First, there is evidence that the quality of a vowel systematically affects duration. Hirata (2004) had Japanese participants produce disyllabic non-words in a carrier phrase and found that the vowel /e/ tended to be acoustically longer than /o/ and /u/. In addition, Bion et al. (2013) analyzed a corpus of spontaneously produced infant-directed speech and found that low vowels were acoustically longer than high vowels.

Japanese vowels are also acoustically shorter in fast speech than slow speech, all else being equal. Hirata (2004) had participants produce Japanese sentences

(including non-words) at three different speech rates - slow, normal, and fast speech - and found that as the speech rate quickened, the vowels became acoustically shorter.

There is evidence that the prosodic position of a sound influences the duration of a vowel, as well. There are various prosodic phrase types in Japanese - utterances are made up of intonational phrases (IPs), which are, in turn, made up of accentual phrases (APs) - and a vowel's position relative to these phrasal units affects its duration. Bion et al. (2013) found that in spontaneous infant-directed speech, vowels are acoustically longer when followed by an intonational phrase boundary, but acoustically shorter when followed by a word boundary that is not an intonational phrase boundary. Martin et al. (2016) calculated the average mora duration in various prosodic positions in spontaneously produced adult- and infant-directed speech. They found that the average mora duration increases, moving from more phrase-medial to more phrase-final positions (from phrase-medial, to AP-final, to IP-final, to utterance-final position), which suggests that segments are acoustically lengthened phrase-finally.

Some work has also shown that neighboring sounds can influence the duration of a vowel. For example, several studies have found that vowels tend to be acoustically longer before a geminate than a singleton consonant (Fukui, 1978; Han, 1994; Kawahara, 2006). Other work has suggested that accented vowels tend to be acoustically longer than unaccented vowels (Hirata, 2004).

Finally, although these factors have not been studied in Japanese, work in other languages suggests sounds may be acoustically lengthened at the beginning of a phrase, in addition to phrase-finally (Keating et al., 2004; Rakerd et al., 1987), that

sounds may be acoustically shorter in function words rather than content words, and that other features of neighboring consonants, for example voicing, may affect the duration of the target vowel (House, 1961; Luce and Charles-Luce, 1985; Umeda, 1975; Van Santen, 1992).

In summary, there are many factors that influence the acoustic production of Japanese vowels other than their phonological length. This suggests that there is systematic variability that could be factored out, so a normalization strategy could be effective. We will test how effective this strategy could be in learning and processing the vowel length contrast.

Chapter 3: Top-down information vs. Normalization on Spontaneous Speech

In this chapter, we computationally implement top-down information and normalization strategies, as discussed, and test their effectiveness on separating short and long vowels. We provide some of the first evidence that, just like on controlled lab speech, top-down information accounts are helpful at disambiguating short and long vowels. However, contrary to expectations, we find that normalization, at least as implemented, is not helpful at separating short and long vowels. This finding forms the basis for further study of why normalization does not work in Chapter 4, and how top-down information could be helpful in acquisition in Chapter 5.

3.1 Top-down information is effective for the Japanese vowel length contrast

We first test how helpful using contextual information as top-down information can be in the acquisition and processing of the Japanese vowel length contrast, by testing to what extent it helps separate short and long vowels in spontaneous speech. To implement this top-down information strategy, we include contextual factors as

additional independent variables in the logistic regression categorization model, as described previously. The idea is that the model will learn regularities between contextual factors and the phonological length of a vowel; that is, it will use these factors in a top-down fashion to bias categorization decisions. We compare various logistic regression models that make use of higher-level contextual factors to the baseline logistic regression that only uses unnormalized duration and formants.

3.1.1 Data

The data we use come from the RIKEN Japanese Mother-Infant Conversational Corpus (R-JMICC) (Mazuka et al., 2006). It is spontaneously produced child-directed speech. Mazuka et al. (2006) collected the data by recording the speech of 22 mothers who visited the lab with their 18- to 24-month old children. The mothers first played with their child with picture books for about 15 minutes. They then played with their child with toys for about 15 minutes. Finally, a female experimenter came into the room and talked to the mother. The mothers' speech in the first two parts, where they interacted only with their child, was labelled as child-directed speech. The mothers' speech in the third part, where they interacted with the experimenter, was labelled as adult-directed speech. The corpus consists of about 14 total hours of speech, and is labelled for both phonetic and prosodic information.

We extracted information about each of the vowels produced by the mothers, but excluded singing, coughing, devoiced vowels, diphthongs, and any segments that

the researchers could not transcribe. We also excluded any vowels that were not labelled with prosodic information. This left 92003 total vowels, 30035 of which were in the adult-directed section of the corpus and 61968 of which were in the child-directed section of the corpus. All of the analyses we report were run on the child-directed part of the corpus; however, we also ran these analyses on the adult-directed parts (the results are described in Appendix A) and did not find substantial differences in model performance.

We extracted both acoustic information and contextual information about each vowel, as described below. The list of the features we extracted is also compiled in Table 4.1.

3.1.1.1 Acoustic cues

- **Duration**: We extracted the duration of each vowel in milliseconds.

- **Formants**: We extracted the first three formants, and used these as direct acoustic cues to vowel length. While duration is thought to be the primary acoustic predictor of vowel length in Japanese, previous work has shown that spectral information can improve categorization performance (e.g. Arai et al., 1999; Kinoshita et al., 2002; Lehnert-LeHouillier, 2010), so we included the formants in these analyses as direct predictors of vowel length. The formants were automatically extracted using Praat (Boersma, 2001) in previous work on this corpus (Antetomaso et al., 2017) and we used the formant values at the midpoint of the vowel.

R-JMICC Spontaneous
Vowel Quality
Speaker
Previous Sound
Following Sound
Prosodic Position of Word (2 factors)
Prosodic Position of Vowel (12 factors)
Accented
Previous Sound Duration (Speech Rate)
Following Sound Duration (Speech Rate)
Condition (Toys or Books)
Part of Speech

Table 3.1: The full set of contextual factors available for the R-JMICC dataset, with factors that were included in the normalization upper-bound shown in bold (as described in the sections on normalization methods). For the R-JMICC corpus, these are taken from the linear regression normalization method, which outperformed the neural network normalization method.

3.1.1.2 Contextual Factors

In addition to extracting acoustic information, we also extracted contextual information about each vowel that have been shown to be relevant for normalization or top-down information accounts:

- **Vowel quality**: This was a categorical variable that took one of five values (/a/, /e/, /i/, /o/, /u/) and was taken from the coding of what the mother said.

- **Speaker**: This was a categorical variable, with 22 different possible speaker values.

- **Neighboring sounds**: We extracted the identity of the previous sound and the following sound (both categorical variables), as labelled by the phonetic transcription. This was marked as '#' if the vowel was preceded by silence.

Because the vowel length contrast is thought to be learned later than other types of contrasts (Sato et al., 2010), it is reasonable to assume that infants can make use of the other contrasts in their language to learn vowel length.

- **Prosodic position**: We represented prosodic position in three different ways. First, we extracted a categorical variable that ranged from 1 to 4, which indicated whether the word that the vowel occurred in was not phrase-final at all (1), was AP-final (at the end of an accentual phrase) (2), was IP-final (at the end of an intonational phrase) (3), or was utterance-final (4). Second, we extracted a second categorical variable that ranged from 1 to 4, which indicated whether the word that the vowel was in was not phrase-initial (1), was AP-initial (2), was IP-initial (3), or was utterance-initial (4). Third and finally, we extracted a vector of length 12, which represented the prosodic position of the vowel itself in a bit more detail. Namely, each element of the 12-long vector was a binary categorical variable, with three elements of the 12 elements corresponding to whether the vowel itself was word-initial, word-medial, word-final, three to whether the vowel itself was AP-initial, AP-medial, AP-final, three to whether the vowel itself was IP-initial, IP-medial, IP-final, and three to whether the vowel itself was utterance-initial, utterance-medial, utterance-final. That is, while the first two categorical variables represented the prosodic position of the word the vowel was in, and would, thus, have the same value for every vowel in a given word, the vector represented the prosodic position of the vowel itself.

- **Accented?**: This was a binary variable that took a value of 1 if the vowel was accented and 0 if it was not.

- **Speech rate**: We extracted the duration of the immediately preceding and the immediately following sounds, as proxies for speech rate. If the vowel was immediately preceded (or followed) by silence, we did not use the duration of the silence, but instead used the average preceding (and following) sound duration from all other vowels that were not preceded (or followed) by silence.

- **Condition of the vowel**: This was a categorical variable with a value of 'B' if the vowel occurred when mother and child were playing with books and 'T' if it occurred when mother and child were playing with toys. We include this to account for the possibility that the mothers' speech was consistently different (e.g. more or less clear) while playing with books than toys.

- **Part of speech**: This was a categorical variable taken from the annotation in the corpus. In our simulations, we either use full part-of-speech information, or simplified part-of-speech information, which only considers the distinction between function and content words. We vary this because we want our results to be applicable to language acquisition. Infants show evidence of distinguishing function vs. content words using acoustic correlates as early as birth (Shi and Werker, 2001; Shi et al., 1999), so it is relatively likely that they can make use of this knowledge in learning the contrast. However, it is less clear that they could make use of full part-of-speech information for this task, as cross-linguistic evidence suggests that infants have much this knowledge only after

Japanese infants have learned the vowel length contrast (Höhle et al., 2004; Mintz, 2006; Shi and Melançon, 2010). That being said, He and Lidz (2017) show evidence that infants know the distinction between nouns and verbs as early as 12 months, so while infants might not have complete part-of-speech information, they may be able to use more than just the distinction between function and content words for acquiring the vowel length contrast. Testing function vs. content word distinctions in addition to full part-of-speech allows us to determine whether our qualitative results hold true regardless of what infants know.

3.1.2 Methods

To test the efficacy of using contextual information in a top-down fashion, we compare the results of four models - divided into three types of models. The baseline model is a logistic regression that learns to predict short/long from only a vowel's absolute duration and formant values (Baseline). The next two models are logistic regressions that learn to predict short/long from contextual factors listed previously and in Table 4.1, in addition to absolute acoustic cues (Acoustic and Top-Down Information Models). The first of these makes use of all of the contextual factors listed in Table 4.1, with part-of-speech simplified to just indicate whether the word was a function or content word. The second of these makes use of all of the contextual factors, including detailed part-of-speech, exactly as annotated in the corpus. Finally, we test how much signal just the contextual factors provide, by running a

logistic regression model that learns to categorize vowels as short/long using only the contextual factors, without any access to acoustic information (Top-Down Information Model Without Acoustics). As we have seen, the absolute acoustics are relatively bad predictors of vowel category. Studying the results of this model will allow us to understand how much of the work context does. That is, it will reveal how many vowels can be identified just by the context they occur in, without even turning to acoustic information, or, in other words, how much information is lost when acoustics are removed.

We split the dataset into a training subset (90% of the data) and a test set (the remaining 10% of the data), keeping the proportions of short and long vowels equal in the two sets. The training and test sets consisted of the same tokens for all of the simulations run in this chapter, though what information about the vowel was input into the model varied depending on which type of model was being run.

Once the logistic regression equations were estimated from the training set, we simply applied each equation to the vowels in the unseen test set to make a prediction about whether that vowel was short or long, as described previously. We compared the models' predictions to the true labels, which allowed us to assess how effective this strategy could be at separating short and long vowels on naturalistic speech. We report two types of evaluation metrics for each tested model.

First, we report overall categorization accuracy, which is simply the percentage of all of the vowels in the test set that the model categorized correctly. Relatedly, we report categorization accuracy on just the short vowels in the test set and categorization accuracy on just the long vowels in the test set.

Second, we report the Bayesian Information Criterion (BIC) for each model, computed over the training set. The BIC is a common metric used to select between different models (Schwarz, 1978). The benefit of the BIC is that it balances how well the model works (the likelihood of the model given the data) with how complicated the model is (how many parameters it uses), so it will prefer simpler models, all else being equal. The BIC is calculated as follows and lower values are better:

$$\text{BIC} = -2 * ln(L) + k * ln(n) \tag{3.1}$$

where L is the likelihood of the model given the data, k is the number of parameters, and n is the number of samples.

We ran each model ten times and averaged performance across these 10 runs.

3.1.3 Results

The results from this analysis on child-directed speech are summarized in Table 3.2. The corresponding results on adult-directed speech can be found in Table A.1 in Appendix A. They are not discussed in the main text, but the results are qualitatively similar.

3.1.3.1 Baseline Model

Our baseline model simply used absolute duration and formant values to predict the vowel category of a sound. This model reached an overall accuracy of 91.1%. It correctly categorized 99.1% of short vowels, and 12.2% of long vowels. It had a

Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Baseline	91.1	99.1	12.2	28716
Top-down information (with simplified POS)	95.2	98.8	59.0	15193
Top-down information (with POS)	95.7	98.8	63.9	13106
Top-down information (with POS, no acoustics)	94.5	98.6	54.0	16301

Table 3.2: Summary of top-down information results from the R-JMICC dataset.

BIC of 28716. Because 90.9% of vowels in the R-JMICC corpus are short, this model performs comparably to a model that simply categorizes every incoming vowel as short, and has failed to learn anything meaningful about the distinction between short and long vowels.

3.1.3.2 Acoustic and Higher-Level Contextual Information Model

The following models used contextual factors as direct predictors to category membership, in addition to using absolute duration and formant values. When part-of-speech was simplified to the distinction between function and content words, the model reached an overall accuracy of 95.2%, correctly classifying 98.8% of short vowels and 59.0% of long vowels. The BIC was 15193. When we included full part-of-speech information, the model achieved an overall accuracy of 95.7%, correctly classifying 98.8% of short vowels and 63.9% of long vowels. The BIC was 13106. Including additional part-of-speech information led to performance improvements, but both models substantially outperformed the baseline model.

3.1.3.3 Higher-Level Contextual Information Model Without Acoustics

In the final model, we only used contextual factors (including full part-of-speech information) as direct predictors of category membership. That is, the model did not have access to any acoustic information and could only make use of information about how likely each category is to occur a priori. Even without any acoustic

information, this model achieved an overall accuracy of 94.5%, correctly classifying 98.6% of short vowels and 54.0% of long vowels. The model BIC was 16301. That is, although there was a slight dip in performance when we removed acoustic information, top-down information models can still perform well even without any acoustic information, suggesting a large role for context.

3.1.4 Discussion

In these analyses, we investigated the hypothesis that infants and adults learn and process the Japanese vowel length contrast by combining bottom-up acoustic cues with top-down expectations about which category is likely to occur in a particular context. To implement this hypothesis, we included contextual factors listed in Table 4.1 as direct predictors of category membership in the logistic regression model (in addition to absolute acoustic cues), and compared its performance against a model that only makes use of absolute acoustic cues as predictors.

We found that including these additional contextual factors as predictors drastically improved accuracy and lowered BIC scores, suggesting that this method does quite well at separating short vowels from long vowels. Given the relatively small set of factors we used - for example, the only word-level information we used was part-of-speech - it is quite impressive that the model achieved this level of performance, and it suggests that this may be a hypothesis worth pursuing as a way that infants could learn and adults could process the Japanese vowel length contrast.

In fact, although excluding acoustic information did hurt performance, a model

relying on contextual information alone still performs very well. Even without *any* acoustic information, this model can correctly identify nearly all short vowels and more than half of all long vowels. This illustrates just how much signal there is in contextual information.

Of course, these are supervised models that have much more information available to them than infants learning language. But it does reveal that there is signal in the input that infants could be using, and we return to this in Chapter 5, where we propose a particular way that infants could make use of this information to learn phonetic contrasts. Certainly this work shows that top-down information could be very useful in adult speech perception. In the following section, we test the efficacy of normalization on this same problem of separating short and long vowels, and show that it does not perform nearly as well.

3.2 Normalization is ineffective for the Japanese vowel length contrast

In this section, we test to what extent normalization can help in the acquisition and processing of the Japanese vowel contrast. To test whether normalization could be helpful, we compare models that use normalized acoustic cues to models that use unnormalized acoustic cues. If normalization is helpful, we would expect normalized cues to better categorize vowels as short or long than unnormalized cues.

3.2.1 Data

The data are exactly the same from the first analysis, but the contextual factors listed in Table 4.1 are normalized out of the acoustics (as described below in the Methods section), instead of being included as independent predictors in the logistic regression categorization model. The same training and test sets are used as in the previous analysis, which allows us to directly compare results.

3.2.2 Methods

In testing the efficacy of normalization on spontaneous speech, we implement and test two normalization methods. First, we apply methods from previous work (Cole et al., 2010; McMurray and Jongman, 2011; Nearey, 1990) to the Japanese vowel length contrast, by using linear regression to normalize out systematic variability from vocalic acoustic cues. Second, we implement normalization using a neural network, which has the advantage over past implementations that it can represent more powerful, non-linear normalization functions. We then input the unnormalized or resulting normalized acoustic cues into a logistic regression model to categorize the vowels as either short or long.

3.2.2.1 Normalization Implementation

We use either unnormalized or normalized acoustic cues as predictors of vowel length. Using unnormalized cues simply involves representing the absolute acoustic cues, so this section will focus on how we implement normalization. The basic idea

underlying both of the implementations we use is to learn a function that predicts acoustic features (duration and formants) of a vowel from the context that a vowel occurs in (i.e. vowel quality, speaker, prosodic position). Once we learn this function, we can make a prediction about a vowel's duration and formants based on everything we know about where it occurs. We can then use the residuals, or the difference between how long we expect the vowel to be given all of the factors and how long it actually is, to represent a normalized version of this vowel. That is, we have excluded the influence of contextual factors and have recoded the acoustic cues in terms of their difference from expected values. Once we learn this equation from the training set, we recode both the training set and the test set in normalized terms. We use two different methods for representing the function between contextual factors and acoustic cues. The first - linear regression - is applied from Cole et al. (2010), McMurray and Jongman (2011), and Nearey (1990), and follows from previous work. The second uses neural networks to implement the function from contextual variables to acoustic cues. The benefit of this method is that it can represent non-linear interactions between the contexts and the acoustic cues, and therefore, is more powerful than a linear regression. It is important to note that there are other normalization implementations (e.g. Dillon et al. (2013)) that may function differently. In this paper, we limit our analyses to two normalization methods, but we discuss how our results generalize to other ways of normalizing in Section 6.3.

Linear Regression as Normalization Following previous work, we first use linear regression to factor out systematic variability (Cole et al., 2010; McMurray and Jongman, 2011; Nearey, 1990). Linear regression models represent a relationship between a continuous dependent variable and a set of independent variables. In this particular case, we try to estimate an equation that can predict what the acoustic features (duration and formants) of a vowel should be from its context. Each of the factors (e.g. vowel quality, speaker, prosodic position from Table 4.1) is weighted and combined linearly to yield a prediction. That is, given the factors $x_1, x_2, ..., x_n$, linear regression models take the form:

$$\text{acoustic cue} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_n * x_n \tag{3.2}$$

Learning this function involves learning an intercept (β_0), as well as a weight for each cue ($\beta_1 ... \beta_n$). The data it learns from consist of the information we want to factor out of the acoustic cues, as well as the known acoustic cue values of the vowel, and weights are learned so as to minimize the error in predicting the duration of the vowel. Once we learn this equation from the training set, we recode both the training set and the test set in normalized terms - by taking the difference between the predicted acoustic cues and the actual acoustic cues.

Neural Networks as Normalization The linear regression models we use do not include interactions, even though previous work did. This is because our analyses use a total of 23 contextual factors, so considering all possible interactions would be

computationally difficult. To test the possibility that our linear regression without interactions was insufficient to handle spontaneous speech, we also implemented normalization using a neural network. We train a neural network on the training set to predict the duration and formants of a vowel token from its context. Once we have a trained neural network, we use it to predict expected acoustic cues for each vowel, subtract them from the vowel's true acoustic values, and input this into a logistic regression model.

We use a simple feed-forward neural network. We use five-fold cross validation on the training data to tune parameters of the neural network. We manipulate the number of hidden layers, the batch size, the number of nodes in the hidden layers (either keeping this constant for all of the layers or decreasing the number of nodes progressively deeper into the network), learning rate, number of epochs, and regularization factors. We choose the parameters that minimize average mean squared error on the training set. We then recode both the training set and the test set in normalized terms - by taking the difference between the predicted acoustic cues and the actual acoustic cues.

3.2.2.2 Logistic Regressions

To test the efficacy of normalization, we compare seven total logistic regression models, which can be grouped into three types of models. The first model is the baseline, which as before uses absolute (unnormalized) duration and formants to predict category membership (short or long). We compare the baseline against

three models that make use of normalized acoustic cues obtained by running a linear regression normalization, as well as three models that make use of normalized acoustic cues obtained by running neural network normalization. That is, for each type of normalization (linear regression and neural network), we run three models. First, we regress out all of the contextual factors listed in Table 4.1 with part-of-speech in simplified form (i.e. function/content word distinctions). Second, we regress out all of the contextual factors listed in Table 4.1 including full, detailed part-of-speech information. In both of these models, the normalization function is trained completely independently of the subsequent logistic regression. That is, the normalization function is not trained to maximize categorization performance. For the third and final model, however, we choose the subset of contextual factors from Table 4.1 that maximizes categorization performance, which gives us an estimate of the upper bound on normalization performance. This is useful because it is possible that we are wrongly including some factors in the first three models and underestimating the efficacy of normalization. Running this model allows us to see what the best normalization performance could be. We run these three models with both linear regression and neural networks, which yields six total models that we compare against our baseline.

3.2.3 Results

A summary of the results from infant-directed speech is presented in Table 3.3. Corresponding results from adult-directed speech are present in Table A.2 in

Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Unnormalized baseline	91.1	99.1	12.2	28716
Linear Regression Normalization	91.2	99.5	8.3	30774
All factors with simplified part-of-speech				
Linear Regression Normalization	91.2	99.6	7.6	30990
All factors with full part-of-speech				
Best Linear Regression Normalization	91.2	99.0	13.6	28122
Neural Network Normalization	91.1	99.8	5.1	32356
All factors with simplified part-of-speech				
Neural Network Normalization	91.1	99.7	5.8	31738
All factors with full part-of-speech				
Best Neural Network Normalization	91.2	99.0	13.4	28188

Table 3.3: Summary of normalization results on R-JMICC corpus.

Appendix A.

3.2.3.1 Unnormalized Model

The baseline model is identical to the baseline model from the previous analysis and uses unnormalized duration and formants as predictors of category membership, without running any linear regression models. As a reminder, this logistic regression model reached an overall accuracy of 91.1%. It correctly classified 99.1% of short vowels and 12.2% of long vowels. Its BIC was 28716.

3.2.3.2 Linear Regression Normalization Models

The following models used linear regression to normalize duration and formants, and used normalized duration as predictors of category membership. When all of the contextual factors with simplified part-of-speech (function vs. content word) were regressed out, the model had an overall categorization accuracy of 91.2%, correctly classifying 99.5% of the short vowels and 8.3% of the long vowels. It had a BIC of 30774. The set of factors used accounted for 26.8% of the variance in duration, 23.0% of the variance in F1, 40.2% of the variance in F2, and 8.1% of the variance in F3. When all of the contextual factors including full part-of-speech information were regressed out, the model had an overall categorization accuracy of 91.2%, correctly classifying 99.6% of the short vowels and 7.6% of the long vowels. It had a BIC of 30990. The set of factors used accounted for 27.8% of the variance in duration, 23.1% of the variance in F1, 40.3% of the variance in F2, and 8.3% of the

variance in F3. Finally, the upper bound normalization performance included the following five contextual factors: speaker, whether the vowel itself was word-final, whether the vowel itself was AP-final, whether the vowel itself was IP-final, and whether the vowel itself was utterance-final. This model had an overall accuracy of 91.2%, and correctly classified 99.0% of the short vowels and 13.6% of the long vowels. It had an overall BIC of 28122. The set of factors that resulted in the best categorization performance accounted for 11.7% of the variance in duration, 3.6% of the variance in F1, 3.3% of the variance in F2, and 3.8% of the variance in F3.

3.2.3.3 Neural Network Normalization Models

We then used neural networks to normalize duration and formants. When all of the contextual factors with simplified part-of-speech (function vs. content word) were normalized out, the model had an overall categorization accuracy of 91.1%, correctly classifying 99.8% of short vowels, and 5.1% of long vowels. The BIC was 32356. When all of the contextual factors, including fully detailed part-of-speech information, were normalized out, the model reached an overall categorization accuracy of 91.1%, correctly classifying 99.7% of short vowels, and 5.8% of long vowels. The BIC was 31738. Finally, the best categorization performance arose from normalizing out the following factors from the acoustics: whether the vowel itself was word-final, whether the vowel itself was AP-initial, whether the vowel itself was AP-final, and whether the vowel itself was utterance-final. This model had an overall accuracy of 91.2%, and correctly classified 99.0% of the short vowels and 13.4% of

the long vowels. It had an overall BIC of 28188.

3.2.4 Discussion

Previous work has argued that normalization can be helpful in acquisition and processing (Cole et al., 2010; Dillon et al., 2013; McMurray and Jongman, 2011); however, our results on Japanese vowel length did not support this hypothesis. We compared the Japanese vowel length categorization performance of a logistic regression model that used unnormalized acoustic cues to the performance of various logistic regression models that used normalized acoustic cues. We considered two different normalization implementations, and three different instantiations of normalized cues for each. The first normalized all available contextual factors, with simplified part-of-speech information (i.e. whether the word containing a vowel was a function or content word). The second normalized all available contextual factors, including detailed part-of-speech information. The third and final normalized out the subset of contextual factors that led to best categorization performance. Crucially, in the first two models, as in past work, normalization was not optimized to give the best categorization. The final model considered categorization performance in choosing how to normalize, giving it the best possible chance to succeed.

The main finding was that, at its best, normalization resulted in only a modest improvement in accuracy and BIC, regardless of which implementation we used. Although the overall accuracy of all of the models is quite high, just guessing that all of the vowels were short would result in similar results. Normalization never

improved accuracy, but improved the BIC from 28716 for the unnormalized version to 28122 for the best normalized version. While this does constitute improvement, it is only modest improvement and a listener would need to learn precisely which factors they should normalize out. Of course, it is possible that results would be better on a larger corpus with more information about the contextual factors. We used previous and following sound duration as a proxy for speech rate, while other measures of speech rate might lead to better performance, and we return to this possibility in the discussion in Chapter 6. However, given how prevalent normalization is in the field, the results are surprisingly bad and call into question the efficacy of normalization (as implemented), at least in this task.

Although it is difficult to directly compare this degree of improvement to the improvement shown in past studies merely on the basis of accuracy, past studies that have implemented and tested normalization reported that normalization resulted in an increase in performance from 28.63% to 54% and 83.3% to 92.9% respectively (Cole et al., 2010; McMurray and Jongman, 2011). In comparison, in this work, the overall accuracy did not change depending on whether cues where unnormalized or normalized, and the long vowel accuracy increased from 12.2% to 13.6% - a much weaker increase in performance than has been observed previously.

There are a number of reasons why we may have observed this discrepancy from previous literature, the main one being that we tested normalization on spontaneous speech, while previous work has considered synthesized or carefully controlled lab speech. In the next chapter, we apply the same analyses from this section to controlled lab speech and show that normalization then works, consistent with past

work, which suggests that the difference between spontaneous and controlled lab speech does indeed explain at least some of the difference in results.

3.3 Summary

This chapter presented work that computationally implemented top-down information and normalization accounts, and tested their relative efficacy on separating short vowels and long vowels. The results showed that top-down information accounts successfully separated short vowels from long vowels even on naturalistic speech, but that normalization (as implemented here) did not, in contrast to what has been found on controlled lab speech. The following two chapters will delve into each of these findings in more detail. Chapter 4 will focus on normalization, and will ask why we observe this discrepancy between our work and previous work. Through a series of simulations and mathematical analyses, we will show that normalization is effective on lab speech, but not effective on naturalistic speech, and we will identify one property of naturalistic speech that could explain differences. In Chapter 5, we will turn more directly to acquisition and propose an account of how infants might learn which dimensions of their language are contrastive, taking advantage of the top-down information signal that was shown to be present in this chapter.

Chapter 4: The discrepancy in results arises from differences between controlled lab speech and naturalistic spontaneous speech

Previous results found normalization to be helpful; however, our results from Chapter 3 were surprising in that they showed that normalization was unhelpful - even when the process was fully supervised. In this chapter, we explore why we see this difference in results, and show that it has to do with differences between controlled speech and naturalistic speech.

4.1 Controlled vs. Naturalistic Speech

The biggest difference between previous work and our own is that most previous work has explored normalization on controlled and carefully enunciated lab speech, but our work looked at normalization on spontaneously produced speech. To bring these results more in line with each other, we apply the same normalization analyses we used on the R-JMICC Spontaneous Speech corpus to Japanese lab speech. What we find is that the same normalization process that was not helpful on spontaneous speech is helpful on Japanese lab speech, suggesting that the discrepancy in results between our work and previous work arises from differences between spontaneous and controlled lab speech.

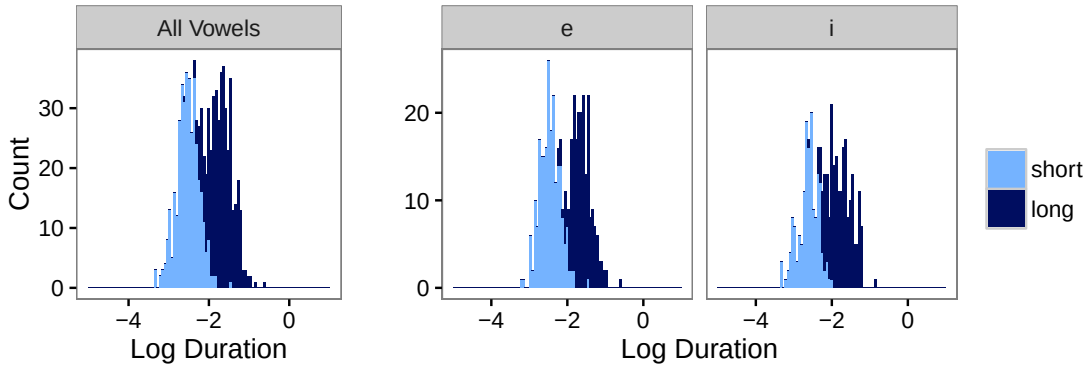


Figure 4.1: Distribution of Werker Read IDS vowels (by log-duration). Values displayed are logs of the vowel durations in seconds. As a result, log-durations will be negative whenever the vowel is less than a second long.

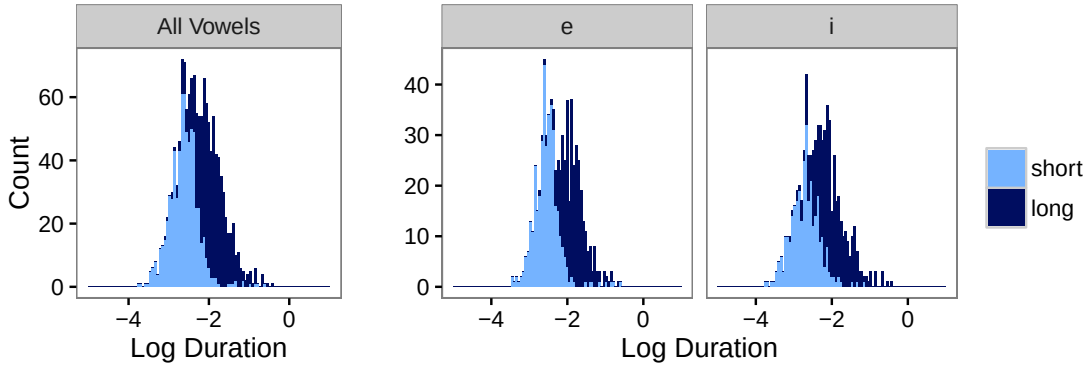


Figure 4.2: Distribution of Werker Spontaneous IDS vowels (by log-duration). Values displayed are logs of the vowel durations in seconds. As a result, log-durations will be negative whenever the vowel is less than a second long.

4.1.1 Data

The data we use here is Japanese infant-directed speech from Werker et al. (2007). The data consist of 10 mothers teaching their 12-month-old infants a set of 16 nonce CVCV words, while looking at picture books together. This interaction included both a reading task, in which mothers were asked to read sentences containing the nonce words with pictures of the novel object (Werker Read dataset - Figure 4.1), as well as a spontaneous speech task, in which mothers were asked to describe

a scene that contained the novel object, using the nonce word as much as possible (Werker Spontaneous dataset - Figure 4.2). The nonce words were made only using /i/ and /e/ as critical vowels, so the data do not contain any annotated instances of /a/, /o/, or /u/, unlike the R-JMICC corpus. The data were collected in the NTT Communication Science Laboratories in Keihanna, Japan and were labelled by trained phoneticians.

These data were much more similar to datasets that had previously been used to study normalization, as the datasets are carefully controlled laboratory speech. Even the Spontaneous subset of the data was relatively controlled, as the researchers decided what the nonce words were and, therefore, what sounds each target vowel was likely to occur next to.

We extracted information about each of the vowels produced by the mothers, excluding any segments that the researchers could not annotate with certainty. The read speech data consisted of 798 vowels, of which 381 (47.7%) were phonemically short vowels and the remaining 417 (52.3%) were phonemically long vowels. The spontaneous speech data consisted of 1382 vowels, exactly half of which were phonemically short and half of which were phonemically long. Similarly as for the R-JMICC data, the information we extracted was either used as an acoustic predictor or as a contextual factor to be normalized out.

Werker Read	Werker Spontaneous
Vowel Quality	**Vowel Quality**
Speaker	**Speaker**
Previous Sound	**Previous Sound**
Following Sound	Following Sound
Prosodic Position	**Prosodic Position**
F0	**F0**

Table 4.1: The full set of contextual factors available for the Werker Read and Werker Spontaneous datasets. Factors that were included in the normalization upper-bound are shown in bold (as described in the sections on normalization methods).

4.1.1.1 Acoustic cues

As before, we used the duration of the vowel in milliseconds and the first three formants, as direct predictors of vowel length. These acoustic cues were either represented unnormalized as in the corpus, or underwent normalization through linear regression.

4.1.1.2 Contextual Factors

We also extracted all of the contextual information that the original researchers had labelled on this dataset. The set of factors available for the Werker data is largely a subset of what was available for the R-JMICC dataset, with the exception that fundamental frequency (F0) was available for the Werker data while it was not extracted for R-JMICC. In addition, the labeling for prosodic position information was much simpler for the Werker data than for R-JMICC data, as will be described below. We collected the following pieces of information about each of the extracted vowels.

- **Vowel quality**: This was a categorical variable that took one of two values (/e/ or /i/) and was taken from the coding of what the mother said.

- **Speaker**: This was a categorical variable with one of 10 different possible values.

- **Prosodic position**: Prosodic position took one of four values: 'Independent Word,', if the vowel occurred in a free-standing word, or 'Sentence Initial,' 'Sentence Medial,' or 'Sentence Final,' depending on whether the syllable the vowel occurred in was first, last, or in the middle of the sentence. This was controlled in the Werker Read data.

- **Neighboring sounds**: We extracted the previous and following sound. Unlike in the R-JMICC data, these were controlled to always be consonants.

- **Fundamental frequency**: We extracted the F0 at the vowel's midpoint.

4.1.2 Methods

We use linear regression to implement normalization. We did not use neural networks because they require large amounts of data, which we do not have for controlled lab speech, and because they performed worse than linear regression normalization models in Chapter 3. As before, we divided the data into a training set and a test set, which consisted of 90% and 10% of the data, respectively. We trained the linear regression model to learn a function between contextual variables and acoustic cues, and then took the residuals to obtain normalized cues on both

the training and test sets. Then, either unnormalized or normalized cues were input into the logistic regression model to classify each vowel as short or long.

We ran three total logistic regressions for both the Werker Read and Werker Spontaneous datasets. The baseline used absolute acoustics to categorize vowels as short or long. The second model normalized the acoustic cues for all contextual factors available for the Werker data. The third and final model chose the subset of the available contextual factors that maximized categorization performance, as determined by cross-validation, again providing us with an upper bound on how useful normalization could be.

Results are averaged across 10 runs. We again report overall accuracy, accuracy on short vowels, accuracy on long vowels, as well as BIC, for which lower values are better.

4.1.3 Results

The results are summarized in Table 4.2. We first consider the efficacy of normalization on the Werker Read speech, before considering Werker Spontaneous Speech.

4.1.3.1 Werker Read Speech Data

Unnormalized Model The unnormalized model achieves 91.4% overall accuracy on the Werker Read speech. Although this is a similar overall accuracy to the R-JMICC spontaneous data, this corpus is much more balanced than the R-JMICC

corpus. In the Werker Read speech, about 47.7% of the used vowels are short, compared to 90.9% in the R-JMICC corpus, so a strategy of simply categorizing every vowel as short (or long) will not yield as good results on the Werker Read Speech as on the R-JMICC corpus. The unnormalized model correctly classifies 89.7% of short vowels and 92.9% of long vowels, achieving a BIC of 246.

Normalized Models When we normalized out all available factors, the model's overall accuracy is 86.1%, and it correctly classifies 83.9% of the short vowels and 88.1% of the long vowels. Its BIC is 399. That is, normalizing all available factors does not improve performance. When we instead choose the best subset of factors, the model no longer factors out the effect of the following consonant, and shows a boost in performance. It achieves an overall accuracy of 95.1%, a short vowel accuracy of 92.3%, a long vowel accuracy of 97.6%, and a BIC of 105.

4.1.3.2 Werker Spontaneous Speech Data

Unnormalized Model The unnormalized model achieves 82.9% overall accuracy on the Werker Spontaneous speech, and correctly classifies 90% of the short vowels and 75.7% of the long vowels. It achieves a BIC of 1072. Again, in the Werker Spontaneous speech, exactly 50% of the used vowels are short, so the unnormalized model substantially outperforms one that simply guesses that each vowel is short, unlike on the R-JMICC corpus.

Data	Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Werker Read	Unnormalized	91.4	89.7	92.9	246
	Normalized (all factors)	86.1	83.9	88.1	399
	Normalized (best factors)	95.1	92.3	97.6	105
Werker Spontaneous	Unnormalized	82.9	90.0	75.7	1072
	Normalized (all factors)	78.5	85.9	71.1	1219
	Normalized (best factors)	90.0	92.9	87.1	869
R-JMICC Spontaneous	Unnormalized	91.2	99.1	12.2	28716
	Normalized (all factors)	91.2	99.6	7.6	30990
	Normalized (best factors)	91.2	99.0	13.6	28122

Table 4.2: Comparison of normalization results on R-JMICC spontaneous speech corpus and Werker controlled laboratory data. The Werker speech corpus had a read component and a spontaneous component, but even the spontaneous component was relatively controlled by the experimenters, as the experimenters provided nonce words for the parents to teach their children.

Normalized Models The model that normalizes out all available contextual factors listed previously and in Table 4.1 achieves an overall accuracy of 78.5%, correctly classifying 85.9% of short vowels and 71.1% of long vowels. When we allow subsequent categorization results to drive which subset of contextual factors are included in normalization, the model achieves an overall accuracy of 90.0%, and correctly classifies 92.9% of short vowels and 87.1% of long vowels. It achieves a BIC of 869. Similarly to the Werker Read speech, normalizing out all of the factors does not help, but depending on what factors are normalized out, normalization can help - and substantially.

4.1.4 Discussion

In this section, we applied the same linear regression analyses that we applied to the R-JMICC spontaneous speech corpus to the Werker corpus. The idea was to test whether we would see similar normalization results as were previously reported, when we used data that more closely resembled that used in previous work. We found that normalization could help on the read speech, as well as the Werker spontaneous speech, even though it did not help when all available contextual factors were factored out.

That is, on the same contrast in the same language, normalization was helpful on carefully controlled lab speech, even though it was unhelpful on naturalistic, uncontrolled spontaneous speech (R-JMICC). This suggests that normalization may be ineffective on naturalistic spontaneous speech.

Another interesting finding was that the Werker Spontaneous speech patterned similarly to the Werker Read speech, instead of the R-JMICC Spontaneous speech. The overall results were worse on the Werker Spontaneous speech than on the Werker Read speech; however, normalization was helpful on the Werker Spontaneous speech, but not the R-JMICC Spontaneous speech. One reason for this could be that duration seems to be used differently by speakers in the R-JMICC data versus the Werker data. In particular, in comparing Figure 2.1 to Figures 4.1 and 4.2, it seems that the contrast is being produced differently in the two datasets, such that duration is a much better cue for vowel length in the Werker data than the R-JMICC data. There is less overlap between the short and long vowel categories in the Werker data: there is a duration such that all vowels that are acoustically longer than it are reliably long vowels. In the R-JMICC data, however, this is not the case: some short vowels are as acoustically long as the most acoustically lengthened long vowels. Another possible reason for this is that unlike in the R-JMICC dataset, the contents (and contexts that particular sounds occurred in) were fairly controlled by the experiments in both the Werker Read and Spontaneous datasets. In the following sections, we will discuss how this could affect normalization performance.

It is important to emphasize that although both the R-JMICC and Werker Spontaneous speech datasets are referred to as spontaneous speech, they differed quite substantially in nature. That is, not all spontaneous speech corpora are comparable. In particular, in the Werker Spontaneous speech, mothers were producing nonce words that were created by researchers, were instructed to teach their infants, and were given pictures to describe. In the R-JMICC Spontaneous speech, mothers

were given toys and books, but were given very little instruction, so they were free to talk about anything. It is important to keep these types of distinctions in mind when developing and comparing performance across various spontaneously produced speech datasets.

Overall, the simulations we have presented have disentangled normalization and top-down information accounts and evaluated their relative efficacy on relatively naturalistic, spontaneously produced speech. Our results from Japanese vowel length suggest that while top-down information accounts are extremely useful even on spontaneous speech, results that argue for normalization only hold for controlled laboratory speech and do not generalize to the type of spontaneous speech that listeners hear. These results force us to scrutinize the role normalization can play in learning and processing, as well as the ways in which the primary cue for a distinction can shift based on the domain of speech.

In the following two sections, we consider what properties of spontaneous speech cause normalization to be ineffective. We provide simulations, followed by a theoretical analysis demonstrating that a listener that makes use of normalization will be impeded if sound categories in their input differ in the types of contexts they are likely to occur in.

4.2 Simulation: Imbalance between categories in the contexts they occur in can hurt normalization

We showed that normalization can help reduce category overlap between Japanese short and long vowels when applied to controlled lab speech, but not when applied to spontaneous speech. What are the properties of spontaneous speech that make normalization ineffective?

In this section, we provide simulations that reveal that one property of spontaneous speech that seems to play an important role is the fact that categories do not occur uniformly across contexts in spontaneous speech, as they do in controlled lab speech. That is, imbalances in where categories occur - precisely the type of signal that is helpful in top-down information accounts - can hurt normalization. We provide an example from the Werker controlled lab speech, in which we take advantage of one contextual factor - the following sound - that is not balanced between short vowels and long vowels. In particular, the consonants /g/, /s/, and /z/ (three of the eight consonants used in the study) each followed either only short vowels or only long vowels. Even within following consonants that occured both with short and long vowels, there were large imbalances in which vowels occured with which consonants. These types of imbalances are uncommon in carefully controlled lab speech, where researchers ensure that each vowel occurs in each context - but are extremely common in spontaneous speech, which has phonotactic constraints and phonological alternations. We previously showed that when the effect of the follow-

ing consonant was one of the contextual factors normalized out, normalization hurt on Werker Read speech, but when it was not normalized out, normalization helped. Here we show that this is because of the large imbalance observed between short and long vowels, by artificially balancing the dataset and showing that normalizing out the following consonant becomes helpful once it is balanced.

4.2.0.1 Methods and Data

The data we use come from the Werker dataset, as described previously. We test the efficacy of normalization (implemented via linear regression) on various subsets of the Werker Read speech data. We limit normalization to one contextual factor - the following consonant.

The first dataset is simply the full dataset (Full). As described previously, some of the consonants in the dataset exclusively follow either short vowels or long vowels (i.e. (/g/, /s/, and /z/). To create the second dataset, we remove all vowel tokens that precede one of these consonants and test the efficacy of normalization on this partially balanced dataset. The remaining consonants (/b/, /d/, /k/, and /p/) are all much more likely to follow one of the vowel categories than the other. For example, /k/ is twice as likely to follow short vowels than long vowels, even though it co-occurs with both. Therefore, to create the third dataset, we randomly remove enough tokens such that each following consonant is preceded by the same number of short and long vowels (Fully Balanced). The Fully Balanced dataset most resembles typical controlled lab speech corpora, as it completely controls for which vowels

occur with which consonants. For each of these three datasets, we test the efficacy of normalizing out the effect of the following consonant, by seeing whether normalized or unnormalized cues result in a better separation between short and long vowels. To ensure that differences in normalization efficacy between datasets are not due to changes in overall proportions of short/long vowels, or due to differences in dataset size, we create two additional control datasets: Partially Balanced Control and Fully Balanced Control. To create these datasets, we randomly remove the same number of short vowels and long vowels from the full dataset as are removed in the Partially Balanced and Fully Balanced datasets, but remove them randomly and uniformly from all contexts, instead of removing them based on the following consonant. We run normalization using linear regression on each of these five datasets, and test whether normalization is helpful on each of them.

4.2.0.2 Results

The results are summarized in Table 4.3. Normalizing for the effect of the following consonant is ineffective on the Full dataset: unnormalized cues result in 91.4% overall accuracy, while normalized cues result in 82.7% overall accuracy. However, normalization was more effective on the Partially Balanced dataset, which removed all vowel tokens that preceded a consonant that only occurred either with long vowels or with short vowels. Unnormalized cues result in 90.1% accuracy, while normalized cues result in 92.6% accuracy. Finally, normalization was even more effective on the Fully Balanced dataset: unnormalized cues resulted in 90.1% accuracy,

Data	Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Full	Unnormalized	91.4	89.7	92.9	246
	Normalized	82.7	79.5	85.7	525
Partially Balanced	Unnormalized	90.1	87.2	92.9	219
	Normalized	92.6	92.3	92.9	241
Fully Balanced	Unnormalized	90.1	87.2	92.9	172
	Normalized	93.8	92.3	95.2	112
Partially Balanced Control	Unnormalized	91.4	89.7	92.9	181
	Normalized	81.5	80.8	82.1	377
Fully Balanced Control	Unnormalized	90.1	89.7	90.5	151
	Normalized	83.5	79.5	87.1	325

Table 4.3: Results from balancing how often short/long vowels precede different sounds in the Werker Read Speech corpus. Results indicate that the more balanced the corpus, the better normalization performs.

while normalized cues brought the accuracy up to 93.8% accuracy. That is, each step of removing imbalances in the data resulted in improvements in normalization performance. In fact, when we completely balanced the dataset, normalization was effective. Just reducing the size of the tested dataset or changing the relative proportion of short/long vowels was not enough to explain this effect, as normalization was still ineffective on the Partially Balanced Control and Fully Balanced Control datasets.

4.2.0.3 Discussion

In this section, we explored why normalization is unhelpful on spontaneous speech. One difference between spontaneous speech and lab speech is that sound categories in spontaneous speech often differ in the contexts they are likely to occur in, while in lab speech, researchers specifically control where sounds occur to make sure that the dataset is fully balanced. We took advantage of one contextual factor within the Werker data for which this was not true, and found that when there were imbalances in a particular context, normalization hurt, but when we artificially balanced the context, normalization was helpful. That is, listeners relying on a normalization strategy when their input contains strong imbalances between categories would be hurt, unless they could somehow learn that they should not normalize for factors that are imbalanced. In this particular case, that would mean learning to normalize for the previous consonant, but not the following consonant.

If category imbalances across contexts were the only factor impeding nor-

malization in our analyses, then we would expect a similar manipulation to make normalization effective on the R-JMICC data. However, in further analyses (not described in detail here), we were unable to show that balancing contextual factors on the spontaneous R-JMICC data made normalization effective. This suggests that although contextual imbalances of this type constitute one key difference between lab speech and spontaneous speech, they are not the only reason that normalization is ineffective on spontaneous speech but not lab speech. Another possibility is that duration is less of a primary cue to vowel length in spontaneous speech than lab speech, and this could make normalization ineffective.

That being said, this simulation points to an interesting interaction between normalization and top-down information accounts, because the imbalances that are harmful for normalization are precisely the imbalances that are helpful for top-down information accounts. That is, when there is signal in the input that is helpful for top-down information accounts, normalization suffers. In the following section, we delve into this interaction in more detail.

4.3 Mathematical Analysis: Contextual category imbalances can hurt normalization, but systematic variability does not hurt top-down information accounts

We have seen that there are two ways that context affects sound production: it affects how likely a particular sound category is to be produced a priori, and, once that is decided, it affects what acoustic realization that sound category is likely to

have. As a result, there are also two main ways that listeners might make use of contextual information when processing or learning the sounds of their language. They could either make use of it to normalize the acoustics, or they could make use of it as top-down information that biases their category perception directly. Thus far, we have shown that in the case of Japanese vowel length, top-down information accounts are robust even on naturalistic speech, but that normalization is not effective on naturalistic speech.

The previous simulation points to an interesting interaction between normalization and top-down information accounts. In particular, it suggests that signal in the input that is helpful for top-down information accounts may be harmful for normalization accounts. In this section, we provide a theoretical analysis about how listeners relying on each of these two strategies will fare depending on the kinds of information sources that are present in their input, including what pitfalls they might encounter. We ultimately show that a listener relying on a normalization strategy when their input contains imbalances in categories across contexts may be misled, consistent with our previous simulation, while a listener who relies on a top-down information strategy when their input contains systematic variability resulting from context will not be. Overall, the results in this section suggest that top-down information strategies are much more robust to various types of input than normalization strategies are.

4.3.0.1 Contextual category imbalances can impede normalization

In this section, we consider how a listener relying on a normalization strategy will fare when their input contains imbalances in category membership - of the type that are helpful in top-down information accounts. In particular, we show that a listener relying on normalization when their input contains imbalances in category membership may be misled.

We begin by recapping what inference task we assume the listener is performing, and how exactly we implement normalization. As discussed in Chapter 2, we use a logistic regression categorization model, which involves calculating the relative probability that a particular vowel is long (or short) as follows where d refers to duration, f_1, f_2, and f_3 refers to formants and all β's refer to learned weights in the logistic regression.

$$P(\text{long}|d^{norm}, f_1^{norm}, f_2^{norm}, f_3^{norm}) = \frac{1}{1 + e^{\beta_0 + \beta_1 d^{norm} + \beta_2 f_1^{norm} + \beta_3 f_2^{norm} + \beta_4 f_3^{norm}}} \quad (4.1)$$

That is, this relies on having normalized duration and formants to categorize a particular vowel as phonemically short or long. There are a number of ways that normalization can be implemented. In our upcoming analysis, we focus on one particular way of normalizing, which is simple, but commonly used - namely, the linear regression implementation discussed in Chapters 2-3. This particular normalization strategy essentially works by subtracting from each vowel's acoustic cue the mean acoustic cue of all vowels that occur in the same context. All sounds that are longer

than expected based on their context will have positive normalized acoustic cues and all sounds that are shorter than expected will have negative normalized acoustic cues, regardless of what their absolute acoustic cues were. Again, this is only one way that normalization can be implemented, but the analysis of this simple case gives us insight into why normalization might not work and likely generalizes to many other normalization methods as will be discussed in Chapter 6. We turn to the analysis now.

In order for normalization to be helpful, we would expect normalization to push the means of the short and long vowel categories apart. To study when normalization is or is not helpful, we derive an equation that quantifies how the distance between category means changes as a result of normalization (as implemented by linear regression). The mean of short vowels before normalization, $\mu^{\text{unnorm}}_{l=\text{short}}$, is the average of the mean duration of short vowels in each context short vowels occur in, weighted by how many of all the short vowels occur in that context. In the following equation, $N_{l=\text{short},c=j}$ is the number of short (l = short) vowels in context j ($c = j$), $N_{l=\text{short}}$ is the total number of short vowels, and $\mu_{l=\text{short},c=j}$ is the mean duration of short (l = short) vowels in context j ($c = j$).

$$\mu^{\text{unnorm}}_{l=\text{short}} = \sum_j \frac{N_{l=\text{short},c=j}}{n_{l=\text{short}}} \mu^{\text{unnorm}}_{l=\text{short},c=j} \tag{4.2}$$

An analogous equation holds for the mean of long vowels before normalization, $\mu^{\text{unnorm}}_{l=\text{long}}$. We can then compute a closed form value for the means of the short and long vowel categories after normalization with linear regression - $\mu^{\text{norm}}_{l=\text{short}}$ and $\mu^{\text{norm}}_{l=\text{long}}$,

respectively. Each vowel token is normalized by taking the difference between that vowel's acoustic cue and the average acoustic cue of vowels that occur in that vowel's context. Once we obtain closed form values for the mean acoustics of short and long vowels pre- and post-normalization, we can derive the following equation, which shows how the difference between short and long vowel means changes as a result of normalization. This allows us to describe under what conditions category means will move closer together or farther apart as a result of normalization. Of course, the success of categorization depends not just on the difference in means, but how large this difference is compared to the variance. But in the simplest case we are considering, where normalization applies an additive mean shift without changing the variance, it is clear that normalization will hurt performance when the means become closer together. See Appendix B for a full derivation of this equation.

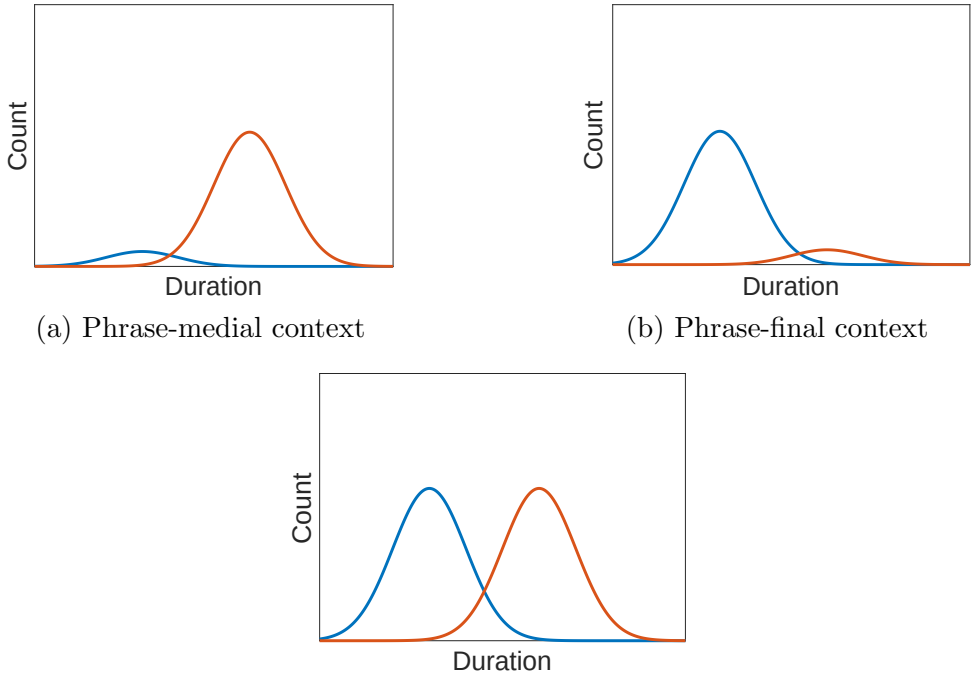
$$\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right) = \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}\right]\left[\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{long},c=j} + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{short},c=j}\right] \quad (4.3)$$

In this equation, $N_{l,c}$ is the number of vowels of length l in context c and $\mu_{l,c}$ is the mean of vowels of length l in context c. The first term in the sum, $\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}$, corresponds to the difference between the fraction of all short vowels that occur in the jth context and the fraction of all long vowels that occur in the jth context. The second term in the sum, $\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{long},c=j} + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu^{\text{unnorm}}_{l=\text{short},c=j}$, is a weighted average between the mean of the long vowels in this context weighted by the proportion of vowels in this context that are long and the mean of the short

vowels in this context weighted by the proportion of vowels in this context that are short. The product of these two terms is summed over all contexts. When the value of $\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right)$ from Equation 4.3 is greater than zero, this means that normalization has pushed the categories apart and when this value is less than zero, this means that normalization has pushed the categories closer together. This equation reveals that if there are imbalances in the contexts in which different categories are likely to occur in, then a listener relying on normalization alone may be misled. To illustrate why, consider a context that is dominated by long vowels (i.e. there are more long vowels than short vowels in this context). For such a contextual factor, we would typically expect the first bracketed term (of two) in Equation 4.3 to be negative. This is because it is likely that the proportion of all long vowels that are in this context, $\frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}$ is greater than the proportion of all short vowels that are in this context, $\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}}$ (although this need not be the case if, for example, there are many more long vowels than short vowels overall: $N_{l=\text{long}} > N_{l=\text{short}}$). In this long-dominated context, the second bracketed term (of two) would be relatively large for the following reason. Most of the vowels in this context are long (by virtue of it being a long-dominated context), so $\frac{N_{l=\text{long},c=j}}{N_{c=j}}$ will be relative large and $\frac{N_{l=\text{short},c=j}}{N_{c=j}}$ will be relatively small. The second bracketed term (of two) then consists of putting a higher weight on the long vowel mean than on the short vowel mean, which will push this value more towards the long vowel mean (and thus higher). Taking the product, the value within the sum will be a relatively large negative number for long-dominated contexts. Conversely, in a context that is dominated by short vowels (i.e. there are more short vowels than long vowels in this

context), we would typically expect the first term to be positive, and the second term to be relatively small, due to a heavier weighting on the short vowel mean than on the long vowel mean (which will push the value towards shorter durations). Taking the product, the value within the sum will be a relatively small positive number for short-dominated contexts. Overall, then, we would expect the sum over all contexts to be negative, since, as we saw, the negative summands should be relatively large, and the positive summands should be relatively small. At a high level, this means that imbalances in sound categories across contexts (i.e. large differences in the relative proportion of short and long vowels within particular contexts) can lead to normalization bringing the category means closer together, rather than farther apart.

Another way to think about this is that vowels are normalized relative to the context they occur in, by subtracting the mean of all of the vowels in that vowel's context from that vowel's own acoustic values. An imbalance between short and long vowels in a particular context will cause the mean of that context to be artificially decreased or increased, respectively. All else being equal, in a context that consists of a majority of long vowels, the mean duration will be artificially lengthened, so the normalized cues will be artificially low. A parallel effect will cause the normalized cues for short-dominated contexts to be artificially higher than expected. That is, vowels in contexts that are majority long will be shifted towards shorter durations, and vowels in contexts that are majority short will be shifted towards longer durations, which will push the short and long vowel distributions together. Essentially, the problem is that imbalances in where categories occur



(a) Phrase-medial context (b) Phrase-final context

(c) Overall

Figure 4.3: A toy example demonstrating how imbalances in the relative proportions of different categories across contexts can hurt normalization. In particular, imbalances can make it hard to estimate the normalization function, because they artificially shift the mean acoustic cue value in a context, which can hurt normalization methods that rely on shifting acoustics based on mean acoustic values across different contexts.

make it hard to estimate a proper normalization function.

Consider again the toy example in Figure 2.2, repeated in Figure 4.3. In this toy example, there are short vowels and long vowels and only two contexts. Note that the acoustics of the short and long vowels do not change across contexts – the average short vowel and long vowel durations are not shifted. However, there is a large imbalance between phonemically short and phonemically long vowels in particular contexts, such that there are many more long vowels than short vowels in phrase-medial position, and many more short vowels that long vowels in phrase-final position. This will cause the mean duration in phrase-medial context to be

much higher than the mean duration in phrase-final context. A listener relying on normalization would try to normalize and would actually increase the amount of within-category variability present in the speech stream and push the categories together. Overall, differences between sound categories, in terms of the contexts they are likely to occur in, can impede a listener who relies on normalization strategies.

To be clear, this analysis only reveals that normalization will be problematic for factors that can also be useful for top-down information accounts - e.g. neighboring sounds, prosodic position, but not for speech rate or speaker, which only affect how a sound is produced and not which sound is likely to be produced. That being said, normalization will be ineffective whenever it is difficult to estimate a normalization function, and there is reason to believe that this might be the case for factors like speech rate too. In particular, it has been shown that factors like speech rate seem to acoustically lengthen long vowels more than short vowels. In its current form, then, normalization will be incorrect for factors like speech rate, because it only estimates one normalization function for both short vowels and long vowels, instead of using a different function for each category. We return to this point and what it tells us about the efficacy of normalization in the General Discussion in Chapter 6.

4.3.0.2 Systematic variability does not impede top-down information accounts

While a listener relying on normalization when there is signal in the input for top-down information accounts will be misled, the opposite does not hold. A listener

relying on a top-down information strategy when there is systematic variability to be normalized in their input will not be misled relative to a listener who simply relies on the acoustics. The model making use of contextual information has access to all the information that the baseline absolute acoustics model does (and more), so it will necessarily perform at least as well. It can always learn to put no weight on contextual factors, and implement exactly the baseline model. Therefore, no matter what the acoustics are like, using top-down information will never mislead a learner more than a listener only relying on acoustic information. More strongly, a listener only relying on contextual information without any access to acoustics cannot be misled by systematic variability in the signal, precisely because it makes no use of acoustic information. Therefore, a listener who makes use of top-down information as a contextual bias in learning and perception will avoid pitfalls that a listener making use of normalization may encounter (as long as they trained on the right distributions).

4.3.0.3 Discussion

In this section, we provided a mathematical analysis showing that listeners who make use of a normalization strategy may suffer when there are imbalances in category membership of the type that are useful for top-down information accounts. However, the opposite is not true - listeners that make use of top-down information accounts will not be hurt by systematic variability in the signal.

Category imbalances are extremely common in natural language, due to phono-

tactic constraints, phonological alternations, historical reasons, and more. Our mathematical analysis shows that listeners who rely on normalization strategies may suffer when their input contains these types of category imbalances. Therefore, for factors that affect which sound category is likely to be produced, normalization is not an effective way to deal with context in processing and especially sound category learning for learners who cannot yet separate categories. Instead, a listener would be much better off making use of top-down information, which is immune to systematic variability in acoustics. Overall, in order to make a claim that listeners do use normalization in order to learn and process sounds, it will become important to explain how listeners can overcome the problems presented by contextual imbalances in category membership.

4.4 Summary

In this chapter, we delved in to the surprising finding that normalizing for systematic acoustic variability did not help separate short and long vowels. We showed that the discrepancy in results between our work and previous work arises because we study naturalistic speech, while past work has studied controlled lab speech. We ran the same analyses that we ran on Japanese naturalistic data on Japanese controlled lab data, and showed that while normalization can help on controlled lab data, it does not on naturalistic data. We then identified one difference between naturalistic data and controlled lab data that could cause different performance depending on the data type. In particular, we showed through simulations

and a mathematical analysis that imbalances in the relative proportion of short and long vowels across contexts can hurt normalization. These sorts of imbalances tend to be meticulously controlled in lab data, but are extremely common in naturalistic data due to phonotactics, phonological alternations, word properties, and other systematic regularities.

In the next chapter, we turn more directly to the question of how infants learn which dimensions of their language are contrastive, and propose a new learning account that takes advantage of these category imbalances, and provide evidence that supports it.

Chapter 5: A Learning Story: Naturalistic Data Support Distributional Learning Across Contexts

In this chapter, I turn more directly to the question of learning. In the previous chapter, we showed that an infant whose input contains imbalances would be better off using top-down information strategies than trying to normalize out systematic variability. With that in mind, we propose a new account for how infants learn which dimensions are contrastive - that takes advantage of the imbalances that have been shown to be very common in speech data. We provide initial support for this learning account, continuing to use vowel length as our test case.

As has been extensively discussed, what infants learn about vowel length differs by the language they are exposed to. For example, French infants learn that vowel length is not contrastive in their language (Jakobson and Lotz, 1949), Japanese infants, of course, learn that it is (Han, 1962), and Dutch infants learn that some vowel qualities in their language contrast by duration, but others do not (Booij, 1999).

Researchers have long thought that infants learn that a dimension is contrastive by tracking the acoustic distribution of speech sounds in their language and looking for bimodal distributions: if the distribution is bimodal, the dimension is

contrastive, and if the distribution is unimodal, the dimensions is not contrastive (Maye et al., 2002). This distributional learning account has received experimental and computational support from simplified speech (e.g. Maye et al., 2002; Vallabha et al., 2007), but is insufficient for naturalistic speech of the type that listeners mostly hear (Bion et al., 2013). In particular, in order for this account to be successful, it must be the case that distributions are unimodal when a dimension is non-contrastive, and bimodal when it is contrastive, and we have shown this is not necessarily the case in naturalistic speech. Distributions along contrastive dimensions can, and often are, unimodal (Bion et al., 2013), and, as a result, this account cannot explain why e.g. Japanese, but not French, infants learn that length is contrastive.

In this chapter, we present a new idea for how learners could use distributions to learn phonetic categories, inspired by work demonstrating that (i) infants are sensitive to higher-level contextual information (e.g. Feldman et al., 2013b; Thiessen, 2007), and (ii) there is real signal in higher-level contextual information for separating short and long vowels, as shown in Chapter 3. Using vowel length as a test case, we show that, unlike for previous distributional learning accounts, this story makes critical predictions that hold true on naturalistic speech. In particular, we show that it can qualitatively explain why Japanese, French, and Dutch infants learn what they do about vowel length.

The remainder of the chapter is organized as follows. In the first section, we lay out what the learning proposal is, and then we present analyses comparing French and Japanese, as well as contrastive vs. non-contrastive dimensions in Dutch.

These analyses show that there is signal in naturalistic speech that could support this learning account.

5.1 Proposal: Distributional Learning Across Contexts

The proposal that we explore here is that infants track acoustic distributions across contexts, and learn that a dimension is contrastive if the shape of the distribution along that dimension varies substantially across different contexts. As a concrete example, infants might break down the overall distribution of vowels into different distributions depending on, for example, what the preceding sound was[1], then compare the shape of those distributions, and learn that length is contrastive if those distributions vary substantially in shape.

This proposal builds off of Maye et al. (2002)'s version of distributional learning in that infants are tracking acoustic distributions to determine contrastiveness. It strays from their version of distributional learning, however, because this account does not require a bimodal distribution in the data. Rather, the account proposes that infants are looking to see whether multiple distributions (from different contexts) are shaped differently (e.g. more or less skewed right). Every distribution an infant observes could be unimodal, but as long as they are differently shaped unimodal distributions, the infant could still learn the contrast.

The reasoning behind this proposal is that the shape of the distribution will vary across contexts because the relative proportion of different sound categories will

[1] As will be discussed, in practice, I will consider more elaborate contexts, but I use preceding sound as an example here.

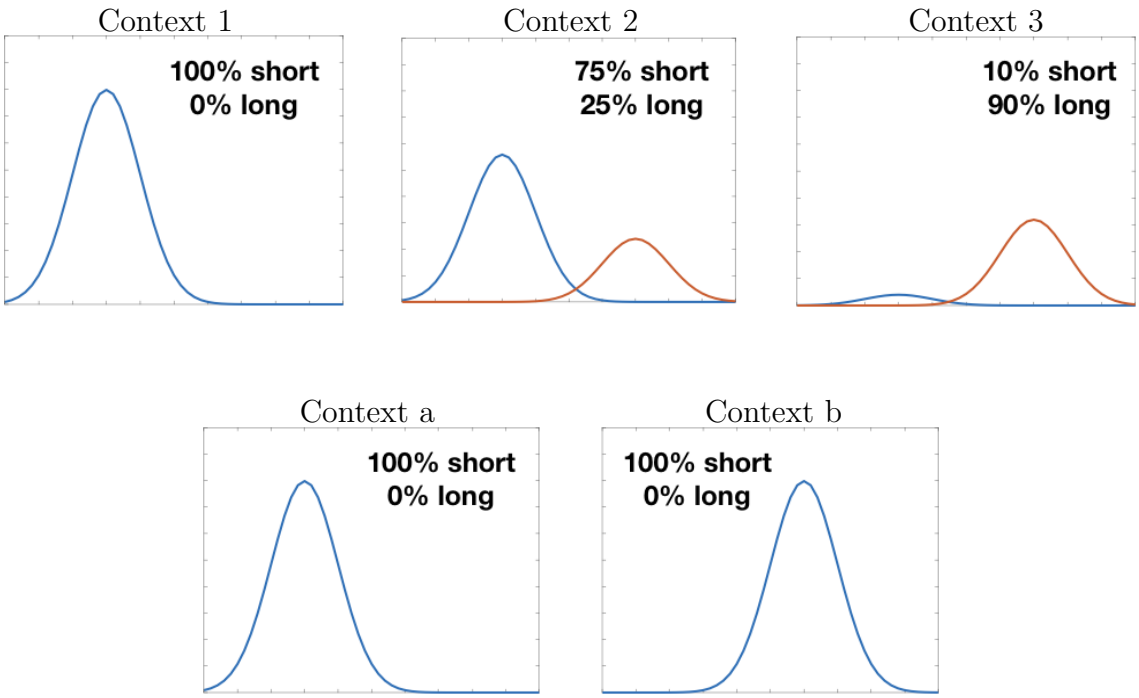


Figure 5.1: A toy example demonstrating how changes in the relative proportion of short and long vowels across contexts could lead to more extreme changes in distribution shape in a language with a contrast (top row) than one with no contrast (bottom row).

change across contexts, and different sound categories have different distributions. The relative proportion of different sound categories could change across contexts, because of phonotactics, phonological alternations, word properties, or other systematic regularities. These could cause the shape of the overall acoustic distribution to change across contexts, but only if there are multiple sound categories along that dimension.

In Figure 5.1, we present a toy example demonstrating this idea. In this figure, the top row represents different contexts in a language that has a contrast along this plotted dimension. The bottom row represents different contexts in a language that does not have a contrast along this dimension. In a language with

a contrast, different relative proportions of short and long vowels will cause the distribution shape to change. For example, the overall distribution in Context 1 is skewed right, whereas the distribution in Context 3 is more skewed left, even though both combined distributions will be unimodal. In a language without a contrast, we might expect the acoustic distribution to shift around a bit based on systematic contextual variability; however, we might not expect the distribution shape to change as much as when there is a contrast, because the distribution will be made up of the same category of vowels throughout. This leads to our idea that tracking distributions across contexts could be revealing about the existence of multiple categories.

Just as for standard distributional learning accounts, in order for this account to be successful, certain properties must hold of naturalistic speech. It must be the case that the distribution shape varies more across contexts in languages with a contrast than in a language without this contrast. It must also be the case that, within a language, the shape varies more when there is a contrast than when there is not. Through a number of analyses, we will show that these properties hold for vowel length in all the corpora we consider. In Analysis 1, we first show that, as our account predicts, there are more extreme distribution shape changes in Japanese (which has a length contrast) than French (which does not). This holds true both when we define a sound's context to be a combination of its quality, prosodic position, and neighboring sounds (e.g. phrase-medial /o/ vowels in frame t_k), as well as when we define a sound's context to be its word frame. We then turn our attention to Dutch, in which some, but not all, vowel qualities have vowel length contrasts. Using

similar analyses, we again show that within Dutch, the distribution shape changes more when we consider the subset of vowels that do contrast length than when we consider the subset of vowels that do not contrast length.

5.2 Analysis 1: French vs. Japanese

In this section, I test whether this learning account can qualitatively explain why Japanese infants learn that vowel length is contrastive in their language, while French infants do not, by testing whether the necessary signal is present in speech. That is, we test whether the acoustic distribution shape changes more substantially across contexts in Japanese than in French, and show that, as predicted by this account, it does.

5.2.1 Data

We use one Japanese corpus and one French corpus for our analyses. The corpora both consist of spontaneously-produced, adult-directed speech, but differ in the details of how the speech was elicited, as will be described below. Ideally, these analyses would use corpora of speech produced to infants who are learning the vowel length contrast; however, we do not have access to an annotated IDS corpus (for a language without a contrast) that parallels our Japanese IDS corpora, and being able to perform fair comparisons between the two languages is critical. In addition, results from Chapter 3 and Appendix A suggest that top-down information accounts performed similarly on ADS and IDS, so we do not have strong reason to believe

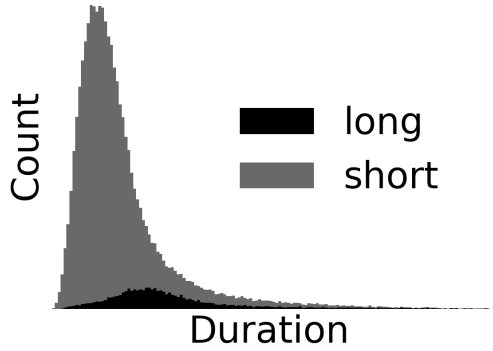


Figure 5.2: Overall distribution of short and long vowels in Japanese adult-directed speech (from CSJ). Similarly to the R-JMICC IDS corpus, only 10.9% of vowel tokens are phonemically long.

that the results would differ substantially if we used IDS instead, but future work should examine this more thoroughly.

5.2.1.1 The Corpus of Spontaneous Japanese (CSJ)

We use the Corpus of Spontaneous Japanese (Maekawa, 2003) for our Japanese corpus. This is a large corpus of spontaneously produced adult-directed speech. Around 90% of the speech consists of spontaneously produced monologues, and the remaining 10% consists of spontaneous dialogues. The monologues consist of speakers talking about their academic field, or more casual topics (e.g. what their most delightful memory is, what they would do if they lived on a deserted island, etc.) in a relaxed environment in front of a small group. The dialogue portions consist of (i) speakers discussing the content of the monologues, (ii) speakers engaged in task-oriented dialogue, where they need to guess the fees associated with a celebrity appearance, or (iii) speakers engaged in free conversation with the

experimenter.

The full corpus consists of 7.5 million words, or around 650 hours, of spontaneous speech, but only a portion of the full corpus was manually annotated for the segmental information required for our analyses. As a result, our analysis focuses on the core portion of the CSJ. The core portion includes speech by 75 speakers, and consists of about 500,000 words or 44 hours of spontaneous speech (of which about 15 hours are dialogues). The core portion of the speech corpus consists of 811,731 total vowel tokens of which 722,968 (or 89.1%) are phonemically short and 88,763 (or 10.9%) are phonemically long. Just as in the R-JMICC IDS corpus, the overall distribution is unimodal along the duration dimension (see Figure 5.2). Analyses of the speech reveal that it has the expected properties of spontaneous speech, including phonetic reductions, disfluencies, informal language, faster speaking rate, and so forth (Maekawa, 2003).

Transcriptions

The core portion of the corpus was annotated at the segment and word levels by the corpus creators (Maekawa, 2003). The segmental transcriptions are basically phonemic, but some phonetic information was also included (e.g. information about vowel devoicing, timing of closure release in stops, and palatalization of consonants before /i/, and more). Researchers provided phonetic transcriptions of speech from which phoneme labels were created. These phoneme labels were automatically aligned to the speech using an Hidden Markov Model-based alignment procedure. These automatic alignments were then hand-corrected by expert human labellers.

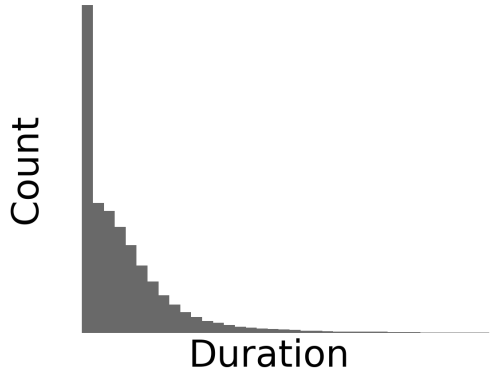


Figure 5.3: Overall distribution of vowels along the duration dimension in French adult-directed speech (from NCCFr). The duration measurements in this corpus were relatively imprecise, so the tall peak could arise because vowels of shorter durations could not be represented.

5.2.1.2 Nijmegen Corpus of Casual French (NCCFr)

We use the Nijmegen Corpus of Casual French (Torreira et al., 2010) for our French corpus, which like the CSJ is a corpus of spontaneously produced adult-directed speech. Unlike the CSJ, however, the NCCFr consists exclusively of conversational speech. The corpus consists of a total of about 36 hours of spontaneous speech, and involves forty-six French speakers (24 male, 22 female) engaged in conversation with their friends. The corpus consists of 132,037 total vowel tokens, whose distribution along the duration dimension is plotted in Figure 5.3.

To get the recordings, twenty-three confederates were recruited, and asked to bring two close friends of the same sex with them to the recording. The confederate knew what the purpose of the recordings was (i.e. phonetic analysis of spontaneous speech), but the two friends did not, and the confederate's role was to elicit natural speech, by continuing the conversation by raising familiar topics. For technical

reasons, the two friends' speech was recorded, but the confederate's speech was not recorded. Each session lasted about 90 minutes, and included three types of speaking tasks. In the first portion of the session (∼10-30 minutes), the experimenter informed the group that the confederate's microphone was not working, and had the confederate leave the room with them to repair it. The speech of the other participants (who stayed in the room) was recorded during this time. In the second portion of the session, the confederate re-entered the room and joined the conversation. Topics were left entirely up to the participants, and included discussion about upcoming exams, an ongoing strike, and travel plans. In the third and final session (∼40 minutes), speakers chose five questions about social and political issues from a list and had to negotiate an answer to each question, which they wrote down. The corpus creaters compared the speech to journalistic speech and analyses confirmed that this corpus did indeed consist of casual speech (Torreira et al., 2010). The NC-CFr involved more disfluencies, more repeated words, more casual and slang words, more swear words, and fewer uses of double negation (an indicator of formal speech in French) than the journalistic speech.

Transcriptions

The corpus was orthographically transcribed by two professional transcribers. The corpus was transcribed at the segmental level by Martine Adda Decker (p.c. with M. Ernestus, January 14, 2019). The precise details of how the corpus was transcribed are not known; however, other phonetic research has been done using these transcriptions (e.g. Torreira and Ernestus, 2010).

5.2.1.3 Information Extracted from Corpora

We extract a number of pieces of information from the two corpora.

Acoustic cues

Unlike in the analyses in Chapters 3-4, we only extract the duration of each vowel in milliseconds, and do not extract formant values. Unfortunately, the French corpus had relatively imprecise measures of duration, and did not represent differences in duration that were shorter than 50ms. In order to make the comparison between French and Japanese fair, we rounded all durations in the Japanese corpus to this same degree of precision as in French, which did not have a big effect on the shape of the Japanese distribution shown in Figure 5.2.

Contextual factors

In addition to extracting acoustic information, we also extracted contextual information about each vowel. We extracted all contextual information that was available for both corpora.

- **Vowel quality**: This was a categorical variable that took one of five values (/a/, /e/, /i/, /o/, /u/) for Japanese, and one of the following values for French: /a/, /e/, /i/, /o/, /u/, /ɛ/ /ə/, /œ/, /ø/, /y/, /ɔ/, nasalized /ɛ/, nasalized /a/, and nasalized /ɔ/.

- **Prosodic position**: We represented prosodic position with four indicator values: (1) whether the vowel was word-initial or not (only the first vowel in a

Analysis	What contextual information was used
Analysis 1a	Neighboring sounds + prosodic position + vowel quality (e.g. phrase-final [o] vowels in frame [s_#]
Analysis 1b	Word frame (e.g. vowels in frame [k_ko])

Table 5.1: Description of what contextual information was used in each analysis. We only include contextual information that infants have been shown to have access to by the time they learn the vowel length contrast.

word counted), (2) whether the vowel was word-final or not, (3) whether the vowel was phrase-initial or not, and (4) whether the vowel was phrase-final or not.

- **Neighboring sounds**: We extracted the identity of the immediately previous sound and the immediately following sound (both categorical variables), as labelled by the phonetic transcription. This was marked as '#' if the vowel was preceded by silence. Again, vowel length contrast is thought to be learned later than other types of contrasts (Sato et al., 2010).

- **Word frame**: We extracted the word frame that the vowel occurred in. For example, one word frame could have been [b_i_ru], which would include both [biːru] and [biru]. The word frame included quality information for each sound in the word, but did not include any length information about the sound in question or any of the other sounds in the word frame.

5.2.2 Methods

5.2.2.1 Analysis Description

We perform two different comparisons between French and Japanese. Both aimed to test whether distribution shape changes across contexts were more extreme in Japanese than in French, but differed in how they defined context (Table 5.1).

In the first analysis (Analysis 1a), we considered the context of the sound to be: (i) the vowel quality of the vowel, (ii) the prosodic position of the vowel, and (iii) the immediately preceding and immediately following sounds. That is, one "context" could be word and utterance-medial /o/ vowels that were preceded by /t/ and followed by /k/. We chose these contextual factors based on a combination of what was readily available from the corpus annotations for all corpora, what had been used in the analyses in Chapter 3, as well as what infants are thought to know about by the time they learn about the length contrast.

In the second analysis (Analysis 1b), we considered the context of the sound to be the word frame that it occurred in without any length information (i.e. with long and short sounds represented identically).

Other than this, the two analyses were identical. In the results we present, we took every context that had at least 50 vowel tokens in it[2], so as to minimize both sparsity issues and computation time. We then computed the distance between the shape of the distribution in each context as compared to every other context that met threshold in a pair-wise fashion. That is, if there are 100 contexts with

[2]As will be discussed, the same results hold for every way of choosing contexts that we tested.

more than 50 vowel tokens in them, each of those contexts will be compared against the remaining 99. We computed the average pairwise distance for each language, and also plotted the overall distribution that the distances took in both Japanese and French. We used Earthmover's distance as the distance metric between two distribution shapes, as described in the following section.

5.2.2.2 Earthmover's Distance

Earthmover's distance, also known as Wasserstein or Kantorovich-Monge-Rubinstein distance, is a metric that is commonly used in computer science and statistics to measure the distance between two distributions (Rubner et al., 1998; Villani, 2008).

Informally, it is often talked about in terms of two piles of dirt, which represent the two distributions being compared. In this context, Earthmover's distance can be thought of as minimum cost of turning one earth pile into the other, where cost corresponds to a combination of the amount of earth being moved as well as the distance it has to be moved. In other words, the distance is the minimum average distance a piece of dirt will have to be moved in order to turn one pile into the other.

More formally, the Earthmover's distance, W, between two distributions, $\mathbb{P}_X$ and $\mathbb{P}_Y$ is defined as follows:

$$W\left(\mathbb{P}_X, \mathbb{P}_Y\right) := \min_{\mathbb{P}_{X,Y}} \mathbb{E}\left[\|X - Y\|\right], \tag{5.1}$$

$$\sum_Y \mathbb{P}_{X,Y} = \mathbb{P}_X, \sum_X \mathbb{P}_{X,Y} = \mathbb{P}_Y$$

where $\mathbb{P}_{X,Y}$ is the joint distribution whose marginals are equal to $\mathbb{P}_X$ and $\mathbb{P}_Y$. The optimization (minimization) is over every possible joint distribution $\mathbb{P}_{X,Y}$ (with the correct marginals). We can think of each of these distributions as one solution that tells us what "dirt" in the first distribution corresponds to what "dirt" in the second distribution (that will allow us to transform one distribution into the other). In other words, if there is a piece of "meta-dirt," that we need to place in the first distribution, $\mathbb{P}_X$, and the second distribution, $\mathbb{P}_Y$, we can think of X in $\mathbb{P}_{X,Y}$ as the location where that piece of dirt gets placed in the first distribution and Y as the location where that piece of dirt gets placed in the second distribution. The requirement that the marginals equal $\mathbb{P}_X$ and $\mathbb{P}_Y$ is to ensure that we only consider solutions in which the way we have placed the dirt actually creates the distributions we are considering. Then, for each of these possible solutions/mappings between the dirt in the first distribution and the dirt in the second distribution, we can calculate how far away, on average, the location where we place our meta-dirt in our first distribution is from the location where we place it in our second distribution. This corresponds to the expectation in Equation (5.1). The Earthmover's distance between two distributions, then, is the minimum of those average distances over all possible solutions.

As a concrete example, consider two identical piles of dirt (distributions), where half the mass is at 0 and half the mass is at 1. By virtue of these distributions being identical, they have the same shape. One possible solution would be to "move" all of the dirt from 0 to 1, and all of the dirt from 1 to 0. However, a simpler solution would be to just keep all of the dirt where it is. As a result, the Earthmover's distance

will be 0, because at minimum, we do not have to move any dirt, even though there are solutions where we will have to move it more. In sum, Earthmover's distance is a metric of how much we have to alter one distribution to make it into another, and we use this metric to compare whether there are greater distribution shape changes across contexts in Japanese than in French.

5.2.3 Results

The results comparing Japanese and French are shown in the boxplots in Figure 5.4 (in which, context = quality + prosodic position + neighboring sounds) and Figure 5.5 (in which, context = word frame). Each point in these graphs represents how different in shape the acoustic distributions are between a pair of contexts, with greater differences in shape indicated by higher Earthmover's distance. Both the average and median pairwise Earthmover's distance is higher in Japanese than French for both Analyses 1a and 1b. In Analysis 1a (context consists of the quality, prosodic position, and neighboring sounds), the average Earthmover's distance is 42.9 in Japanese, but 25.2 for French. The median Earthmover's distance is 27.7 in Japanese, but 19.7 in French. In Analysis 1b (context is simply the vowel's word frame), the average Earthmover's distance is 36.2 in Japanese, and 26.7 in French. The median Earthmover's distance is 21.0 in Japanese, and 17.9 in French.

However, more strikingly, the French and Japanese boxplots differ in their overall shape, in that the Japanese plots have longer tails (more outliers extending towards larger values of Earthmover's distance). This means that there are more

French vs. Japanese Results
(Context = Vowel quality + Prosodic position + Neighboring Sounds)

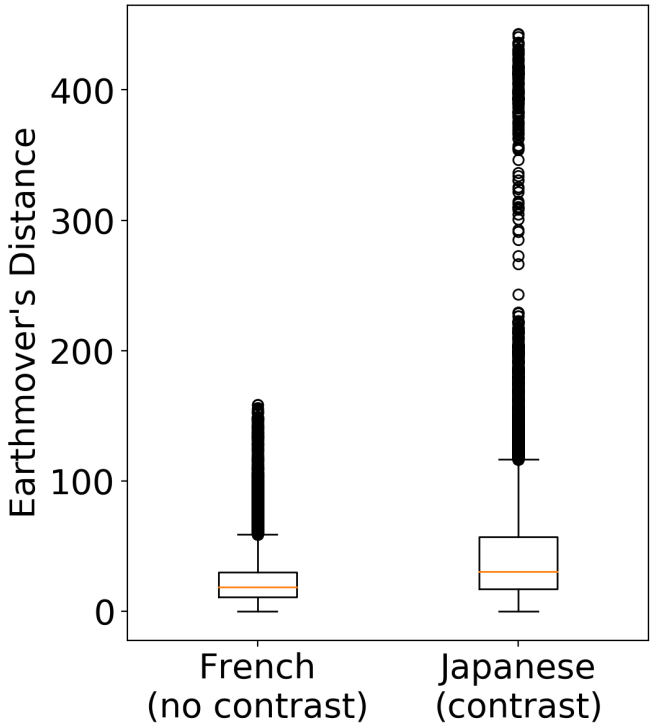


Figure 5.4: Results from Analysis 1a, which compared French and Japanese. Context is defined as a combination of vowel quality, prosodic position, and neighboring sounds. Each point represents how different the acoustic distributions are between a pair of contexts, with greater differences indicated by higher Earthmover's distance. The Japanese plot has a much longer tail, which means that there are more pairs of contexts with extreme distribution shape changes in a language that has a contrast along this dimension.

Context	Percent Long	Count	Frequency Rank
Phrase-initial, word-final /e/ vowels	64.7	1357	18
Phrase-initial and final /a/ vowels	56.7	255	95
Phrase-initial and final /e/ vowels	87.9	244	100

Table 5.2: Information about a subset of the Japanese contexts that drive the long tail in Analysis 1a (Figure 5.4 above). "Count" refers to the number of vowel tokens in this context in the portion of the CSJ that we used for this analysis. "Frequency rank" indicates the rank of the context when all contexts are ordered in terms of frequency (i.e. a rank of 1 would mean this was the most frequent context observed).

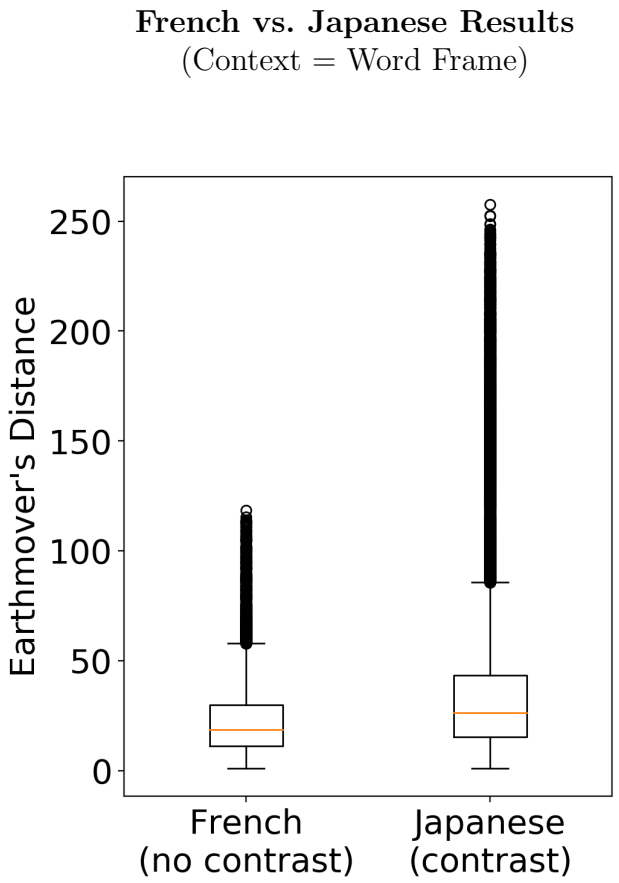


Figure 5.5: Results from Analysis 1b, which compared French and Japanese. Context is simply defined as the word frame the sound occurred in. Greater differences in distribution shape are indicated by higher Earthmover's distance.

Word Frame	Percent Long	Count	Frequency Rank
#_e_#	68.1	2465	5
#y_u_#	92.8	1375	12
#an_o_#	51.9	981	20
#y_o_#	75.7	569	29
#_i_#	30.9	538	32
#_e_to#	81.0	385	38
#_a_#	41.5	328	52
#_u_#	30.0	70	171

Table 5.3: Information about the Japanese contexts that drive the long tail in Figure 5.5 above.

pairs of contexts that have extreme distribution shape changes in Japanese, a language that has a contrast along the duration dimension, than in French, a language that does not. This overall qualitative pattern holds for both ways of defining context, though the extreme Earthmover's distances are much higher for Analysis 1a than 1b. In particular, for Japanese, some context pairs exceed Earthmover's distances of 400 in Analysis 1a, but only 250 in Analysis 1b. Meanwhile, in Analysis 1a, the French context pairs reach Earthmover's distances of almost 200, compared to 125 in Analysis 1b.

Overall, in both cases, we observe that the distribution shape changes are much more extreme when a contrast is present than when it is not. These results suggest that there is signal present in naturalistic speech that support the proposed learning account in both ways of relying on contextual information.

Further analysis of the contents of the Japanese tails reveals that these tails are largely driven by a few contexts that have a different shape from all of the remaining contexts. These driving contexts (contexts that repeatedly and consistently result in Earthmovers distances above 250 in Analysis 1a or above 150 in Analysis 1b when paired with other contexts) tend to be contexts that consist mostly of phonemically long vowels. That is, they are contexts in which long vowels are extremely overrepresented relative to their overall 10% base rate of occurrence. For example, in Analysis 1a, one driving context is /e/ vowels that are both phrase-initial and phrase-final (i.e. utterances that consist only of the vowel quality [e]). 87.9% of vowels in this context are phonemically long, while only 12.1% are phonemically short. Another driving context is [a] vowels that form their own utterances. In this

context, 56.7% of vowels are phonemically long, while 43.3% are phonemically short. In Analysis 1b, the following word frames are among the driving contexts (where the underscores surround the vowel in question, whose quality is known, but length is unknown): [_e_] (in which 68.1% of vowels are phonemically long), [y_u_] (in which 92.8% are long), [an_o_] (in which 51% are long), [y_o_] (in which 75.7% are long), [_a_] (in which 41.5% are long), and [_e_to] (in which 81.0% are long). These results reveal that not only do we get the predicted pattern, but that this pattern appears for the hypothesized reason. Contexts differ in the relative proportion of short and long vowels when there is a contrast, which causes the distribution shape to change more across contexts when there is a contrast than when there is not.

5.2.4 Discussion

In this section, we showed that critical predictions that our learning account makes hold true on naturalistic speech. In particular, in two different ways of defining 'context,' Japanese exhibits much more extreme distribution shape changes across contexts than French. Therefore, our results are among the first to be able to qualitatively explain why Japanese infants might learn that length is contrastive, while French infants do not (despite the fact that both overall distributions are unimodal). This account is promising because the contexts that seem to drive this effect are very frequent, so they make up a large proportion of the tokens and would, arguably, be quite noticeable to infants (assuming these results replicate on IDS). Most importantly, unlike past accounts, this account works for naturally produced

speech.

This learning account takes advantage of precisely the signal that the top-down information account used to achieve good categorization performance in Chapter 3. In particular, this account relies on there being different relative proportions of short and long vowels across different contexts, which could arise for grammatical reasons (e.g. phonotactics, alternations, and so forth) or non-grammatical reasons (e.g. historical reasons). Because the signal is driven by these contexts in which long vowels are overrepresented relative to their low overall base rate of occurrence, this account makes the strong prediction that these types of contexts are necessary for phonetic learning to happen. This prediction could be tested relatively easily. For the signal we are currently testing for, we need to have access to speech corpora that have been transcribed and aligned at the sound level. But, to test for the existence (and sufficient frequency) of these driving contexts, we could simply use orthographically annotated corpora, which are much more prevalent.

In this case of Japanese, most of these critical contexts arise because of specific word types/exclamations. That is, there seem to be a few word types with long vowels that either do not have short vowel minimal pairs, or, do, but outnumber them drastically. Even when we define context as a combination of the vowel quality, neighboring sounds, and prosodic position, the signal we are picking up on seems to come from particular word types/exclamations. In fact, some of the signal in Analysis 1a comes from the same words as in Analysis 1b. For example, [a] and [e] that form an entire utterance provide some of the signal in Analysis 1a, while the frames [_a_] and [_e_] provide some of the signal in Analysis 1b. Despite the fact that

they are the same word types/exclamations, they result in more extreme distance measures in Analysis 1a than 1b because by not including word information, these frames are only picked up when they occur as an entire utterance. These vowels are substantially acoustically lengthened when they occur as an entire utterance, so they deviate even more strongly from the typical context shape than when we consider all instances of these words. This explains why the tail extends to much longer distances in Analysis 1a and 1b.

Two notes on the status of the word in these analyses: first, while it is true that words are the source of this signal in Japanese, this need not be the case, and the source of the signal could vary across contrasts and languages. For example, to differentiate [n] vs. [ŋ] in English, the fact that [ŋ] does not occur word-initially could serve as signal.

Second, despite the fact that we build in word-level information to both of the analyses (i.e. whether a sound is word-initial or word-final in Analysis 1a and the word frame in Analysis 1b), the results from Analysis 1a suggest that listeners might be able to pick up on these contrasts without full word information. This is because the driving contexts in Analysis 1a, shown in Table 5.2, consist of vowels that form entire utterances (and are surrounded by pauses), so Japanese infants could notice that the shape of the distribution changes substantially in some contexts having only partial word-form knowledge.

In summary, this analysis compared two different languages which differ in how length is treated, and showed that distribution shape changes were greater for a language with a contrast than without. Crucially, distribution shape changes that

arise from different relative proportions of short and long vowels seem to be larger than distribution shape changes that stem from systematic acoustic variability.

Of course, in order for this learning account to work, it would need to explain, within one language, which dimensions infants learn are contrastive or not. In the next section, we perform this analysis, showing that within one language, places where contrasts exist show much more extreme shape changes than places where contrasts do not exist.

5.3 Analyses 2 and 3: Dutch

As discussed, we would ideally also like to show that within a language it is possible to pick out the contrastive vs. non-contrastive dimensions by looking for extreme distribution shape changes. Direct comparisons between different acoustic dimensions (e.g. duration vs. formants in French, where the former does not contrast categories but the latter does) are difficult because different dimensions have different scales, making it unclear how to make the necessary comparisons.

We overcome this issue in this section, by taking advantage of a property of Dutch vowel length. In Dutch, some pairs of vowel qualities also contrast in length, but others do not. In particular, while /ɑ, aː/, /ɛ, eː/ and /ɔ, oː/ are said to contrast in duration, /ɪ, i, u, y/ are said not to (Nooteboom and Slis, 1972; Swingley, 2019). This is a slightly different case than in Japanese, because the contrasts are also accompanied by vowel quality changes; however, the contrast exists and infants learn about it (Dietrich et al., 2007; Van der Feest and Swingley, 2011). In this

section, we use two Dutch corpora (one infant-directed and one adult-directed) to show that the distribution shape changes more substantially across contexts when we consider the subset of vowels that do exhibit a vowel length contrast than when we consider the subset of vowels that do not exhibit a vowel length contrast.

5.3.1 Data

5.3.1.1 Ernestus Corpus of Spontaneous Dutch (ECSD; Dutch ADS)

We use the Ernestus Corpus of Spontaneous Dutch (Ernestus, 2000) for one of our Dutch corpora. The corpus consists of adult-directed, conversational speech that was spontaneously produced by twenty different speakers. The speakers were all male, and were between the ages of 21 and 55 years old. The corpus consists of about 15 hours of recorded speech, which includes 153,200 word tokens and 9,035 word types. The speech was elicited as follows. Pairs of colleagues or friends talked with each other in a soundproof room in sessions that lasted about 90 minutes each. The sessions had two portions. In the first half, the pairs of speakers spoke freely about any conversation topic they wanted. The experimenter was present, but only participated if they needed to enliven the conversation. In the second half, the pairs of speakers engaged in a task-oriented conversation, in which they negotiated about the purchase of camping goods. As with the other corpora used in this chapter, the speech was casual in nature, and exhibited many properties of spontaneously produced speech, despite the slightly artificial setting of the recording. It is worth noting that four word types accounted for 8.2% of all of the word tokens, and two

of them had long vowels in them (*maar* 'but' and *nee* 'no').

Transcriptions

Professional transcribers created a orthographic transcription of the interactions, which was manually aligned to the speech. The corpus was also phonetically transcribed; however, unlike with the past corpora, the phonetic transcription was created via forced alignment, not manually. The forced alignment procedure took as input the speech files, the orthographic transcriptions, a lexicon containing multiple possible pronunciations for all word types, and acoustic models for the phones, and output a phonetic transcription that was aligned to the speech. More details about how the forced alignment worked can be found in Schuppler et al. (2011).

The forced aligned annotations were validated by Schuppler et al. (2011), who performed the alignment. Because there was no manually annotated portion of the ECSD that could be used to validate the transcriptions, the researchers evaluated the quality of the forced aligner using the spontaneous portion of the IFA Dutch Spoken Language Corpus, a similar corpus that was manually annotated. They ran the exact same forced aligner on that corpus and compared the output of the forced aligner against the manual annotations. The authors observed a 14% discrepancy between the manual annotations and forced-aligned annotations which is in the range of human disagreement. However, because they did not directly validate durational information, it is currently unclear how accurate annotations of the start and end points of the phones are. This could introduce some noise into our analyses, as it will have consequences on how accurate the vowel durations are, as well as how

accurately we can determine which word a vowel belong to, and future work should replicate these findings once cleaner data become available.

5.3.1.2 Swingley Dutch Infant-Directed Speech

We also tested the predictions of the account on a corpus of Dutch infant-directed speech collected by Fikkert (1994) and Levelt (1994). The annotated portion of this corpus is small: it contains a total of 300 utterances, with a total of only 1296 vowel tokens. However, the data more closely resembles the type of data that infants learn from, so we chose to analyze it.

The corpus consists of naturalistic longitudinal observations. Experimenters did home visits every two weeks for about 7-15 months and recorded their spontaneous interactions with the children. The full corpus consists of speech directed at 11 children; however, the portion we use only includes speech directed to one of them (Catootje), who was aged 1;10. The speaker was a native Dutch-speaking mother, but was not the child's mother, though a caregiver was usually present for the interactions. The spontaneous speech was produced in the context of playing or discussing picture books in sessions that lasted between 30-45 minutes.

Transcriptions

The corpus was transcribed at the word level, and time-aligned phonetic annotations were created by Dan Swingley (Swingley, 2019). Given the transcriptions, the speech toolkit HTK (Young et al., 2002) was used to estimate the boundaries

of the phones using the HVITE forced-alignment tool. The output of the forced-alignment tool was manually corrected in Praat (Boersma, 2001) by D.S. a speaker of Dutch. The boundaries of the phones as well as the labels were corrected, as needed.

In the shareable format of the corpus, the word-level transcription was not time-aligned to the speech. Using the word-level transcriptions as well as the time-aligned phonetic transcriptions, I manually aligned the word-level transcription to the speech, based on the location of the phones. Although I am not a speaker of Dutch, this was a straight-forward task because the mapping between the word-level information and sound-level information was apparent.

5.3.2 Methods

The methods were largely similar to those used in the French vs. Japanese comparisons, but with slight modifications based largely on corpus size. We subsetted the Dutch ADS and IDS data sets into two portions each, that we then compared within corpus. For both corpora, the first subset consisted of all /ɑ, aː, ɛ, eː, ɔ, oː/ vowels (which are thought to contrast in duration). The second subset consisted of all /ɪ, i, y, u/ vowels (which are not thought to contrast by duration) (Nooteboom and Slis, 1972; Swingley, 2019). The IDS analyses effectively only included /ɪ, i, u/ vowels in the second subset, because there were only a couple of /y/ vowels in the whole corpus. In the following sections, we outline the corpus-specific design decisions we made.

5.3.2.1 Analysis 2: Dutch ADS

For the analyses on the adult-directed speech, we again defined context in two ways: in Analysis 2a, we defined context as a combination of the vowel quality, prosodic position, and neighboring sounds. In Analysis 2b, we defined context as the word frame of the vowel. We then chose contexts that had over 50 vowel tokens in them. As before, we did pairwise one-by-one comparisons between the distribution shape in each context and every other context that met the threshold of 50 vowel tokens. We again used Earthmover's distance to quantify the distribution shape change. We computed the overall average Earthmover's distance, and plotted the overall distribution of pairwise distances observed for each subset of vowels.

For Dutch, there were more vowels in the "contrastive" subset than the "non-contrastive" subset. This means that there were relatively more contexts that met the 50 vowel token threshold in the contrastive case than the non-contrastive case. To try to address this issue, we also ran versions (i) where we included the same number of contexts for both the Dutch IDS and the Dutch ADS, and (ii) where we artificially decreased the size of the (larger) contrastive subset. All alternatives resulted in the same qualitative findings as those we report here.

5.3.2.2 Analysis 3: Dutch IDS

Unlike for the other analyses, for the analyses on Dutch IDS, we only defined context in one way: word frames. We did not run an analysis looking at context as a combination of the vowel quality, neighboring sounds, and prosodic position, due

to the small dataset size.

We chose all word frames that had at least 10 vowel tokens in them. As before, we did pairwise one-by-one comparisons between the distribution shape in each context and every other context that met threshold. We used Earthmover's distance, computed the overall average Earthmover's distance for each subset of data, and plotted the distribution of distances observed.

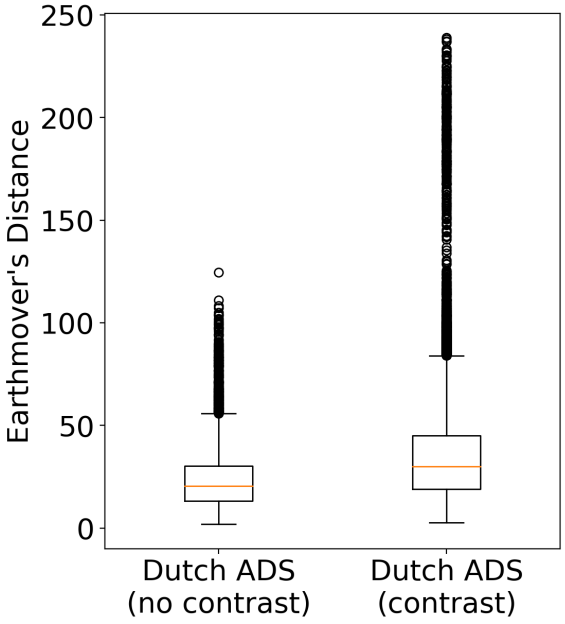
5.3.3 Results

The results comparing contrastive and non-contrastive vowels in Dutch mirror the results observed when comparing French to Japanese.

5.3.3.1 Dutch ADS

The results from Dutch ADS (Analysis 2) are shown in Figure 5.6. As before, the average and median Earthmover's distance is larger when there is a length contrast than when there is not. For Analysis 2a (context = quality + prosodic position + neighboring sounds), the average Earthmover's distance is 40.9 for the contrastive subset, but 29.0 for the non-contrastive subset. The median Earthmover's distance is 25.8 for the contrastive subset and 21.2 for the non-contrastive subset. For Analysis 2b (context = word frame), the average Earthmover's distance is 35.4 for the contrastive subset of vowels, but 30.7 for the non-contrastive subset. The median Earthmover's distance is 29.3 for the contrastive subset of vowels, but 26.3 for the non-contrastive subset.

Context = Vowel quality + Prosodic position + Neighboring Sounds



Context = Word Frame

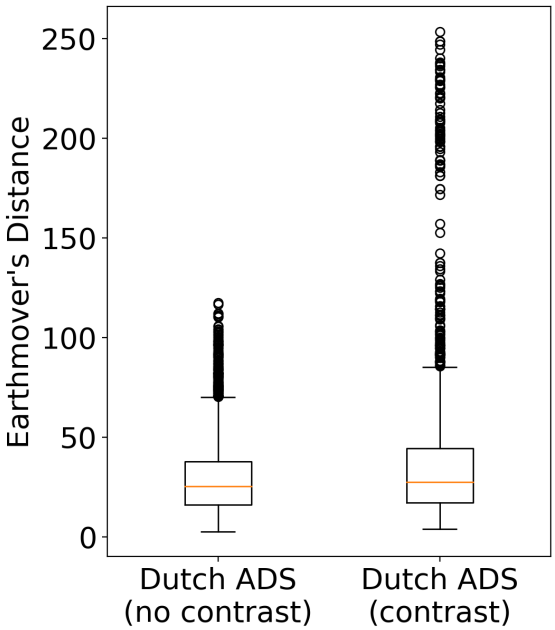


Figure 5.6: Distribution shape changes across contexts in Dutch ADS

In addition, just as in French/Japanese, the boxpot corresponding to the contrastive subset has a much longer tail than the boxplot corresponding to the non-contrastive subset, in both ways of defining context. The one difference that we do observe between this analysis and the previous analysis is that the tail for Analysis 2a (Dutch ADS) does not extend as far as for Analysis 1a (French/Japanese). In this analysis, the highest distance is around 250, while it reached 400 in the French/Japanese analysis. One reason for this could be that length contrasts have a different status in Dutch vs. Japanese. In Dutch, there are also quality differences, while they are strictly length contrasts in Japanese. Another reason could be that the Japanese analysis was picking up on exclamations which were particularly acoustically long.

Overall, however, we observe the same pattern of results as before, but within one language instead of between languages. This is real signal that Dutch listeners could pick up on to learn about the vowel length contrast in Dutch.

5.3.3.2 Dutch IDS

The results from the Dutch infant-directed speech corpus (context = word frames) are shown in Figure 5.7. The results are, again, consistent with our learning account. In particular, the average Earthmover's distance is higher for the vowels that contrast in length than for vowels that do not. The average Earthmover's distance is 49.0 for the contrastive subset, but 27.8 for the non-contrastive subset. The median Earthmover's distance is 27.8 for the contrastive subset, but 23.8 for

Context = Word Frame

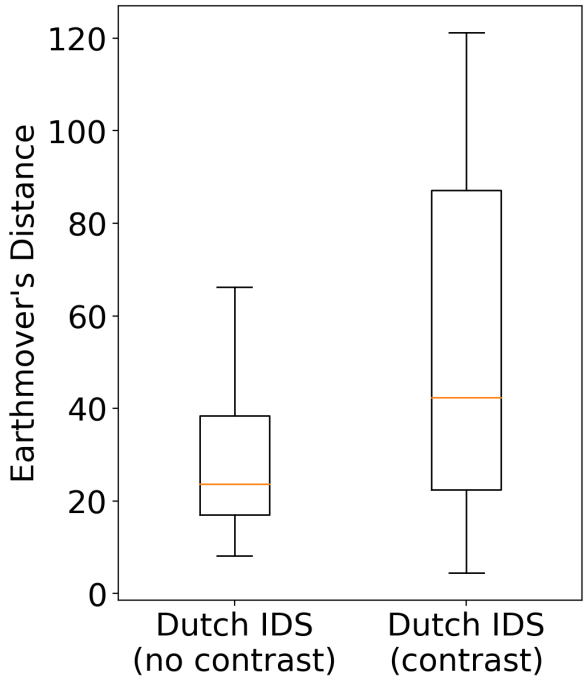


Figure 5.7: Distribution shape changes across contexts in Dutch IDS

the non-contrastive subset.

In addition, as before, the tail for the contrastive vowels is longer than the tail for the non-contrastive vowels. The scale is different than before (the largest distance is 120 rather than 250 or even 400); however, the same pattern that differentiates the non-contrastive from the contrastive vowels exists. That is, even in very small data sets, the necessary signal for infants to learn the contrast appears in naturalistic speech corpora. Taken together with the other analyses, this is promising evidence for this account that infants learn the contrasts of their language by tracking the acoustic distributions *across contexts*, and searching for substantially large changes in shape.

5.3.4 Discussion

In this section we showed that in Dutch, just as in French vs. Japanese, the distribution shape varies more across contexts in the subset of vowels that do contrast in length than in the subset of vowels that do not. These results provide further evidence that this is a relative stable property of contrastive dimensions that infants may be able to exploit.

The Dutch analyses are an important complement to the Japanese/French analyses, because they move us toward understanding what signal might be present for an infant to learn one particular language's contrasts. These results show that, even within one language, there is signal that infants could exploit: there are differences between vowel qualities that do contrast in length and those that do not. However, it is very important to emphasize that our results are not intended to illustrate the step-by-step process that infants follow to learn the phonetic contrasts of their language in an unsupervised fashion. Rather they are meant to show that there are real, reliable, and robust differences between cases where there is a phonetic contrast and cases where there is no phonetic contrast that are detectable in naturalistic speech. For example, we demonstrate that the subset of Dutch vowels with a length contrast has more extreme distribution shape changes across contexts than the subset of Dutch vowels without a length contrast. This is important because it is one of the first times that these two subsets have been differentiated on naturalistic speech. However, infants would not necessarily know how to group these vowels before they learn about the contrast. Future work will need to be done

to turn this into a full-fledged unsupervised learning account, and we speculate on future directions of this sort in Chapter 6. Nonetheless, the fact that we can even show a difference between the contrastive and non-contrastive subsets on naturalistic speech and that this difference replicates regardless of what language, corpus, speaking style, type of context, or context-pair inclusion criterion we use, is already a step forward.

5.4 Summary and Discussion

In this chapter, I presented a new idea for how infants could learn that a particular acoustic dimension of their language is contrastive, and provided initial support that the signal necessary for this learning story is present in naturalistic speech. I showed that it was present between languages, and could explain why infants of one language (Japanese) learn there is a contrast, while infants of another language (French) do not. In addition, I showed that the subset of Dutch vowels that do have a length contrast differ from the subset that does not, within one language. Overall, these results constitute one of the first accounts that can explain language-specific phonetic learning. Unlike many past learning accounts, this account has support from naturalistic speech.

At the core of the account is the idea that changing relative proportions of sound categories across contexts will lead to different distribution shapes across contexts. Our results suggest that this is, to some extent, a property of contrastive dimensions that infants could pick up on. This result did not need to happen. First,

it could have been that systematic acoustic variability caused large differences in distribution shape across contexts in French, just as in Japanese. Second, these driving contexts (that consist primarily of a non-dominant category) did not need to be so prevalent in the speech data. In particular, if all of the contexts that deviated from 90% short vowels and 10% long vowels were infrequent, this result would not have arisen. Instead, this result arose and it arose in *any way* we tried running the analysis - changing the language, corpus, speech style, context type, context inclusion criteria, etc. all led to the same results.

These analyses raise many open questions. I will discuss some of the more practical issues here, and will leave more speculative discussion about e.g. how this relates to further phonetic learning, what work needs to be done to turn this into an unsupervised learning account, and whether listeners can actually do this for the General Discussion in Chapter 6.

5.4.1 Do these results generalize to infant-directed speech?

With the exception of the Dutch infant-directed speech corpus, the analyses used adult-directed speech corpora because they were the only large, naturalistic corpora that had the necessary phonetic transcriptions and alignments. Future work should certainly test these predictions on IDS when the necessary corpora become available; however, in the meantime, I speculate that these results will hold true on IDS just like on ADS. First, comparing Dutch ADS and Dutch IDS provides some initial support that these results will hold. In order for this pattern to hold in IDS

more generally, we must (i) still observe a long tail for the Japanese IDS corpus, and (ii) not observe a long tail for the French (or other non-contrastive) IDS corpus.

The fact that we observe a long tail for the Japanese corpus stems from the fact that there are contexts or word frames that primarily have long vowels in them. Therefore, we would expect a long tail in the Japanese IDS data as well, as long as these contexts exist in the IDS. There has been a lot of discussion of differences between ADS and IDS for phonetic learning (e.g. Cristia and Seidl, 2014; De Boer and Kuhl, 2003; Eaves Jr et al., 2016; Kirchhoff and Schimmel, 2005; Kuhl et al., 1997; McMurray et al., 2013), but these tend to discuss differences in the acoustics (e.g. expanded vowel spaces) and focus less on differences in what is said. The fact that the top-down information account models performed so well in Chapter 3 suggests that the necessary contexts do exist in IDS, just as in ADS. As has been explained before, the existence of these contexts does not imply that the tail will still be present, but it seems fairly reasonable that it would still exist based on the logic of this account.

What seems less clear is whether there would still be no tail in French. If, for example, French mothers elongate certain vowels more in some contexts than others in IDS, but not ADS, then this could lead to an analogous tail in French. One example is if French mothers elongate vowels that make up an entire utterance much more than they elongate other vowels (if they elongate all vowels equally then all of the distributions will be shifted and this should not make a big difference in shape). This seems relatively likely to be the case; however, even if they do this, if Japanese mothers do the same thing, then we would still expect to see a longer

tail for Japanese than French, even if both tails were longer for IDS than ADS. Combined with the Dutch ADS vs. IDS results, it seems likely that the pattern would hold in IDS.

5.4.2 Implications beyond language acquisition

We have discussed these results in the light of acquisition, but they have implications for language change and language universals as well. At a high level, these results highlight a critical connection between phonetic learning and phonotactic constraints (or other systematic regularities that dictate which contexts particular sound categories tend to occur in). As a result, this account makes predictions about which contrasts will or will not persist over time. In particular, it predicts that phonemic contrasts will disappear if the contrasting sounds start occurring in identical contexts. Put slightly differently, it predicts that languages without phonotactic constraints (or in which sound category makeup is identical across contexts) cannot exist, unless there are other cues present that infants could use to learn phonetic contrasts. These predictions are testable both by looking at historical language data, as well as by in-lab experiments testing iterative generational learning, and future work should investigate them.

5.4.3 Corpus sizes

It is important to note that the full datasets we used for the analyses were always larger for the contrastive cases than for the non-contrastive cases. The

CSJ (spontaneous Japanese corpus) is larger than the NCCFr (spontaneous French corpus), and because there are more vowel qualities in Dutch that do contrast in length than that do not contrast, the Dutch contrastive dataset was always larger than the non-contrastive dataset. This is potentially problematic - the long-tail effect is driven by a small number of contexts, and so, if these contexts did not occur in the dataset, then we would not observe the pattern. Having a smaller dataset size for the non-contrastive cases, increases the possibility that these contexts exist in speech even for French, but were simply not represented in the dataset. We were aware of this problem, and corrected for it. We both (i) artificially balanced the sizes of the data sets (by randomly sampling vowel tokens from the larger corpus), and (ii) considered more contexts in the non-contrastive case than in the contrastive case. We observed the effect regardless. To some extent, this gets rid of the problem by making it a more fair comparison; however, there is still a possibility that the relevant contexts were simply not present in the non-contrastive corpora when they *are* available to infants who get access to much more data. Of course, having more data for Japanese would also probably increase the number of driving contexts there. As a result, it seems unlikely that this would explain the whole effect; however, this is an issue that is worth keeping in mind.

Beyond just the different ways of dealing with imbalanced corpus sizes, there are additional decisions to make when choosing which contexts to include in the analyses. For example, we could include contexts that have more than N vowels in them (where N could be any arbitrary number), or we could simply include the top M contexts. Again, the results hold true regardless. In fact, any procedure will yield

the long tail in Japanese as long as at least some of the driving contexts described in Tables 5.2 and 5.3 are present in the data used, which seems likely based on their counts and frequency ranks. This relatively consistency is promising for this account, as it suggests that this effect is relatively robust.

That is, no matter how we run this analysis - no matter the language, the corpus, the speech style, the context type, the context inclusion criteria - the results come out the same. There is more distribution shape change across contexts when there is a contrast than when there is not on naturalistic speech. This is strong support for a new learning account of how infants learn what dimensions of their language contrast. We speculate further about this learning account in Chapter 6.

Chapter 6: General Discussion

6.1 Summary of this Dissertation

This dissertation asked how contextual information (like who spoke a sound, what the neighboring sounds where, and where in an utterance a sound occurred) can help in separating overlapping categories, focusing both on the problem of adult speech perception and on infant phonetic learning. In Chapter 3, we tested the efficacy of two ways of using context in helping to disambiguate overlapping categories. We tested top-down information accounts, where listeners make use of context to bias their expectations of what category they will observe, as well as normalization accounts, where listeners use context to help factor out systematic variability. Although well-studied, these ideas have been somewhat conflated in past work and have rarely and with limited success been applied to naturalistic spontaneous speech. In this dissertation, we further explored these two ideas, trying to overcome these issues with past work. We disentangled these two ideas and carefully studied the relative contribution of each of them to the listener's task, applying them to spontaneous speech. Our simulations showed that a top-down information strategy is effective even on spontaneous speech, but that normalization is not. This result was surprising given that normalization has been found to be effective in the past.

In Chapter 4, we resolved this discrepancy, by showing that normalization was helpful when we ran our same analyses on simplified controlled lab data - of the type generally studied in the normalization literature - rather than on naturalistic spontaneous speech. We then provided simulations and a mathematical analysis showing that normalization may be ineffective when there are context-specific category imbalances - precisely of the type that are useful for top-down information accounts. This suggests that a learner whose input contains imbalances through phenomena such as phonotactic constraints and phonological alternations, is better off using context to bias their perception in a top-down fashion rather than normalizing it out, at least as we have implemented these strategies here.

This led us, in Chapter 5, to propose a new account for how infants might learn that a particular dimension contrasts in their language that does just that. This account was inspired both by distributional learning accounts (Maye et al., 2002), and by findings that infants are sensitive to higher-level contextual information (e.g. Feldman et al., 2013b; Thiessen, 2007). We suggested that infants could track the acoustic distributions of speech sounds across contexts, and learn that a particular dimension is contrastive if the distribution shape varies substantially across contexts along that dimension. We showed that, unlike past accounts, this account makes predictions that hold true on naturalistic speech, and that it can qualitatively explain why Japanese, French, and Dutch infants learn what they do. Importantly, it does so on highly variable, naturalistic speech.

In what follows, we end with a discussion of how we might expect these results to generalize (or not) to other problems in speech perception and learning, where

this leaves normalization in the literature, what the implications of and next steps for testing the proposed learning account are, as well as the importance of testing ideas on spontaneous data in addition to controlled lab data.

6.2 How generalizable are these results?

The results presented in this dissertation are directly applicable to the Japanese vowel length contrast, but this is just one test case where context has been argued to play a role, due to overlap in perceptual cues. To what extent do our results generalize to other overlapping category problems in phonetic learning and speech perception? We discuss this, in turn, for: (i) top-down information being helpful in separating short and long vowels in Chapter 3, (ii) normalization being unhelpful for separating short and long vowels in Chapter 3, as well as (iii) there being signal that supports our distributional learning across contexts account.

6.2.1 Top-down information is helpful for separating short and long vowels

It is relatively likely that our results showing the efficacy of top-down information accounts would generalize to other tasks in speech perception and phonetic learning. In both of these areas, there is already ample evidence that top-down information is useful (though mostly from controlled lab data). In addition, there is evidence that there are systematic regularities in which contexts sounds of all types occur (reviewed in the top-down information background section). Therefore,

although the generality of this result would need to be confirmed on more realistic data, we speculate that these results would generalize to other cases.

6.2.2 Normalization is ineffective for separating short and long vowels

The picture is a bit more complicated for our results on normalization. Our analysis reveals that normalization is ineffective when it is difficult to estimate the normalization function. As we have shown here, it will be difficult to estimate normalization functions for contextual factors that would be helpful for top-down information accounts (i.e. when there are regularities in which categories occur in which contexts). In addition, as will be discussed more extensively in Section 6.3, it will also be difficult to estimate one good normalization function for contextual factors that affect the productions of different categories differently (e.g. a contextual factor that acoustically lengthens long vowels more than they acoustically lengthen short vowels). To the extent that people are dealing with contextual factors that do not fall into one of these classes, normalization could very well help for the tasks of speech perception and phonetic learning. In particular, it is possible that in the case of Japanese vowel length, there is sufficient signal via top-down information to distinguish most short/long minimal pairs without attending to the acoustic duration at all, so that in conversational speech, the durational contrast is mostly neutralized. It may be the case that normalization is ineffective for contrasts with low functional load (like Japanese vowel length), but more effective for contrasts with high functional load, where speakers must produce a perceptible contrast in order to

be understood. We, nonetheless, speculate that the ineffectiveness of normalization will generalize to many other contrasts, as naturalistic speech is full of top-down information, which helps predict which sound will be uttered, even without hearing the acoustics of the sound, but hurts normalization. However, further work will need to be done to study the extent to which these findings generalize to other contrasts within the domains of speech perception and phonetic learning.

6.2.3 Distribution shape changes can signal contrastive dimensions

The analyses we performed in support of our learning account in Chapter 5 were already more general than those in Chapter 3 and 4, because they considered multiple languages. However, they still nonetheless all focused on vowel length. This was done, in part, because this was a natural extension from the work in Chapter 3. Beyond that, vowel length was a natural choice because there is one main, clear and agreed-upon cue to the length contrast (i.e. duration) that is readily available in any corpus that is annotated at the segmental level (this is not true for other cues like formants, which need to be measured after transcriptions are made). However, future work will test whether this result generalizes, considering both a broader range of languages and corpora, and a broader range of contrast types.

Will this result generalize to other corpora and languages when we still consider vowel length? This result holds for corpora in French, Japanese, and Dutch, and we predict that it should generalize to other corpora types and languages as well. As discussed, one of the most important tests will be whether it generalizes to

infant-directed speech. Beyond that, considering the test cases of English (which is similar to Dutch, in that it has secondary length distinctions), as well as Spanish will be informative, and phonetically annotated adult-directed corpora are available for these languages (e.g. Buckeye Corpus (Pitt et al., 2005) and the Nijmegen Corpus of Casual Spanish (Torreira and Ernestus, 2012), which parallels NCCFr). One of the issues in comparing French and Japanese is that the vowel inventories of these two languages are quite different: Japanese has 5 vowel qualities, while French has 10-15. Because vowel quality was part of how we defined context in some of our analyses, this created (probably harmless) differences between French and Japanese, in that there were more context types in French. Turning to Spanish instead of French would allow us to avoid this problem, because Spanish has the same exact five vowel qualities as Japanese, allowing for fairer comparisons. In addition, turning to other languages would also allow us to test whether the contexts that drive the long tail still primarily arise from word types, or whether phonotactics and alternations can also create this signal in the data.

Will this result generalize to other contrasts? In principle, there is no reason why it should not. The logic remains the same, even when we turn to other contrasts: where there are multiple categories along a dimension, there should be changes in relative category proportions across contexts because of the language's frequent word forms, phonotactics, and other properties. Because different categories have different distributions and locations in acoustic space, this will change the distribution in ways that might be visible and detectable from the speech signal.

Future work will investigate whether this is the case, starting with contrasts

that, like length, are signaled by one primary acoustic dimension. For example, it would be interesting to compare a language that contrasts [s]-[ʃ] against one that does not, along centroid frequency (the primary cue to this contrast). From the corpora that will become available to us, English vs. Spanish, in which [ʃ] only appears in loanwords, might be a good case to consider. Another contrast that would be worth considering is [r] vs. [l] along the third formant, and Japanese vs. English could serve as a good test case for that. Other cases like nasalization would also be interesting to consider eventually, but there is less of an agreed upon primary cue that listeners are thought to use. Nonetheless, comparing French against English vowels would be an interesting test case for nasalization.

There are two types of contrasts that would be especially interesting to consider. First, it would be interesting to look at a contrast where the categories are much more balanced in their overall frequency. Vowel length is unique in that one of the categories occurs so much less frequently than the other. When we looked at vowel length, the driving contexts sometimes consisted of over 75-80% long vowels, even though long vowels constitute only 10% of all vowel tokens. When the categories are more balanced to begin with, contexts that overrepresent particular categories will be more similar in make-up to the most common contexts, so it is worth seeing whether this result continues to hold there.

Second, it will be interesting to compare languages where a particular contrast is meaningfully contrastive in one, but allophonic (and common) in the other. Comparing nasalization in French vs. English would be a good test case for this because French has contrastive nasal and oral vowels, but these are allophonic in English:

vowels are predictably nasalized before nasal consonants. For English, this means that the relative proportion of nasal and non-nasal vowels will vary substantially across contexts: vowels before nasals will be primarily nasal, and all other vowels will be primarily oral. This is precisely the type of signal that this account picks up on, so it will be interesting to see how the signal differs between contrastive vs. allophonic cases.

One possible outcome is that the signal will be different for the contrastive vs. allophonic vs. the non-allophonic cases that we have considered. For the allophonic case, the relative proportion of different categories will be much more consistent across contexts: almost all of the vowels before nasals will belong to one category (nasals) and almost all of the vowels before other sounds will be non-nasals, and this could potentially result in different signal.

The more likely outcome, however, is that the allophonic cases will be indistinguishable from the contrastive cases. This is not particularly problematic because even English infants need to learn about nasal vs. oral vowels in order to represent rules about them and to produce the correct segment in the right environment. If our learning account cannot differentiate phonetic contrasts from allophonic segments, then a full learning account would need to invoke proposals about how infants learn allophonic rules (e.g. Peperkamp et al., 2006) on top of our current learning account. The accounts that have been proposed for how infants learn about allophones are actually quite similar in nature to the account proposed here. In particular, Peperkamp et al. (2006) proposed that infants might learn that two sounds are allophones by measuring the discrepancy in context probabilities for each pair of segments, based

on the observation that these segments will be in complementary distribution (i.e. will occur in mutually exclusive contexts).

It is worth noting that this would result in a two-stage process, where infants first learn the phonetic categories of their language, before learning the phonemes of their language. This sort of two-stage process (as proposed in Peperkamp et al. (2006)) has been criticized in the past by Dillon et al. (2013). In particular, Dillon et al. (2013) have argued that the initial phonetic categories cannot be accurately learned without some model of phonology, and if the outcome of the first stage of the process (i.e. phonetic learning) is wrong, the second stage (i.e. phonemic learning) will fail. This is an important argument to keep in mind as we further develop this learning story. One possible counterargument is that I, too, agree that the output of past phonetic learning models has been insufficient, but it is possible that the input to the second stage may be better using our phonetic learning account than previous phonetic learning accounts. In fact, the goal of this work is to improve upon those models, and this might make Peperkamp et al. (2006)'s proposal more effective. Another possibility to pursue is trying to build in more phonology into our proposal. For example, perhaps there is a way to not only measure how much distribution shape changes across contexts, but to also combine this with a mechanism to understand why the extreme distribution shape changes occur. It is possible that there is a way to distinguish cases where the extreme distribution shape changes happen because of a predictable rule (indicating an allophonic contrast) or not (indicating a phonemic contrast), but we leave this for future work.

6.3 The Status of Normalization

The idea that normalization plays a role in processing and acquisition is a widely held idea, but our results bring up important issues with it that complement other problems discussed in earlier work Johnson (1997, 2006); Pierrehumbert (2002). In order to normalize well, it is important to be able to estimate the correct normalization function. Failing to do so can actually increase the amount of variability and overlap between categories, rather than reduce it, as we saw in the toy example in Figure 2.2. The results in this dissertation show that one obstacle to estimating the normalization function well is the fact that different sound categories occur in different contexts with different probabilities. For example, short and long vowels differ in what consonants they are likely to precede, which makes factoring out systematic acoustic variability from the following consonant difficult. This is a problem for any contextual factor that both affects which target sound is likely to occur a priori, and systematically affects a target sound's acoustics (i.e. any contextual factor that would be a good top-down predictor of vowel category). Therefore, this is not a problem for sources of systematic variability like speech rate or gender: short (or long) vowels are no more likely to occur in fast speech than slow speech, or in speech by men than women (and vice versa), so speech rate and speaker gender are unlikely to be informative about whether a particular vowel is short or long.

That being said, there are other problems with normalization that would also affect factors like speech rate and gender. Context can affect how some categories are produced more than others. For example, long vowels might be acoustically

lengthened more than short vowels in slow speech. In its current form, normalization cannot handle these types of sources of systematic variability because it uses one normalization function across all categories. If different categories are actually differentially affected, then the learned normalization function is guaranteed to be wrong for some tokens, and this may increase variability and overlap rather than reduce it.

Variation due to speech rate, gender and other speaker characteristics is also a challenge for top-down prediction, since these effects change the realization of an acoustic token but do not alter the prior distribution on what tokens will be present (women pronounce sibilants differently than men, but they are not *a priori* more likely to produce /s/ vs /ʃ/) Strand and Johnson (1996). These cases indicate that top-down prediction cannot operate alone, but must be supplemented with some other mechanism; in principle, this could be either normalization or adaptation. However, the adaptation account is supported by experimental evidence regarding other kinds of variability between speakers, such as dialect. Even young infants are capable of learning to adapt to dialect variation given sufficient evidence (e.g. van Heugten and Johnson, 2014). Normalization is not a good account of this learning process, since it would require complex phonological alternations such as vowel shifts to be normalized directly in the acoustic space rather than learned phonologically (Elsner et al., 2013). Since an adaptation mechanism would also be able to cope with variability due to speech rate or gender, studies of these factors do not provide independent evidence that normalization must also be present.

The question that remains is whether normalization could be altered to over-

come these issues and be effective for categorization and acquisition. Here, we discuss possible changes; future work should investigate whether these changes make normalization effective.

With regards to adult speech perception, one possibility would be for listeners to learn different normalization functions for each category type (i.e. one normalization function for short vowels and another for long vowels). This would also allow the process to take into account category imbalances. However, if the listener is equipped with one normalization function for short vowels and one normalization function for long vowels, they will not know which function to use until they have already categorized the sound, so normalization would not, in this case, be helpful during the categorization process (only after). Another possibility would be that listeners build separate normalization functions for separate categories, but average them during categorization, weighting them by the relative proportion of each category type. A third, and perhaps most likely, possibility is that instead of comparing the overall context means to determine if there is systematic variability between contexts (as the current normalization implementation does), listeners could compare the short vowel mean in one context to the short vowel mean in another context, as well as the long vowel mean in one context to the long vowel mean in another context. Once they have done so, they could normalize only when they notice differences between the means *within each category type*. We saw in the toy example in Figure 2.2 that category imbalances can artificially increase (or decrease) the mean, making it seem like the sounds in a context are systematically acoustically lengthened (or shortened). However, this problem would be avoided if the short and long

vowels were treated differently. Of course, this idea would only be possible for adults who have learned to categorize and know the distinction between short vowels and long vowels. These three ideas have not yet been tested, and it is currently unclear that they would increase the efficacy of normalization on spontaneous speech, but future work should investigate them.

The issues with normalization become even more problematic when considering acquisition. The learner does not yet know the distinction between short and long vowels, and cannot take into account category imbalances. As a result, they will necessarily be applying inaccurate normalization functions, which may actually increase category overlap rather than reduce it. In addition, throughout the dissertation, we saw that normalization performance depended on the precise set of factors being normalized out. Therefore, a learner would have determine which factors to factor out - and would need to learn that some factors that systematically affect acoustic productions should be factored out, but other factors that similarly affect acoustic productions should not be factored out. These issues complicate the view that normalization is helpful in language acquisition.

Overall, although normalization has received a lot of support in the literature, there is actually little to no evidence suggesting that this is a strategy that could be helpful for acquisition and processing naturalistic speech. A lot of the evidence that has been used to argue for normalization is also consistent with a top-down information strategy, which, unlike normalization, was shown to be effective here, as well as adaptation accounts. In addition, normalization has mostly been tested on controlled lab speech, rather than the speech that listeners primarily hear and

learn from. We showed here that these results from lab speech do not necessarily generalize to naturalistic speech (and did not in the case of Japanese vowel length). This work calls into question the role that normalization could play in acquisition and processing. Future work should work to see if amending the normalization process helps, but in its current form, there is more evidence against normalization than for it. In order to stand by the idea that normalization helps disambiguate overlapping categories, it is critical to find some evidence that normalization - in any form - is actually effective in separating categories when applied to spontaneously produced speech.

That being said, the fact that we show that normalization may not lead to better separation between short and long vowels does not imply that listeners do not normalize. If it is the case that listeners process their input by normalizing acoustics relative to context, then our results indicate that listeners are overcoming even more overlap between short and long vowels than represented in Figure 2.1. Our results show that normalization is unlikely to be the solution to the overlapping categories problem.

Finally, normalizing via a learned normalization function is only one way to implement the idea of factoring out systematic variability, and other alternatives may be more effective for processing and acquisition.

One alternative would be that instead of normalizing by learning an explicit normalization function (implemented here as a linear regression or neural network), listeners might rely on a alternative set of acoustic cues that are more invariant than those that are typically measured and described. For example, researchers

have argued that instead of absolute duration, listeners might rely on the ratio of a vowel's duration to a neighboring vowel's duration, or the ratio of a vowel's duration to its word's duration as an acoustic cue to vowel length in Japanese (Hirata, 2004). Other cues like the ratio of first and second formants to third formant values, as well as ratios between stop closure and previous vowel duration have been argued to be helpful for perception of vowel qualities and stops, respectively (e.g. Monahan and Idsardi, 2010; Port and Dalby, 1982). The idea is that categories may not overlap as substantially along these other acoustic cues as they do along those that are typically discussed (like individual formants or duration). These accounts are promising: MEG experiments have shown that the auditory cortex is sensitive to ratios of these sorts (Monahan and Idsardi, 2010), and analyses have shown that these can be clear cues to category membership (Hirata, 2004; Monahan and Idsardi, 2010), though these analyses have been of controlled lab speech rather than naturalistic speech. Preliminary work of ours found that these alternative acoustic cues did not help for the Japanese vowel length contrast, but it is possible that other untested acoustic cues would, and future work should be done to systematically study this class of ideas as applied to naturalistic speech.

A second alternative to normalization of the type discussed in this dissertation would be an adaptation strategy, which reduces systematic variability without having to calculate an explicit normalization function (Kleinschmidt and Jaeger, 2015). It does so by essentially learning a separate mapping between acoustics and linguistic category for each context observed. This avoids the need to learn a precise normalization function, but can still overcome systematic variability by treating

each context separately. Currently, we run into data sparsity issues, as this requires splitting already small datasets by context (even more so than in the analyses in Chapter 5); however, it is promising to continue pursuing, as it does not encounter any of the issues that normalization does, and can explain the experimental findings that have been used to argue for normalization.

6.4 The Status of Distributional Learning Across Contexts

In Chapter 5, we presented a new account of learning which dimensions are contrastive that seems promising, but more work will need to be done to show that this is actually a viable learning account.

First and foremost, as discussed previously, it will be important to make sure that these results generalize more broadly. From there, it will be important to work out a particular learning mechanism that could make use of this signal, as well as think about how this early phonetic learning feeds into later learning. Finally, it will be important to verify that humans can actually perform these types of computations. We discuss various portions of this, in turn, in the sections that follow.

6.4.1 Relationship with Other Phonetic Learning

Beyond learning which dimensions are contrastive, infants must learn how many categories that dimension contrasts. For example, while Japanese has two vowel lengths (short and long), some languages, like Estonian, have three vowel

lengths, and infants need to identify what kind of language they are in, once they have identified that a particular dimension is contrastive. In addition, infants learn the particular location of the categories in acoustic space (i.e. what has typically been operationalized as the mean and variance of a category). Finally, they need to be able to categorize particular tokens into those categories. Assuming that the presented learning account is true, how could the remainder of this learning happen? In the following sections, I will speculate on how the various components of learning could take place given that an infant has learned what dimensions contrast in their language.

6.4.1.1 Learning how many categories exist along a contrastive dimension

It is common for a particular dimension to contrast more than two categories. Although languages with three vowel lengths are relatively rare, common dimensions like voice onset time can contrast multiple categories. Therefore, once an infant has learned that a particular dimension is contrastive, they may need to learn that there are more than two categories along it. One possibility is that this is learned simultaneously with learning the category parameters (i.e. something akin to means and variances), which will be discussed in the following section. However, it could also be learned separately, and I present one idea about how this learning could happen here.

In particular, in the analyses comparing French and Japanese, we saw that the

contexts driving the signal in the input consisted of contexts in which long vowels were overrepresented relative to their base rate of occurrence. These contexts were different in shape from the vast majority of the other contexts. It is possible that this might be extended in a three-way contrast (that consists, hypothetically, of Category A, Category B, and Category C). In particular, in a three-way contrast, perhaps there are some contexts consisting primarily of Category A tokens, some contexts consisting primarily of Category B tokens, and other contexts consisting of Category C tokens. In this case, on the surface, we might observe a long tail that resembles what we observed in the Japanese case. However, unlike in the two-way contrast case, the tail would include contexts with distributions that are different in shape from the vast majority of contexts, but also different from each other. That is, while in the two-way length contrast, there seem to be two main types of distribution shape differentiated in the tail, in a three-way contrast case, there might be three main types of distribution shape differentiated in the tail (those that primarily have Category A, those that consist primarily of Category B, and those that consist primarily of Category C). Infants might consider not just whether there are big changes in distribution shape, but also which contexts are driving them, and if they are, in turn, similar to each other. If they notice that the driving contexts themselves are different in shape from one another, they could possibly learn that there are as many categories as there are "different category shapes."

6.4.1.2 Learning the category parameters

Once infants have learned that a particular dimension is contrastive, they must also learn where the categories are in acoustic space. One possibility is that infants learn how many categories exist along a dimension (as described in the previous section), and then learn where those categories are. Another possibility is that infants learn the category parameters and number simultaneously.

Possibility 1: Infants learn how many categories a dimension contrasts and then learn the category parameters

The first possibility is that, once they have learned how many vowels exist, they then use that information to learn where the categories are. Within this possibility, one option is that, once infants know how many categories exist, finding the actual categories is trivial - simply clustering the categories into two, three, or more groups based on the overall acoustics could arrive at the correct solution (or nearly correct solution that can be corrected over time).

Another possibility is as follows: if it turns out that there are, in fact, two main types of distribution shape in the tail in a two-way contrast (one that consists primarily of Category A, and another that consists primarily of Category B), but three main types of distribution shape in the tail in a three-way contrast (one that consists primarily of Category A, another that consists primarily of Category B, and still others that consist primarily of Category C), then it is possible that the location and shape of the distribution in each of these context types could clue a

learner in to the category parameters. That is, the distribution in contexts consisting primarily of Category A will likely have parameters (e.g. mean and variance) that are fairly similar to those of Category A, whereas distributions in contexts consisting primarily of Category B will have parameters that more closely resemble those of Category B itself. Therefore, if an infant notices that there are multiple types of distribution shape, then they could use the parameters of these contexts as initial category parameters. Of course, there would be some noise in these estimates; however, because there is so much variability to begin with in how each category is produced, it might not hurt the learner too much.

Possibility 2: Infants simultaneously learn the number of categories and category parameters

Another possibility is that they learn the number of categories at the same time as the parameters. Here, I will outline one possible way that this could happen, inspired entirely by Balaji et al. (2019). This will also move us in the direction of a full-fledged unsupervised learning model that explicitly takes into account distribution shape changes across contexts.

The key idea is as follows: having introduced the idea of comparing distribution shapes across contexts, a new way of conceptualizing the category learning problem is trying to identify the set of categories that best explains the observed variability and change in distribution shapes across contexts. For example, one possible category set is one single category with e.g. mean 100 and variance 20; another possible category set is two categories - one with mean 100 and variance 20, and

another with mean 200 and variance 40. With this in mind, if successful, a French infant would land on a one-category solution that has parameters corresponding to the short category in French. A Japanese infant, on the other hand, would arrive at a two-category solution that has the parameters corresponding to short and long vowels in Japanese. Two categories would allow them to explain why there are such large distribution shape changes.

One way this can be thought about comes directly from a paper that tries to define a variant of Earthmover's distance that is not as sensitive to category imbalances across contexts (Balaji et al., 2019). The paper points out that Earthmover's distance is affected by what the relative proportion of categories is across contexts: even if two distributions contain identical categories, if one distribution is 50%/50% short and long vowels, and the other has 90%/10%, then the Earthmover's distance between them will be large. This is *precisely* the signal that our learning account picks up on; however, for some applications in computer science, this property is problematic and Balaji et al. (2019) propose a new variant of Earthmover's that corrects for this.

At this point, it would be helpful to introduce their formal definition of what they term "normalized Wasserstein distance" and what we will refer to as "normalized Earthmover's distance." We will then explain how this could relate to the learning.

Let $\mathbf{G}$ be a set of generator functions with k components defined as $\mathbf{G} := [\mathbf{G}_1, ..., \mathbf{G}_k]$. We can think of this as the underlying category set we have been discussing, where each component corresponds to one category. One possible G has

one (short vowel) component (e.g., G = [$\mathcal{N}$(100, 20)]), and another possible G has two components (e.g., G = [$\mathcal{N}$(100, 20), $\mathcal{N}$(200, 40)], corresponding to a short vowel category and a long vowel category).

Next, let $\mathbb{P}_{\mathbf{G},\pi}$ be a mixture probability distribution for a random variable X, where X = $\mathbf{G}_i(Z)$ with probability π_i for $1 \leq i \leq k$. Z is assumed to have a normal distribution. This is essentially defining a probability distribution over surface acoustic cues, X, in a particular context. In this context, π dictates what the relative probability of each of the categories is. In one context, we might generate the surface acoustic duration from the short vowel category distribution 50% of the time, and generate a sound from the long vowel category distribution 50% of the time. In this case, π = [0.5, 0.5] and G = [$\mathcal{N}$(100, 20), $\mathcal{N}$(200, 40)]. In another context, we might generate a sound from the short vowel category distribution 90% of the time, and generate a sound from the long vowel category distribution 10% of the time. In this case, π = [0.9, 0.1], and G = [$\mathcal{N}$(100, 20), $\mathcal{N}$(200, 40)]. In a third case, we might only have one vowel category, so we will generate our acoustic value, X_i from that category. In this case, π = [1.0] and G = [$\mathcal{N}$(100, 20)]. That is, $\mathbb{P}_{\mathbf{G},\pi}$ is simply a way of defining a probability distribution over what acoustic cues, X, we will see based on the categories that exist in our language, and their relative probability in that context.

Then, normalized Earthmover's distance, $\tilde{W}$, is defined as follows.

$$\tilde{W}\left(\mathbb{P}_X, \mathbb{P}_Y\right) := \min_{\mathbf{G},\pi^{(1)},\pi^{(2)}} W\left(\mathbb{P}_X, \mathbb{P}_{\mathbf{G},\pi^{(1)}}\right) + W\left(\mathbb{P}_Y, \mathbb{P}_{\mathbf{G},\pi^{(2)}}\right) \tag{6.1}$$

where Earthmover's distance, W, is defined as in Equation (5.1).

That is, in comparing two distributions, we consider a particular category set, $\mathbf{G}$, a particular relative proportion of each of those categories in the first distribution, $\pi^{(1)}$, and a particular relative proportion of each of those categories in the second distribution, $\pi^{(2)}$. For each combination of those, we calculate the Earthmover's distance between the first distribution and $\mathbb{P}_{\mathbf{G},\pi^{(1)}}$, which we sum with the Earthmover's distance between the second distribution and $\mathbb{P}_{\mathbf{G},\pi^{(2)}}$. The final normalized Earthmover's distance is the minimum of those sums, over all $\mathbf{G}$, $\pi^{(1)}$, $\pi^{(2)}$.

This will allow us to treat a Japanese distribution from a context that is 90% long and 10% short, and a distribution from a context that is 10% short and 90% long as similar in shape, because instead of comparing these distributions against one another (as we are currently doing), we will compare them against intermediate distributions that both have the same underlying category set, but explicitly represent that the tokens in the first distribution mostly come from the long category, and the tokens in the second distribution mostly come from the short category.

This could potentially relate to learning sound categories, because we can think of learning as choosing the $\mathbf{G}$ that best explains the change in distribution shape across contexts, where the $\mathbf{G}$ that best does this is the one that minimizes distance. One issue that arises is that based on the way it is currently defined, if we are comparing G's with one component to G's with two components, then the G with two components will always necessarily reduce the distance. Therefore, we would need to correct for this (e.g. with a prior for fewer categories) in a full version; however, this could be one way to turn the signal that we have observed

into a learning account. Japanese infants learn that there are two categories because that solution best explains the distribution shape changes observed across contexts, whereas French infants learn that there is one category because that solution explains the lack of distribution shape changes well, without need to add in a second category.

In practice, this is unlikely to work on naturalistic speech. There is systematic acoustic variability that could affect the results, so it would be necessary to build in something that could allow for some shifts in context arising for other reasons than the underlying category structure. Nonetheless, it is interesting to think about as we move towards building unsupervised models, as it provides a new framework and way of thinking about learning, where learning consists in explicitly trying to explain these differences in shape.

6.4.1.3 Learning to categorize individual sounds

The last portion of the learning problem consists of learning to categorize individual tokens, given knowledge about where the categories are. We saw in Chapter 3 that we can categorize particular vowels fairly accurately using contextual information in a top-down fashion, once categories are known, and the categorization would presumably be even better once specific word identity information was included. Therefore, once children know the location of the sound categories in acoustic space, they could use a combination of the acoustics and systematic differences in the contexts different sounds occur in to start predicting their category membership. This could be updated over time - after all, children are thought to continue

making categorization and production errors for a number of years after initially learning about the distinction. It is worth noting that children will be helped out once they know the category parameters because they will know that there is an extreme frequency imbalance between short and long vowels.

6.4.2 Can people do this?

Once this effect is further verified in other corpora, languages, and contrasts, it will be important to show that humans can actually perform these computations, and that they could use them to learn about phonetic categories. This account requires that listeners keep track of acoustic distributions across many contexts - some of the analyses use more than 150 contexts, though most use fewer and these results still hold if we reduce the number of contexts even further. In addition, this account requires that listeners make relatively fine-grained comparisons between those distributions. Although this is not computationally trivial, this account is similar to other accounts (e.g. normalization accounts and adaptation accounts) in terms of what it requires of the listener. For example, normalization and adaptation accounts both require that listeners compute distributions across many contexts, and normalization accounts additionally require that listeners be able to learn specific functions that map from one distribution to another. In addition, the fact that listeners can easily perform tasks like accent adaptation (e.g. Bradlow and Bent, 2008; Maye et al., 2008; Norris et al., 2003) and speaker identification suggests that they can compute and compare relatively complex acoustic distributions. Regardless, if

further analyses continue to suggest that this is a promising account, future work should experimentally test whether listeners learn different things about phonetic contrasts when they are presented with distributions that change in shape across contexts vs. distributions that are relatively stable across different contexts.

6.5 Adaptation as another top-down information account of acquisition

The results presented in this dissertation are promising for the role of top-down contextual information in acquisition. Results from Chapter 3 indicated that using contextual information in a top-down fashion results in relatively accurate categorization of short and long vowels. Furthermore, in this work, we only included a small set of contextual factors, and adding in additional factors could help separate short and long vowels even further. In particular, with the exception of part-of-speech, our work did not include any word type or word-level information, which has been argued to be helpful in the past (e.g. Feldman et al., 2013a). This suggests that using contextual information as top-down information or to guide expectations could be extremely helpful in adult speech perception.

In Chapter 5, we began laying out one possible learning account for how this contextual information could be used to learn phonetic contrasts; however, there are other ways that top-down information could be used in acquisition accounts, and, here, we focus on another promising account: adaptation.

An adaptation strategy builds a separate mapping from acoustics to categories

for each context encountered. In doing so, it has access to information about which categories are more/less likely to occur within a particular context. Therefore, it is possible that within particular contexts, the short and long vowel categories are more separated than they appear overall. For example, short vowels and long vowels might be well-separated when they occur in phrase-final position and preceded by a particular consonant. If the distribution is bimodal along the duration dimension in a particular context like this one, then the learner could learn that there are two categories along the duration dimension via a process of distributional learning (Maye et al., 2002), and then generalize this to other contexts where the distinction is less clear.

In particular, our results have shown that there seems to be something about carefully enunciated Japanese speech that reliably elicits different durations for short versus long vowels. While most of the input infants hear is highly-variable and spontaneous, infants do sometimes get some exposure to clearer instances of speech through e.g. hearing their parents read them books. Therefore, it is, in principle, possible that children learn which words have which vowels precisely by listening to social situations in which their parents speak carefully.

This type of adaptation strategy, in which children learn about the distinction in a particular context and generalize it, is a particularly promising strategy, as it provides both a way to take advantage of top-down expectations of category membership, as well as a way to remove systematic variability. By building a separate mapping from signal to category for each context, it has access to information about top-down information about which categories are more or less likely to oc-

cur in a particular context, and, therefore, can account for results that have been used to argue for top-down information accounts. At the same time, experimental results that have been used to argue for normalization functions can also be explained by adaptation strategies, as these results show that listeners account for the fact that sounds are produced differently in different contexts, but cannot dissociate whether listeners do so via an explicit normalization function, or by building a separate model for each context. As a result, adaptation accounts can explain the experimental findings that have been used to argue in favor of both top-down information accounts and normalization accounts in a unified way. However, again, infant-directed speech corpora are generally quite small, and it is difficult to test adaptation accounts without running into data sparsity issues.

In sum, although we have provided one possible learning account, there are other ways that infants could make use of contextual information, including but not limited to adaptation accounts, and future work should continue to investigate these various possibilities by implementing them computationally and testing how effective they are at learning the contrast between short and long vowels, as well as other contrasts.

6.6 Controlled Lab Speech vs. Naturalistic Speech

Finally, the results presented in this dissertation reiterate once again that there is a crucial distinction between controlled laboratory speech and spontaneously produced naturalistic speech. Essentially all of our understanding of speech perception

comes from work on carefully controlled and carefully enunciated laboratory speech, but almost all of our experience as listeners comes from messy, variable spontaneous speech. These two types of speech differ quite substantially from each other in nature, both in how the speech is produced, as well as the content of the speech. Indeed, where tested, many of the ideas developed on controlled lab speech have been shown to be ineffective on spontaneous speech. Previous work has shown that top-down information accounts developed and tested on carefully controlled or synthesized speech do not generalize to spontaneously produced lab speech Antetomaso et al. (2017). The current work shows that normalization is helpful on lab speech, but ineffective on spontaneously produced speech.

There is obviously a great deal of value that comes from working on speech where various factors are controlled for and isolated. In addition, listeners can process synthesized and controlled lab speech effortlessly, so our theories must be able to handle clear, enunciated speech, in addition to more naturalistic daily speech. However, what is critical is for ideas generated and tested on this lab speech to then be applied to spontaneous speech, to make sure that researchers are working on the same problem that listeners are solving.

There has certainly been some research starting to look at spontaneous speech, especially with the development of hand-annotated child-directed speech corpora such as from Mazuka et al. (2006). As we have discussed, Antetomaso et al. (2017) applied the model from Feldman et al. (2013a) to spontaneous Japanese speech, showing that the model's success did not generalize to spontaneous speech. Other work has also investigated spontaneous speech corpora both in the case of the over-

lapping categories problem Narayan et al. (2017); Swingley and Alarcon (2018) and more widely Guevara-Rukoz et al. (2018); Ludusan et al. (2016,0); Martin et al. (2016). However, it is still not prevalent, and our work aligns with previous work in revealing that studying spontaneous speech is critical for ensuring our ideas apply to naturalistic listening situations, and that we are at the point where we can make scientific progress by doing so.

6.7 Conclusion

In conclusion, this dissertation studied how contextual information could be useful for phonetic learning and speech perception. We first compared the relative efficacy of two ways of using context to help in phonetic learning. The first involved making use of contextual information as top-down information to guide expectations about what category was likely to be heard. The second involved factoring out systematic acoustic variability that resulted from the context a sound was produced in. These ideas have been conflated and almost entirely studied on controlled lab speech, not naturalistic speech. In this work, we showed that, for the case of the Japanese vowel length distinction, a top-down information strategy is effective even on spontaneous speech, but, contrary to previous findings, normalization is not. We then resolved this discrepancy in findings, by demonstrating that the same normalization procedure is helpful on lab speech - the focus of most previous studies - but ineffective on spontaneous speech - the focus of our study. We provided simulations and a mathematical analysis showing that normalization may be ineffective when

there are context-specific category imbalances - precisely of the type that are useful for top-down information accounts. Finally, we provided a potential solution for how infants learn about phonetic contrasts. This account explicitly takes advantage of the context-specific category imbalances that we show exist in the input. We provide evidence that this account can qualitatively predict what Japanese, French, and Dutch infants learn about vowel length, making it one of the first accounts to be able to do so. Taken together, these results force us to come to a more nuanced and precise view of the role that context might play in speech perception and phonetic learning.

More generally, they show that results from controlled lab speech do not necessarily generalize to naturalistic speech and that using realistic speech data matters. The field has tended to focus on whether listeners do or do not use particular strategies, making the key assumption that these strategies would be effective for learning and processing naturalistic speech. This was true for distributional learning in the early 2000s (which infants have been shown to do in the lab, but is now widely understood to be insufficient for naturalistic speech), and prior to this dissertation, it was also largely true for normalization and top-down information accounts. Simplifications are crucial for science; however, if the simplifications change the learning problem, then they can lead the field astray. This dissertation demonstrated the need to study naturalistic speech, so that we can make sure we, as researchers, are solving the same problems that listeners do.

Appendix A: Normalization and Top-Down Information Results on Japanese Adult-Directed Speech (R-JMICC)

This appendix reports results from applying the normalization and top-down information categorization models to R-JMICC adult-directed speech, rather than the R-JMICC infant-directed speech which was discussed in the main text. For more information about what each of the reported models is, see the data and methods section in Chapter 3. As discussed, the results are similar to those from the IDS reported in Chapter 3.

A.1 Top-Down Information Results

The results are summarized in Table A.1.

A.1.0.1 Baseline Model

Our baseline model simply used absolute duration and formant values to predict the vowel category of a sound. This model reached an overall accuracy of 90.5%. It correctly categorized 98.9% of short vowels, and 11.9% of long vowels. It had a BIC of 14820. The BIC for the ADS results is lower than for the IDS results, because the ADS portion of the data set is smaller, but overall the results are comparable

Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Baseline	90.5	98.9	11.9	14820
Top-down information (with simplified POS)	95.8	98.5	70.4	8331
Top-down information (with POS)	95.8	98.5	70.4	8328
Top-down information (with POS, no acoustics)	95.1	98.4	64.6	9105

Table A.1: Summary of top-down information results from the R-JMICC dataset.

to those from IDS, and reveal that the model has not learned anything about the distinction between short and long vowels (due to the high base rate of short vowels).

A.1.0.2 Acoustic and Higher-Level Contextual Information Model

The following models used contextual factors as direct predictors to category membership, in addition to using absolute duration and formant values. When part-of-speech was simplified to the distinction between function and content words, the model reached an overall accuracy of 95.8%, correctly classifying 98.5% of short vowels and 70.4% of long vowels. The BIC was 8331. When we included full part-of-speech information, the model achieved an overall accuracy of 95.8%, correctly classifying 98.5% of short vowels and 70.4% of long vowels. The BIC was 8328. Similarly for the IDS, both models substantially improved the baseline, though an even greater percentage of long vowels is classified correctly on ADS than on IDS (where long vowel accuracy reaches around 64%. Another difference we observe is that for ADS, the results are almost identical regardless of how part-of-speech information is used, while including more detailed part-of-speech information causes performance to go up on IDS. Overall, however, we observe the same pattern of results for ADS and IDS: using contextual information in a top-down information improves categorization performance.

A.1.0.3 Higher-Level Contextual Information Model Without Acoustics

In the final model, we only used contextual factors (including full part-of-speech information) as direct predictors of category membership. That is, the model did not have access to any acoustic information and could only make use of information about how likely each category is to occur a priori. Even without any acoustic information, this model achieved an overall accuracy of 95.1%, correctly classifying 98.4% of short vowels and 64.6% of long vowels. The model BIC was 9105. Similarly as for IDS, even without any acoustic information, the model can successfully classify a large portion of vowels just based on where the sound occurred.

A.2 Normalization Results

A summary of the results is presented in Table 3.3. We only ran a subset of the models on the ADS. In particular, we compared the baseline unnormalized model against two linear regression normalization models that used all contextual factors, but either used simplified part-of-speech information or full part-of-speech information. These results show that normalization is ineffective for the R-JMICC ADS, just as it was on the IDS.

Model	Accuracy	Short Accuracy	Long Accuracy	BIC
Unnormalized baseline	90.5	98.9	11.9	14820
Linear Regression Normalization All factors with simplified part-of-speech	90.2	99.7	2.4	16306
Linear Regression Normalization All factors with full part-of-speech	90.2	99.7	1.7	16348

Table A.2: Summary of normalization results on R-JMICC corpus (ADS portion).

A.2.0.1 Unnormalized Model

The baseline model is identical to the baseline model from the previous analysis and uses unnormalized duration and formants as predictors of category membership, without running any linear regression models. As a reminder, this logistic regression model reached an overall accuracy of 90.5%. It correctly categorized 98.9% of short vowels, and 11.9% of long vowels. It had a BIC of 14820.

A.2.1 Linear Regression Normalization Model Results

The following models used linear regression to normalize duration and formants, and used normalized acoustics as predictors of category membership. When all of the contextual factors with simplified part-of-speech (function vs. content word) were regressed out, the model had an overall categorization accuracy of 90.2%, correctly classifying 99.7% of the short vowels and 2.4% of the long vowels. It had a BIC of 16306. When all of the contextual factors including full part-of-speech information were regressed out, the model had an overall categorization accuracy of 90.2%, correctly classifying 99.7% of the short vowels and 1.7% of the long vowels. It had a BIC of 16348.

A.3 Summary of R-JMICC ADS vs. IDS Results

Overall, the results on adult-directed speech qualitatively mirror the results on infant-directed speech. The biggest difference appears to be that the top-down information model achieves slightly better categorization performance on the adult-

directed speech, though the ADS portion of the dataset is smaller so it is difficult to compare. Regardless of which type of data we use, we observe that using contextual information as top-down information helps separate short and long vowels, but normalizing for systematic contextual variability does not (at least as we implement it).

Appendix B: Derivation for Equation 4.3

Notation

$N_{l,c}$ - The number of vowels of length l (phonemically short or long) in context c

$\mu_{l,c}$ - The mean duration of vowels of length l (phonemically short or long) in context c

v_i - The unnormalized duration of vowel token i

Derivation

The following derives Equation 4.3, which characterizes how the means of two categories will move relative to one another as a result of normalization when top-down expectations are present. First, we write out what the unnormalized average duration of long vowels and short vowels is, beginning with long vowels. The average duration of long vowels is simply the sum of every long vowel's duration, divided by the total number of long vowels:

$$\mu_{l=\text{long}}^{\text{unnorm}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} v_i \tag{B.1}$$

Summing every long vowel's duration is equivalent to summing every long vowel's duration in every context, and then adding up the sums from each context, which

can be written as

$$\mu_{l=\text{long}}^{\text{unnorm}} = \frac{1}{N_{l=\text{long}}} \sum_{j} \sum_{i:\, l_i=\text{long}, c_i=j} v_i \tag{B.2}$$

We then multiply in $\frac{N_{l=\text{long},c=j}}{N_{l=\text{long},c=j}}$ to obtain:

$$\mu_{l=\text{long}}^{\text{unnorm}} = \frac{1}{N_{l=\text{long}}} \sum_{j} \frac{N_{l=\text{long},c=j}}{N_{l=\text{long},c=j}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i \tag{B.3}$$

The value $\frac{1}{N_{l=\text{long},c=j}} \sum_{i:\, l_i=\text{long},c_i=j} v_i$ is simply $\mu_{l=\text{long},c=j}^{\text{unnorm}}$. This is because we are summing the durations of all long vowels in context j and then dividing that by the total number of long vowels in context j, which is equivalent to the mean duration of long vowels in context j. This gives us the following equation for the unnormalized average duration of long vowels:

$$\mu_{l=\text{long}}^{\text{unnorm}} = \sum_{j} \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}} \mu_{l=\text{long},c=j}^{\text{unnorm}} \tag{B.4}$$

Similarly, the unnormalized average duration of short vowels is

$$\mu_{l=\text{short}}^{\text{unnorm}} = \sum_{j} \frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} \mu_{l=\text{short},c=j}^{\text{unnorm}} \tag{B.5}$$

Next, we compute the normalized average duration of long vowels and short vowels, starting with long vowels. To normalize a particular vowel's duration based on the context it occurs in, we take that vowel's unnormalized duration, v_i, and subtract from it the average duration of vowels in that context. The average duration of vowels in that context can be calculated by taking the sum of all short vowel durations in that context, adding that to the sum of all long vowel dura-

tions in that context, and then dividing the total sum (of short vowel and long vowel durations) by the total number of vowels in that context. This difference, $\left(v_i - \frac{1}{N_{c=c_i}}\left(\sum_{k:\, l_k=\text{long}, c_k=c_i} v_k + \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k\right)\right)$, is the normalized duration value for one particular long vowel, i. We can then take the sum of this value from each long vowel token, and divide by the total number of long vowels to obtain the average long vowel normalized duration:

$$\mu^{\text{norm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \left(v_i - \frac{1}{N_{c=c_i}}\left(\sum_{k:\, l_k=\text{long}, c_k=c_i} v_k + \sum_{k:\, l_k=\text{short}, c_k=c_i} v_k\right)\right) \tag{B.6}$$

Multiplying everything out yields

$$\begin{aligned}\mu^{\text{norm}}_{l=\text{long}} = \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} v_i - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k \\ - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k\, l_k=\text{short}, c_k=c_i} v_k\end{aligned} \tag{B.7}$$

The first term in Equation B.7 is summing the unnormalized durations of all long vowels and dividing by the total number of long vowels there are, so this first term is equivalent to the mean unnormalized duration of long vowels. Therefore, we can rewrite Equation B.7 as:

$$\begin{aligned}\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k \\ - \frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k\, l_k=\text{short}, c_k=c_i} v_k\end{aligned} \tag{B.8}$$

From here, we will rewrite both the second and third terms of Equation B.8, and we

make an aside here to show how. Consider first the second term of Equation B.8. In it, we are summing over all long vowels. Just as in the transition from Equation B.1 to Equation B.2, we can rewrite this as summing over all long vowels in a particular context, and then summing over these contexts, which yields (B.9). To get from (B.9) to (B.10), notice that the term, $\frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k$ in the inner sum will be the same for every long vowel in context j, so this term will be repeated exactly $N_{l=\text{long}, c=j}$ times.

$$\frac{1}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}} \frac{1}{N_{c=c_i}} \sum_{k:\, l_k=\text{long}, c_k=c_i} v_k = \frac{1}{N_{l=\text{long}}} \sum_{j} \sum_{i:\, l_i=\text{long}, c_i=j} \frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k \tag{B.9}$$

$$= \frac{1}{N_{l=\text{long}}} \sum_{j} N_{l=\text{long}, c=j} \frac{1}{N_{c=j}} \sum_{k:\, l_k=\text{long}, c_k=j} v_k \tag{B.10}$$

Using the same logic for the third term, we can, therefore, rewrite Equation B.8 as follows:

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_{j} \frac{1}{N_{c=j}} \frac{N_{l=\text{long}, c=j}}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{long}, c_i=j} v_i - \sum_{j} \frac{1}{N_{c=j}} \frac{N_{l=\text{long}, c=j}}{N_{l=\text{long}}} \sum_{i:\, l_i=\text{short}, c_i=j} v_i \tag{B.11}$$

As before, the mean duration of long vowels in a particular context is equivalent to the sum over all long vowel durations in that context, divided by the total number of long vowels in that context. Writing this out notationally will help us rewrite

Equation B.11:

$$\mu^{\text{unnorm}}_{l=\text{long},c=j} = \frac{1}{N_{l=\text{long},c=j}} \sum_{i:\, l_i=\text{long},c_i=j} v_i \tag{B.12}$$

Using Equation B.12, we can rewrite Equation B.11 as

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_j \frac{N_{l=\text{long},c=j}}{N_{c=j}} \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}} \mu^{\text{unnorm}}_{l=\text{long},c=j} - \sum_j \frac{N_{l=\text{long},c=j}}{N_{c=j}} \frac{N_{l=\text{short},c=j}}{N_{l=\text{long}}} \mu^{\text{unnorm}}_{l=\text{short},c=j} \tag{B.13}$$

Factoring out $\frac{N_{l=\text{long},c=j}}{N_{c=j} N_{l=\text{long}}}$ gives us the following equation for the mean normalized duration of long vowels,

$$\mu^{\text{norm}}_{l=\text{long}} = \mu^{\text{unnorm}}_{l=\text{long}} - \sum_j \left[\frac{N_{l=\text{long},c=j}}{N_{c=j} N_{l=\text{long}}} \left(N_{l=\text{long},c=j} \mu^{\text{unnorm}}_{l=\text{long},c=j} + N_{l=\text{short},c=j} \mu^{\text{unnorm}}_{l=\text{short},c=j} \right) \right] \tag{B.14}$$

Similarly, the mean normalized duration of short vowels is

$$\mu^{\text{norm}}_{l=\text{short}} = \mu^{\text{unnorm}}_{l=\text{short}} - \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{c=j} N_{l=\text{short}}} \left(N_{l=\text{short},c=j} \mu^{\text{unnorm}}_{l=\text{short},c=j} + N_{l=\text{long},c=j} \mu^{\text{unnorm}}_{l=\text{long},c=j} \right) \right] \tag{B.15}$$

Up until this point, we have calculated the mean unnormalized duration of long vowels and short vowels, as well as the mean normalized duration of long vowels and short vowels. We can subtract the average unnormalized short vowel duration from the average unnormalized long vowel duration to obtain a measure of how far apart the two vowel categories are before normalization. Similarly, we can subtract the average normalized short vowel duration from the average normalized long vowel duration to obtain a measure of how far apart the two vowel categories are after normalization. To compare whether the means of the two categories move

closer together or farther apart after normalization, we can calculate the value of $\left(\mu^{\text{norm}}_{l=\text{long}} - \mu^{\text{norm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right)$, by simply plugging in the relevant terms from above.

$$
\begin{aligned}
&\left(\mu^{\text{norm}}_{l=\text{long}} - \mu^{\text{norm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) \\
&= \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) - \left(\mu^{\text{unnorm}}_{l=\text{long}} - \mu^{\text{unnorm}}_{l=\text{short}}\right) \qquad \text{(B.16)} \\
&\quad - \sum_j \frac{N_{l=\text{long},c=j}}{N_{c=j} N_{l=\text{long}}} \left(N_{l=\text{long},c=j}\mu^{\text{unnorm}}_{l=\text{long},c=j} + N_{l=\text{short},c=j}\mu^{\text{unnorm}}_{l=\text{short},c=j}\right) \\
&\quad + \sum_j \frac{N_{l=\text{short},c=j}}{N_{c=j} N_{l=\text{short}}} \left(N_{l=\text{long},c=j}\mu^{\text{unnorm}}_{l=\text{long},c=j} + N_{l=\text{short},c=j}\mu^{\text{unnorm}}_{l=\text{short},c=j}\right) \\
&= \sum_j \left[\left(\frac{N_{l=\text{short},c=j}\mu^{\text{unnorm}}_{l=\text{short},c=j} + N_{l=\text{long},c=j}\mu^{\text{unnorm}}_{l=\text{long},c=j}}{N_{c=j}}\right)\left(\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}\right)\right] \\
&\qquad \text{(B.17)}
\end{aligned}
$$

This gives us Equation 4.3 from the main text:

$$
\begin{aligned}
&\left(\mu^{norm}_{l=\text{long}} - \mu^{norm}_{l=\text{short}}\right) - \left(\mu^{unnorm}_{l=\text{long}} - \mu^{unnorm}_{l=\text{short}}\right) = \\
&\qquad \sum_j \left[\frac{N_{l=\text{short},c=j}}{N_{l=\text{short}}} - \frac{N_{l=\text{long},c=j}}{N_{l=\text{long}}}\right]\left[\frac{N_{l=\text{long},c=j}}{N_{c=j}}\mu_{l=\text{long},c=j} + \frac{N_{l=\text{short},c=j}}{N_{c=j}}\mu_{l=\text{short},c=j}\right] \text{(B.18)}
\end{aligned}
$$

We can then study whether this value is positive or negative. This value will be positive when the difference between the normalized means is greater than the difference between the unnormalized means (i.e. when normalization is effective and reduces the overlap between categories). Likewise, this value will be negative when normalization is ineffective and actually increases the overlap between categories.

As stated in the main text, this equation reveals that when different categories

differ in the contexts that they are likely to occur in, then normalization may actually increase the amount of overlap between different categories.

References

WA Ainsworth. Durational cues in the perception of certain consonants. *Proceedings of the British Acoustical Society*, 2:1–4, 1973.

WA Ainsworth. The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, 17(2):103–109, 1974.

J Sean Allen, Joanne L Miller, and David DeSteno. Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 113(1): 544–552, 2003.

Stephanie Antetomaso, Kouki Miyazawa, Naomi Feldman, Micha Elsner, Kasia Hitczenko, and Reiko Mazuka. Modeling phonetic category learning from natural acoustic data. In *BUCLD 41: Proceedings of the 41st Annual Boston University Conference on Language Development*, 2017.

Keith S Apfelbaum and Bob McMurray. Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, 22(4):916–943, 2015.

Takayuki Arai, DM Behne, Peter Czigler, and Kirk Sullivan. Perceptual cues to vowel quantity: Evidence from Swedish and Japanese. In *Proceedings Of the Swedish Phonetics Conference (Fonetik)*, volume 81, pages 8–11, 1999.

Richard N Aslin. Some developmental processes in speech perception. *Child Phonology: Perception & Production*, 1980.

Richard N Aslin, Peter W Jusczyk, and David B Pisoni. Speech and auditory processing during infancy: Constraints on and precursors to language. 1998.

Mark E Auckland, Kyle R Cave, and Nick Donnelly. Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review*, 14(2):332–337, 2007.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Normalized Wasserstein distance for mixture distributions with applications in adversarial learning and domain adaptation. *arXiv preprint arXiv:1902.00415*, 2019.

Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, 15(4):600–609, 2003.

Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617, 2004.

Moshe Bar and Shimon Ullman. Spatial context in recognition. *Perception*, 25(3): 343–352, 1996.

Elika Bergelson and Daniel Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012.

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

Catherine T Best, Gerald W McRoberts, Rosemarie LaFleur, and Jean Silver-Isenstadt. Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant behavior and development*, 18(3):339–350, 1995.

Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, 1982.

Ricardo AH Bion, Kouki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka. Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLOS ONE*, 8(2):e51594, 2013.

Paul Boersma. Praat: A system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

Geert Booij. *The phonology of Dutch*. Oxford University Press, 1999.

Victor J Boucher. Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. *Perception & Psychophysics*, 64(1):121–130, 2002.

Susan J Boyce, Alexander Pollatsek, and Keith Rayner. Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):556, 1989.

Ann R Bradlow and Tessa Bent. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729, 2008.

Roger W Brown and Donald C Hildum. Expectancy and the perception of syllables. *Language*, 32(3):411–419, 1956.

Hui Chen, Naoto Yamane, N. X. Rattasone, Katherine Demuth, and Reiko Mazuka. Japanese infants are aware of phonemic vowel length in novel words at 18 months. In *BUCLD 40: Proceedings of the 40th Annual Boston University Conference on Language Development*, 2016.

Jennifer Cole, Gary Linebaugh, Cheyenne Munson, and Bob McMurray. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38(2):167–184, 2010.

Alejandrina Cristia and Amanda Seidl. The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, 41(4):913–934, 2014.

Thomas H Crystal and Arthur S House. Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, 88(1):101–112, 1990.

Delphine Dahan, Sarah J Drucker, and Rebecca A Scarborough. Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, 108 (3):710–718, 2008.

Jodi L Davenport and Mary C Potter. Scene consistency in object and background perception. *Psychological Science*, 15(8):559–564, 2004.

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

Christiane Dietrich, Daniel Swingley, and Janet F Werker. Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences*, 104(41):16027–16031, 2007.

Laura C Dilley and Mark A Pitt. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11):1664–1670, 2010.

Brian Dillon, Ewan Dunbar, and William Idsardi. A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive Science*, 37(2): 344–377, 2013.

Baxter S Eaves Jr, Naomi H Feldman, Thomas L Griffiths, and Patrick Shafto. Infant-directed speech is consistent with teaching. *Psychological Review*, 123(6): 758, 2016.

Rebecca E Eilers and Fred D Minifie. Fricative discrimination in early infancy. *Journal of Speech and Hearing Research*, 18(1):158–167, 1975.

Peter D Eimas. Speech perception in early infancy. In *Infant perception: From sensation to cognition*, pages 193–231. Elsevier, 1975.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 42–54, 2013.

Mirjam Theresia Constantia Ernestus. Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface. 2000.

Naomi H Feldman, Thomas L Griffiths, Sharon Goldwater, and James L Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751, 2013a.

Naomi H Feldman, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438, 2013b.

Paula Fikkert. *On the acquisition of prosodic structure*. [Sl: sn], 1994.

Hiroya Fujisaki and Osamu Kunisaki. Analysis, recognition, and perception of voiceless fricative consonants in Japanese. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):21–27, 1978.

Hiroya Fujisaki, Kimie Nakamura, and Toshiaki Imoto. Auditory perception of duration of speech and non-speech stimuli. *Auditory analysis and perception of speech*, pages 197–219, 1975.

Seiji Fukui. Perception for the Japanese stop consonants with reduced and extended durations. *Onsei Gakkai Kaihou*, 59:9–12, 1978.

William F Ganong. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1):110, 1980.

Hiromu Goto. Auditory perception by normal Japanese adults of the sounds [l] and [r]. *Neuropsychologia*, 1971.

RT Green and MC Courtis. Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, 1966.

Adriana Guevara-Rukoz, Alejandrina Cristia, Bogdan Ludusan, Roland Thiollière, Andrew Martin, Reiko Mazuka, and Emmanuel Dupoux. Are words easier to learn from infant-than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science*, 42(5):1586–1617, 2018.

Mieko S Han. Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America*, 96(1):73–82, 1994.

Mieko Shimizu Han. *Japanese Phonology: An Analysis Based Upon Sound Spectograms*. Kenkuysha, 1962.

Angela Xiaoxue He and Jeffrey Lidz. Verb learning in 14-and 18-month-old english-learning infants. *Language Learning and Development*, 13(3):335–356, 2017.

James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5):3099–3111, 1995.

James Hillenbrand, Michael J Clark, and Terrance M Nearey. Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, 109(2):748–763, 2001.

Yukari Hirata. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, 32(4):565–589, 2004.

Miwako Hisagi, Valerie L Shafer, Winifred Strange, and Elyse S Sussman. Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain research*, 1360:89–105, 2010.

Kasia Hitczenko, Reiko Mazuka, Micha Elsner, and Naomi H Feldman. How to use context to disambiguate overlapping categories: The test case of Japanese vowel length. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, volume 40, 2018.

Barbara Höhle, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, 5(3):341–353, 2004.

Arthur S House. On vowel duration in English. *The Journal of the Acoustical Society of America*, 33(9):1174–1178, 1961.

Toshiko Isei-Jaakkola. Lexical quantity in Japanese and Finnish. *Unpublished doctoral dissertation*, 2004.

Roman Jakobson and John Lotz. Notes on the French phonemic pattern. *Word*, 5 (2):151–158, 1949.

Keith Johnson. Speech perception without speaker normalization: An exemplar model. *Talker variability in speech processing*, pages 145–165, 1997.

Keith Johnson. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of phonetics*, 34(4):485–499, 2006.

Allard Jongman, Ratree Wayland, and Serena Wong. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3): 1252–1263, 2000.

Shigeto Kawahara. A faithfulness ranking projected from a perceptibility scale: The case of [+ voice] in Japanese. *Language*, pages 536–574, 2006.

Patricia Keating, Taehong Cho, Cécile Fougeron, and Chai-Shune Hsu. Domain-initial articulatory strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, pages 143–161, 2004.

Keisuke Kinoshita, Dawn M Behne, and Takayuki Arai. Duration and F0 as perceptual cues to Japanese vowel quantity. In *Seventh International Conference on Spoken Language Processing*, 2002.

Katrin Kirchhoff and Steven Schimmel. Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, 117(4):2238–2246, 2005.

Dave F Kleinschmidt and T Florian Jaeger. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148, 2015.

Patricia K Kuhl, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Björn Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608, 1992.

Patricia K Kuhl, Jean E Andruski, Inna A Chistovich, Ludmilla A Chistovich, Elena V Kozhevnikova, Viktoria L Ryskina, Elvira I Stolyarova, Ulla Sundberg, and Francisco Lacerda. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686, 1997.

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

Heike Lehnert-LeHouillier. A cross-linguistic investigation of cues to vowel length perception. *Journal of Phonetics*, 38(3):472–482, 2010.

Clara Levelt. *On the acquisition of place*. PhD thesis, Rijksuniversiteit Leiden Leiden, 1994.

Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358, 1957.

Leigh Lisker and Arthur S Abramson. The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*, volume 563, pages 563–567. Academia Prague Prague, Czech Republic, 1970.

Paul A Luce and Jan Charles-Luce. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, 78(6):1949–1957, 1985.

Bogdan Ludusan, Alejandrina Cristia, Andrew Martin, Reiko Mazuka, and Emmanuel Dupoux. Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America*, 140(2):1239–1250, 2016.

Bogdan Ludusan, Reiko Mazuka, Mathieu Bernard, Alejandrina Cristia, and Emmanuel Dupoux. The role of prosody and speech register in word segmentation: A computational modelling perspective. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 178–183, 2017.

Kikuo Maekawa. Corpus of Spontaneous Japanese: Its Design and Evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

Virginia A Mann and Bruno H Repp. Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Attention, Perception, & Psychophysics*, 28(3):213–228, 1980.

William Marslen-Wilson and Paul Warren. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, 101(4):653–674, 1994.

Andrew Martin, Yosuke Igarashi, Nobuyuki Jincho, and Reiko Mazuka. Utterances in infant-directed speech are shorter, not slower. *Cognition*, 156:52–59, 2016.

Dominic W Massaro and Michael M Cohen. Phonological context in speech perception. *Attention, Perception, & Psychophysics*, 34(4):338–348, 1983.

Jessica Maye, Janet F Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, 2002.

Jessica Maye, Richard N Aslin, and Michael K Tanenhaus. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, 32(3):543–562, 2008.

Reiko Mazuka, Yosuke Igarashi, and Ken'ya Nishikawa. Input for learning Japanese: RIKEN Japanese Mother-Infant Conversation Corpus. *The Technical Report of the Proceedings of the Institute of Electronics, Information and Communication Engineers*, 106(165):11–15, 2006.

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

Bob McMurray, Kristine A Kovack-Lesh, Dresden Goodwin, and William McEchron. Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, 129(2):362–378, 2013.

Bob McMurray, Ani Danelz, Hannah Rigler, and Michael Seedorff. Speech categorization develops slowly through adolescence. *Developmental psychology*, 54(8): 1472, 2018.

James M McQueen, Dennis Norris, and Anne Cutler. Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5):1363, 1999.

George A Miller, George A Heise, and William Lichten. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5):329, 1951.

Joanne L Miller. Effects of speaking rate on segmental distinctions. *Perspectives on the Study of Speech*, pages 39–74, 1981.

Joanne L Miller and Alvin M Liberman. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, 25(6):457–465, 1979.

Joanne L Miller, Francois Grosjean, and Concetta Lomanto. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41(4):215–225, 1984.

Joanne L Miller, Timothy B O'Rourke, and Lydia E Volaitis. Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, 54(3-4):121–137, 1997.

FD Minifie, PK Kuhl, and EM Stecher. Categorical perception of/b/and/w/during changes in rate of utterance. *The Journal of the Acoustical Society of America*, 62(S1):S79–S79, 1977.

Toben H Mintz. Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, pages 31–63, 2006.

Philip J Monahan and William J Idsardi. Auditory sensitivity to formant ratios: Toward an account of vowel normalisation. *Language and cognitive processes*, 25 (6):808–839, 2010.

Elliott Moreton and Shigeaki Amano. Phonotactics in the perception of Japanese vowel length: Evidence for long-distance dependencies. In *EUROSPEECH*, 1999.

Ryoko Mugitani, Ferran Pons, Laurel Fais, Christiane Dietrich, Janet F Werker, and Shigeaki Amano. Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, 45(1):236, 2009.

Chandan Narayan. Developmental perspectives on phonological typology and sound change. *Origins of Sound Change: Approaches to Phonologization*, pages 128–146, 2013.

Chandan Narayan, Andrew Peters, and Vanessa Woldenga-Racine. Fragile phonetic contrasts in longitudinal infant-directed speech: Implications for infant speech perception. In *BUCLD 42: Proceedings of the 41st Annual Boston University Conference on Language Development*, 2017.

Terrance Nearey. Vowel space normalization in synthetic stimuli. *The Journal of the Acoustical Society of America*, 63(1), 1978.

Terrance Nearey. The segment as a unit of speech perception. *Journal of Phonetics*, 1990.

Rochelle S Newman and James R Sawusch. Perceptual normalization for speaking rate: Effects of temporal distance. *Attention, Perception, & Psychophysics*, 58 (4):540–560, 1996.

Rochelle S Newman, Sheryl A Clouse, and Jessica L Burnham. The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3):1181–1196, 2001.

Sibout Govert Nooteboom and Iman Hans Slis. The phonetic feature of vowel length in Dutch. *Language and Speech*, 15(4):301–316, 1972.

Dennis Norris, James M McQueen, and Anne Cutler. Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238, 2003.

Stephen E Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3:519–526, 1975.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41, 2006.

JM Pickett and Louis R Decker. Time factors in perception of a double consonant. *Language and Speech*, 3(1):11–17, 1960.

Janet Pierrehumbert. Word-specific phonetics. *Laboratory Phonology*, 7, 2002.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The Buckeye Corpus of Conversational Speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95, 2005.

Linda Polka and Janet F Werker. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human perception and performance*, 20(2):421, 1994.

Linda Polka, Connie Colantonio, and Megha Sundara. A cross-language comparison of/d/–/ð/ perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, 109(5):2190–2201, 2001.

Robert F Port and Jonathan Dalby. Consonant/vowel ratio as a cue for voicing in English. *Attention, Perception, & Psychophysics*, 32(2):141–152, 1982.

Robert B Post, Robert B Welch, and Kristin Caufield. Relative spatial expansion and contraction within the Müller-Lyer and Judd illusions. *Perception*, 27(7): 827–838, 1998.

Brad Rakerd, William Sennett, and Carol A Fowler. Domain-final lengthening and foot-level shortening in spoken English. *Phonetica*, 44(3):147–155, 1987.

Caitlin Richter, Naomi H Feldman, Harini Salgado, and Aren Jansen. Evaluating low-level speech features against human perceptual data. In *Transactions of the Association for Computational Linguistics*, 2017.

David Rose and Paola Bressan. Going round in circles: Shape effects in the Ebbinghaus illusion. *Spatial Vision*, 15(2):191–203, 2002.

P Rubin, M Turvey, and Van Gelder. Initial phonemes are detected faster in spoken words than in nonwords. *Haskins Laboratories Status Report on Speech Research*, 1976.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision*, pages 59–66. IEEE, 1998.

Yutaka Sato, Yuko Sogabe, and Reiko Mazuka. Discrimination of phonemic vowel length by Japanese infants. *Developmental Psychology*, 46(1):106, 2010.

James R Sawusch and Rochelle S Newman. Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Attention, Perception, & Psychophysics*, 62(2):285–300, 2000.

Thomas Schatz, Naomi Feldman, Sharon Goldwater, Xuan Nga Cao, and Emmanuel Dupoux. Early phonetic learning without phonetic categories–Insights from machine learning. 2019.

Barbara Schuppler, Mirjam Ernestus, Odette Scharenborg, and Lou Boves. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, 39(1):96–109, 2011.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6 (2):461–464, 1978.

Rushen Shi and Andréane Melançon. Syntactic categorization in French-learning infants. *Infancy*, 15(5):517–533, 2010.

Rushen Shi and Janet F Werker. Six-month-old infants' preference for lexical words. *Psychological Science*, 12(1):70–75, 2001.

Rushen Shi, Janet F Werker, and James L Morgan. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11–B21, 1999.

Elizabeth A Strand and Keith Johnson. Gradient and visual speaker normalization in the perception of fricatives. In *KONVENS*, pages 14–26, 1996.

Quentin Summerfield. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5):1074, 1981a.

Quentin Summerfield. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5):1074, 1981b.

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

Daniel Swingley. Learning phonology from surface distributions, considering Dutch and English vowel duration. *Language Learning and Development*, pages 1–18, 2019.

Daniel Swingley and Claudia Alarcon. Lexical learning may contribute to phonetic learning in infants: A corpus analysis of maternal Spanish. *Cognitive Science*, 2018.

Daniel Swingley and Richard N Aslin. Lexical competition in young children's word learning. *Cognitive Psychology*, 54(2):99–132, 2007.

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

Dejan Todorović. Context effects in visual perception and their explanations. *Review of Psychology*, 17(1):17–32, 2010.

Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

Francisco Torreira and Mirjam Ernestus. Phrase-medial vowel devoicing in spontaneous French. In *11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, pages 2006–2009, 2010.

Francisco Torreira and Mirjam Ernestus. Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology*, 2(2):331–353, 2011a.

Francisco Torreira and Mirjam Ernestus. Vowel elision in casual French: The case of vowel /e/ in the word c'était. *Journal of Phonetics*, 39(1):50–58, 2011b.

Francisco Torreira and Mirjam Ernestus. Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica*, 69(3):124–148, 2012.

Francisco Torreira, Martine Adda-Decker, and Mirjam Ernestus. The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3):201–212, 2010.

Joseph C Toscano and Bob McMurray. Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, 74(6):1284–1301, 2012.

Sandra E Trehub. The discrimination of foreign speech contrasts by infants and adults. *Child Development*, pages 466–472, 1976.

Sho Tsuji and Alejandrina Cristia. Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, 56(2):179–191, 2014.

Teruaki Tsushima, Osamu Takizawa, Midori Sasaki, Satoshi Shiraki, Kanae Nishi, Morio Kohno, Paula Menyuk, and Catherine Best. Discrimination of english /r-l/ and /w-y/ by japanese infants at 6-12 months: Language-specific developmental changes in speech perception abilities. In *Third International Conference on Spoken Language Processing*, 1994.

Noriko Umeda. Vowel duration in American English. *The Journal of the Acoustical Society of America*, 58(2):434–445, 1975.

Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278, 2007.

Suzanne VH Van der Feest and Daniel Swingley. Dutch and English listeners' interpretation of vowel duration. *The Journal of the Acoustical Society of America*, 129(3):EL57–EL63, 2011.

Marieke van Heugten and Elizabeth K Johnson. Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, 143(1):340, 2014.

Jan PH Van Santen. Contextual effects on vowel duration. *Speech Communication*, 11(6):513–546, 1992.

Robert R Verbrugge, Winifred Strange, Donald P Shankweiler, and Thomas R Edman. What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, 60(1):198–212, 1976.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Richard M Warren. Perceptual restoration of missing speech sounds. *Science*, 167 (3917):392–393, 1970.

Sarah C Wayland, Joanne L Miller, and Lydia E Volaitis. The influence of sentence articulation rate on the internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, 92(4):2465–2465, 1992.

Sarah C Wayland, Joanne L Miller, and Lydia E Volaitis. The influence of sentential speaking rate on the internal structure of phonetic categories. *The Journal of the Acoustical Society of America*, 95(5):2694–2701, 1994.

Janet F Werker and Richard C Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63, 1984.

Janet F Werker, John HV Gilbert, Keith Humphrey, and Richard C Tees. Developmental aspects of cross-language speech perception. *Child development*, pages 349–355, 1981.

Janet F Werker, Ferran Pons, Christiane Dietrich, Sachiyo Kajikawa, Laurel Fais, and Shigeaki Amano. Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, 103(1):147–162, 2007.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The HTK book. *Cambridge University Engineering Department*, 3:175, 2002.