UNIVERSITY OF CALIFORNIA, MERCED

Hierarchical Temporal Structure and Convergence in Development, Performance, and Bilingualism

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

in

Cognitive and Information Sciences

by

Sara Schneider

Committee in Charge:

Professor Chris Kello, Chair
Professor Ramesh Balasubramaniam
Professor Michael Spivey
Professor Anne Warlaumont

2021

The dissertation of Sara Schneider is approved, and it is acceptable in quality and form for publication in microfilm and electronically:

_____

Professor Chris Kello, Ph.D., Committee Chair
Cognitive and Information Sciences
University of California, Merced

_____

Professor Ramesh Balasubramaniam, Ph.D., Committee Member
Cognitive and Information Sciences
University of California, Merced

_____

Professor Michael Spivey, Ph.D., Committee Member
Cognitive and Information Sciences
University of California, Merced

_____

Professor Anne Warlaumont, Ph.D., Committee Member
Department of Communication
University of California, Los Angeles

University of California, Merced

2021

This dissertation is dedicated to my spouse, Sean, for fully supporting me from start to finish on this quest.

I also dedicate this dissertation to my grandfather, the original Dr. Schneider, for setting the academic precedent for me to follow.

Finally, this dissertation is dedicated to my grandmother, Elaine, for instilling in me strong guiding senses of persistence and discipline.

Table of Contents

List of Figures

List of Tables

Acknowledgments

There are many people I wish to express gratitude towards, the first and foremost being my advisor, Chris. Thank you for always being supportive and understanding of my goals, prioritizing time for our meetings, and helping me to secure funding for my research and everyday life. Thank you also to my committee members, Spivey, Ramesh, and Anne, for always being eager to support me, provide advice, and answer questions. I would not have been able to do this without your guidance.

I am also grateful for my external mentors, including Jennifer Pardo, for all of her contributions, advice, and patience with this dissertation (namely the third chapter). Thank you to Rick Dale, who helped introduce me to cognitive science and provided inspiration and advice for projects. I also wish to thank the members of my alma mater community, the University of Cincinnati, for allowing me to become involved in experimental research as an undergraduate and for first showing me how much fun it can be to gather experimental data.

My thanks are extended to Gina Pretzer and Lukas Lopez for spending so many hours teaching me the ins and outs of the infant database used in the second chapter, helping me train research assistants, and for always being one phone call away. Likewise, thank you to Katie Coburn for her many hours teaching me how to write and analyze statistics in R. Your guidance was invaluable.

I am indebted to my research assistants who have spent hundreds of hours aiding me with these projects, some over the course of years. Countless thanks go to those including, but absolutely not limited to, Zulema Avalos, Carol Gavilan, Alexis Luna, Jeffrey Mai, Silvano Mendiaz, Jimel Mutrie, Harmanjit Singh, Gilbert Sepulveda, Cosme Soto Alvarez, Aileen Velasquez, and Justin Wong. These studies would not have been possible without your tireless efforts. Thank you more broadly to the many students who kept me inspired over the years.

Thank you to my lab mates, Adolfo Ramirez-Aristizabal, Camila Alviar Guzman, Daniel Schloesser, and Ketika Garg, for providing me with companionship, encouragement, and snacks. Further, thank you to the faculty members and other fellow graduate students within and outside of my program, including Shannon Proksch and Taylor Fugere, for making this a wonderful experience and for always offering the best advice. My gratitude also goes to my friends who always offered to proofread my work and provide emotional support.

Many thanks go to my family for their continuous support throughout my life, including my parents, sister, and many grandparents; I can always count on you to lend a listening ear and offer whatever support possible. Thank you to my spouse, Sean, for sticking by me throughout my undergraduate and graduate career, providing limitless love. Thanks to Mufasa, Leo, and Kobe for being the best workplace associates throughout 2020.

Finally, thanks to the National Science Foundation for their support with the following projects (grant numbers 1529127, 1633722, and 1229033).

Curriculum Vitae

Sara Schneider
sschneider2@ucmerced.edu

Education

The University of California, Merced 2016 – Present

Ph.D., Cognitive and Information Sciences

M.S., Cognitive and Information Sciences

The University of Cincinnati 2011 – 2015

B.S., Psychology (Magna Cum Laude)

B.A., Philosophy (Magna Cum Laude)

Certificate, Deaf Studies

Academic Employment

University of California, Merced

Laboratory Assistant 2015 – 2016

Cognition and Action Laboratory
PI: Rick Dale

Cognitive Mechanics Laboratory 2016
PI: Christopher Kello

Graduate Student Researcher 2018-2019

Emergence of Communication Laboratory
PIs: Christopher Kello, Anne Warlaumont

Teaching Experience

Instructor of Record

University of California, Merced

Introduction to Cognitive Science (Online) Summer 2020

Teaching Assistantships

University of Cincinnati

Statistics and Research Methods in Psychology — Fall, 2012; Fall, 2013
Professors: Kenneth King; Julie Weast

Sensation and Perception — Spring, 2013
Professor: Michael Richardson

University of California, Merced

Mind, Brain, Computation — Fall, 2016
Professor: Christopher Kello

Research Methods — Spring, 2017; Spring, 2018
Professors: Rick Dale; Stephane Beaudin

Introduction to Cognitive Science — Fall, 2017
Professor: Michael Spivey

Introduction to Ethics — Fall, 2019
Professor: David Jennings

Publications

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*. *23*(4). 845-857.

Lopez, L., Moorman, K., Schneider, S., Baker, M., Holbrook, C. (2020). Morality is relative: Anger, disgust and aggression as contingent responses to sibling versus stranger harm. *Emotion*. *21*(2), 376-390.

Schneider, S., Pardo, J. S., Coburn, K. M., Kello, C. (under revision). The effects of prosody and prosody convergence on performance in the Montclair Map Task.

Poster Presentations

Schneider, S., Shockley, K., & Tolston, M. (2014, June). The effect of discontinuous transitions in story structure on postural sway dynamics. International Society for Ecological Psychology. Oxford, OH, USA.

Schneider, S., Ramirez-Aristizabal, A., Kello, C. (2016, November). Complexity matching across Spanish and English. Society for Computers in Psychology. Boston, MA, USA.

Schneider, S., Ramirez-Aristizabal, A., Gavilan, C., Kello, C. (2018, July). Complexity matching across Spanish and English. International Workshop for Language Production. Nijmegen, The Netherlands.

Schneider, S., Pardo, J. S., Coburn, K. M., Kello, C. T. (2019, November). Prosody, prosodic convergence, and performance in the Montclair map task. Psychonomics. Montréal, Canada.

Oral Presentation

Schneider, S. (2017, February). Complexity matching across Spanish and English. Dynamics of Language workshop. Santa Barbara, CA.

Fellowships

Summer Bridge Program	2016
NSF Research Traineeship (RT) for Intelligence Adaptive Systems	2020 –2021

Extracurricular Training

NSF RT for Interdisciplinary Computational Graduate Education	Spring 2018
Dissertation Bootcamp	Winter, Summer 2019
Reflecting on Anti-Racist Pedagogy Workshop	Fall 2020

Honors and Awards

Phi Beta Kappa	2015
Student Marshal	2015
Outstanding Teaching Award	2018

Academic Service

Coordinated CIS Interview Weekend	2017 – 2019
Graduate Pedagogical Association Secretary and Vice President	2018 – 2020

Abstract

Regardless of the language spoken, prosodic features of speech (including variations in amplitude and rhythm) are important for conveying emotional states and semantic meanings. Such differences in prosodic structure can allow for monotone versus demonstrative speech. Previous research has shown that features of speech including prosody often converge when pairs of speakers interact. However, few studies have investigated how convergence applies to speakers in bilingual conversations, despite bilingualism being prevalent globally. For example, in lingua receptiva situations, each speaker uses a preferred language that the other understands. Further, prior studies have found convergence to be related to rapport and have theorized it to be important for successful communication, but few studies have analyzed how either prosody or prosodic convergence are correlated with various forms of performance. This dissertation aims to contribute a stronger understanding of how hierarchical temporal structure—a novel feature of prosody—and convergence between these structures correlate with performance, as well as monolingual and bilingual conversations (i.e., intra- and inter-language interactions).

Chapter 1 introduces the background behind hierarchical temporal structure and the method of quantification. Likewise, the background and method of measurement are described for the convergence of hierarchical temporal structures, known as complexity matching. Chapter 2 then investigates how infant and adult prosody and volubility correlate with lexical development, as reported by the caregivers using a vocabulary assessment. Likewise, complexity matching and volubility matching were measured between infants and adults, and the relationship between both forms of convergence and vocabulary size is explored. Following this, Chapter 3 examines if either hierarchical temporal structure or speaking rate are related to collaborative task performance with adult pairs. Chapter 4 next tests if any differences for hierarchical temporal structure, complexity matching, or lexical matching were observed by language condition (Spanish only, English only, or one speaker used Spanish only and the other English only). Chapter 5 closes with a discussion of the implications and limitations of the findings, along with proposed future directions of study.

This dissertation, *Hierarchical Temporal Structure and Convergence in Development, Performance, and Bilingualism*, is submitted by Sara Schneider in 2021 in partial fulfillment of the degree Doctor of Philosophy in Cognitive and Information Sciences at the University of California, Merced under the guidance of dissertation committee chair Chris Kello.

Chapter 1

Introduction

1.1 Prosody and Communication

Prosody has been deemed the "music of speech," where it encompasses the way we style our sentences, rather than the specific words spoken (Erekson, 2010). When using identical words, the prosodic style used can drastically alter the meaning of the phrase. For example, the sentence "We're going to the doctor," may be said in a matter-of-fact manner like when giving a reminder, it may be said fearfully, or it may be a parent saying it to an infant using a singsong voice. Prosody thus carries substantial semantic importance and is crucial for successful communication. Three acoustic features comprising prosody are fundamental frequency, intensity, and duration, which are perceived as pitch, loudness, and the length of units of speech (Patel, Niziolek, Reilly, & Guenther, 2011). Notably, different timescales of speech (e.g., syllables, words, and phrases) each contain a measurable fundamental frequency, intensity, and duration, indicating prosodic differences are not limited to a single timescale.

In the following subsections, I first introduce prosody in early, performance-based, and bilingual interactions, followed by a description of a feature of prosody explored throughout this study (hierarchical temporal structure) and how it is quantified. In the following section, I then discuss the convergence or alignment of prosody and other features of speech between speakers (again during early, performance-based, and bilingual interactions), followed by a corresponding description of the convergence between hierarchical temporal structures (complexity matching) and how it is quantified.

1.1.1. Early Interactions

During some of the earliest interactions between caregivers and infants, prosodic structure already shows to be vital. When addressing infants, caregivers typically use more prosodically rhythmic, higher-pitched, and slower-paced speech, known as infant-directed speech (IDS). These features of IDS are largely theorized to aid infants in language acquisition because they help to provide distinctions between vowels and syllables within and across utterances, (Kuhl et al., 1997; Leong, Kalashnikova, Burnham, & Goswami, 2014), making it easier for infants to understand the acoustic boundaries between them (Thiessen, Hill, & Saffran, 2005). One feature of prosody measured throughout this dissertation was hierarchical temporal structure, which is reflective of prosodic exaggeration, or more precisely variations in the amplitude and timing of acoustic events derived from the recorded speech sounds (a detailed description of the background behind hierarchical temporal structure and how it is quantified follows in Sections 1.1.4. and 1.1.5.). Slower speaking rates—which often go hand in hand with IDS—correspond to greater hierarchical temporal structure, which is depicted when plotted as having a steeper slope (Ramirez-Aristizabal, Médé, & Kello, 2018). Providing further confirmation of the relationship between hierarchical temporal structure and IDS, Falk and Kello (2017) demonstrated that mothers using IDS have greater hierarchical

temporal structure than speech directed toward adults, and in turn steeper slopes when plotted. Abney, Warlaumont, Oller, Wallot, and Kello (2017) also found that greater hierarchical temporal structure for utterance onsets (visualized as steeper slopes) corresponded to non-speech-related vocalizations, meaning sounds like crying and coughing, as compared to speech-related vocalizations. These findings together indicate that speech-related vocalizations and adult-directed speech are reflected through flatter slopes, or less hierarchical temporal structure.

I first aim to use this dissertation to have a better understanding of the relationship between language development and hierarchical temporal structure. If hierarchical temporal structure is statistically predictive of language development, it may potentially pave the way for a new and useful methodology for developmental researchers, although studying developmental delays is outside the scope of the present work. To achieve this goal, I determine if (1) adult prosody is informative for infant lexical abilities, and/or if how lexically advanced infants are is linguistically reflected in the adult's speech, (2) lexical ability is reflected through the features of prosody analyzed, where these features could be useful for testing for developmental delays in future studies, and (3) less advanced vocabularies are reflected through greater hierarchical temporal structure.

1.1.2. Performance-Based Interactions

Thus far, I have established that prosody is an important feature of communication that helps allow for the expression of emotions and semantic meanings, or more demonstrative and performative communication. Although prior studies have analyzed relationships between adult prosody and infant language acquisition (Kuhl et al., 1997), I am aware of no studies that have explored how prosody correlates with joint task performance. Since the sex pairing and time course of experiments may generally allow for a variety of changes as participants become familiar with one another and the task (Harrison, Mohammed, McGrath, Florey, & Vanderstoep, 2003), it is unclear how a relationship between prosody and performance may vary, if at all, by these variables.

The next goal of this dissertation is to contribute a novel experiment that provides an initial understanding of how performance is reflected through prosody, particularly when mediated by the effects of time course and sex pairing. More specifically, since effectively conveying information is necessary for communication-based tasks, I was interested in exploring if two measures of prosody, hierarchical temporal structure and speaking rate, are reflective of how well or poorly pairs of adult speakers score on a collaborative difference-finding task (the Montclair Map Task; Pardo et al., 2019), and if the relationship was mediated by either the effects of epoch (beginning, middle, or end of the experiment) or sex pairing (female-female, male-male, or male-female).

1.1.3. Bilingual Interactions

While the literature so far has concentrated on different forms of "performance" for infants and adults, how prosody varies based on the language spoken has not yet been addressed, although multilingualism is common globally (Grosjean, 2010; Romaine, 2012). A particularly interesting phenomenon of multilingual speakers is that

conversations sometimes occur with each speaker using a different language they are more comfortable with, but that both understand (ten Thije, 2013). It is currently unclear how prosodic structure might vary, if at all, during such circumstances. As described in the following sections, hierarchical temporal structure is measured independently of specific words spoken, allowing for the languages to be compared. One study has tested for hierarchical temporal structure across languages during professional TEDx talks, and found it was exhibited across the six languages analyzed (Kello et al., 2017), confirming prosodic structure is prevalent regardless of the language spoken. However, this speech was outside the context of a naturalistic conversation and situated in monolingual scenarios.

Another goal of this dissertation is therefore to expand on the existing literature by exploring if prosody varies depending on the language spoken in Spanish-English bilingual conversations. To test this question, bilingual speakers engaged in three language conditions—only English, only Spanish, or an inter-language condition where each speaker used only one of the two languages—while discussing three conversational topics: music, movies, or television.

1.1.4. Hierarchical Temporal Structure and Allan Factor Analysis

A comprehensive description of hierarchical temporal structure and Allan Factor (AF) analysis are discussed in the current and following sections. I next discuss the literature surrounding the convergence of speakers' hierarchical temporal structures, or complexity matching. I then describe my goals for studying prosodic convergence in the contexts of infant language development, task performance, and bilingual conversations, analogously to prosody in Sections 1.1.1 through 1.1.3.

I define hierarchical temporal structure as a measure of prosody because prior studies have shown it to be related to duration and intensity (i.e., speaking rate and prosodic exaggeration), two of at least three crucial features for prosody (the third being fundamental frequency, which still needs to be analyzed by future studies in relation to hierarchical temporal structure). That is, faster speaking rates, a direct measure of prosody, corresponded to less hierarchical temporal structure, and slower speaking rates to greater structure (Ramirez-Aristizabal et al., 2018). Less hierarchical temporal structure corresponds to flatter Allan Factor (AF) functions, or slopes (see Section 1.1.5.). Providing further evidence that hierarchical temporal structure is a feature of prosody, it has shown to be reflective of prosodic exaggeration in adult-directed versus infant-directed speech, where flatter AF slopes corresponded to adult-directed speech, and steeper slopes to infant-directed speech (Falk & Kello, 2017). More monotone speech used during formal presentations also corresponds to flatter AF functions, compared to conversational dialogues (Kello et al., 2017).

Across all the following experiments, hierarchical temporal structure in speech was examined, which can be broadly understood as the amount of variability in nested clustering in speech at different timescales of analysis. Analyzing different timescales is relevant to prosody because prosody may theoretically be measured at different timescales of interest, given that the duration, intensity, and pitch of different sounds may vary within units across smaller to larger timescales (e.g., the features of these syllables

can differ, as could the features of words). Thus, neither prosody nor hierarchical temporal structure are limited to a single level of analysis.

Hierarchical temporal structure was analyzed using AF analysis, a method for generally quantifying the variability of peak events. These binary peaks were derived from the original acoustic signal, where a peak occurred only if a sound occurred. This method is beneficial because most traditional measures analyze one or two levels of speech, such as syntax, whereas measuring hierarchical temporal structure captures multiple levels simultaneously, providing us with a fuller picture of the speech signal.

AF analysis was introduced to the cognitive sciences by Lowen and Teich (1996), who measured hierarchical nested clustering (or temporal structure) in neuronal spikes. This methodology was later adopted by Abney et al. (2014) to measure hierarchical temporal structure in acoustic speech energy. In a study based on infant-directed and adult-directed speech, Falk and Kello (2017) determined that acoustic durations of linguistic units across multiple levels of analysis—including syllables, words, and phrases—were correlated with the degree of nested clustering of acoustic energy. Furthermore, they found that greater prosodic exaggeration used in infant-directed speech is displayed as greater nested clustering of acoustic speech energy, which can be visualized as a steeper AF slope when plotted.

Variability in the nested clustering of peak events is measured in the amplitude and timing of the events. No sounds correspond to zero events, whereas the occurrence of a short sound corresponds to one event. Therefore, spoken sounds require many events, although the exact number depends upon the timescale of interest. Though this relationship still needs to be explored in future studies since AF analysis is not designed to quantify pitch, one could imagine glottal pulses, or the rate of opening and closing the vocal folds (Smith & Patterson, 2005), correlating with peak events. In other words, a glottal pulse could correspond to a peak event. Since the frequency of glottal pulses determines pitch (Harris & Nelson, 1993), where faster pulses are associated with higher pitches, the number of peak events within the windows of time analyzed may reflect pitch in addition to amplitude and timing.

Peaks at the shortest timescale will typically contain very little clustering or variability, and peaks at the longest timescale contain a lot (for example, see the plot within Figure 1 where the lowest timescales on the left contain the least amount of clustering). This is because the sounds at the shortest timescales of speech, like phonemes, cannot be meaningfully broken down further, whereas longer units of speech like phrases can be divided into sentences, words, and syllables. When hierarchical temporal structure is plotted, minimal variability in values across timescales corresponds to flatter AF slopes and more monotonous speech, such as someone giving a TEDx talk versus engaging in a conversation (Kello et al., 2017), or when using adult-directed versus infant-directed speech (Falk & Kello, 2017).

1.1.5. Quantifying Hierarchical Temporal Structure Using Allan Factor Analysis

To quantify hierarchical temporal structure, the Hilbert transform—a method for calculating the amplitude envelope—for each audio file was first downsampled to 11 kHz to reduce the quantity of data, and then passed through two thresholds to ensure that the

files were sufficiently sparse. The first threshold allowed only one peak to occur within ±5 milliseconds of other peaks, serving as a type of low-pass filter, where this set the second threshold to have a maximum of 200 peaks per second. The second threshold required one peak for every 200 samples on average. The amplitude threshold derived from the second threshold was automatically set relative to each recording. Again, any amplitudes below the threshold were set to zero, helping to eliminate background noise (see the teal line in Figure 1). The thresholds were set to create sparse enough peak events that allowed for long occurrences of both peaks and silence per recording, thereby allowing for clusters of peak events across timescales. These threshold rules were consistent across recordings so that they were not individually altered. Further, both of these threshold rules have been used in prior literature surrounding AF analysis in speech (Kello et al., 2017; Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020) and in the present studies, meaning that the thresholds have been held consistent.

As displayed in Figure 1, each series of peak events was partitioned into non-overlapping windows of different sizes, where the window sizes corresponded to each given timescale (T), and each timescale contained half as many windows as the next shortest timescale, so that if the largest timescale contained four windows, the next shortest timescale would contain eight. AF variance was quantified at 11 timescales ranging from milliseconds to seconds, with those values depending slightly based on the file lengths, and the number of peak events (N) were counted within each window for each timescale. AF variance at each timescale T was computed by taking the sum of the squared differences between neighboring window counts (N and N_{t+1}) and dividing by twice the mean. More clustering of peak events resulted in larger squared differences. Therefore, AF variance measured the average degree of clustering (or structure) in peak events at each timescale (or temporal hierarchy), meaning AF variance was a function of timescale T. If events are clustered across timescales, then $A(T) > 1$ and increases with each larger timescale. If events are random or evenly distributed, then $A(T) \approx 1$ across timescales (AF analysis does not distinguish between random and periodic events).

The degree of clustering or variance across timescales was quantified by fitting a regression line to each AF function in log-log coordinates. Across the first two chapters, the AF regression lines were fit only to the five longest timescales, and in the third this was also true, but the regression line was also fit to the six shortest timescales because we tested for differences in matching between the lines. We only used the longest timescales in the other two chapters so that the slopes of the regression lines computed from speech signals roughly corresponded to larger units of speech, like phrases and words, and therefore contained more reliable variability to be analyzed. Having less speech in the signal can notably produce a steeper regression line, or AF slope, because less available peak clustering amplifies the variability detected within infrequent clusters of peaks. However, if the quantity of speech is held relatively constant, then we can argue that steeper AF slopes correspond to more emphasis and prosodic variability such as that found in IDS (Falk & Kello, 2017), and conversely that shallower slopes correspond to more monotone speech (Kello et al., 2017).

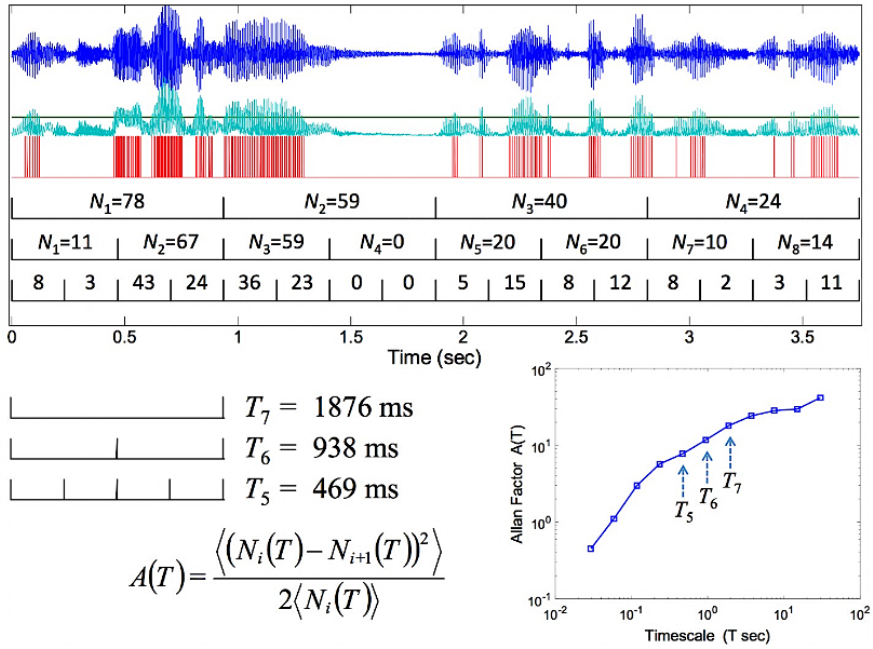

Figure 1. An illustration of the acoustic event analysis derived from Kello et al. (2017). A sample 3.5 second waveform is presented in blue at the top, followed by the teal Hilbert envelope. The envelope provides the peak threshold (the green line), where only amplitudes falling above this threshold may be converted into red peak events. The number (*N*) of peak events are shown inside each window (*T*) for three different timescales (2*T*), timescales 5, 6, and 7, out of 11 timescales. The AF equation and plot of the output are also shown at the bottom.

1.2. Convergence and Communication

In monolingual adult interlocutors, it is common for interacting features of speech to become aligned, ranging from small units of speech like phonemes (Pardo, 2006) to larger units like prosodic exaggeration (Abney et al., 2014). The theory of interactive alignment argues that convergence of such linguistic representations must converge in order for communication to be successful (Garrod & Pickering, 2004; Pickering & Garrod, 2004a, 2006). For instance, if speakers unwittingly discuss different concepts while under the impression that they are discussing the same concept, information exchange will be awkward and generally unsuccessful until the discrepancy in understanding has been resolved.

1.2.1. Convergence and Early Interactions

Some studies have begun to test if the theory of interactive alignment extends to infant-adult interactions, and have primarily found that adults converge features of their speech towards their infant's, but less so vice versa (Ko, Seidl, Cristia, Reimchen, & Soderstrom, 2016). However, infants are not entirely passive in these interactions either.

The age at which convergence between speakers begins remains a relatively open question, so aiming to answer this was one of my first goals surrounding convergence, where infants aged between 3 and 18 months were analyzed.

My first goal of studying convergence is to better understand the process of interactive alignment by analyzing convergence in infant-adult interactions. Because prosody is important for expressing emotional states and semantic meanings, and infant-directed prosody is broadly related to infant language acquisition, and convergence is required for successful communication, my next goal was to test whether prosodic convergence (measured as the hierarchical temporal structure between speakers, also known as complexity matching) between infants and adults was either potentially reflective of, or caused by, the infant's lexical ability. In other words, I wanted to determine if more lexically advanced infants displayed greater convergence, as would be more similar to convergence between adult pairs, and if so, at what age this effect began to occur.

1.2.2. Convergence and Performance-Based Interactions

Greater convergence tends to correspond to stronger feelings of rapport (Valdesolo, Ouyang, & DeSteno, 2010), meaning speakers who sound more alike in various ways tend to feel more amiable towards each other. This has led some researchers to study the effects of different forms of convergence on different types of task performance, but these findings have been mixed, with greater convergence sometimes leading to better performance, sometimes worse, or no relationship at all (e.g., Manson, Bryant, Gervais, & Kline, 2013; Reitter & Moore, 2014). To my knowledge, only one study by Fusaroli, Abney, Bahrami, Kello, and Tylén (2013) has documented that complexity matching increases over time. The authors found their measure of performance ("collective benefit," measured as the ratio of dyad's slope versus the slope of the more sensitive member in the pair) did not correlate with matching until the second session, meaning that the relationship between convergence and performance did not occur immediately.

The next goal was to use this dissertation to help determine if convergence was related to performance as a function of the prosodic measure of convergence, time course, or sex pairing. More specifically, since prosody and alignment both broadly correlate with successful interactions, and convergence correlates with rapport (with both of these features being important for task performance generally), I test if complexity matching is broadly related to performance on a difference-finding task, the Montclair Map Task. I aim to help broadly resolve discrepancies between studies by measuring two forms of prosodic convergence—complexity matching and speaking rate matching—to test whether the effects are similar or diverge, and how any potential effects may be moderated by other variables, including sex pairing and time course of the experiment. Extending from the research by Fusaroli et al. (2013), a final goal is to test if the temporal analyses replicate, where convergence increased over time, during our goal-oriented task.

1.2.3. Convergence and Bilingual Interactions

Most studies of alignment and convergence are between monolingual adult speakers despite the global prevalence of bilingual speakers. In particular, few studies have analyzed alignment across bilingual speakers during inter-language situations, where each speaker uses a different language that both understand (e.g., Ni Eochaidh, 2010). This is largely because it is more difficult to assess convergence when speakers use predominantly separate lexicons. However, understanding convergence in bilingual conversations is crucial for having a complete understanding of how the features of speech interact when the direct matching of words is largely unavailable. Since multiple levels of linguistic representation are necessarily not aligned in bilingual conversations, it seems that the theory of interactive alignment would predict for there to be less matching in inter-language conditions, as compared to intra-language conditions, where this would be possible. The final primary goal of this dissertation will therefore be to explore convergence during intra- and inter-language conversations.

1.2.4. Complexity Matching of Hierarchical Temporal Structures

For each of the following studies, complexity matching was measured as the strength of the correlation between AF slopes, derived from the quantified hierarchical temporal structures (Abney et al., 2014, 2017; Kello et al., 2017; Ramirez-Aristizabal et al., 2018). The concept of complexity matching stems from theoretical work on complex networks and their interactions. West, Geneston, and Grigolini (2008) theorized that when complex networks interact by sending events back and forth to each other, the exchange of information is maximized when the networks share a common power law dynamic. Exchange of information was defined in terms of the mutual impact on each other's peak event dynamics, and the power law was defined in terms of the time intervals between events, which partly describes their patterning in time.

West et al.'s analysis of complex networks was purely theoretical and abstracted away from human behavior, but other researchers applied their ideas to empirical studies of conversations and interpersonal coordination (e.g., Abney et al., 2014; Marmelat & Delignières, 2012; Schmidt & Richardson, 2008). The theory is that human behavior should conform to principles of complex networks to the extent that these principles hold across scales, and information exchange can be measured in the coupling of coordination dynamics, as a reflection of underlying dynamics in complex networks. Unlike measurements of one-to-one matching, with complexity matching parameters of the ensemble statistics of behavior are predicted to converge, particularly when those statistics are power law distributed. The theory of complexity matching can be applied by treating interacting people as interacting complex networks.

Abney et al. (2014) first proposed complexity matching as a theoretically motivated measure of coordination dynamics in speech that generalizes over phase and other temporal relations in speech dynamics. Abney et al. converted speech waveforms into temporal series of peak events that capture hierarchical temporal structure in the nested clustering of acoustic speech energy. The timescale of small clusters roughly corresponds with phonemes, larger timescales with syllables and words, even larger

timescales with phrases and sentences, up to the largest timescales that cover prosodic exaggeration and conversational turns. Clustering across timescales reflects the hierarchical nesting of phonemes within syllables, syllables within words, words within phrases, and so forth (for direct evidence of this, see Falk and Kello, 2017). Hierarchical temporal structure as displayed through AF slopes are indirect measures of power law dynamics to the extent that AF functions follow straight lines in log-log coordinates, and complexity matching can be measured in terms of conversational convergence in the exponent estimates for each speaker in a conversation.

Abney et al. (2014) found evidence for complexity matching in conversations about friendly topics that allowed speakers to find common ground, but there was no evidence of matching in argumentative conversations about topics of disagreement. This study provided some of the first insights into complexity matching in speech, based on AF analysis. The methodology of complexity matching has been replicated in follow-up studies including Abney et al. (2017) and Schneider et al. (2019).

Notably, we could extend the concept of prosody out from the individual's speech to the dyad's collective speech (Cummins, 2018), even when speakers don't use identical words. Conversational turns may then be considered as part of the joint prosodic timing and rhythm between speakers, and would thus be influential for complexity matching. For example, in asymmetric conversations where one speaker dominates the interaction and there are few turns, complexity matching and other forms of convergence may not be detected. This issue of asymmetry motivated us to use the Montclair Map Task (Pardo et al., 2019) rather than the HCRC Map Task Corpus (Anderson et al., 1991) because the latter task had a leader-follower dynamic in which the leader primarily spoke across interactions, whereas the Montclair Map Task allowed both participants to engage more evenly. Future studies should continue to explore the relationship between matching and turn-taking, but turn-taking seems to be an influential mediator of matching.

Correlations are useful for determining the strength of matching because of their built-in baseline of zero. In other words, a chance relationship of matching corresponds to a correlation of zero, removing the need for surrogate analyses. The one disadvantage of using correlations is that different measures of performance sometimes must be quantified during a specific period of time (here being either by month as infants age, or trial during a difference-finding task) rather than over time because it is not possible to quantify a correlational value per trial. We therefore computed matching per month or trial by taking the absolute value of the differences between each speaker's AF slopes, where values nearing zero corresponded to the strongest possible matching and those toward one the least possible matching. We were then able to correlate these absolute difference measures of matching with their respective scores and test whether matching varied by trial or month. Other complementary forms of matching were analyzed in each experiment, but those are described within the separate methodology sections.

1.3. Summary of Dissertation

Prosody and prosodic convergence are together reflective of speaker's emotional states, semantic meanings, feelings of rapport, and successful communication. However, these measures have gone relatively untapped in the performative literature, despite their

potential to be highly informative for our understanding of effective interactions. Most of the literature surrounding convergence has also been focused on monolingual speech despite the global prevalence of bilingual conversations. Therefore, in this dissertation I analyze two measures of speech in three diverse contexts. These are hierarchical temporal structure (a measure of prosody), and the convergence between hierarchical temporal structures, known as complexity matching (a measure of prosodic convergence). The three contexts are infant language acquisition, task performance, and bilingualism. Each study also pairs hierarchical temporal structure and complexity matching with other forms of speech or convergence.

In Chapter 2, infant lexical ability is correlated with prosody and prosodic convergence, as well as volubility and volubility convergence. These additional measures of talkativeness were relevant because many prior studies test for differences in volubility when assessing infants or children with and without developmental disabilities (Rowe, 2012). Furthermore, volubility provided a more traditional measure whereas hierarchical temporal structure is more novel. I test if the effects of these measures are mediated by infant age, where I expect convergence to become stronger, infants to become more talkative, and hierarchical temporal structures to become flatter as infants advanced in age and linguistic ability.

Chapter 3 extends from the performative analyses between infant-adult pairs to adult-adult pairs. Here, I test whether pairs' average score on a difference-finding task is reflected in two measures of prosody, hierarchical temporal structure and speaking rate, as well as the strength of convergence for each measure. Speaking rate is of interest since how fast one speaks can influence how well one performs (Haake, Hansson, Gulz, Schötz, & Sahlén, 2014). Intuitively, if a speaker is conveying information as quickly as possible, some information may be not registered properly. Conversely, slow and monotonous speech may cause a listener to discontinue paying attention, thereby also not allowing information to be effectively processed. Since prior experiments surrounding different forms of convergence and task performance have reported inconsistent findings within and across studies, analyzing complementary measures of prosody may help elucidate how and why such discrepancies occur.

In Chapter 4, I move away from analyses of performance to study how the theory of interactive alignment extends to monolingual and bilingual conversations. I test if either complexity matching (a measure of physical acoustic convergence) or lexical matching (a measure of non-physical representations of the meanings of words) is affected by the use of two languages (Spanish and English) during two language conditions (inter- and intra-language conversations). In other words, pairs were asked to have three naturalistic conversations about simple topics when speaking English only, Spanish only, or a Mixed condition where one partner spoke Spanish while the other spoke English. These novel methods allow me to study convergence while working around the unavoidable issue of the lexicons not matching due to differences in lexicons.

Following this, in Chapter 5 I first summarize the primary findings from Chapters 2 – 4 and interpret what these findings mean for each other, and the broader literature surrounding prosody and prosodic convergence in the contexts analyzed. Several future directions based on the limitations and interpretation of the studies are then proposed, followed by a conclusion of the full dissertation.

Chapter 2

Hierarchical Temporal Structure and Language Acquisition

2.1. Preface

In this chapter, I investigate the relationships between vocalization and vocalization convergence on infant language learning to determine whether infant's lexical development at 18 months is reflected through infant-adult interactions at 3, 6, 9, and 18 months. Using naturalistic at-home recordings, vocalizations produced by adults and infants in the environment were encoded to test how two features of vocalizations (hierarchical temporal structure and volubility), and the strength of convergence between each, correlates with infant's lexical abilities (measured as the number of words the caregivers reported the infant could produce at 18 months). Against my initial expectations, only the adult's hierarchical temporal structure correlated with lexical ability and no convergence was found between infant-adult vocalizers (or "speakers"). The lack of convergence across ages may have been due to the highly naturalistic environment, where infants and adults may or may not have been interacting with one another although vocalizing nearby, or our sample of infants may not yet be converging with adults on these measures. These findings overall suggest that hierarchical temporal structure may be a better moderator of lexical ability than convergence, but additional research is needed to better confirm this hypothesis. This research is not under review and has not been published in any journals.

2.2. Introduction

It has been well established that caregivers use infant-directed speech (IDS) when addressing infants, with IDS entailing a typically slower paced (Song, Demuth, & Morgan, 2010), higher pitched (Golinkoff, Can, Soderstrom, & Hirsh-Pasek, 2015), and more prosodically variable (Trainor, Austin, & Desjardins, 2000) form of speech, as compared to adult-directed speech (ADS). IDS often relates to better language acquisition under the theory that it promotes distinctions between different vowels and syllables (Kuhl et al., 1997; Leong et al., 2014; Peterson & Barney, 1952; Thiessen et al., 2005). Some literature has also focused on the infant's own early vocalizations because they are theorized to relate to later language ability and developmental issues. Much of this work has focused on consonant use (Fasolo, Majorano, & D'Odorico, 2008; Stoel-Gammon, 1991; Watt, Wetherby, & Shumway, 2006) and some on exploring volubility, or the rate of vocalizing per some unit of time (Iyer, Denson, Lazar, & Oller, 2016). However, few studies have tested if a relationship exists using acoustic features of infant vocalizations.

We therefore first analyze correlations between two measures of infant and adult speech and the infant's vocabulary at a later age to assess if either speaker's vocalizations statistically predict the infant's measurable vocabulary size. These features are *vocalization durations* and *hierarchical temporal structure*, with the former being a more traditional measure of volubility and the latter a newer measure of speech acoustics. If either feature is related to vocabulary for adults, this might suggest adults could alter their

speech to better aid the infant, i.e. by generally speaking more or using IDS more often. Conversely, rather than the adult's speech influencing the infant's, it is also possible that adults alter their speech in response to their infant's abilities, where they might for example verbalize more or use more advanced speech when addressing infants with stronger lexical abilities. For infants, a correlation could suggest that these may be meaningful measures for developmental studies, including those which test for verbal delays, although developmental delays are outside the scope of this study.

In addition to the individual patterns of speech, several studies surrounding adult pairs have explored what is sometimes known as convergence, synchrony, or alignment between various features of speech (and bodily movements). The widely accepted theory of interactive alignment argues that for communication to successfully occur between speakers, their linguistic representations must be shared across multiple interacting levels (Garrod & Pickering, 2004; Pickering & Garrod, 2004a, 2004b). Examples of alignment have been demonstrated such as phonetic convergence, where a pair's phonemes become more alike while interacting as compared to before their conversation, and to a lesser extent following their interaction (Kim, Horton, & Bradlow, 2011b; Pardo, 2006). Another example is language style matching, in which pairs converge on the same categories of words when digitally messaging one another (Yilmaz, 2015). Some studies have tested for convergence in infant-adult pairs and found that adults converge to the infants more than vice versa, but infants are still active in these interactions (Abney et al., 2017; Ko et al., 2016).

More convergence between adult pairs also broadly relates to stronger feelings of rapport toward one another, meaning that pairs who either speak or move more similarly rate their partner more positively than those who do not converge (Manson et al., 2013; Valdesolo et al., 2010). This phenomenon also extends to infant-adult pairs in their movements. In Cirelli, Einarson, and Trainor (2014), infants were held by an assistant and bounced in synchrony with another experimenter to a musical beat. Those bounced in synchrony behaved more altruistically than those bounced asynchronously towards the experimenter when they later pretended to accidentally drop an object and need help retrieving it. Analogous results were found in Cirelli, Wan, and Trainor (2014), but the altruistic task changed to the infants being given several toys and being asked by the experimenter if they could have one.

Since convergence is meaningful for successful interactions and is related to stronger feelings of rapport, some researchers have explored its relationship with autism spectrum disorder and turn-taking dynamics in infant-adult pairs. Warlaumont, Richards, Gilkerson, and Oller (2014) analyzed the timing of infant and caregiver vocalizations, and reported that infants with autism tend to produce less vocalizations overall, less speech-related vocalizations ("speech-related" meaning non-cry, non-vegetative, and non-laugh vocalizations), and lead interactions less than typically-developing infants (Warlaumont et al., 2014). Meanwhile, adults are more likely to produce responses contingent upon speech-related than non-speech related vocalizations, together creating a weaker social feedback loop. The typical expectation is therefore that more convergence would relate to better language ability, but the findings for both adults and infants have been mixed, suggesting that the relations between interactive alignment and language acquisition (and other forms of task performance in adult pairs) are complex and context

dependent. In another study, Jaffe, Beebe, Feldstein, Crown, and Jasnow (2001) found greater vocal rhythm coordination between infants and adults at 4 months was associated with higher cognitive scores. Further, an analysis of day-long recordings for infants who were hard of hearing reported that the number of conversational turns between infants and adults at two years was positively correlated with stronger language abilities when the infant was two and three years old (Ambrose, VanDam, & Moeller, 2014). Expanding on such research, the present study thus secondly aims to explore if convergence in either the vocalization durations or hierarchical temporal structures are found between infant and adult vocalizations, and if so, if convergence at earlier ages relates to the infant's vocabulary later on.

In sum, the goals of the present study are to (1) analyze the relationships between infant's and adult's volubility and hierarchical temporal structures on infant vocabulary, (2) test for convergence between infant's and adult's levels of volubility and also their hierarchical temporal structures, and (3) analyze the relationships between both forms of convergence and vocabulary. The following sections will first describe the literature surrounding adult vocalizations and infant language acquisition, followed by an analogous section on infant vocalizations and language acquisition. We then discuss vocalization convergence in adults and infants, and its relationship with language acquisition in a proceeding section. Our experimental methodology follows, and we finish with a discussion of our findings and how they impact the current theories surrounding infant language acquisition, convergence, and "performance" more broadly.

2.2.1. Adult Vocalizations and Infant Language Acquisition

It is intuitive that a greater quantity of adult speech—or greater volubility—relates to better infant and child language abilities, such as having stronger vocabularies (Rowe, 2012; Sultana, Wong, & Purdy, 2020). On the most extreme spectrum, we remember from cases of neglect like Genie how detrimental extreme isolation and silence can be for developmental growth (Curtiss, 1977). However, simply listening to a greater quantity of speech alone is not enough to acquire language, which is why an infant cannot be left in front of a radio or television and be expected to learn how to speak (Golinkoff et al., 2015). In one elucidating study, Rowe (2012) found that increased volubility was most important for language learning at 18 months, but by 30 months, the diversity of the parent's vocabulary became more important than volubility alone. This demonstrates that both the quantity and quality of the caregiver's speech matter for language development.

Another notable quality of the caregiver's speech is IDS. While some debate exists regarding the functionality or purpose of IDS and the mechanisms involved (e.g., McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013), many researchers have found that this style of speech is useful for aiding infants in language acquisition and comprehension (Golinkoff et al., 2015; Graf Estes & Hurley, 2013). This is at least in part because the prosodic emphasis of IDS differs from ADS, with IDS having more hierarchical temporal structure or effectively prosodic stress, even when reading the same stories and singing the same songs (Falk & Kello, 2017). This increased emphasis on vowels and syllables has been theorized to help infants distinguish between speech

sounds, allowing parents to actively aid in or help direct language learning, whether intentionally or not (Thiessen et al., 2005).

From the comprehension perspective, Kuhl et al. (1997) found that when using IDS, mothers across the United States, Sweden, and Russia stressed their vowels, simplifying the auditory process of creating vowel categories or distinctions between them (such as the differences between "heed," "hid," and "head"; Peterson & Barney, 1952). Another study found that changes in ADS were syllable-dominant modulations, whereas IDS modulations were stress-dominant, meaning that more prosodic stress is used when producing syllables in IDS (Leong et al., 2014). In other words, adults can hear a monotone voice in ADS and still distinguish between sounds, whereas this would be a considerably more challenging task for an infant. These findings suggest that while other reasons for IDS use likely also exist and are sometimes concurrent—such as expressing IDS to display one's emotional state (Trainor et al., 2000) or gaining the infant's attention (Zangl & Mills, 2007)—caregivers aid in their infant's language development. However, whether or not IDS or ADS is used may notably be reflective of the infant's own vocalizations and abilities. Since adults eventually transition to using ADS instead of IDS, meaning that ADS is used towards more advanced speakers, ADS use towards infants seems more probable for those with stronger vocabularies.

Some researchers have demonstrated the benefits of IDS for infant vocalization production via computational models. In one such study, de Boer and Kuhl (2003) used recordings of mothers speaking to either their infants using IDS or ADS to test which type of speech is more easily learnable. The model's goal was to learn the positions of three vowel categories (/i/, /a/, and /u/) from a set of corresponding target words ("sheep," "sock," and "shoe"), spoken in either IDS or ADS. A learning algorithm called "expectation maximization of a mixture of Gaussians," was used to implement statistical pattern recognition to learn to classify the different datasets, or establish the distribution of each vowel category in acoustic space. The authors acknowledged that the algorithm was less powerful than real human learning capabilities, but despite this disadvantage, the vowel positions were still learned more accurately when the model was fed the target words in IDS compared to ADS.

In a complementary study, a computational model learned how to produce basic words from three languages based only on caregiver responses. Howard and Messum (2014) had English, German, and French speakers act as "caregivers" to an unsupervised learning model, which acted as an infant learner (named Elija) who could produce movements that corresponded to sounds. Elija first learned to produce sounds based on his motor outputs alone and was then allowed to interact with a caregiver. The caregivers were instructed to close their eyes, imagine they were speaking to a human infant, and only respond to Elija when it felt natural to do so, which Elija detected. Although not explicitly instructed to do so, the caregivers typically responded to more speech-like sounds and would reformulate the utterances into better formed versions of words in their first language. Elija then retained any motor patterns that produced sounds the caregivers responded to and discarded the ignored sounds. After several repetitions of these interactions, the caregivers were allowed to try to teach Elija a few words of their choosing. By detecting the sounds produced by the caregiver and mapping it onto previous outputs, and perhaps going through some repetitions of this final process, Elija

was able to say at least a few simple words in each language. This study was meaningful because the caregiver's responses to Elija's sounds and motor outputs allowed the model to begin learning three distinct languages, further suggesting that adult inputs are important for infant language production.

In sum, caregiver responses are beneficial for infant language comprehension and production. Although Falk and Kello (2017) determined differences exist in the hierarchical temporal structures of IDS and ADS, they did not explore the impacts of hierarchical temporal structure on vocabulary or other measures of language ability. As mentioned, it is also possible that the infant's vocabulary may conversely affect the adult's prosodic structure, meaning that more advanced infants may be addressed using ADS instead of IDS. The present study therefore collectively expands upon these theories by exploring the association between hierarchical temporal structure and lexical acquisition. As a complementary measure, adult and infant volubility are also analyzed in relation to the infant's vocabulary strength. If either volubility or hierarchical temporal structure are related, this may indicate that caregivers could either try to speak more often or with either more or less prosodic exaggeration to better assist their infant. (Changes as small as using different toys to interact with can increase adult volubility; Sosa, 2016).

2.2.2. Infant Vocalizations and Infant Language Acquisition

Testing for a relationship between infant vocalizations and later language capabilities is of interest to developmental researchers in large part because vocalizations provide a noninvasive way to predict developmental delays. This a particularly promising area of research because a relationship between vocalizations and abilities has been broadly shown to exist (Marchman & Fernald, 2008; Roe, 1975; Stoel-Gammon, 1991). While much of this research has focused on developmental delays, it is also interesting to analyze these relationships in typically developing children to better understand what early acoustic skills are reflected in more successful communication later on. Although the literature is slightly mixed, several studies have found that more infant volubility (Camp, Burgess, Morgan, & Zerbe, 1987; Patten et al., 2014) and speech-related vocalizations (Overby, Belardi, & Schreiber, 2020) are related to better vocalization quality or ability. Additionally, Fernald, Perfors, and Marchman (2006) found that faster and more accurate performance by 25-month-old infants on a spoken word recognition task was related to better grammatical and lexical development from 12 to 25 months. Language development was reported using the MacArthur-Bates Communicative Development Inventory, a standard measure of assessment also utilized in the present study to measure vocabulary.

In contrast to typically-developing infants, those born with autism spectrum disorder (Patten et al., 2014; Warlaumont, Richards, Gilkerson, & Oller, 2014) or later diagnosed with apraxia of speech (Overby et al., 2020) tend to be less voluble and produce less speech-related vocalizations. However, infants born prematurely tend to be as voluble as typically-developing infants, though they produce speech-related vocalizations less consistently (Oller, Eilers, & Steffens, 1994). As described earlier, a repercussion of such speech-related issues is that parents are less inclined to respond to non-speech-related vocalizations, or those with a poorer quality, creating fewer

opportunities for such infants to engage in feedback loops that are important for language development (Warlaumont et al., 2014). We therefore expect for our measure of infant volubility to be related to vocabulary, where more volubility correlates with stronger vocabularies.

While volubility and vocalization type are more popular measures of vocal development, few studies to our knowledge have analyzed hierarchical temporal structure, with two exceptions being Abney et al. (2017) and Ritwika et al. (2020). In Abney et al. (2017), hundreds of day-long, naturalistic recordings of infants and their environments ranging in age from 11 days to 2 years and 5 months were created. Vocalizations and the silence between them were considered to be binary events, where "interevent intervals" were the temporal durations between events. Diverging slightly from the present study, hierarchical temporal structure was measured using interevent intervals derived from the original audio file, rather than using the original audio file itself. Greater hierarchical temporal structure—or when plotted this may be seen as steeper slopes—corresponded to non-speech-related vocalizations. With the findings from Falk and Kello (2017) also in mind (where IDS corresponded to greater hierarchical temporal structure), it seems possible that less advanced infant vocalizations correspond to more hierarchical temporal structure, or steeper slopes, by both speakers. We therefore expect to find that for both infants and adults, greater volubility and less hierarchical temporal structure will relate to stronger infant vocabularies.

2.2.3. Convergence in Adult and Infant Vocalizations

Across numerous studies, pairs of adult speakers have often converged or become aligned across interacting levels of linguistic representations when communicating (Pickering & Garrod, 2004a). Such alignment has been demonstrated as syntactic convergence, where the pair's use of active or passive phrases (Bock, 1986) or other syntactic structures like tree depth and sentence length (Xu & Reitter, 2016) converge. Another example of convergence has been observed at the phonetic level (Pardo, 2006), where speaker's phonemes were judged by independent listeners as sounding most like their partner's when interacting with one another, as compared to how their phonemes sounded before the task. Following alignment during the conversation, speakers phonetically converged the next most strongly post-task when not interacting, suggesting convergence can linger. Beyond these two forms, lexical (Brennan & Clark, 1996; Fusaroli et al., 2012) and categorical convergence (Manson et al., 2013; Yilmaz, 2015) are other commonly studied examples. Convergence has been theorized to be crucial for successful communicative interactions (Garrod & Pickering, 2004) and is also related to stronger feelings of rapport or positivity towards the other speaker (Valdesolo et al., 2010).

Moving past this more established work, some studies have analyzed infant-adult convergence as it might relate to predicting developmental delays. While the author found no studies specifically proposing so, such experiments could provide a timeline as to when and how interactive alignment begin to occur. It is possible that vocalization convergence begins at a young age since infants have been recorded becoming more coordinated in their behaviors with their parents beginning around 3-6 months (Feldman,

2007), with more primitive behaviors such as sleep cycle and heart rate perhaps becoming synchronized between mother and fetus in the third trimester (Feldman, 2006). Some weaker levels of pitch convergence have been found between parents and infants as early as 3 months, and as infants become older and more linguistically advanced, convergence becomes stronger (McDaniel, 2020). As mentioned earlier, convergence is even important for establishing rapport for infant-adult pairs, in that moving in synchrony together to a musical beat makes infants behave more altruistically, i.e. by picking up an object the experimenter pretended to drop and needed help retrieving it (Cirelli, Einarson, et al., 2014) or offering a toy (Cirelli, Wan, et al., 2014). This suggests that even for young infants, convergence of different forms plays an early and important role in parental interactions.

In addition to this work, caregivers have been recorded altering features of their own vocalizations to become more similar to their infant's. In one naturalistic study, infants ranging from 12-30 months and their mothers were recorded across an average of 3.85 days (Ko et al., 2016). When the mothers responded to their child's vocalizations, positive correlations were found between the child's and mother's pitch. On the other hand, this effect lessened when the child responded to the mother, suggesting that mothers adapt their speech more to their children's vocalizations versus their children adapting their vocalizations to their mother's speech. However, the children did still make some small adaptations to their voices to become more similar to their mother's, suggesting they do play an active role in early interactions.

Abney et al. (2017) also analyzed naturalistic data which analyzed a few forms of vocalization convergence. This included cluster-based coordination, or the amount of coordination between binary temporal events derived from the original speech signals, also known as hierarchical temporal structure. Rate-based coordination was also measured as convergence in volubility between the infants and adults. In line with Ko et al.'s findings, as infant age increased caregivers adjusted their hierarchical temporal structure to match their infant's structure, with this relationship being asymmetric. For rate-based coordination (or volubility convergence), both the infant and adult converged their volubility rates toward the other, thus again demonstrating that infants play an active role in their early interactions with their caretakers.

These results show the importance of infant-caregiver convergence in vocalizations, and while infants are less advanced compared to adult speakers, they are still at least somewhat active in their interactions with caregivers. These findings also reflect how adults adapt features of their speech to converge toward the infants' vocalizations. Like Abney et al. (2017), the present study tested for convergence in hierarchical temporal structure and volubility, although there were differences in how hierarchical temporal structure was quantified. The following section discusses the potential importance of convergence between infants and adults for infant language acquisition.

2.2.4. Vocalization Convergence and Infant Language Acquisition

Considering that convergence is important for a stronger sense of rapport, and the ability to achieve more symmetric convergence seems to be a meaningful milestone for

children, it seems probable that convergence would be influential for different forms of performance, namely task performance in adults and language ability in infants. To better understand the existing literature, we must first discuss adult pairs because more research on convergence and performance has been situated in these speakers performing joint tasks. Perhaps due to the diverse nature of the tasks, task goals, or levels of analysis, the findings have been mixed, with some studies reporting positive, negative, or no correlations. For instance, Fusaroli et al. (2012) found mixed results within the same experiment. Participants separately viewed two sets of Gabor patches and independently select the set which contained a smaller, different patch. If they agreed they progressed to the next trial, but if they disagreed, they spoke to come to a joint decision. Expressions of confidence, such as "to think," or "to be sure," were used to analyze linguistic convergence, or the probability of each participant transitioning to using their partner's expression following their use of it. Joint or collaborative performance was measured as the collective benefit of the dyad, meaning that the dyad performed better together than the more sensitive member of the pair would have alone. On a trial-by-trial basis, linguistic convergence positively correlated with collective benefit, but when the full repertoire was analyzed, linguistic convergence negatively correlated with collective benefit. This suggests that task-relevant vocabularies are related to better joint actions, but other types of linguistic alignment are not, indicating that only some vocabularies may be used to optimize the relationship between convergence and performance.

Analogously to adult pairs, the literature on infant-adult convergence and language development have also presented mixed results. The more common finding overall is that increased convergence or synchrony of different kinds, such as emotional convergence (Yirmiya et al., 2006) and movement synchrony (Fitzpatrick et al., 2016), relate to typically developing infants, or those with stronger language abilities. Quigley, McNally, and Lawson (2016) demonstrated one such example where infants at either high or low risk for autism (based on whether or not the infant's sibling had autism) were recorded interacting with their mothers during 20-minute sessions across 3 and 12 months, and a follow-up visit at 18 months. By 12 months, both pitch and prosodic intensity converged for infant-adult pairs for typically developing infants. However, for infants with autism, only pitch convergence was found during the 18-month visit, suggesting that the process of convergence may take more time to develop. As also mentioned, Jaffe et al. (2001) and Ambrose et al. (2014) respectively found greater vocal rhythm coordination and greater conversational turns at earlier ages were correlated with stronger cognitive and language abilities at later ages.

In contrast to these findings, older children with and without autism have been found to syntactically (Hopkins, Yuill, & Keller, 2016) and lexically (Branigan, Tosi, & Gillespie-Smith, 2016) align their speech with others similarly. Even in infancy, convergence has surprisingly not always been found to be related to language capability. A study by Seidl et al. (2018) recorded infants with and without a high risk for autism over the course of a day, and found pitch and volubility convergence were similarly low for both groups. Although neither measure related to autism spectrum disorder, volubility and turn count alone did correlate, where typically developing infants produced greater numbers of words and turns throughout the day. The reasoning behind these mixed findings is unclear, but the mixed nature of the findings suggests that how convergence

relates to language development is complex and may only sometimes be reflected in alignment between the two speakers.

2.2.5. Current Experiment

In the present study, we explored the relationships between two features of adult and infant vocalizations, the convergence between these features, and the infant's vocabulary at 18 months. These features were volubility, as quantified by the vocalization duration for each speaker, and hierarchical temporal structure, an acoustic measure of the vocalizations. Vocabulary was parentally reported using the MacArthur-Bates Communicative Development Inventory, a standard method for quantifying language ability. We tested for these effects when the infant was 3, 6, 9, and 18 months old, although the vocabulary assessment was at 18 months only, since this is an age at which infants can often say several words. Based on the existing literature, we expected for more volubility and less hierarchical temporal structure to relate to better lexical abilities, while the expected association between convergence and language ability were less clear.

2.3. Methods

2.3.1. Participants

Sound data were collected from 29 infants (17 male, 12 female) at 3, 6, 9, and 18 months of age, as part of a larger study (Pretzer, 2019; Ritwika et al., 2020) in which recordings from 58 participants were collected (11 did not complete the study). Only participants who completed the entire study were analyzed here. Further, only participants whose audio files had been fully encoded by researchers were analyzed (described in Section 2.3.2.4.). If the duration of total audio produced by either the infant or adult alone was less than 20 seconds, the audio file was omitted, leaving 216 of 348 files, and discarding 132 files, or nearly 38% of the data.

Participants were recruited from the San Joaquin Valley of California through word of mouth, community events, pre-birth hospital tours, and flyers hung around local areas, including the university campus, Merced College, and a local hospital. Demographic information was collected as part of the larger study but was not analyzed here. To give a sense of the participant demographics, out of the 29 infants analyzed, four were born in 2015, 12 in 2016, 12 in 2017, and one in 2018. Six of them were exposed to at least some Spanish around the household, and one of these six infants was also exposed to German, and another Portuguese on occasion. Caregivers reported 15 infants being Hispanic/Latino or Hispanic/Latino/white, nine white, four Asian/white, and one Hispanic/Latino/Native American/Black. Caregivers were an average of 32.3(±8.7) years old with a minimum age of 15, maximum of 64, and median of 32. Caregivers were relatively evenly split among socioeconomic status (seven reported a household income of less than $30,000, seven were in between $30,000-60,000, six were between $60,000-90,000, and nine reported over $90,000).

2.3.2. Procedures

2.3.2.1. Recording Procedure

Families were lent Language ENvironment Analysis (LENA) recorders (https://www.lena.org/) and at least two vests that contained pockets to secure the recorders. Research assistants and other experimenters involved on the larger project physically provided these materials to the families, or materials were mailed to the participants. These recorders captured and classified all of the sounds in the environment, including the infant's vocalizations and any adult vocalizations. Caregivers were instructed to have the infant wear the vest containing the recorder from 8 AM or earlier, until 7 PM or later, requiring a minimum of 10 hours of audio data. Caregivers were allowed to pause the recording for privacy reasons as needed for a maximum of an hour but were otherwise encouraged to not pause the recording. We would also remove portions of the audio files if participants forgot to pause the recording and retroactively decided they wanted it removed for privacy reasons. Caregivers were compensated with $20 for the 3-month recording, $30 for the 6-month, $40 for the 9-month, $60 for the 18-month, and an additional $40 during a final visit in which they were given a summary of LENA's breakdown of what vocalizations occurred during each recording session.

2.3.2.2. Vocabulary Assessment: The MacArthur-Bates Communicative Developmental Inventory

During the last recording at 18 months, caregivers were asked to complete three questionnaires on the same day as the recording: the MacArthur-Bates Communicative Developmental Inventory (MCDI), the Ages and Stages Questionnaire (ASQ), and the StimQ questionnaire (the latter questionnaires were also completed during the three prior recording sessions). The MCDI is the only relevant questionnaire for the current study, and is a widespread method of measuring infant vocabularies and language abilities (e.g., Fernald et al., 2006; Lopez, Walle, Pretzer, & Warlaumont, 2020; Makransky, Dale, Havmose, & Bleses, 2016), sometimes used for assessing children with disabilities (Heilmann, Ellis, Evans, & Hollar, 2005; Luyster, Lopez, & Lord, 2007). For this questionnaire, caregivers were instructed to review 22 categories with 680 total available words, such as "animal (real or toy)," which lists specific words, like "dog." If the infant could say the word or something similar that carried the correct semantic meaning, the caregivers were to fill in the corresponding bubble. Researchers then counted the number of bubbles filled in to gauge the infant's vocabulary strength. It is possible caregivers may have either overreported or underreported their child's vocabulary, but this has previously been shown to be a reliable measure of assessment (Skarakis-Doyle, Campbell, & Dempsey, 2009).

2.3.2.3. Audio File Selection

The LENA Pro software first processed the audio recordings and classified several types of sounds in the environment: infant vocalizations, vocalizations by other

children, male adult speech, female adult speech, silence, noise, electronic sounds, and vocalizations overlapping with another sound source. Using the infant vocalization classification, we automatically extracted the number of infant vocalizations contained within five-minute durations across the entire file. We then checked that these segments were a minimum of 30 minutes apart from one another. Researchers also checked that the LENA software did not misinterpret other sounds for the infant—such as another child in the environment—when extracting the number of infant vocalizations, which could lead to few real vocalizations. Finally, we checked that the infant's mouth was not blocked for the duration of the file, such as when feeding. If the files were less than 30 minutes apart, if the software misinterpreted the audio, or if the infant's mouth was obstructed, we chose the next most voluble five-minute file until there were three files per infant.

2.3.2.4. Encoding Procedure

Researchers were trained to encode infant and adult vocalizations in each file, marking the start and stop times of each utterance, along with the type of vocalization produced. The infant's vocalizations were broken down into four categories: canonical, non-canonical, reflexive laughing, and reflexive crying. As described by Buder, Warlaumont, and Oller (2013), canonical vocalizations are those which contain a consonant, vowel, and speech-like timing, whereas non-canonical sounds contain more primitive speech-like vocalizations, which may, for instance, only contain a vowel sound. No vegetative sounds were encoded, which include hiccups, burps, and other bodily sounds.

All adults in the environment were classified as one speaker, or in other words, adult speakers were not differentiated. Teenagers or older children who were acting as caregivers to the infant were also encoded as adults. Adult speech was categorized first by whether or not their utterances were directed towards the infant. Regardless of directionality, all spoken words were transcribed. If an adult was speaking to the infant, the type of speech used was encoded, such as prohibitive speech and imitation. Teenagers or older children who were acting as caregivers to the infant were also encoded as adults. Researchers were unaware of the participant's identification when encoding these files, helping to ensure they were not biased by knowledge of the infant's age at the time of the recording.

For the present study, only the start and stop times of each utterance for both infants and adults were utilized due to the varying sparseness of the audio files. Participants were only selected for analysis here if researchers had fully encoded all 12 of each infant's audio files (three five-minute files deriving from recordings at 3, 6, 9, and 18 months).

2.3.2.5. Data Preprocessing for Allan Factor Analysis

Since the recorders only contained one stream of audio and we were interested in analyzing the infants' and adults' vocalizations separately (to the extent that the vocalizations were not overlapping, where any overlaps were uncontrolled for and a limitation of the present study), we used the start and stop times from each encoded

utterance to create two separate streams of audio for each speaker in MATLAB R2020b (9.9.0.1467703). Any adults in the environment were treated as a single speaker since they were not encoded separately, and any sounds in the audio stream that did not correspond to the designated speaker were muted. The total duration of the infants' and adults' vocalizations were then calculated.

2.3.2.6. Hierarchical Temporal Structure and Allan Factor Analysis

As described in detail within Chapter 1, hierarchical temporal structure in speech was measured using Allan Factor (AF) analysis. Hierarchical temporal structure was quantified as the variability of acoustic events (or degree of nested clustering) across timescales. To remind the reader, clusters of peaks at the shortest timescale contain very little clustering or variance, whereas those at the longest timescale contain a lot. This is because the shortest units of speech provide minimal clusters as they are already at their smallest state, but longer units of speech like phrases can be broken down into sentences, words, and phonemes. Little clustering or variance across timescales corresponds to flatter AF slopes and more monotone speech (Kello et al., 2017), whereas steeper slopes relate to more prosodically variable speech. This methodology is beneficial for use because it uniquely allows speech variability to be quantified across timescales simultaneously, regardless of how linguistically advanced the speaker is. The description for how hierarchical temporal structure is measured using AF analysis is also described in Chapter 1, with the only specific note for the present study being that we used audio durations of exactly 5 minutes, and the 11 timescales of AF variance thus ranged from approximately .01 to 13.13 seconds. The AF regression lines were fit only to the 5 longest timescales, ranging from 0.82 to 13.13 seconds.

2.3.2.7. Complexity Matching Measures

Complexity matching was measured as the strength of the correlation between AF slopes, derived from the quantified hierarchical temporal structure at each timescale (Abney et al., 2014, 2017; Kello et al., 2017; Ramirez-Aristizabal et al., 2018). As a more exploratory but relevant analysis, complexity matching was also measured as the strength of the correlation between the durations of vocalizations, measured in milliseconds. Again, correlations are useful for determining the strength of either form of matching because of the automatic correlational baseline of zero, removing the need for surrogate analyses. The aforementioned disadvantage of using correlations is that they require more than a single score per pair, so they cannot quantify the relationship between matching and MCDI score for each pair at each month. We therefore computed matching per month by taking the absolute value of the differences between each speaker's AF slopes, or each speaker's volubility, where values nearing zero corresponded to the strongest possible matching (Schneider et al., 2020). We then correlated these absolute differences with MCDI scores, and whether this form of matching varied by month.

All statistical analyses and plots were conducted using R (version 3.6.3 (2020-02-29) – "Holding the Windsock"), and packages lme4, lmerTest, and ggplot2 (Bates,

Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2014; Wickham, 2016, respectively).

2.4. Results

2.4.1. Summary of Hypotheses and Analyses

In the following sections, we first tested for differences in infant and adult AF slopes by the infant's age to determine if any changes to hierarchical temporal structure occurred. We next tested if either speaker's AF slopes related to the infant's MCDI score, a measure of verbal ability (M = 64.67 words, SD = 64.36, see Figure 2). Steeper adult AF slopes were initially hypothesized to predict MCDI score, since steeper slopes correspond to more infant-directed speech (Falk & Kello, 2017), which has been proposed across many studies to aid in language acquisition (e.g., Kuhl et al., 1997; Leong, Kalashnikova, Burnham, & Goswami, 2014). However, since adult-directed speech corresponds to flatter slopes, it was also possible that adults may use more adult-directed speech towards more verbally advanced infants, or more verbally advanced infants may simply have caregivers who use more adult-like speech. For infants, we hypothesized that slopes would become flatter over time since this corresponds to more adult-like speech. Similarly to these analyses for AF slope, we next tested for differences in vocalization durations by both speaker types (infant or adult) and infant age. Vocalization durations were of interest because they had the potential to be unbalanced between speakers and reflective of volubility. We also tested if volubility related to MCDI score, and expected that for either speaker at any month, greater volubility or longer vocalization durations would predict higher scores.

Analogously to the above tests, we next investigated if complexity matching in AF slopes or volubility matching in vocalization durations occurred between speakers, and if this varied by infant age. Finally, we tested if either form of matching related to MCDI scores. We expected increased matching in AF slopes and volubility to relate to better MCDI scores, under the theory that vocalizing more similarly could reflect stronger verbal abilities. Infants who speak in more adult-like patterns presumably have stronger vocabularies than those who do not, and caregivers using more variable patterns of speech is suggestive of attending to and perhaps aiding the infant in their language learning (whether consciously or not).

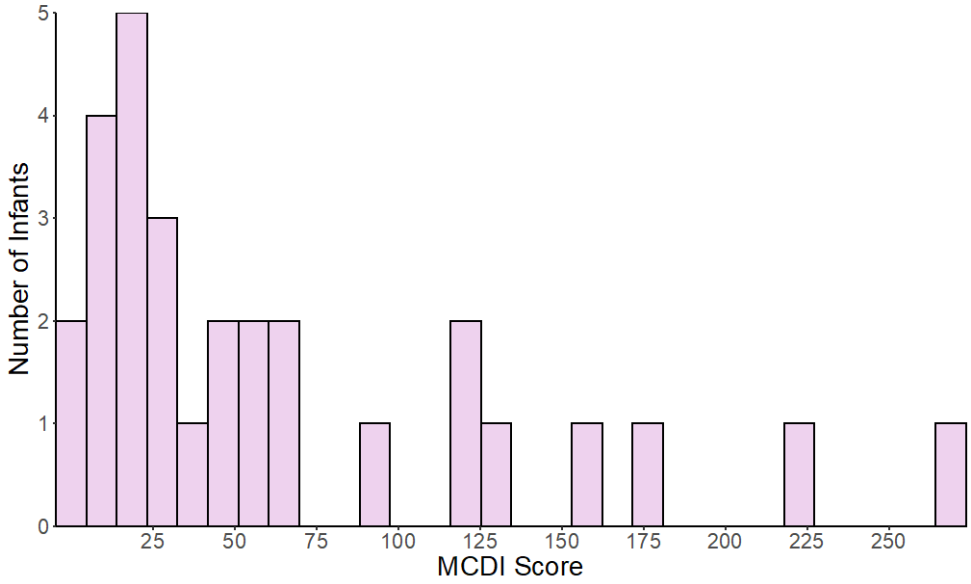

Figure 2. MCDI score was plotted as a histogram, where each infant had one score which was reported at 18 months.

2.4.2. Allan Factor Analysis, Volubility, and MCDI Score

We were first interested in testing if there was a relationship between AF slope and volubility to determine if less speaking could lead to increased clustering, or steeper AF slopes. Using two linear mixed effects regressions (LMER), we predicted volubility on AF slope for infants and for adults, with participant identification as the random effect. This supported our hypothesis that steeper AF slopes would correlate with less volubility (adults: $\beta = -0.66$, $t(213.60) = -12.76$, $p < .001$, infants: $\beta = -0.51$, $t(214) = -8.65$, $p < .001$). Despite this, we know from Falk and Kello (2017) that when utterances are held constant, adults display steeper AF slopes when using infant-directed speech (versus adult-directed speech), suggesting that prosodic variability is influential on this measure nonetheless. These effects did not vary for adult speakers by infant age when the interaction between volubility and month was added to each model (all $p > .05$) but varied slightly for infants (see Figure 3A). For 3-month-old infants only, no relation between AF slopes and volubility was detected, $\beta = -0.01$, $t(205.82) = -1.41$, $p = .16$ (see Figure 3B), and this effect was significantly different from the other 3-month conditions (interaction at 3 and 6 months: $\beta = 0.44$, $t(205.38) = 3.07$, $p = .002$; 3 and 9 months: $\beta = 0.51$, $t(207.13) = 3.06$, $p = .003$; 3 and 18 months: $\beta = 0.46$, $t(207.89) = 2.72$, $p = .007$).

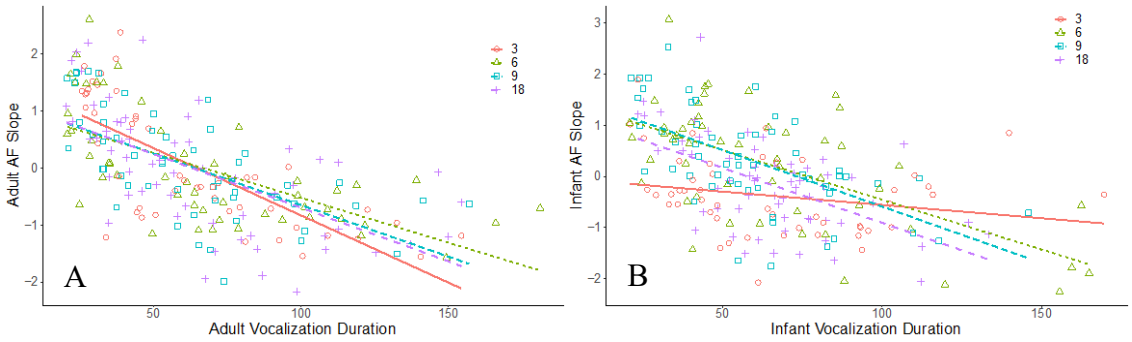

Figure 3. A. Adult vocalization durations (or volubility) predicted adult AF slopes across months. B. Infant vocalization durations predicted infant AF slopes except at 3 months. Note: All plots of regressions shown here do not include error terms for better visibility and because the plots differ minimally from the outputs which include error terms. Volubility was also standardized in the statistical models, but not in the plot for easier interpretation.

We next tested if AF slopes differed by adult and infant speakers using a two-sample t-test. AF slopes significantly differed by adult and infant speakers, $t(418.78) = 3.03$, $p = .003$, with adults having steeper slopes ($M = 0.80$, $SD = 0.30$) than infants ($M = 0.72$, $SD = 0.25$). To determine whether this difference was due to adults speaking less than infants, we ran another two-sample t-test on infant and adult volubility and found no reliable difference, $t(419.13) = 0.23$, $p = .82$. Steeper slopes could therefore reflect more infant-directed speech, since flatter slopes are associated with adult-directed speech, as previously shown in Falk and Kello (2017). Using a one-way repeated measures ANOVA, we found that neither adult nor infant slopes varied reliably by infant age (adults: $F(1,28) = 0.01$, $p = .91$; infants: $F(1,28) = .06$, $p = .81$; see Figure 4), although infants descriptively had flatter slopes at 3 and 18 months. While the meaning of this finding is less clear at 3 months, at 18 months we can speculate that vocalizations are becoming more adult-like or advanced.

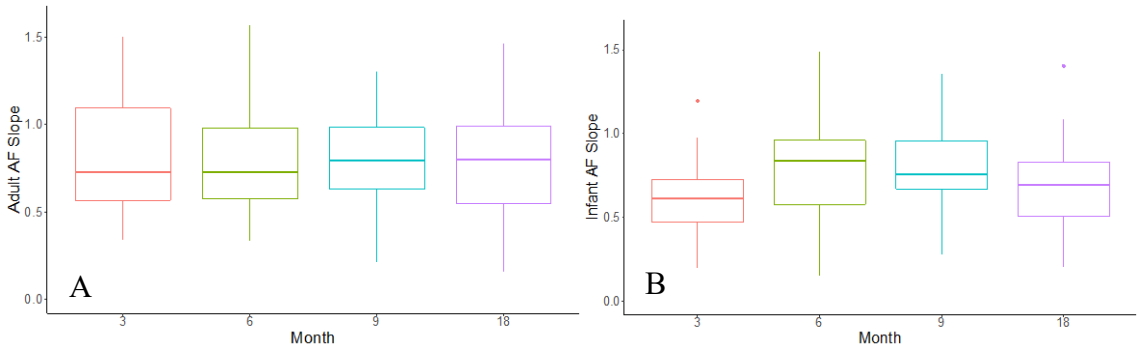

Figure 4. Average AF slope by month for adults (A) and infants (B). Adults had a higher mean slope than infants overall. Infants had flatter slopes at 3 and 18 months compared to 6 and 9 months.

We next tested if either speaker's AF slope was related to MCDI score, where we initially expected that steeper adult slopes and flatter infant slopes would relate to higher scores, although we could imagine reasonable hypotheses being proposed in opposing directions too. We used a linear model here with no random effects, rather than a LMER with random effects, because the model failed to converge using the latter model. Thus, using a linear model with the standardized slope as the predictor value and MCDI score as the predicted value, we found that infant AF slopes were not related to MCDI score across months, $\beta = -3.97$, $t(214) = -0.90$, $p = .37$. We next tested if this effect varied by month by adding an interaction term between slope and month to the model. Infant AF slopes also did not predict MCDI within months, all $p > .05$ (see Figure 5B). However, the relationship trended such that steeper slopes at 3 and 6 months, and flatter slopes at 9 and 18 months, related to better MCDI scores, meaning that the infant's vocalizations might be predictive of their future and current vocabulary at those respective months with more power.

Using the same models with adult rather than infant data, when the data were averaged across months, adult AF slope was marginally related to MCDI score, $\beta = -9.79$, $t(214) = -2.25$, $p = .03$, with flatter slopes correlating with stronger scores. In other words, more monotone adult speech correlated with better scores, which may be reflective of this type of speech affecting infants' vocabularies, since MCDI scores have been reported to be a reliable assessment of language ability in prior literature (Skarakis-Doyle et al., 2009). Although this effect trends the same way across all months, it is not reliable when broken down as such, all $p > .05$ (see Figure 5A). For both models, flatter slopes or more monotone and adult-like speech descriptively related to better scores, and perhaps with more power these effects would become reliable.

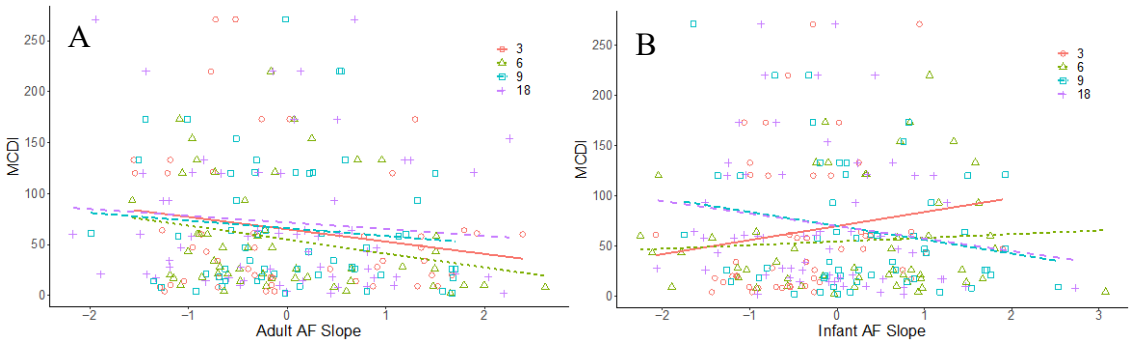

Figure 5. A. Adult AF slope predicted MCDI score when the data were collapsed across all months, but lost power and became nonsignificant when broken down by month. B. Infant AF slope did not statistically predict MCDI score, but the 9- and 18-month slopes trended similarly to the adults' slopes, where flatter slopes related to better scores.

To ensure that the effect of adult AF slope predicting MCDI score was not due to adult volubility, we created a similar linear model with adult vocalization duration predicting MCDI score. No effect was found over all months, $t(214) = 0.43$, $p = .67$.

When broken down by month, the adults' vocalizations when the infants were 3 months differed marginally from 18 months, β = 26.26, $t(208) = 2.04$, $p = .04$, but there was no effect for either month, all $p > .05$ (see Figure 6A). The lack of an effect overall suggests more monotone adult speech related to better infant scores since volubility did not correlate with MCDI scores, but AF slopes did. (Infant volubility also did not relate to MCDI score over all months, β = 1.96, $t(214) = 0.45$, $p = .66$., or by month, all $p > .05$; see Figure 6B.)

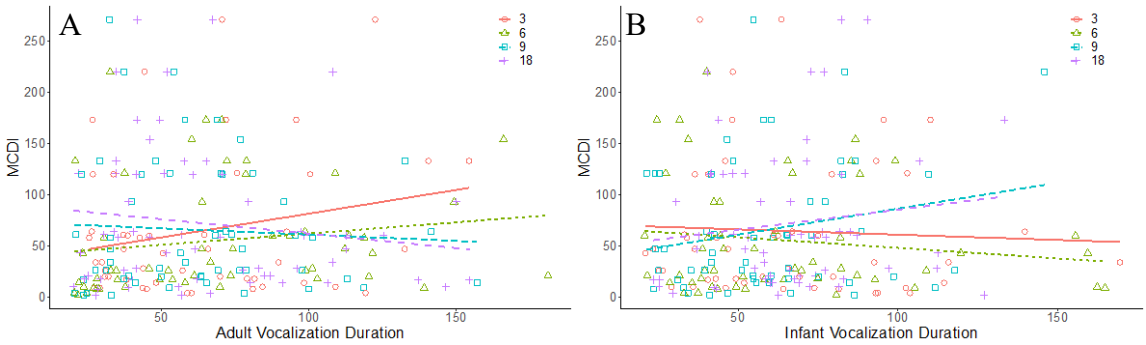

Figure 6. A. Adult vocalization duration (or volubility) did not predict MCDI score when the data were collapsed across all months or broken down by month, although the 3- and 18-month slopes marginally differed from each other. B. Infant vocalization duration did not statistically predict MCDI score.

2.4.3. Complexity Matching, Volubility Matching, and MCDI Score

We anticipated that we would replicate the general effect of complexity matching found in prior studies (Abney et al., 2014; Schneider et al., 2020). We used a LMER to test if this effect was found across all months, with adult slope predicting infant slope and the random effect of participant identification, and found no such effect, β = 0.01, $t(208.46) = 0.14$, $p = .89$. When the interaction between adult slope and month was added to the model, no effects were found by month (all $p > .05$, see Figure 7A), raising the question of at what age matching begins to occur. Likewise, we were also curious if volubility matching could be observed. Using a similar model with adult volubility predicting infant volubility, no matching was observed here either across months, β = 0.03, $t(210.2) = 0.39$, $p = .70$, or within months, all $p > .05$ (see Figure 7B). To the author's knowledge, this is one of the first times complexity matching has not been observed between speakers (other than during argumentative conversations; Abney et al., 2014), suggesting that complexity matching and volubility matching are neither arbitrary measures nor phenomena.

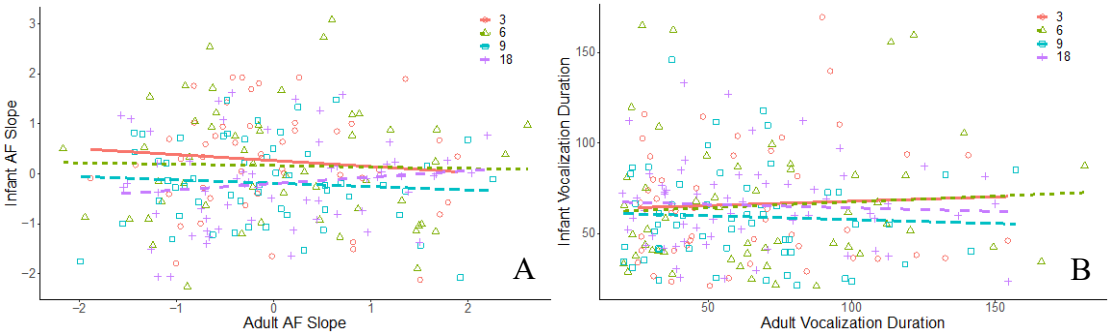

Figure 7. Adult and infant AF slopes (A) and vocalization durations (B) were not correlated at any month, meaning complexity matching and volubility matching were not detected, respectively.

Despite the post hoc knowledge that there was no matching between speakers, we still tested if there was a relationship between matching and MCDI scores, our primary question of interest. To determine the strength of matching at each month, we used the absolute difference between infant and adult slopes, where zero corresponded to the strongest possible matching. We then created a linear model predicting complexity matching on MCDI score without a random effect, because a LMER with a random effect failed to converge. Again going against our initial expectations, there was no relationship between matching and MCDI scores across all months, $\beta = -20.01$, $t(214) = -1.96$, $p = .29$. After adding the interaction term between matching and month, there was no relationship between months either, all $p > .05$ (see Figure 8A). Using analogous models, there was also no relationship between volubility matching and MCDI score across months, $\beta = -0.07$, $t(214) = -0.47$, $p = .64$, or by month, all $p > .05$ (See Figure 8B). These null results were anticipated post hoc since there was no matching at any age, but this went against our original hypothesis that more matching would be predictive of better MCDI scores.

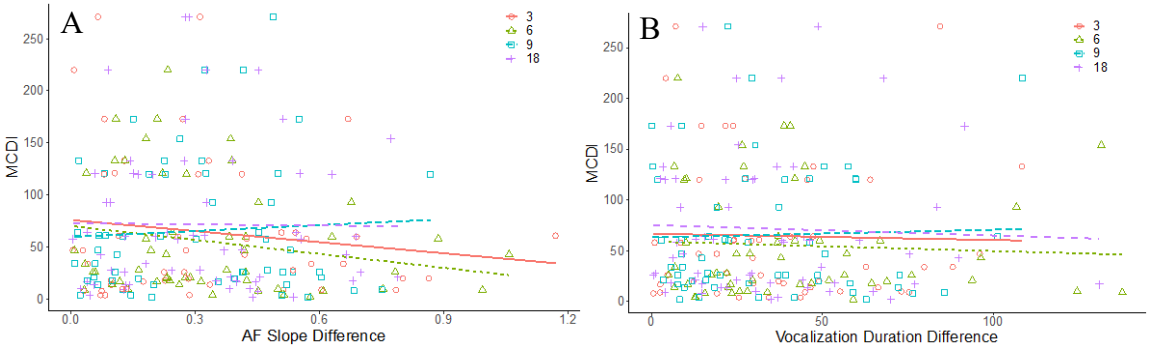

Figure 8. Complexity matching was measured here as the absolute difference between infant and adult AF slopes (A) and absolute difference between infant and adult vocalization durations (B). Neither form of matching significantly correlated with MCDI score.

2.5. Discussion

In this study we examined the effects of adult and infant volubility, volubility convergence, hierarchical temporal structure, and complexity matching (or the convergence between the speaker's hierarchical temporal structures) on the infant's vocabulary strength to determine if these vocalization features or the alignment between them were related to vocabulary size at 18 months. Out of these analyses, only adult hierarchical temporal structure predicted vocabulary. These largely null effects for individual vocalizations were surprising given the amount of literature relating greater volubility (Overby et al., 2020; Warlaumont et al., 2014) and other features of speech (Fernald et al., 2006; Rowe, 2012) to typically developing or stronger later language outcomes. Since the effects of convergence were grounded in mixed findings from prior literature (e.g., Quigley et al. (2016) versus Seidl et al. (2018)), our hypotheses were more exploratory. However, this was notably one of the first times complexity matching—at least as measured by hierarchical temporal structure—was lacking between speakers, except during argumentative conversations between adults (Abney et al., 2014). The general lack of effects may be due to the measures themselves not being predictive of these later outcomes, having too small of a sample size, the naturalistic character of the experiment, or perhaps because the vocabularies were self-reported by caregivers.

2.5.1. Less Adult Hierarchical Temporal Structure Correlates With More Advanced Vocalizations

The primary significant finding was that the adult hierarchical temporal structure marginally predicted vocabulary across all months, where flatter AF slopes or less hierarchical temporal structure correlated with stronger vocabularies. However, this effect vanished when broken down by month, suggesting that this effect was either weak or more data were required. One significant limitation was that no files containing >20 seconds of vocalizations were utilized to help ensure the data were rich enough to be analyzed, but this omitted a large portion of the dataset. Since Falk and Kello (2017) found that flatter AF slopes correspond to ADS, and ADS corresponds to more advanced and adult-like speech, these findings suggest that adults adjusted their prosodic structure to be more similar to more lexically advanced infants who vocalize with flatter slopes. In other words, perhaps once infants begin to sound more adult-like and speculatively use more advanced speech-related vocalizations, adults align their speech to the infant to also become more adult-like.

On the less likely hand, these findings could stand in contrast to the majority of the literature arguing that IDS is more helpful than ADS for language learning (e.g., de Boer & Kuhl, 2003; Golinkoff et al., 2015; Thiessen et al., 2005). Perhaps using more ADS in one's speech models more mature language for the infant to use, allowing for stronger vocabularies. Since hierarchical temporal structure can vary based on volubility, we can imagine one arguing that the AF slopes are flatter because adults spoke more instead of using ADS, but this is unlikely since volubility was not related to vocabulary whereas hierarchical temporal structure was. However, we should again remind the

reader that regardless of either explanation, this effect was marginal, and we are therefore cautious about overinterpreting our findings.

In addition to the marginal nature of these findings, we are also hesitant to overinterpret them since the measure of vocabulary was self-reported by the caregivers. Although this method of reporting vocabulary was found to be reliable in past studies (Skarakis-Doyle et al., 2009), it is still plausible that caregivers may either overreport or underreport their child's vocabulary size, particularly since they filled out the questionnaire without researcher supervision, although they did receive written instructions from a researcher beforehand. Another limitation is that an infant's vocabulary size is not the only measure of language capability, and other qualities of vocalizations including the vocalization type may be informative or relate more strongly to volubility and hierarchical temporal structure (Pretzer, 2019). Other measures of language ability should therefore continue to be analyzed in relation to hierarchical temporal structure and volubility.

2.5.2. Less Infant Hierarchical Temporal Structure Descriptively Relates to Stronger Vocabularies

While infant hierarchical temporal structure did not significantly correlate with vocabulary, the plots of these trends are compelling because for the younger 3- and 6-month-olds, steeper AF slopes descriptively related to stronger vocabularies at 18 months, and the opposite trend occurred for the two older groups (see Figure 5B). This finding is descriptively interesting because it falls in line with findings from Abney et al. (2017), where AF slopes were flatter for more advanced speech-related vocalizations and steeper for non-speech-related vocalizations. These results together suggest that more advanced vocalizations are reflected in flatter AF slopes, whether that may be for ADS (Falk & Kello, 2017), speech-related vocalizations (Abney et al., 2017), or perhaps infant vocabulary. On the other hand, steeper slopes correspond to IDS, non-speech-related vocalizations, and perhaps poorer vocabulary scores. Future studies should continue to investigate if hierarchical temporal structure relates to different language abilities, as it might potentially prove to be another noninvasive measure to predict developmental delays.

2.5.3. Volubility Did Not Significantly Relate to Vocabulary Size

Going against our initial hypotheses, we were surprised to find null effects of adult and infant volubility on vocabulary given both how intuitive this effect is and because it has been demonstrated in prior studies (e.g., Hirsh-Pasek et al., 2015; Sultana et al., 2020). However, the association between vocabulary size and volubility may have been reduced by including all adult and infant vocalizations, where some sounds like infant cries and adult vocalizations directed towards others in the environment ("other-directed speech") may have been less associated with vocabulary size. Specifically, other-directed speech has been excluded from prior studies (Ramírez-Esparza, García-Sierra, & Kuhl, 2014) or found to not be associated with language learning (Weisleder & Fernald, 2013). One insightful explanation proposed by authors including Golinkoff et al.

(2015) and Sperry, Sperry, and Miller (2019) is that both the quantity and quality of adult vocalizations are important, rather than the quantity alone. Rowe (2012) also proposed this rationale: When assessing infants of different ages, volubility was predictive of vocabulary ability in 18-month-old infants, but by 30 months, vocabulary diversity became the better predictor. While these findings still stand somewhat in contrast to our own since our sample of infants were 18 months old and younger, it may still be the case that the quality of the vocalizations mattered more than the quantity of them. It is also possible that there is another measure of volubility or feature of vocalizing that is a better predictor of vocabulary. For example, while we used the total amount of time vocalizing to determine volubility, one could also measure the total number of utterances. It would additionally be useful for future research to focus on the specific types of utterances used, such as only those that were speech-related or non-speech-related. We did not have the power needed to analyze parsed utterances by type, but we can imagine this being informative for later vocabulary size.

2.5.4. Neither Complexity Nor Volubility Matching Related to Vocabularies

Even though adults tend to align their voices to infants more than vice versa (Abney et al., 2017; Ko et al., 2016), we still expected to find some levels of complexity or volubility matching between the pairs, especially by 18 months. In contrast to our predictions neither form of matching was found at any age. The lack of complexity matching was notable since it has been observed across languages and during most friendly adult conversations (Schneider et al., 2020), and only less so during infant-adult interactions (Abney et al., 2017) and not during argumentative conversations (Abney et al., 2014). This finding therefore demonstrates that complexity matching is indeed a meaningful measure of matching that is not guaranteed. Further, this result leaves the question of when complexity matching begins to occur open for future studies to continue exploring, though our preliminary findings did not detect it between 3 and 18 months.

2.5.5. Limitations

Although infant and adult volubility were not reliably different from one another, it is possible the interactions were still rather asymmetric, in that either speaker could have vocalized for the majority of the first half of the recording and the other for the second half, for example. Since the recordings were also in naturalistic environments, it is plausible infants and adults were in the same room and vocalizing at the same time, but not to one another, such as if the caregiver were addressing someone else while the infant vocalized separately. Furthermore, the caregiver was often the primary person being encoded by researchers, but not always; researchers were instructed to encode any adults in the environment. This carried the benefit of including all adult sounds the infant heard, but these sounds were not required to be directed toward them (researchers did parse the data by directionality, but the data were again sparse enough we chose to include all vocalizations). It would also be valuable for future studies to measure interrater reliability for researchers who encoded the vocalization durations.

While using naturalistic recordings had many benefits, future experiments aiming to explore convergence might find it useful for the experiments to be more contrived so that a specific goal may be given, such as by directly instructing the caregivers to interact with the infants. This cannot guarantee the infant will respond in a symmetric manner, but it does guarantee that the caregiver is addressing the infant. Such a study would be useful for continuing to test when convergence between children and parents starts, so as to better inform the theory of interactive alignment. For example, in a study by Newman et al. (2016), mothers and their seven-month-old infants played with a standard set of toys (e.g., plush animals, a baby doll) while their vocalizations were recorded. Once the infants turned two years old, the parents completed a standard vocabulary assessment using the MCDI, similarly to the present study. Although the number of words spoken by the parents to the infants was analyzed, hierarchical temporal structure and measures of convergence were not, leaving such a study open for future research.

2.5.6. Conclusion

In sum, only the adult's hierarchical temporal structure or style of speech showed a relation to the infant's vocabulary size, but the infant's hierarchical temporal structure trended such that a correlation may potentially exist for infants too. Conversely, neither complexity nor volubility matching related to lexical ability. These findings together suggest that an individual's vocalizations are a better predictor of later vocabulary ability than the matching between features of each person's vocalizations. However, the limitations of our study should encourage future researchers to explore these questions, including what vocalization features are predictive of later language learning and at what age interactive alignment begins to occur.

Chapter 3

Hierarchical Temporal Structure and Performance in the Montclair Map Task

3.1. Preface

In this chapter, I assess the relationships between speech and speech convergence on task performance to determine if stronger or poorer performance is reflected through emotive (or performative) features of speech. Using audio data from the Montclair Map Task Corpus, I test whether performance on a difference-finding task correlates with two measures of prosody (hierarchical temporal structure and speaking rate), or the strength of convergence between each measure (complexity matching and speaking rate matching, respectively), as functions of epoch and sex pairing. I found that male-male pairs spoke faster to each other and had more speaking rate matching than female-female or male-female pairings, but there was no effect of pair sex on hierarchical temporal structure or complexity matching, indicating these are distinct measures of analysis. Only male-male pairs showed a relationship between better task performance, more hierarchical temporal structure, and slower speaking rates, suggesting these pairs performed better when they spoke more slowly with more prosodic exaggeration. Somewhat analogously to the findings from the previous chapter, performance was again better moderated by hierarchical temporal structure than forms of convergence. A version of this manuscript is under review and was co-authored by Chris Kello, Jennifer Pardo, and Kathleen Coburn.

3.2. Introduction

3.2.1. Prosody and Effective Communication

Prosody is essential for effective communication across spoken and signed languages (Brentari & Crossley, 2002; Dachkovsky & Sandler, 2009; Greiser & Kuhl, 1988; Grünloh, Elena, & Michael, 2011; Ito & Mester, 2013). The features composing prosodic structure include speaking rate, amplitude, pitch, rhythm, and intensity (Culpeper, 2011; Erekson, 2010; Zhu & Penn, 2006). Prosodic structure helps inform the listener of the speaker's semantic meanings, intents, emotions (Culpeper, 2011; House, 2007), aids in word recognition (Cutler, Dahan, & van Donselaar, 1997; Kim & Sumner, 2017) and also helps with listener comprehension (Frazier, Carlson, & Clifton, 2006; Speer & Ito, 2009). Prosody therefore seems to act as a communicative tool during verbal exchanges, helping the speaker to convey information to the listener (Erekson, 2010; House, 2007; Selting, 2010).

Despite prosody's importance for successful communication, little research to the author's knowledge has analyzed features of prosody for individual adult speakers in relation to performance on collaborative tasks. However, as described in the previous chapter, many studies have examined features of adult prosody in relation to infant language acquisition. The majority of this literature has found young infants prefer the more prosodically exaggerated speech used when addressing infants as compared to

adults (Cooper & Aslin, 1990), and that this form of speech helps infants learn vowel categories and word boundaries (de Boer & Kuhl, 2003). For example, Thiessen, Hill, and Saffran (2005) found that infants can differentiate between artificial words and part words (or the syllables between words boundaries, including "ty ba" in "pretty baby") after being exposed to them in infant-directed speech, but not in adult-directed speech. This form of speech may help to grab an infant's attention because it is less monotone than adult-directed speech (Falk & Kello, 2017; Kello et al., 2017). Another study by Haake, Hansson, Gulz, Schötz, and Sahlén (2014) found 5- to 6-year-old children performed better on a grammatical test when it was spoken to them using slower speaking rates. On the other hand, our own findings from Chapter 2 indicated that more monotone adult speech was related to better speech ability in infants, suggesting that these findings may be somewhat mixed. While these studies lie in the context of infants and children, the findings overall suggest that prosody is important for some forms of understanding and learning, which may potentially also be applicable to adults, such as how more prosodically exaggerated speech is less monotone and thus perhaps more reflective of greater task engagement. On the other hand, more monotone speech may indicate that the speakers are either bored or taking the task very seriously.

3.2.2. Measures of Prosody and Performance

Since the use of prosody is important for expressing one's meanings and emotions clearly, we predicted that the use of prosody would relate to performance. In the present study, we first examined two different measures of prosody, *hierarchical temporal structure* and *speaking rate*, and how they related to performance on a collaborative task that required verbal communication. The former is a more novel measure of variability in the amplitude and timing of speech as seen through the nested clustering of acoustic speech energy, whereas the latter is a more traditional measure of analyzing words per second. The data come from a previously published study of dyadic speech interaction (Pardo et al., 2019) in a variant of the well-known HCRC Map Task (Anderson et al., 1991), which involved one participant guiding another through a map, where the leader was shown a visible path but the follower was not. Since we have only thus far drawn analogies from infant studies, and these findings were somewhat mixed, we expected that either monotone or prosodically exaggerated speech would relate to task performance, although we were unsure of the directionality of any potential effects. We tested if these relationships depended on the epoch (beginning, middle, or end of the experiment) or the sex pairing of the interlocutors (male-male, female-female, or mixed male-female) since prior analyses of the present corpus (Pardo et al., 2019) and other studies (Manson et al., 2013) have found sex differences, and testing for the effects of practice is standard in the behavioral sciences (Harrison et al., 2003; McGregor, Eden, Arbisi-Kelm, & Oleson, 2020; Wesnes & Pincock, 2002).

As described in the following sections, we also tested whether the pairs' speaking rates and Allan Factor (AF) functions converged or became aligned, and whether the degree of convergence related to map task performance. Like hierarchical temporal structure and speaking rate, we tested if any such relationships with performance depended on the pair sex or epoch. Sex differences have been found on the HCRC map

task corpus and during other conversations, motivating these analyses (Bilous & Krauss, 1988; Leaper & Ayres, 2007; Mulac, 1989; Pardo et al., 2019; Vogel, Lopes, & Esposito, 2017).

Hierarchical temporal structure and speaking rate were chosen in part because each reflects differing but complementary features of prosody. AF analysis relates to prosodic exaggeration as shown through variability in hierarchical temporal structure caused by changes in emphasis and amplitude, whereas overall speaking rate may vary independently of changes in emphasis and amplitude (although both measures were averaged over the duration of the conversation). For example, one could increase or decrease their speaking rate while remaining monotone, or one could speak either loudly or quietly without changing their speaking rate. Further, speaking rate is a more direct measure of analysis. In contrast, the use of AF analysis allows us to analyze prosodic variation across timescales, providing a wider scope of analysis than, say, amplitude alone, though measuring amplitude would be interesting for future studies.

3.2.2.1. Hierarchical Temporal Structure

Hierarchical temporal structure, or the nested clustering of acoustic speech energy, was measured using Allan Factor (AF) analysis, the details surrounding which are described in the introductory chapter. To summarize this methodology, Falk and Kello (2017) found that the prosodic exaggeration of infant-directed speech manifests as greater hierarchical temporal structure. More specifically, they found that the variability in acoustic durations of linguistic units—including syllables, words, phrases, and speaking rate—was correlated with the degree of hierarchical temporal structure as measured by AF analysis, but only the longer timescales were analyzed here. This is to say that hierarchical temporal structure a feature of prosody that is reflective of prosodic exaggeration. Along similar lines, Kello et al. (2017) found that speakers in a conversation displayed greater hierarchical temporal structure than those giving a TEDx talk for an audience, where speech is more monotone. These results served as the basis for using AF analysis in the present study to assess hierarchical temporal structure, or prosodic exaggeration.

3.2.2.2. Speaking Rate

Speaking rate was measured as the mean number of linguistic units produced per time unit, such as the mean number of words or syllables per second. It can vary depending on factors like the degree of speaker excitation (Koolagudi & Krothapalli, 2011; Yildirim et al., 2004), and faster speaking rates may also result in less variability (Dellwo & Wagner, 2003; Ramirez-Aristizabal et al., 2018). That is, when speaking faster, more syllables can be packed into phrases, reducing the number of phrases overall (D'Imperio, Gorka Elordieta, Prieto, & Vigário, 2005; Jun, 2003) because faster speaking rates may come at the cost of less accurate and shortened phonemes where more short utterances can fit into a phrase than long utterances (Hilton, Schüppert, & Gooskens, 2011). While faster speech may be more monotone, it has often been found to correspond with increased social attractiveness, liveliness, and confidence (Buller, LePoire, Aune, &

Eloy, 1992; Jiang & Pell, 2017; Mariooryad, Kannan, Hakkani-Tür, & Shriberg, 2014; Weiss & Burkhardt, 2010). This may be because faster speakers are perceived as being more decisive and energetic, but this effect may also be due to the reduced variability in speech, lessening the cognitive load of the listener (Francis & Nusbaum, 1996). However, speaking quickly is not always beneficial because speaking slower can aid in listener comprehension (Robinson, Sterling, Skinner, & Robinson, 1997).

3.2.3. Convergence and Performance

In spoken interactions, behaviors ranging from postural sway (Shockley, Santana, & Fowler, 2003), to gaze (Richardson, Dale, & Tomlinson, 2009), to various levels of speech (Bock, 1986; Pardo, 2006), often converge between interlocutors. This phenomenon is also frequently referred to as alignment, matching, and coordination. Different forms of prosodic convergence have also been observed in prior studies (Gorisch, Wells, & Brown, 2012; Manson et al., 2013; Schneider et al., 2020). For instance, individuals have varying baseline speaking rates, but when two individuals speak together, their speaking rates often converge toward their mean rate (Giles & Baker, 2008). Further, in early verbal interactions between mothers and their infants, mothers often adjust features of their speech to converge toward their children (Abney et al., 2017; Ko et al., 2016). Other forms of convergence in speech also exist, such as phonetic convergence (Nielsen, 2011; Pardo, 2006) and syntactic convergence (Bock, 1986; Healey, Purver, & Howes, 2014), which respectively measure the alignment of phonemes and syntax between speakers.

The communicative role of convergence in verbal interactions is open to investigation (Falk & Kello, 2017; Giles, 1973; Giles & Baker, 2008; Staum Casasanto, Jasmin, & Casasanto, 2010), but convergence across interacting levels of representations has been theorized to be crucial for successful communication (Garrod & Pickering, 2004; Pickering & Garrod, 2006). For example, at the level of semantic representations, if one speaker is referring to a “note” as in a musical note, but the other is referring to a “note” in a textbook (Ullmann, 1974), their semantic representations are not aligned, and the interaction would be confusing until the representations became aligned and participants were colloquially on the same page.

There is also evidence for the involvement of convergence in either establishing, or perhaps resulting from, common ground and increased rapport. Rapport is a term used to express positive sentiments between people, such as feelings of connectedness, closeness, or prosocial behavior (Valdesolo et al., 2010). When pairs or groups of people experience heightened feelings of rapport, researchers have theorized that they display increased behavioral coordination (Bernieri, 1988; Lakin & Chartrand, 2003; Vacharkulksemsuk & Fredrickson, 2012), and while the literature has been somewhat mixed (e.g., Carmody, Mateo, Bowers, & McCloskey, 2017), previous studies have mostly supported this hypothesis. In a study of Japanese telephone conversations, greater convergence in pitch and intensity were correlated with greater speaker engagement and amiability (De Looze, Scherer, Vaughan, & Campbell, 2014). Another study examined hierarchical temporal structure convergence, or complexity matching (Abney et al.,

2014), and found convergence across timescales only during affiliative or friendly conversations, but not during argumentative conversations.

Evidence for prosodic convergence as a communicative function has led some researchers to investigate whether aspects of prosodic convergence, as well as other types of convergence, might relate to performance when speakers must work together towards a common goal (Coco, Dale, & Keller, 2018; Fusaroli et al., 2013; Pardo, 2006; Pardo et al., 2013; Pardo, Jay, & Krauss, 2010; Reitter & Moore, 2014; Weber, Christiansen, Indefrey, & Hagoort, 2018). The results of prior studies have been generally mixed, suggesting that relationships with performance, if they exist, are subtle and may depend on a variety of contextual factors yet to be determined. In an example where more convergence was related to better performance, Reitter and Moore (2014) analyzed syntactic alignment and score in the classic HCRC Map Task Corpus (Fay et al., 2018; Mushin, Stirling, Fletcher, & Wales, 2003; Reitter & Moore, 2014). As briefly described, an instruction giver (leader) verbally communicated information about a path on a map to an instruction receiver (follower). The instruction giver was shown a path on their map, but the instruction receiver was not shown a path on their otherwise identical map, and the pair could not see each other's maps (Anderson et al., 1991). Therefore, the instruction giver verbally guided the instruction receiver from start to finish on the map. More syntactic alignment between speakers correlated with better dyadic performance on the task. In another study, male-male pairs, but not female-female pairs, who had more co-laughter counts and proportions were more likely to cooperate on a prisoner's dilemma task (described in more detail in the following section; Manson et al., 2013).

Using less standard methods of analyses, Fusaroli et al., (2013) analyzed a study (Bahrami et al., 2010) in which pairs of participants viewed Gabor patches and independently selected the patch that was different from the rest. If the pair disagreed about which patch was different, they discussed the trial until coming to a joint agreement. Speech convergence in their discussions was measured by complexity matching, or convergence in the temporal features of speech across a range of timescales. This convergence was found to positively correlate with dyadic visual discrimination, as did convergence in confidence ratings (Fusaroli et al., 2012). Fusaroli and Tylén (2016) also conducted a similar study and found alignment of repeated prosodic entropy across pairs, as measured by cross-recurrence quantification analysis (CRQA), which is a method for essentially quantifying nonlinear cross-correlation patterns. Thus, while there is evidence for a positive relationship between task performance and convergence, performance may be better reflected in measures of convergence like complexity matching or CRQA, that go beyond alignment to capture other kinds of non-linear, interactive relationships.

In contrast, convergence sometimes exhibits a *negative* correlation with measures of task performance. For example, Ireland and Henderson (2014) studied how dyads negotiate over instant messenger about what to do on an imaginary vacation, with priorities for the vacation given in advance, and cash incentives for winning the negotiations. The words spoken by participants were placed into nine categories using the Linguistic Inquiry and Word Count (LIWC) text analysis program, which places words into meaningful categories organized by various aspects of social, affective, and cognitive processes (Francis & Pennebaker, 1993; Tausczik & Pennebaker, 2009). Dyads

who aligned their language styles in terms of LIWC categories were less likely to come to an agreement with their partner. In a differing movement-based study, Coco, Dale, and Keller (2018) asked pairs of individuals to determine if they were viewing the same picture or not without seeing each other's picture. When dyads freely engaged in conversation, broadening their visual attention through *less* gaze alignment led to better task performance, perhaps using more of a divide-and-conquer strategy.

These broad differences across studies are likely due to engaging in different tasks or analyzing different levels of convergence. Within studies, some differences in convergence and its relation to task performance may also change over the course of the interaction. Some studies have found that convergence increases over time (Fusaroli et al., 2013; Natale, 1975; Pardo, 2006), while others have observed a decrease (Matarazzo, Wiens, Saslow, Dunham, & Voas, 1964; Ray & Webb, 1966). For example, Pardo (2006) found phonetic convergence in a slightly modified version of the HCRC map task, and this convergence was observed in the first half of the conversation, but was stronger in the second half. Additionally, prosodic alignment across a range of timescales was found to increase over time in a task where participants were to determine which Gabor patch was different from the rest (Fusaroli et al., 2013). At a much longer timescale, Ray and Webb (1966) analyzed convergence in the lengths of questions and answer at news conferences with President John Kennedy, and found that across the course of several years, the correlation between question and answer length increased. However, when the authors analyzed the same correlation within conferences, the correlation decreased from the first, to the second, to the third segment of each conference (also see Levitan and Hirschberg, 2011).

3.2.4. Measures of Convergence: Complexity Matching and Speaking Rate Matching

In the present study, complexity matching was measured as the degree of correlation between two speaker's AF slopes. Thus, this measure reflects the degree of convergence in prosodic exaggeration between paired speakers. Described in the introductory chapter, this convergence is referred to as *complexity matching* (Abney et al., 2014), where in theory, more alignment should be indicative of better information exchange (West et al., 2008), more successful communicative interactions (Garrod & Pickering, 2004; Pickering & Garrod, 2006), and stronger feelings of rapport (Valdesolo et al., 2010). While complexity matching has been observed across friendly adult conversations (Schneider et al., 2020), it has not been shown in argumentative conversations (Abney et al., 2014) or in recordings of adult and infant speakers, as shown in Chapter 2. We therefore expected more complexity matching to be correlated with higher scores or better performance. However, since other studies have found divergence to sometimes be more beneficial for performance depending on the task and level of analysis (Coco et al., 2018; Ireland & Henderson, 2014), we could reasonably imagine this effect going in the opposite direction. Due to the built-in nature of correlations, where a value of zero is equivalent to matching up speakers by chance, no surrogate analyses were required.

We measured speaking rate matching as the degree of convergence between each talker's speaking rates, or words per second. Like other levels of speech, speaking rates often converge when people converse (Giles & Baker, 2008; Kousidis et al., 2008), and it has been hypothesized that convergence in speaking rates may enable convergence at other levels of speech (Fusaroli & Tylén, 2012; Garrod & Pickering, 2009; Pickering & Garrod, 2004b). For example, Manson et al. (2013) asked triads of participants to interact freely for 10 minutes, and then afterwards each participant played a one-shot prisoner's dilemma game with each of their two partners. Participants were given $3 from the experimenter if they chose to defect, keeping their money for themselves, whereas if everyone chose to cooperate and transfer their money, everyone's payoff would double. Cooperators who encountered defectors lost their money. Therefore, a higher payoff indicated more mutual cooperation. After completing the game, participants rated how warm and competent they felt their partner was using a slider as a measure of rapport, and convergence was measured in each pair's speaking rate. Results showed that speaking rate, but not rapport, predicted increased cooperation. This indicates that speech rate matching could reflect cooperation, enable friendlier exchanges, and perhaps reduce the cognitive load of the interaction. Like the correlational measure of complexity matching, no surrogate analyses were required to determine that no speaking rate matching would correspond to a value of zero.

3.2.5. Current Experiment

In the present study, we investigated relationships between communicative performance and the two measures hierarchical temporal structure and speaking rate. We also analyzed convergence of these measures over the course of the spoken interactions. We investigated these phenomena in naturalistic conversations elicited during the Montclair Map Task (Pardo et al., 2019), which is a variant of the HCRC Map Task (Anderson et al., 1991). The HCRC map task corpus has been a boon to studies of speech communication and performance, but the asymmetrical nature of the task results in unbalanced conversational contributions. That is, those in the instruction giver (leader) role typically produce twice as much speech as those in the instruction receiver (follower) role (Pardo, Jay, & Krauss, 2010; Pardo et al., 2019). To resolve this issue, Pardo et al. (2018) modified the task so that both participants had a path, and the maps differed in the composition of the landmarks. The new goal of the Montclair Map Task was to identify the differences between the participants' two maps and to indicate the locations of the missing landmarks on their own maps (see Figure 9). Further, unlike the original task where one participant gave instructions and the other received instructions, participants were assigned identical roles, which engaged them more evenly in their interactions.

We also investigated the degree to which relationships between performance, our measures of prosody, and convergence vary as a function of the measure (speaking rate versus hierarchical temporal structure), pair sex (male-male, female-female, or mixed), and amount of practice (epoch). To our knowledge, only one other study has documented the time course of prosodic convergence (Fusaroli et al., 2013), and no prior studies have examined how adult performance and prosody co-vary over time, or vary depending on pair sex. Given that the phenomenon of convergence is commonly theorized to be related

to communication accommodation (Gallois & Giles, 2015), and that performance on nearly any task tends to improve over time (Feinstein, Brown, & Ron, 1994; Roediger & Karpicke, 2006; Sawyer et al., 2011), we expected both forms of convergence to increase over time.

Regarding pair sex, males and females have been found to have both different (Anders, Hovy, & Søgaard, 2015; Holmes, 1995; Leaper & Ayres, 2007; Puts, Hodges, Cárdenas, & Gaulin, 2007) and similar (Hancock & Rubin, 2014; Mehl, Vazire, Ramírez-Esparza, Slatcher, & Pennebaker, 2007) qualities to their speech (Palomares, 2009), although such findings are not always straightforward or consistent (Byrne, Dillon, & Tran, 1994; Leaper & Ayres, 2007). For example, Leaper and Ayres (2007) conducted a meta-analysis and found that males speak more than females overall, but only when the female was a stranger or a spouse, and not if she was a friend. This difference was strongest in task-oriented as opposed to free-form conversations. In other studies, males have been found to sometimes dominate conversations by speaking more than females (Brescoll, 2011; Leaper & Ayres, 2007; Puts et al., 2007). Notably for our dataset, Pardo et al. (2019) found that males spoke more than females overall, and both greater verbosity and increased speaking rate were related to decreased performance among same-sex male pairs. However, pairs with female members showed the opposite trend, where faster speech was associated with higher scores. These effects suggest that pair sex may be a factor in convergence and its possible relationship to communicative performance.

3.3. Methods

3.3.1. Participants

Speech data were analyzed from a study by Pardo et al. (2019) in which 48 pairs of participants were recruited from Montclair State University (mean age = 21, *SD* = 2.8 years). Participants were randomly paired together in one of three sex pairings—female-female (FF), male-male (MM), or a mixed female-male condition (MX)—totaling 16 pairs in each condition. All participants reported being native English speakers, and 26 of them reported proficiency in a second language, but the experiment was entirely in English. Participants were not acquainted with one another prior to the experiment, and they were compensated $20 for participation. Demographic and other information about participants can be found in Pardo et al. (2019). The Allan Factor (AF) analyses reported in the introduction required recordings to be at least 3.5 minutes long, and as a result five pairs were omitted from analyses, as four or more of six trial recordings were <3.5 minutes long. Of the remaining 43 pairs (14 FF, 16 MM, 13 MX), 25 trials were omitted for being <3.5 minutes long.

3.3.2. Procedures

3.3.2.1. Conversational Task

Participants engaged in a modified version of the well-known Edinburgh HCRC map task (Anderson et al., 1991), called the Montclair Map Task (Pardo et al., 2019). The

original goal of the Edinburgh map task was to have pairs engage in a leader-follower interaction where the leader verbally guided the follower through a drawn map, and the path could only be seen by the leader. Because the asymmetric task goals made studying convergence more difficult, Pardo et al. (2019) designed the Montclair Map Task to be like the original Edinburgh map task in that pairs were still given similar but slightly different maps, and they were required to speak to resolve discrepancies between their maps. However, in the Montclair Map Task, pairs had the different goal of finding the differences in landmarks on their maps, and were both provided with the same path as a reference point (for example, see Figure 9), which meant that neither participant should necessarily act as a leader or a follower. The Montclair Map Task was thus a more symmetric task than the original Edinburgh HCRC map task.

Each pair of participants sat in a soundproof booth at a small table with a divider that prevented visual interaction between them. Dyads were evenly paired by sex among three combinations: male-male (MM), female-female (FF), and mixed female-male (MX). Pairs were given six sets of maps to engage in six difference-finding trials, and the maps each had five shared and five distinct landmarks, or a total of five shared and ten different landmarks between the two maps. For each map, both participants were instructed to determine and locate the five landmarks that were missing from their own maps but were present on their partner's maps. Each map was printed on an 8.5" x 11" piece of paper and participants could not see each other's maps. The assignment and ordering of map pairs was randomized. We grouped the six maps or trials into three epochs (beginning, middle, and end), with two trials per epoch, to test for differences over time. The pair's goal was to find all five differences per map, or 10 differences altogether. Once a missing feature was found, the participant used a pencil to draw where they believed the absent landmark was located on their partner's map, using verbal communication only. Each pair decided without researcher intervention when they had found all the missing features for each set of maps and when to move on to the next trial. Additional details regarding the materials, equipment, and recordings can be found in Pardo et al. (2019).

The trials were recorded as one continuous waveform, so each trial was extracted by hand and analyzed separately. Recordings contained little crosstalk or noise between the channels separating partner vocalizations, and the Audacity™ software was used to remove what slight noise there was. A noise profile was chosen based on crosstalk examples selected from a visual and auditory review of the speech waveform by the researchers, and the noise profile chosen was applied to the entire waveform, and therefore across all trials, to ensure the selection process was minimally subjective but still adequately filtered.

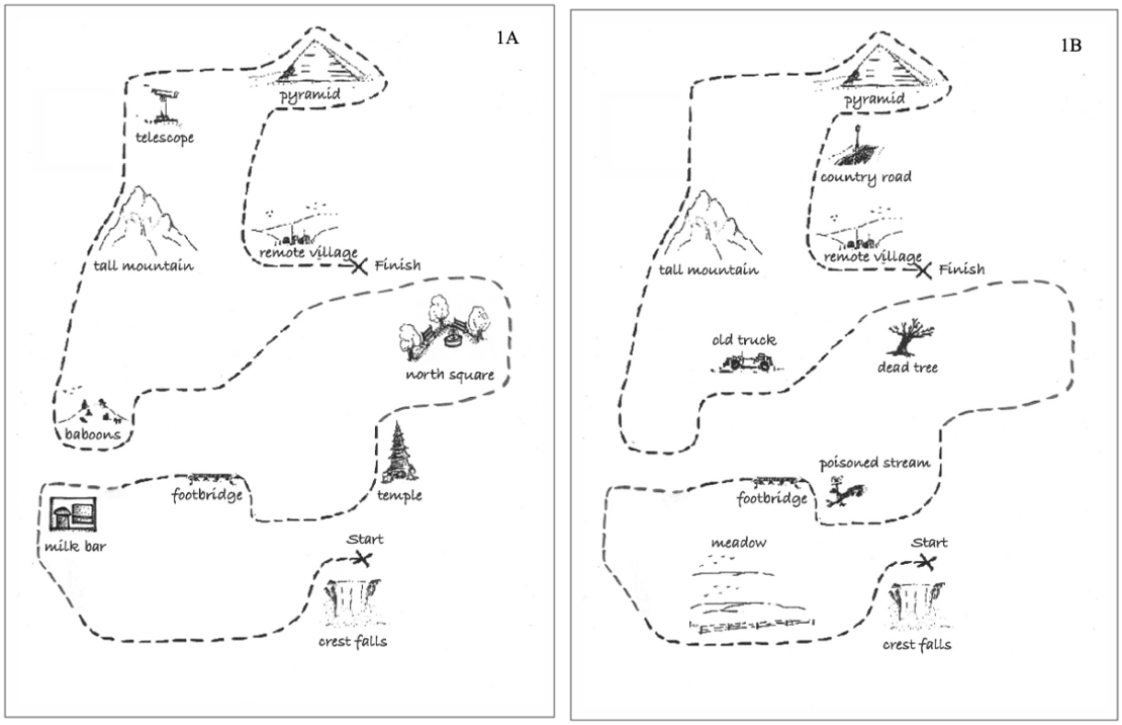

Figure 9. An example of one set of maps, where one participant would be assigned map A and the other map B. Five features are shared between the maps: the pyramid, remote village, tall mountain, footbridge, and crest falls. Ten features are different between the two maps. In A, the landmarks missing are the country road, old truck, dead tree, poisoned stream, and meadow. In B, these are the telescope, baboons, north square, temple, and milk bar. All sets of maps contained the same number of shared and different features.

3.3.2.2. Task Performance Scoring

Two independent raters scored performance on each map based on five categories: *correctly marked* location of missing feature (20 points); incorrect mark *adjacent* to the correct location and on same side of path (15 points); incorrect mark *near* the correct location or on other side of path (10 points); incorrect mark *distant* from a correct location (>2 inches, 5 points); and no corresponding mark (0 points). The maximum score for each map was thus 100 points since there were five missing features, and the final participant's scores used the average of both raters' scores. Pardo et al. (2019) reported inter-rater reliability to be high (r = 0.84). This scoring methodology is analogous to that originally used by the aforementioned Anderson et al. (1991), who superimposed a 1cm-by-1cm grid and scored the participants based on how closely the actual and drawn paths matched.

3.3.2.3. Speaking Rate and Hierarchical Temporal Structure Measures

Speaking rate was measured as words per second during each trial, including time spent pausing during a speaking turn, but not time between turns. As described in detail in Chapter 1, hierarchical temporal structure was quantified using AF analysis (Kello et al., 2017; Ramirez-Aristizabal et al., 2018; Schneider et al., 2020); please return there for a more informative description. However, to provide a brief reminder, more hierarchical temporal structure as seen by steeper slopes relates to more prosodically variable speech, such as when engaging in a conversation as opposed to giving a TEDx talk, where little hierarchical temporal structure corresponds to monotone speech and flatter AF slopes (Kello et al., 2017). This methodology is useful because it concurrently allows variability in speech to be measured across timescales rather than only studying one level of analysis. Like Chapter 2, the present audio durations were exactly five minutes long, so the 11 timescales of AF variance ranged from approximately .01 to 13.13 seconds, and the AF regression lines were fit only to the five longest timescales, also ranging from 0.82 to 13.13 seconds.

3.3.2.4. Speaking Rate Matching and Complexity Matching Measures

Speaking rate matching was measured as the strength of correlation between paired speaking rates. As also described in Chapter 1, complexity matching was measured as the strength of correlation in the slopes of AF functions (Abney et al., 2014, 2017; Kello et al., 2017; Ramirez-Aristizabal et al., 2018). As a reminder, the use of correlation has the advantage of obviating the need for surrogate analyses, or a control group, because a chance relationship necessarily corresponds to zero correlation. However, when analyzing the relationship between either complexity matching or speaking rate matching and the dyad's average score on the task, we needed a measure of complexity matching for each trial to correspond with the scores that were observed. Correlations can only be computed across trials, so we could not relate them to performance on each separate trial. Instead, we computed per-trial measures of matching by taking the absolute value of the difference between speaking rates or slopes for each trial. We then correlated these absolute differences with measures of performance, and tested whether absolute differences varied with epoch or pair sex.

All statistical analyses and plots were conducted using R (version 3.6.3 (2020-02-29) – "Holding the Windsock"), and packages lme4, lmerTest, and ggplot2 (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2014; Wickham, 2016, respectively), with the exceptions of the left two plots for Figure 10, which were created using MatLab_R2020B (9.9.0.1467703).

3.4 Results

3.4.1. Summary of Hypotheses and Analyses

In the following sections, we first tested if task performance was related to hierarchical temporal structure, as reflected by AF slopes, to see if steeper slopes (or

greater hierarchical temporal structure, consistent with greater prosodic exaggeration) corresponded to better task performance. Because prosody reflects emotional states like feeling enthused or uninterested, and hierarchical temporal structure is reflective of prosodic structure (Falk & Kello, 2017; Kello et al., 2017), we expected greater hierarchical temporal structure to relate to better scores. As a second measure of prosody, we also tested if faster speaking rates were related to performance, as might be predicted if faster speech relates to either better information exchange by rapidly producing words efficiently, or worse information exchange by rushing through the task. Both hypotheses were possible and therefore entertained.

As complementary analyses, we also tested if convergence in either AF slopes (complexity matching) or speaking rates (speaking rate matching) related to performance. Convergence was of interest in part because it has been theorized to be required across levels of representation in order for successful communication to occur (Garrod & Pickering, 2004; Pickering & Garrod, 2004a, 2004b). Also, stronger degrees of convergence are often related to stronger feelings of rapport (Manson et al., 2013; Valdesolo et al., 2010), which may additionally be indicative of task performance. In other words, pairs who converge more either tend to feel more positively towards the other (although the causality of this effect is unknown), and we therefore hypothesize that greater complexity and speaking rate matching will be related to higher scores. We first determined if convergence effects replicated previous studies, and next if the degrees of either form of matching were related to task performance. For all analyses, we tested whether the effects were dependent upon epoch or pair sex to check if the mixed results of past studies may be explained by the complex relationships between task performance, hierarchical temporal structure, speaking rate, and both forms of matching. While we did not have clear predictions about the directionality of effects when broken down by pair sex or epoch, it seemed possible that these conditions would be influential for matching strength.

3.4.2. Basic Findings by Epoch and Pair Sex for Participant Accuracy, Hierarchical Temporal Structure, and Speaking Rate

Participants were relatively accurate in finding the missing landmarks: 74.5% scored 90 points or more across all 6 maps, with only 9.7% scoring 80 points or less (M = 91.5 points; SD = 9.5). The average amount of time spent completing each trial was almost 6 minutes (347 seconds), ranging from approximately 3.5 to 13 minutes (212 and 770 seconds, respectively). Using a one-way within-subjects ANOVA, score was found to increase over time, with an average of 89 (±10) points in the beginning epoch and 93 (±9) points in the middle and end, $F(2, 160) = 15.94$, $p < 0.001$. Time to complete the task did not vary as a function of pair sex, $F(2, 40) = 0.45$, $p = .64$, nor did score vary by pair sex, $F(2, 40) = 0.85$, $p = .44$.

We next tested if AF slopes varied by either epoch or pair sex. For both models the predicted variable was AF slope, the predictor variable was either epoch or pair sex, and the error term was individual participant identification (ID). AF slopes were not significantly different for either epoch or pair sex (all $p > .05$, see Figure 10). The lack of AF effects for epoch or pair sex, compared with the large and reliable effects for speaking

rate, show that they are different prosodic measures that can reflect different aspects of spoken communication. In this case, speaking rate was sensitive to differences in conditions whereas hierarchical temporal structure was not.

The average speaking rate was 3.04 (±.41) words per second. A linear mixed effects regression (LMER) was first used to test for differences in speaking rates by epoch, where epoch was the predictor variable, speaking rate was the predicted variable, and ID divided by epoch was the error term. Results showed that speaking rate generally decreased from the beginning to the end of the experiment (see Figure 10), where speaking rate was fastest in the beginning epoch (beginning: $M = 3.09$, $SD = 0.42$; middle: $M = 3.02$, $SD = 0.41$; end: $M = 2.99$, $SD = 0.40$). The beginning had faster speaking rates than the end, $\beta = 0.18$, $t(152.06) = 3.27$, $p = .001$, and descriptively faster rates than the middle, $\beta = 0.10$, $t(152.19) = 1.69$, $p = .09$. The middle and end were not significantly different from each other, $\beta = 0.09$, $t(152.68) = 1.48$, $p = .14$. In a similar model, to test for pair sex differences, we replaced epoch with pair sex, and changed the random effect to participant ID. Male-male (MM) pairs spoke significantly faster than both female-female (FF) pairs, $\beta = 1.00$, $t(82.61) = 4.74$, $p < .001$, and mixed male-female (MX) pairs, $\beta = 0.71$, $t(83.44) = 3.32$, $p = .001$ (MM: $M = 3.26$, $SD = 0.37$; FF: $M = 2.86$, $SD = 0.35$; MX: $M = 2.99$, $SD = 0.41$). FF and MX pairs did not speak at significantly different rates, $\beta = 0.28$, $t(82.47) = -1.28$, $p = .20$.

AF slopes and speaking rates were negatively correlated across all participants and averaged over trials, $r(84) = -.28$, $p = .008$, which was expected given prior results showing that faster speaking rates correspond to flatter AF functions (Ramirez-Aristizabal et al., 2018). Since MM pairs spoke faster than FF and MX pairs, we were curious if the correlation between AF slope and speaking rate varied by pair sex. We used a LMER with AF slope as the predictor variable, speaking rate as the predicted variable, the interaction between AF slope and pair sex as a fixed effect, and individual participant ID as the random effect (these data were not averaged over trials). Contrary to our hypothesis, no significant effects or interactions between pair sexes were found (all $p > .05$; see Figure 11).

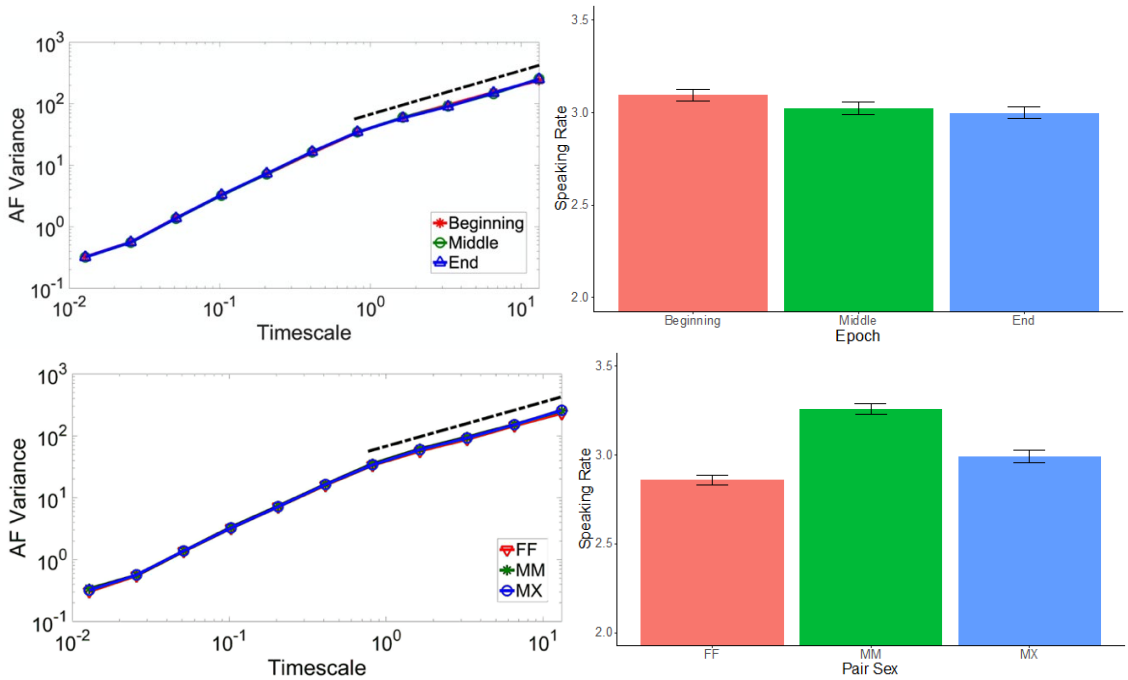

Figure 10. (Left) Mean AF variances are plotted as a function of timescale for the three different epochs and sex pairings in log-log coordinates. All regression lines were fit to the five longest timescales. Mean speaking rates are shown as a function of epoch and pair sex with standard error bars. (Right) Mean speaking rates are shown as a function of epoch and pair sex with standard error bars.

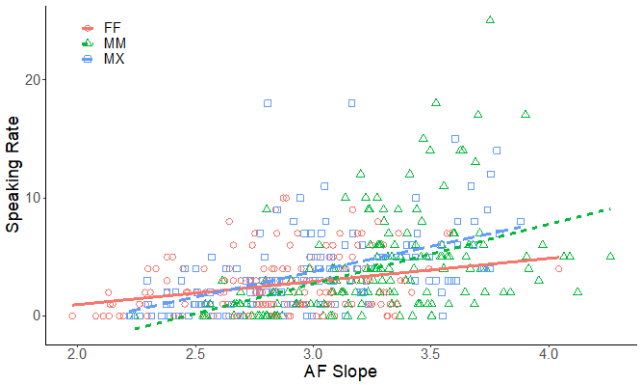

Figure 11. AF slope (the predictor variable) is plotted against speaking rate (the predicted variable) as a function of sex pairing, with regression lines for each condition. Note: All regression plots presented here do not reflect error terms for clearer visibility and because the plots differ minimally from the outputs which include error terms.

3.4.3. Hierarchical Temporal Structure (AF Slope) and Performance

We were interested in testing if hierarchical temporal structure was related to performance because this structure is reflective of prosodic exaggeration or the lack thereof. Since prosody allows speakers to display feelings of excitement, frustration, and boredom, it is important for successful communication, and one can therefore imagine it being important for tasks that require verbal interactions. Hierarchical temporal structure was displayed through AF slopes and performance was measured by the spatial accuracy and number of true differences found. Expectations for the degree of hierarchical temporal structure are based on its correlation with prosodic emphasis, including steeper AF slopes correlating with more prosodic exaggeration and vice versa. Steeper AF slopes have been found when adults speak to infants as compared to other adults (Falk & Kello, 2017) or when having a conversation compared to giving a professional talk for an audience (Kello et al., 2017). Exaggeration could reflect either the excitement of finding and confirming differences or frustration towards poor communication that impedes success, and monotone speech might reflect boredom or a serious tone about completing the task.

We tested our primary question of whether hierarchical temporal structure correlated with average dyad scores first by creating a LMER with AF slope as the predictor variable, score per trial as the predicted variable, the interaction between AF slope and trial as a fixed effect, and dyad as the random effect. A non-significant correlation was found between slope and score, $\beta = 0.58$, $t(430.57) = 0.94$, $p = .35$, but when we switched the interaction term from AF slope and trial to AF slope and epoch, so that the data were broken down into three groups instead of six trials, this new correlation ranged from being marginal to significant (beginning: $\beta = 1.23$, $t(459.05) = 1.97$, $p = .05$; middle: $\beta = 1.96$, $t(457.13) = 2.71$, $p = .007$; end: $\beta = 1.55$, $t(458.54) = 2.49$, $p = .01$; all interactions had $p > .05$; see Figure 12A). This finding suggests that more exaggerated speech correlated with higher scores and that the relationship was relatively consistent across epochs. On the other hand, these findings together indicate that the discrepancy between the models was due to the grouping of trials and initially a lack of power, but future studies should attempt to replicate this effect to ensure that it is consistent across studies.

We next tested whether this correlation depended on sex pairing by replacing the epoch term in the above model with pair sex. No effects were found for FF or MX pairs (all $p > .05$), whereas better performance corresponded with higher AF slopes for MM pairs only, $\beta = 4.07$, $t(456.40) = 7.18$, $p < .001$ (see Figure 12B). MM pairs differed from FF pairs, $\beta = 4.20$, $t(456.78) = 5.16$, $p < .001$, and MX pairs, $\beta = 4.95$, $t(455.90) = 5.08$, $p < .001$. These results show that more hierarchical temporal structure—or prosodically exaggerated speech—was correlated with better performance, but primarily for male-male interactions. This seems to indicate that when male speakers engaged with other males (but not females), their scores were reflected through their speech, where it might be that having more monotone speech displays boredom or a somber attitude and more exaggerated speech displays excitement about engaging in the task itself or doing well.

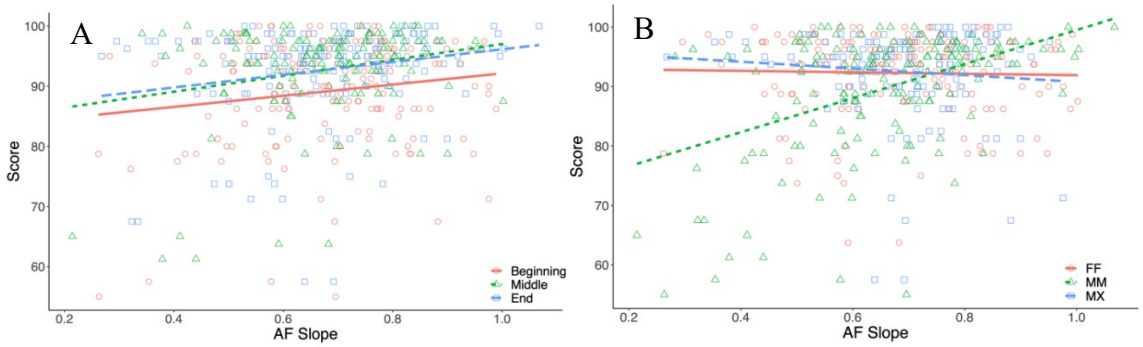

Figure 12. The predictor AF slope is plotted against the predicted averaged dyad score as a function of epoch (A) and pair sex (B), with regression lines for each group. Higher scores were found across epochs, but only for MM pairs with steeper AF slopes, or greater hierarchical temporal structure.

3.4.4. Speaking Rate and Performance

We next tested if speaking rate and score were related where one can imagine the rate of information transferal being important for task performance. More specifically, faster speaking rates could lead to better performance if this meant that more words produced in a shorter duration of time conveyed more useful information, but this may alternatively mean the opposite if faster rates reflect rushing through the experiment, indicative of a speed/accuracy tradeoff.

To test this relationship we created a LMER with speaking rate for the individuals as the predictor variable, score per trial as the predicted variable, the interaction between speaking rate and trial as a fixed effect, and dyad as the random effect. A negative correlation was found with slower speaking rates corresponding to higher scores, $\beta = 1.58$, $t(440.41) = 2.50$, $p = .01$. With the current findings and those from the above section in mind, it appears that speaking rate did display a speed/accuracy tradeoff in rushing through trials, whereas AF slopes generally expressed the excitement of finding and confirming differences. These findings broadly demonstrate that these two measures of prosody are distinct from one another.

We next tested whether this correlation with score varied over time by adding the interaction between epoch and speaking rate into the prior model. Speaking rate was found to correlate with score at the end, $\beta = -1.81$, $t(458.19) = -2.69$, $p = .007$, and marginally at the middle, $\beta = -1.26$, $t(457.43) = -1.83$, $p = .07$, but not at the beginning, $p > .05$ (the beginning was different from the end, $\beta = -2.08$, $t(457.75) = -2.28$, $p = .02$; see Figure 13A). This suggests that the relationship between speaking rate and score grew over time, and slower speaking rates were statistically predictive of higher scores and vice versa by the end of the experiment. This further implies that speaking rate was sensitive to epoch whereas hierarchical temporal structure was not. In other words, the detrimental effect of rushing through trials was strongest at the end, whereas participants expressed their performance through hierarchical temporal structure more consistently over epochs.

Like the prior section, we tested whether correlations with score depended on the pairing of sexes by replacing the epoch term in the above model with pair sex. No effects were again found for FF or MX pairs (all $p > .05$), but there was a strong negative relationship for MM pairs, $\beta = -3.21$, $t(457.02) = -4.57$, $p < .001$ (see Figure 13B). The interactions or differences between MM versus FF ($\beta = 4.04$, $t(456.53) = 3.97$, $p < .001$) and MM versus MX pairs ($\beta = 4.11$, $t(455.72) = 3.74$, $p < .001$) were reliable. Thus, the results from the present and above sections show that slower speaking rates and more exaggerated prosody, or greater hierarchical temporal structure, were correlated with better performance but primarily for male-male interactions.

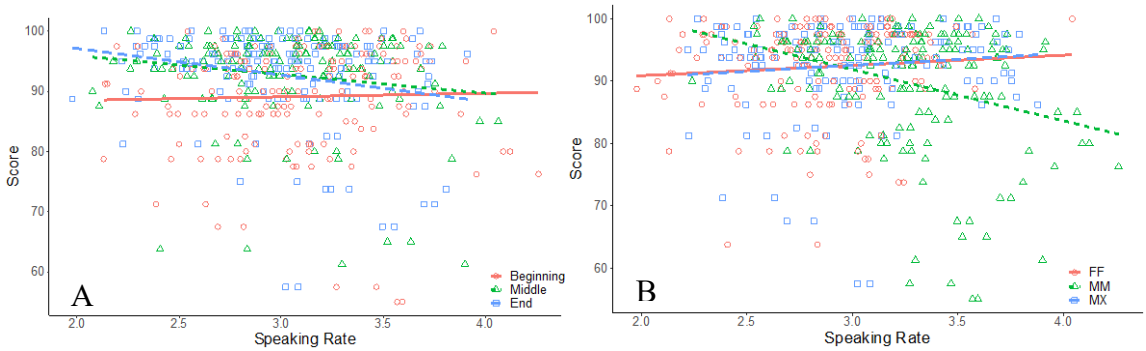

Figure 13. The predictor speaking rate is plotted against the predicted averaged dyad score as a function of epoch (A) and pair sex (B), with regression lines for each group. Higher scores were found across epochs but only for MM pairs with steeper AF slopes, or greater hierarchical temporal structure.

3.4.5. Complexity Matching

We next tested if we replicated the general effect of complexity matching found in prior studies, and expected we would do so (Abney et al., 2014; Ramirez-Aristizabal et al., 2018; Schneider et al., 2020). We used a LMER to predict the standardized AF regression slope for the more talkative speaker of each pair based on the regression slope of their less talkative partner, with dyad as the random effect. The slope of the more talkative speaker was found to significantly correlate with the slope of the less talkative speaker, $\beta = 0.26$, $t(230.34) = 4.41$, $p < .001$, meaning we found complexity matching. This replicates prior studies showing that the degree of hierarchical temporal structure in longer timescales of speech converges during vocal interactions (Schneider et al., 2020).

As in the prior models, we next tested if matching varied over time by adding the interaction between epoch and the predictor AF slope to the above model. Complexity matching was reliable across all three epochs (beginning: $\beta = 0.25$, $t(216.68) = 2.63$, $p = .009$; middle: $\beta = 0.36$, $t(213.73) = 3.55$, $p < .001$; end: $\beta = 0.23$, $t(212.64) = 2.84$, $p = .005$; see Figure 14A), with no marginal differences between epochs, all $p > .05$. In other words, matching was present across epochs and did not significantly vary between them. To test if matching depended on sex pairing, we replaced the interaction between AF slope and epoch with AF slope and pair sex. Convergence was found across the different sex pairings (FF: $\beta = 0.26$, $t(221.03) = 2.63$, $p = .009$; MM: $\beta = 0.25$, $t(226.91)$

= 2.64, p = .009; MX: β = 0.29, t(227) = 2.26, p = .03) and there were no interactions between these conditions, all p > .05 (see Figure 14B). Therefore, complexity matching was found across all epochs and pair sexed conditions.

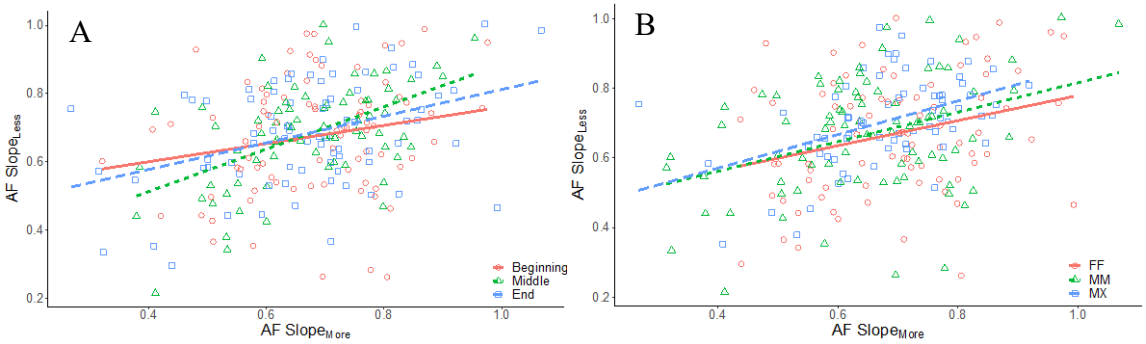

Figure 14. The predictor AF slope (the more talkative speaker) is plotted against the predicted AF slope (the less talkative speaker) as a function of epoch (A) and pair sex (B), with regression lines for each condition.

3.4.6. Speaking Rate Matching

Analogously to the above section on complexity matching, we expected to replicate the general effects of speaking rate matching (Cohen Priva, Edelist, & Gleason, 2017; Manson et al., 2013; Street, 1984). To test for matching, we used a LMER to predict the speaking rate for the more talkative speaker participant of each pair based on the speaking rate of their less talkative partner, with dyad as the random effect. As anticipated, speaking rates converged across the full dataset, β = 0.21, t(228.27) = 3.85, p < .001. Therefore, the effects of speaking rate matching were also replicated. To test if speaking rate matching varied over time, the interaction between epoch and the predictor speaking rate was added into the above model. Like complexity matching, rate matching was reliable across all three epochs (beginning: β = 0.25, t(214.22) = 3.79, p < .001; middle: β = 0.16, t(211.9) = 2.24, p = .03; end: β = 0.16, t(213.7) = 2.19, p = .03; see Figure 15B), and these values did not significantly vary from one another, all p > .05.

Finally, to test if these effects depended upon sex pairing, we again replaced epoch with pair sex. Unlike our findings for complexity matching, there was significant matching for MM pairs only, β = .30, t(226.99) = 3.10, p = .002 (p > .05 for FF and MX pairs), although matching for MM pairs did not interact with, and was therefore not reliably different from, matching for FF or MX pairs (β = 0.16, t(224.01) = 1.25, p = .21; β = 0.14, t(220.15) = 1.03, p = .30). It is descriptively interesting that there was no interaction because the scatter plot shows MM pairs to be different from both MX and FF pairs (see Figure 15B). Further, we previously found that MM pairs spoke faster than the other sexed pairs. To determine if the surprising lack of difference was due to the error term, we conducted an additional linear model where we averaged the speaking rates for each participant over all trials without an error term. Here, we found that convergence interacted between paired sexes and was stronger for MM versus MX pairs, β = 0.53, t(227) = 4.20, p < .001, and for MM versus FF pairs, β = 0.41, t(227) = 2.87, p = .005.

This suggests the discrepancy between the plot and findings was indeed due to the error term, and with the plot and other findings regarding speaking rate in mind, this suggests that MM pairs did indeed converge more than FF or MX pairs.

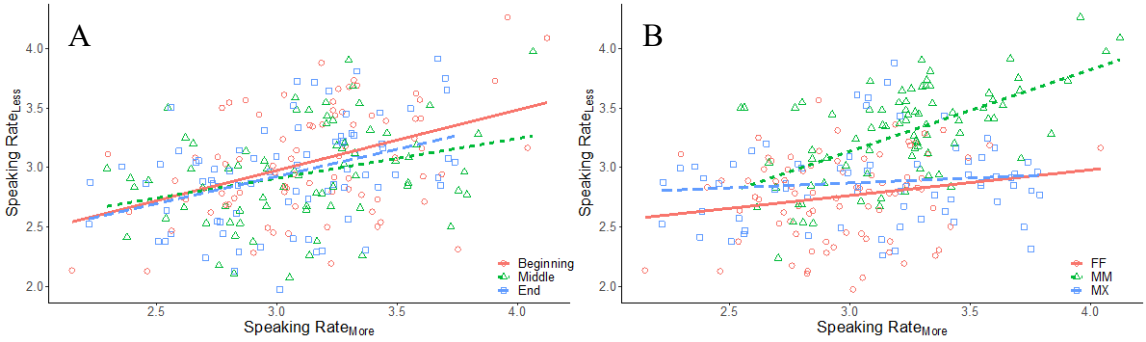

Figure 15. The predictor speaking rate for the more talkative speaker is plotted against the predicted speaking rate for the less talkative speaker as a function of epoch (A) and pair sex (B), with regression lines for each condition.

3.4.7. Measure of Matching

To summarize, we have shown thus far that task performance was reflected in the hierarchical temporal structures and speaking rates of individual male speakers when they were paired together. We have also seen that complexity matching occurred across all sex pairings and epochs, and speaking rate matching was found across all epochs, but only for MM pairs. These findings prepare us to investigate whether *convergence* in either measure is related to performance, where greater convergence is broadly related to greater feelings of rapport (Manson et al., 2013; Valdesolo et al., 2010) and thus may be related to the pair's ability to perform well.

To investigate these hypotheses, we needed a measure of complexity and speaking rate matching for each trial to analyze the average score per trial. Per-trial complexity matching was measured as the absolute difference between partner's AF slopes per trial, and likewise for speaking rate matching (with AF slopes being replaced by speaking rates). Before testing against the relationship to performance, we first checked if the matching measures were correlated by averaging each one over trials per participant, and then correlating the averages over dyads. The average per-trial measures of rate matching and complexity matching were not correlated with each other, $r(231) = -0.06$, $p = .38$, suggesting they reflect different, possibly complementary processes.

3.4.8. Complexity Matching and Performance

To test for a relationship between complexity matching and performance, we first ran a LMER with complexity matching as the predictor variable, average dyad score as the predicted variable, and dyad as the random effect. Complexity matching was not found to be statistically predictive of lower scores ($\beta = -0.75$, $t(208.26) = -1.64$, $p = .10$), although the plots trended such that better performance was reflected in greater

complexity matching. To be consistent with past models, we next tested if the null effect varied over time by adding the interaction between the predictor variable and epoch into the model. Matching was marginally more predictive of score for the beginning epoch, β = -1.25, $t(204.22) = -1.90$, $p = .06$, where the relationship descriptively reversed by the last epoch, β = 0.33, $t(195.06) = 0.45$, $p = .65$, but these effects were not statistically different from one another, $p > .05$ (see Figure 16A). We last tested whether the relationship between matching and performance varied by pair sex by replacing the interaction between matching and epoch with matching and pair sex, but found no significant effects, although more matching in the plot can be seen as relating to higher scores across pairs (see Figure 16B).

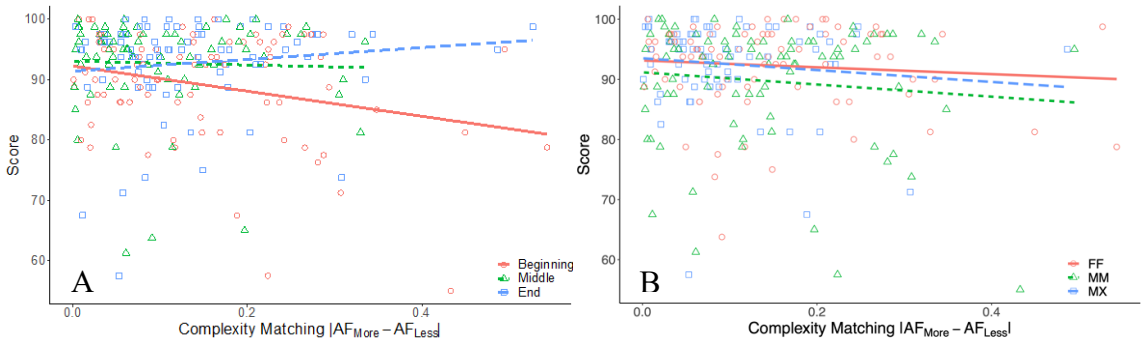

Figure 16. Predictor complexity matching as measured by the absolute difference between AF slopes for each pair are plotted against predicted average dyad score as a function of epoch (A) and pair sex (B), with regression lines for each group. Here a smaller difference indicates greater complexity matching. As with prior plots, pairs were divided into more talkative versus less talkative speakers.

3.4.9 Speaking Rate Matching and Performance

All of the models run here were identical to the above section, but complexity matching was replaced with speaking rate matching. Like complexity matching, speaking rate matching was also not predictive of lower scores (β = 0.74, $t(230.16) = 1.31$, $p = .19$), but the plots trended such that better performance can be seen reflecting less rate matching. Therefore these measures descriptively related to performance in opposing ways. Also in contrast to complexity matching, rate matching was negatively related to lower scores across epochs, but again not significantly so (beginning: β = .90, $t(219.91) = 1.01$, $p = .27$; middle: β = 0.50, $t(217.53) = 0.64$, $p = .53$; end: β = 0.56, $t(213.82) = 0.72$, $p = .47$; see Figure 17A). Together, these findings visually show that in the beginning epoch, there were contrasting—albeit non-significant—effects between speaking rate matching and complexity matching, where more speaking rate matching related to lower scores, and more complexity matching related to higher scores, but by the last epoch both forms of matching related to lower scores. Pair sex was again found to not be related to score, all $p > .05$, but in contrast to the findings for complexity matching, less speaking rate matching in the plot descriptively related to higher scores across sexes (see Figure

17B). Thus, there were no reliable effects overall, but the relationships between matching and performance descriptively varied in opposing directions by both pair sex and epoch.

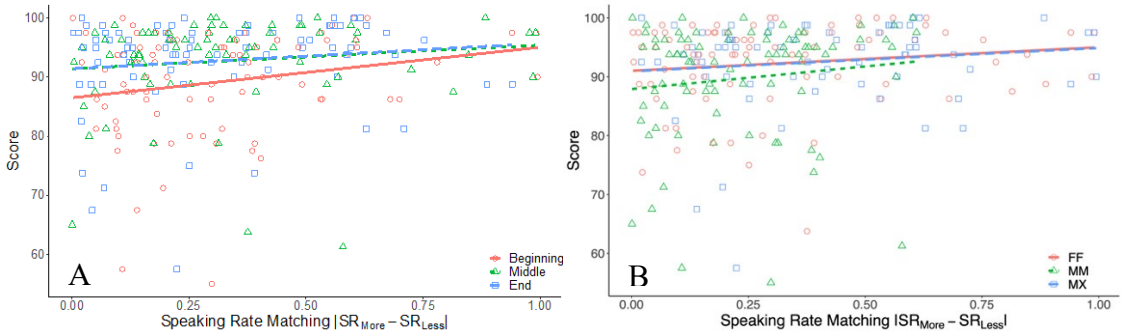

Figure 17. Predictor speaking rate matching as measured by the absolute difference between speaking rates for each pair are plotted against predicted average dyad score as a function of epoch (A) and pair sex (B), with regression lines for each group. A smaller difference is indicative of more matching. As with prior plots, speakers were divided into more talkative versus less talkative with respect to each pairing.

3.5. Discussion

Prosody is an important feature of communication that allows for clarity in meanings and emotional states. However, the relationship between prosody and performance on communicative tasks have gone relatively unexplored in prior studies, but correlations found between features of adult prosody and infant language acquisition are quite common. In contrast, convergence in spoken interactions is often observed, including in the literature surrounding complexity matching, speaking rate matching, lexical matching, and other forms of matching (Falk & Kello, 2017; Garrod & Pickering, 2009; Giles & Baker, 2008). Studies have more often analyzed various forms of convergence and found mixed results where it either aids, hinders, or is not related to different forms of task performance (Carmody et al., 2017; Fusaroli & Tylén, 2016; Ireland & Henderson, 2014; Manson et al., 2013).

In the present study, we examined prosody and prosodic convergence by comparing two different measures, both known to exhibit convergence, but neither bearing an established relationship to joint task performance. We measured hierarchical temporal structure and speaking rate as two different measures of prosody, the former based on variability across timescales, and the latter based on an average tendency. We replicated prior studies in that both measures converged during cooperative spoken interactions (Abney et al., 2014) like other demonstrations of alignment in language interactions (Giles & Baker, 2008; Pickering & Garrod, 2006; Schneider et al., 2020).

3.5.1. Summary and Interpretation of Results

We found that the two measures differed depending on who was interacting. Males spoke faster to each other while communicating to find their map differences, and

their speaking rate matching was stronger, compared with females or mixed pairs. In contrast, no such effect of pair sex was observed for mean AF slopes and complexity matching. These results indicate that our two measures of prosody reflect different and possibly complementary aspects of communication. In further support of this conjecture, AF slopes became steeper and speaking rates became slower as males in same-sex pairs performed the task better (especially in the first epoch), but no such relationship was observed for the other pair types. Together, these complementary patterns indicate that males performed better when they spoke more slowly and with more prosodic emphasis (or greater hierarchical temporal structure). Given that MM pairs were faster to speak overall, and they did not perform better than the other two groups, the lower performing males appeared to speak overly fast, enough to diminish prosodic emphases used to signal communicative information (Pardo et al., 2019; Yuan, Liberman, & Cieri, 2006). Some males may tend to speak faster while trying to dominate the conversation, which would appear to detract from task performance (Brescoll, 2011; Leaper & Ayres, 2007; Puts et al., 2007). These findings are similar to those from Manson et al. (2013) in that MM pairs differed from FF pairs, but our findings differed in that Manson et al. reported co-laughter convergence for MM pairs correlated with cooperation on the prisoner's dilemma task, whereas we found no statistically reliable effect of convergence on performance. Manson et al. (2013) also found speaking rate matching across speakers—though this effect did not appear to have been tested as a function of pair sex—whereas we only found this form of matching for speakers in MM pairs.

The complementary relationship between speaking rate and hierarchical temporal structure was further supported by descriptively (but not significantly) converse relationships between task performance, complexity matching, and speech rate matching. In the beginning epoch, more complexity matching was correlated with better task performance and more rate matching was correlated with worse task performance. By the end, less complexity matching and speaking rate matching correlated with higher scores. While we cannot overinterpret these findings due to their lack of statistical significance, this indicates that more convergence might be helpful in some respects for supporting effective information exchange in collaborative tasks, whereas *less* convergence is helpful in other respects, particularly at the end of a collaborative exchange. With the relatively mixed prior literature also in mind, finding these converse relationships in two different measures of prosody suggests effective communication may benefit from a balance of convergence and independence, so that partners resonate with each other but also independently contribute to the interaction and task (see also Abney et al., 2014; Fusaroli & Tylén, 2016).

3.5.2. Limitations and Future Directions

Convergence is a process that presumably must develop over the course of a spoken interaction when pairs are meeting each other for the first time. In our experiment, each partner needed to first gauge the prosody of the other in order to make adjustments that resulted in convergence. Therefore, we expected complexity and speaking rate matching to develop over epochs, but instead analyses yielded results that were significant across epochs (and were not significantly different from one another). We

speculate that matching developed quickly while participants talked during the first minute or two of the first trial. Our understanding of convergence and its role in communication would benefit from further investigation into the time course of different measures of matching depending on their temporal resolution.

Another direction for future research would be to examine gender roles instead of sex at birth. Gender identity information was not collected for the Montclair Map Task, as this was relatively uncommon at the time of data collection in 2013-2014, but studies have shown that interpersonal dynamics during spoken interactions can depend on gender roles (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Leaper & Ayres, 2007; Yuan et al., 2006), and gender roles are not necessarily tied to sex at birth, e.g. for gender non-binary and transgender participants (Matsuno & Budge, 2017). Future studies could include LBGTQIA+ participants to test whether the present results generalize to other gender identities, and to better examine the role of gender identity in prosody, prosodic convergence, and their roles in communication.

3.5.3. Conclusion

In conclusion, our findings reinforce the impression from prior studies that the relationships between prosody, prosodic convergence, and performance are complex and sometimes conflicting. Our study contributes to resolving prior opposing results by demonstrating how the relationship between speech and task performance depends on the measure of speech and the conditions of interaction. This dependence is important enough that even opposite relationships can be observed at times. We suggest that complementarity may underlie some of the complexity of findings in the literature. That is, performance can benefit from a simultaneous mixture of convergence and independence as a way of balancing the needs for pairs to resonate and adapt to one another, while still maintaining their identities and making independent contributions to the interaction. Our results also show the importance of identity in terms of sex and possibly gender roles, which has not been accounted for in many studies of convergence and performance to date. Theories of spoken communication will likely benefit from incorporating the complex dynamics indicated by the results herein.

Chapter 4

Hierarchical Temporal Structure and Convergence in Spanish-English Conversations

4.1. Preface

In the previous chapters, I tested the relationships between different forms of speech and speech convergence on two forms of "performance," the first being the infant's vocabulary score, and the second being pairs' combined scores on a difference-finding task. For both studies, the prosodic feature of hierarchical temporal structure was found to be a better moderator of performance than the measures of convergence tested. Extending from analyses of performance, this chapter assesses how speech and speech convergence vary by language condition. Spanish-English bilingual speakers were asked to engage in three conversational topics (movies, music, or television) using three language conditions (English only, Spanish only, and an inter-language condition where one participant spoke Spanish and the other English). With these conversations, I tested if hierarchical temporal structure, the convergence between hierarchical temporal structures (complexity matching), or lexical matching varied by language condition, conversational topic, or conversational order. For complexity matching, I also analyzed the timescales of analysis (short or long) within and across speakers. Across speakers, hierarchical temporal structure and measures of matching were consistent for all language conditions, and there were no effects of topic or order. Since this effect occurred when speakers were using different languages, this suggests convergence may be independent of one-to-one word alignment, where alignment may occur at the level of semantic representations. Complexity matching was only found at the long timescales for between-speaker matching, but was found at both timescales within speakers, suggesting hierarchical temporal structure is reflective of one's prosodic style of speech. A version of this chapter was published as an open-access manuscript, permitting unrestricted reuse and reproduction (see Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity Matching and Lexical Matching in Monolingual and Bilingual Conversations. *Bilingualism: Language and Cognition*, *23*(4), 845–857).

4.2. Introduction

More than half of the global population has been estimated to speak more than one language (e.g., Bialystok, Craik, & Luk, 2012; Grosjean, 2010; Romaine, 2012), and native or nearly native fluency can often be achieved when bilingual language learning starts at an early age (Perani et al., 2003). This allows proficient multilingual speakers to switch between languages when conversing with other speakers of the same languages with little difficulty (Fricke & Kootstra, 2016; Toribio, 2004). Sometimes the use of each language is asymmetric between speakers, where both speakers are bilingual in the same languages but each prefers using a different language (e.g., one speaker using Spanish while the other uses English). For example, in immigrant families the elder members may prefer to speak their heritage language whereas the younger members may prefer the

language of their new community, but each can understand what the other says (e.g., Park & Sarkar, 2007). This phenomenon of communicating in two different languages concurrently is known as *lingua receptiva*, and is a relatively normal mode of bilingual communication (Bahtina-Jantsikene & Backus, 2016; Bahtina, ten Thije, & Wijnen, 2013; ten Thije, Gooskens, Daems, Cornips, & Smits, 2017; ten Thije, 2013).

The prevalence of lingua receptiva raises the question of whether principles of language interaction in monolingual speech may apply to bilingual interactions as well. One of the most well-established principles of monolingual interaction is *interactive alignment* (Pickering & Garrod, 2004b), which broadly refers to the alignment or convergence of interacting levels of representation when speakers interact. For instance, one speaker in a conversation may use the more specific word "penny loafer" to refer to a shoe, and the other person may choose to use the same word or the more basic form of it, "shoe" (Brennan & Clark, 1996). If the person chose to converge on the same word usage, alignment would have occurred. Convergence can also occur at other levels of speech including, but not limited to, phonetic features (Kim, Horton, & Bradlow, 2011; Pardo, 2006), syntactic structures (Bock, 1986), and prosodic structure (Abney et al., 2014; De Looze et al., 2014; Xia, Levitan, & Hirschberg, 2014).

In the present study, we investigate speech convergence in open-ended conversations using two measures of matching that can be applied to Spanish-English bilingual speakers using either the same language or two different languages. Conversations were spoken in English, Spanish, or in a "Mixed" condition where one person spoke English and the other Spanish. The study of convergence in Mixed conversations is challenging because there are unlikely to be many direct correspondences between the surface forms of linguistic units or features, given different lexicons are used. To avoid this issue, we use *complexity matching* (Abney et al., 2015, 2014), a recent measure of speech convergence. Complexity matching can be directly applied to acoustic speech signals without any linguistic coding or translation required, allowing it to be equally applicable for measuring convergence in either the same or different languages. We also use a complementary measure of *lexical matching* (Brennan & Clark, 1996; Brennan, Kuhlen, & Charoy, 2018; Garrod & Anderson, 1987; Niederhoffer & Pennebaker, 2002) to determine if the semantic representations of speakers became aligned within and across languages. Our measure of lexical matching can be applied to open-ended conversations without a one-to-one matching of words, but it did require a degree of translation to estimate the semantic correspondences of lemmas used across languages.

We therefore aimed to compare our measures of matching within and across languages using two different measures: an acoustic measure of the physical speech signals produced, and a linguistic measure of the non-physical, semantic representations. Prior studies lead us to predict that both kinds of matching should occur in Spanish-only and English-only conversations, provided that speakers are sufficiently fluent in the language, but we are unsure whether conversations across languages might exhibit weaker signs of matching due to translation, or if matching is more basic to spoken interactions as a form of convergence.

In the following section, we review prior studies of convergence within and across languages, noting the need for measures spanning different levels of analysis that

may be applied more generally to open-ended conversations. We then describe our measures of complexity matching and lexical matching. Our experiment follows, and we end with a discussion of the implications of our results for theories of bilingualism, convergence, and language interaction.

4.2.1. Monolingual and Bilingual Speech Convergence

Whether bilingual pairs use the same or different languages, conversations are coordinated interactions where speakers lead, follow, and echo each other as they exchange ideas. For example, phonetic features like vowel quality and voice onset time often become more similar between conversational partners. This phenomena has been coined phonetic convergence and has been observed in monolingual conversations (Pardo, 2013; Pardo, 2006; Pardo, Gibbons, Suppes, & Krauss, 2012), and has been indirectly explored in at least one bilingual study (Sancier & Fowler, 1997). Phonetic convergence was measured by Pardo (2006) and Pardo et al. (2012) by having listeners judge the similarity of each speaker's phonetic production before, during, and after the conversation, with the most convergence occurring during the conversation but some convergence lingering post-interaction. In another example, Nielsen (2011) found that phonetic convergence depends on factors such as word frequency and voice onset time.

Extending from these monolingual studies, convergence in voice onset time has also been used to measure bilingual phonetic imitation (Tobin, Nam, & Fowler, 2017). For example, Balukas and Koops (2015) analyzed words spoken in conversational interviews by Spanish-English bilinguals from New Mexico. The words analyzed contained an initial /p/, /t/, or /k/ sound in both languages. Spanish was most participant's first language, and participants speaking Spanish were found to have voice onset time values within the normal range for monolingual Spanish. However, when speaking English, the voice onset times fell within the low range of monolingual English. Therefore, participants appeared to adjust phonemes of their non-dominant language towards the dominant language near code-switching events.

At more grammatical levels of language processing, syntactic priming is another form of convergence where speakers tend to produce (Bock, 1986; Healey et al., 2014) and comprehend (Branigan, Pickering, Liversedge, Stewart, & Urbach, 1995) new sentences using syntactic structures recently produced or heard. Syntactic priming is well-established in both monolingual conversations (Hardy, Messenger, & Maylor, 2017) and bilingual conversations since languages often share common syntactic constructions (Hartsuiker, Pickering, & Veltkamp, 2004). This makes syntactic convergence a popular and relatively easy measure to study in both monolingual and bilingual conversations.

While researchers have found evidence for syntactic convergence across languages, the language tasks used are often contrived and sometimes require confederates (Fleischer, Pickering, & McLean, 2012; Hartsuiker et al., 2004), though recent corpus-based studies on bilingual syntactic priming have provided a more naturalistic source of evidence (Gries & Koostra, 2017). For example, Hartsuiker et al. (2004) had Spanish-English bilingual participants talk about cards in English with a bilingual confederate who spoke in Spanish. Participants who heard a passive sentence in

Spanish were more likely to respond using a passive sentence in English, suggesting the integration of syntactic representations across languages.

Speech convergence has been theorized to benefit language interactions by helping to establish common ground (Brennan & Clark, 1996), affiliation (Manson, Bryant, Gervais, & Kline, 2013; Pardo et al., 2012), and better comprehension (Branigan et al., 1995; Schober & Clark, 1989). Convergence may either stem from domain-general processes of imitation (De Looze, Oertel, Rauzy, & Campbell, 2011; van Baaren, Holland, Steenaert, & van Knippenberg, 2003), possibly implemented through links between speech and language perception and production (Buchsbaum, Gregory, & Colin, 2001; Tian & Poeppel, 2012), or on-the-fly processes that arise to support convergence and understanding (Brennan & Hanna, 2009).

Findings of convergence in bilingual speakers suggest that similar mechanisms of matching underlie both bilingual and monolingual conversations (Fricke & Kootstra, 2016; Kroll, Dussias, Bogulski, & Kroff, 2012), which may or may not be symmetric between spoken languages (Blumenfeld & Marian, 2007). This rationale can be explained through the theory of interactive alignment (Garrod & Pickering, 2004; Pickering & Garrod, 2004b) where convergence stems from the alignment of language representations across multiple interacting levels of processing. To the extent that convergence is found across languages, interactive alignment argues that representations from each language must be aligned for communication to be successful. Representational structures may have direct correspondences across languages in some cases, such as shared discourse processes stemming from one culture with multiple languages. In other cases, alignment may require some degree of translation, as between Spanish and English lexicons. Costa, Pickering, and Sorace (2008) note that convergence may be weaker if speakers are less proficient in a language, and one could imagine a similar weakening if speakers need to maintain activation of two languages simultaneously while interacting.

The present study expands the current literature on speech convergence by applying *complexity matching* and *lexical matching* to conversations within and across languages. In the following two sections, we explain each measure and how they can be applied to these language conditions.

4.2.2. Hierarchical Temporal Structure and Complexity Matching

Complexity matching has been recently investigated in both speech and physical movements when pairs interact (Abney et al., 2014; Marmelat & Delignières, 2012; West et al., 2008). As detailed in Chapter 1, West et al. (2008) theorized that when complex networks interact by sending events back and forth to each other, the mutual impact on each other's peak event dynamics was maximal when the networks shared similar time intervals between events. In other words, information exchange was maximal when the network's power law dynamics were shared. While this founding concept was abstracted away from the behavior of humans, Abney et al. (2014) theorized that interactions between people should conform to the principles of complex networks, and found supporting evidence for this theory in conversations of agreement, but not disagreement.

The precondition of power law clustering across languages was supported in a recent study by Kello, Dalla Bella, Médé, and Balasubramaniam (2017). They measured

hierarchical temporal structure in TEDx talks spoken in six different languages, including English and Spanish, and found all sampled languages showed the same power law pattern. This result reflects the fact that speech is always comprised of phonemes nested within syllables nested within words and phrases and so on, regardless of the particular phonemes, syllables and words. Although the authors did not test for matching in this study, the observed similarity across languages suggests that complexity matching is not only possible, but predicted, in bilingual conversations.

The study of convergence in bilingual conversations is an opportunity to advance our understanding of both bilingualism and theories surrounding convergence. Bilingual conversations raise the question of whether convergence may occur in the sounds of speech, even when the phonemes, syllables, and words do not match. The fact that different languages employ different sounds and words does not prohibit the chance for complexity matching to occur, making this measure particularly valuable. Complexity matching has been measured as convergence in the hierarchical clustering of speech sounds, rather than individual units of sound or meaning at specific levels of analysis. As long as clustering is power law distributed across timescales, the theory of complexity matching predicts that the estimated exponents should converge within or across languages.

Complexity matching differs from other measures of matching in conversations in two important ways: (1) Complexity matching measures convergence in statistical ensemble measures, such as hierarchical temporal structure, rather than direct correspondences in linguistic units or features produced; and (2) Complexity matching has the precondition that statistics are power law distributed due to underlying network structures. In the present study, these distinguishing properties lead us to investigate complexity matching in monolingual versus bilingual conversations.

4.2.3. Word Frequency Distributions and Lexical Matching

In contrast to complexity matching, where convergence is measured between each speaker's hierarchical temporal structure based on physical acoustic speech sounds, lexical matching measures convergence between the nonphysical semantic representations of words. Lexical matching is a more traditional measure of convergence, where several past studies have found lexical matching in the words or referents speakers choose throughout conversations (Anderson, Garrod, & Sanford, 1983; Brennan & Clark, 1996; Clark & Brennan, 1991). As briefly mentioned, Brennan and Clark (1996) found that speakers formed "conceptual pacts" by converging on certain words, even if a word was more specific than its more common form, such as "penny loafer" versus "shoe." To measure lexical convergence, the authors quantified the probability of target words being produced when cued from trial to trial.

Lexical convergence has also been found in more open-ended speech exchanges. For instance, Nenkova, Gravano, and Hirschberg (2008) measured lexical convergence in conversations by measuring similarities in the proportions of times interlocutors (or pairs of speakers) used particular words. Levitan, Benus, Gravano, and Hirschberg (2015) also measured the convergence of turn-taking behaviors between speakers using *Kullback-Leibler Divergence*, which measures the degree to which one probability distribution is

contained within another. Levitan et al. (2015) compared Kullback-Leibler Divergence between conversational partners and surrogate pairs, and found less divergence—or more convergence—between partners than surrogate pairs.

While informative, we cannot apply the above methods to measure lexical matching between different languages in the present study because there would not be a clear mapping between their respective lexicons since different words are spoken. In one study, Ni Eochaidh (2010) found lexical convergence across languages by using a highly constrained English-Irish bilingual naming task that allowed for unambiguous mappings between the different referents. There are otherwise few empirical studies testing lexical convergence across languages, although a study by Bortfeld and Brennan (1997) suggests that less experience with a second language may not entirely interfere with the effect—they found lexical convergence to occur equally for language interactions in which speakers were more or less proficient in the language spoken.

In the present study, the phrase "lexical matching" refers to convergence in the frequencies of lemma usage, where lemmas abstract over the surface forms of words, and provide a more consistent basis for translation across languages. For example, variants of the word "run," including "ran," "runs," and "running," would each be counted as one instance of the root word "run." We used a variant of Kullback-Leibler Divergence known as *Jensen-Shannon Divergence* (JSD), which is a symmetric version of Kullback-Leibler Divergence. JSD is normalized so that a value of zero indicates identical probability distributions, whereas a value of one indicates non-overlapping distributions, meaning no lemmas shared by interlocutors. This measure requires a correspondence between words, which is likely to occur when conversational pairs are speaking the same language. To avoid this requirement becoming problematic when two different languages were spoken, we measured overlap by translating the lemmas of one language into those of the other. Lexical matching was then measured as significant differences from a baseline for conversations using the same and different languages. Therefore, lexical matching was based on both direct correspondences in the intra-language conditions and translations based on corresponding semantics in the inter-language conditions.

4.2.4. Current Experiment

In the present study, we investigated convergence in speech during naturalistic conversations in three language conditions using two diverse measures. These conditions were only English, only Spanish, or a Mixed condition where one speaker used English and the other used Spanish. We measured convergence as complexity matching and lexical matching, the former of which is a more novel acoustic measure, whereas the latter is a more linguistically traditional measure. We compared the degrees of matching across these conditions to test if the pure English condition replicated prior findings, or if the theories surrounding convergence could be extended to another language or inter-language condition, as reflected respectively through the pure Spanish and Mixed conditions. Although our sample of speakers came from a relatively homogenous participant pool, we analyzed if being dominant or non-dominant in Spanish and English had an effect on convergence when Spanish was spoken. We last tested whether the strength of complexity or lexical matching was correlated with each other, and we

expected that there would be a positive relationship under the hypothesis that they share a common basis via alignment across interacting levels of speech.

4.3. Methods

4.3.1. Participants

Sixty participants (males = 8, females = 52; mean age = 19.45) were recruited in pairs through the University of California, Merced SONA participant pool for course credit. Two pairs were omitted from analyses due to technical difficulties with the audio recordings, leaving 28 pairs (5 males, 51 females; mean age = 19.35) and a sample size comparable to prior studies that analyzed speech convergence in pairs (Abney et al., 2014; Falk & Kello, 2017; Marmelat & Delignières, 2012; Pardo et al., 2012; Pardo et al., 2019). Only three pairs reported knowing one another prior to the experiment, and these dyads were acquaintances. Participants filled out a questionnaire about their language proficiency and background (see Appendix A). Their native languages were reported as being Spanish (n = 24), English (n = 5), or both Spanish and English (n = 17). One participant listed Punjabi as their second language. Participants rated which language they used most comfortably on a daily basis, and 30 reported English, 14 both Spanish and English equally, 7 Spanish, and 5 had no response. Participants reported their native countries of origin as the United States (n = 39), Mexico (n = 13), El Salvador (n = 2), and both Mexico and the United States (n = 2). See Table 1 for additional details. A Spanish-English bilingual researcher confirmed that all participants were conversationally fluent in both languages.

Table 1.
Average proficiency ratings (with standard deviations in parentheses) for English and Spanish. Participant's self-reported reading, writing, and speaking proficiency scores were rated with a maximal score of 10, corresponding to strongest fluency. Frequency of use expresses how often each language is used weekly, where both languages could be rated as being used every day in a typical week, or 100% use per language. Non-dominant language corresponded to the participant's "second" language or L2, as self-reported on the first questionnaire (the single participant whose L2 was Punjabi was omitted from this table).

	English	Spanish
Age of acquisition	4.5 (3.2)	1.0 (1.6)
Reading proficiency	9.0 (1.2)	8.2 (1.5)
Writing proficiency	8.6 (1.4)	7.1 (1.9)
Speaking proficiency	9.3 (1.1)	8.6 (1.4)
Frequency of use for dominant language	88.2% (20)	77% (24.2)
Frequency of use for non-dominant language	81.8% (25.6)	61.8% (23.6)

4.3.2 Apparatus

Paired participants sat in a small experimental room (8.5’ by 7’) at a table while facing one another (approximately 2.5’ apart). Participants wore two Shure SM10A headset microphones while discussing three conversational topics. The conversations were recorded using these headsets, an M-Audio Mobile Pre-amp, and the Audacity 2.0.2 audio software (www.audacityteam.org).

4.3.3. Procedures

Participants were first asked to read and sign a consent form explaining that they would be participating in an approximately 20-minute-long study, but they could opt out and end the study at any time without penalty. After signing, participants either silenced or turned their cell phones off to avoid any disruptions. Participants next filled out a short questionnaire with demographic and language background questions (see Appendix A). Pairs were informed that they would be having three five-minute conversations with each other about three popular topics: either movies, music, or television (the order was randomized). One of the conversations was to be spoken in all English, one in all Spanish, and the other in a Mixed condition where each speaker was randomly assigned to one of the two languages for the duration of the conversation.

After completing the consent form and questionnaire, pairs were asked to introduce themselves to one another to help ease into the conversations as the researcher tested the audio quality and adjusted the input gain on the microphones so that they were relative to each speaker. Although no specific directions were given to do so, the initial conversations were all in English. To avoid confusion, the conversational and language conditions were written on pieces of paper and hung on the walls each speaker faced during every trial (e.g., Spanish movies). Once the researcher said either “begin” or “start,” participants engaged in each of the conversational trials. The researcher stayed in the room with the participants, but did not engage with them during the conversations and faced away from them towards the monitor to ensure the audio data were properly recorded. After the trials were completed, participants filled out a second questionnaire about how comfortable they felt communicating in each language on a Likert scale from 1-5, with 5 being the most comfortable (see Appendix B).

The conversations were each recorded to an uncompressed stereo WAV file, with the output of one microphone sent to the left channel and the other to the right. Although the input gain level was adjusted for each participant to ensure adequate recording levels while minimizing crosstalk between the microphones, some crosstalk occurred nevertheless. To help remove crosstalk, we used the audio software Audacity. We selected a noise profile based on crosstalk examples chosen manually from a visual display of the speech waveform and the experimenter listened to confirm these examples were crosstalk. The selected noise profile was applied to the whole recording to filter out acoustic energy that resembled crosstalk. For Audacity’s filtering function, the sensitivity parameter was consistently set to 25 and the frequency smoothing parameter to 3.

The Spanish and English speech from each recording was transcribed using TranscribeMe (www.transcribeme.com). Two researchers reviewed all the transcriptions

for quality control. Researchers confirmed that on average, dyads in the Mixed condition only accidentally switched languages about twice, and corrected themselves quickly thereafter. In the Spanish only condition, five dyads made one mistake each, and no participants made any mistakes in the English condition.

4.3.4. Hierarchical Temporal Structure and Complexity Matching

For each speaker, hierarchical temporal structure was measured as the amount of variance across timescales in peak amplitude events, where the events were derived from the original speech signal. A detailed explanation of hierarchical temporal structure and the methodology for measuring it (Allan Factor analysis) can be found in Chapter 1. While it is recommended the reader return to that chapter, to provide a brief summary here, the following operations were performed. Each waveform was converted into a series of acoustic peak events using two thresholds set to hold the number of peaks per sample constant across speakers. A log-log Allan Factor (AF) function was then computed for each event series, which quantifies the change or variance in peak clustering across timescales. AF was computed for 11 timescales ranging from approximately 20 ms to 20 sec, where these values varied slightly depending on minor differences in audio durations due to manually stopping the recording at the five-minute mark. Hierarchical temporal structure was quantified in both the shorter (1-6) and longer timescales (7-11) by fitting a regression line to each half of the AF function. The shorter timescales roughly corresponded to variability in the timing of smaller units of speech, such as phonemes, syllables, and words, whereas the longer timescales roughly corresponded to variability in the larger units of speech, like phrases and sentences.

Convergence between the slopes for each speaker's AF function, known as complexity matching, was measured as the strength of the correlations between slopes. Complexity matching was only observed at the longer timescales, where the durations of the units of speech roughly correspond to prosodic structure (Falk & Kello, 2017). Using correlations to measure complexity matching eliminated the need for surrogate analyses because correlations inherently have a baseline of zero to indicate when there is no linear relationship. A more detailed explanation of complexity matching may also be found above in Chapter 1.

4.3.5. Jensen-Shannon Divergence Analysis and Lexical Matching

Lexical matching in the probability distributions over each pair's word usage was measured. We quantified matching as the overlap in the probability distributions of the frequency of the words used by the speakers in each pair. Words were coded in terms of their underlying lemmatized roots using English and Spanish lemma dictionaries that replaced inflected words with their roots (derived from www.corpora.heliohost.org). If a word was not in the lemma dictionary, its originally transcribed form was preserved. For participants assigned to speak Spanish in the Mixed condition, one Spanish-English bilingual researcher listened to each conversation and translated the individual Spanish lemmas into their closest probable English counterparts. A second bilingual researcher

reviewed these translations and resolved any discrepancies in the translations with the other researcher.

For every lemma spoken, its probability of occurring was computed as its token frequency divided by the total number of lemma tokens used by that person in that conversation. As an illustrative example, Table 1 shows the 20 most frequent lemmas for one example dyad in the English-only language condition, and the same dyad in the Spanish-only condition. The number of unique English words in either the English or Mixed condition was significantly higher (M = 136.63, SD = 32.37) than the number of unique Spanish words in either the Spanish or Mixed condition (M = 116.30, SD = 27.76), $t(162.23) = 4.37$, $p < .001$.

The degree of overlap between each participant's lemma distribution was quantified using Jensen-Shannon Divergence (JSD), which is a symmetric, normalized extension of Kullback-Leibler divergence. The formula to compute JSD is as follows:

$$JSD(A,B) = \frac{1}{2}KL(A \parallel AB) + \frac{1}{2}KL(B \parallel AB)$$
$$KL(P \parallel Q) = -\sum_i P_i \ln\frac{P_i}{Q_i}.$$

JSD is the average Kullback-Leibler divergence (KLD) for each participant's probability distribution, A and B, relative to their combined probability distribution AB. JSD = 1 is reflective of totally non-overlapping frequency distributions, and JSD = 0 means the distributions are identical. Since this analysis was not correlational, JSD values were compared against a baseline to determine if the lemma distributions for a given dyad overlapped more than expected by chance. A surrogate JSD value was determined for each participant in each trial by pairing the individual's lemma frequency distribution with that of all other participants in a different dyad, but in the same language and topic condition. Surrogate JSD values were averaged and compared per dyad.

Table 2.
Two example conversations from one dyad, with one conversation spoken in all English and the other all Spanish. The 20 most frequently used lemmas are listed, and the lemmas spoken by both speakers are bolded.

English				Spanish			
Participant A		Participant B		Participant A		Participant B	
Lemma	Frequency	Lemma	Frequency	Lemma	Frequency	Lemma	Frequency
I	**29**	**be**	**30**	**que**	**18**	**no**	**12**
be	**18**	**I**	**26**	**y**	**16**	gustar	9
it	**18**	**not**	**20**	**de**	**12**	**pero**	**9**
like	**17**	**that**	**16**	sí	11	**que**	**8**
the	**16**	**it**	**16**	**escuchar**	**10**	**me**	**8**
movie	**14**	**the**	**13**	lo	9	en	7
yes	**12**	**have**	**13**	comer	9	mucho	6
of	9	a	12	el	8	entender	6
one	**9**	and	11	yo	8	**de**	**5**
that	**8**	**movie**	**11**	a	8	música	5
not	**8**	**like**	**11**	**no**	**8**	**y**	**5**
have	**7**	really	9	porque	7	su	5
do	**6**	so	9	**pero**	**7**	**escuchar**	**5**
know	6	to	9	ese	7	**ser**	**4**
you	**5**	**one**	**9**	ir	7	o	4
because	5	**but**	**8**	**me**	**6**	se	4
but	**5**	**yes**	**8**	**ser**	**6**	tener	4
they	5	**do**	**7**	estar	6	también	3
many	5	**you**	**7**	todo	5	bien	3
watch	5	good	6	tierra	5	he	3

4.4. Results

4.4.1. Hierarchical Temporal Structure and Allan Factor Analyses

The mean AF functions for each of the three language conditions are shown in Figure 18. The nearly straight lines indicate roughly self-similar nested clustering across timescales. The bend in the functions suggests that clustering was more nested in the shorter timescales, as reflected in a steeper slope to the curve on left side compared with the right.

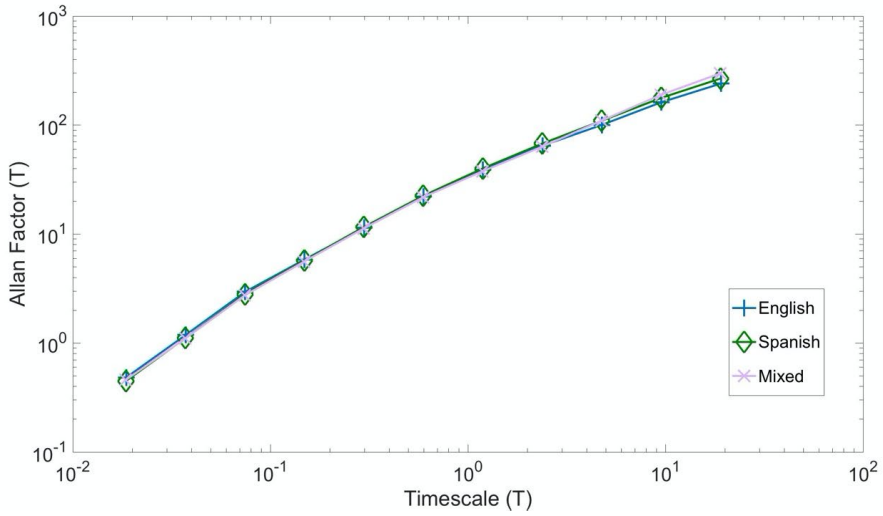

Figure 18. Averaged AF functions displaying the mean amount of hierarchical temporal structure at each timescale for the three language conditions.

We first analyzed if the slopes of AF functions differed by language condition or timescale. We ran a two-way repeated measures ANOVA with language (English, Spanish, Mixed) and timescale (short, long) as independent variables, the slope of the AF function as the dependent variable, and individual participants as the random variable. Slopes did not differ as a function of language, $F(2,324) = 0.41$, $p = .67$, $MSE = 0.01$, but they were steeper in the shorter timescales, $F(1,324) = 122.97$, $p < .001$, $MSE = 4.07$, confirming our visual inspection of Figure 18. No interaction between language and timescale was found, $F(2,324) = 0.18$, $p = .83$, $MSE = 0.006$.

To test for effects of order and conversation type, we ran a three-way repeated measures ANOVA with trial number, conversational topic, and timescale as the independent factors, slope as the dependent factor, and dyad as the random factor. No main effect was found for either trial or topic, and no interaction was found between the three independent factors (all $p > .05$), indicating that the effect of timescale did not influence effects of trial or topic.

4.4.2. Complexity Matching

We next analyzed overall complexity matching in AF functions across both timescales using a linear mixed effects regression (LMER), with one speaker's AF slope predicting their partner's. Dyad was set as the random effect with a random intercept and random slope. A reliable effect of overall complexity matching was found, B = 0.87, $t(52.8) = 18.35$, $p < .001$.

To test how complexity matching related to timescale, the interaction between timescale (short or long) and the predictor AF slope was added to the above model. Matching was reliable in the longer timescales, B = 0.54, $t(87.4) = 6.77$, $p < .001$, but not the shorter timescales, B = 0.18, $t(136.55) = 1.05$, $p = .3$, though the interaction with timescale was only marginally reliable, B = 0.36, $t(126.7) = 1.87$, $p = .06$ (see Figure 19A). In a similar model, we switched the interaction between the predictor slope and

timescale to be slope and language condition, and found complexity matching did not vary as a function of language condition (all $p > .2$; see Figure 19B). In two similar models, with only the interaction term adjusted, we also did not find effects of either conversational order or topic (all $p > .05$).

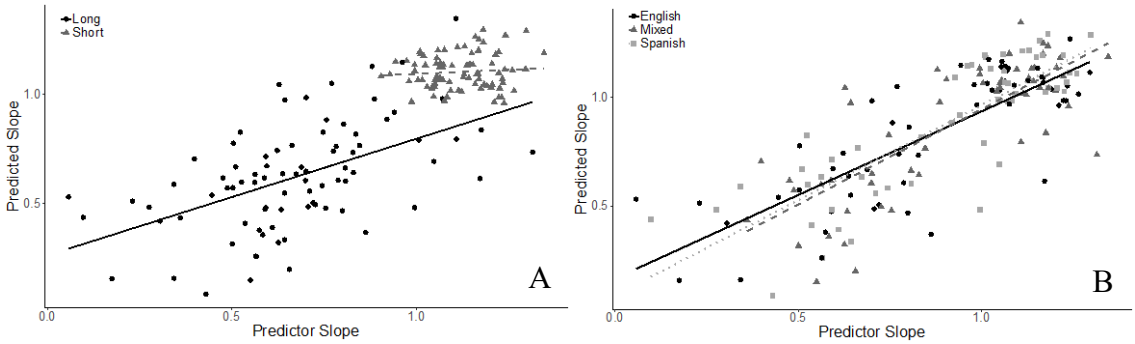

Figure 19. Predictor AF slopes plotted against predicted AF slopes as a function of timescale, either short or long (A), and language condition (B).

The results so far suggest that complexity matching does not require participants to speak the same language or to use the same words for their hierarchical temporal structures to become aligned. Therefore, in theory, complexity matching should be applicable to the same person compared with themselves speaking in the two different conversations because hierarchical temporal structure reflects the prosodic style of a person's speech, regardless of language or conversation. We consider this possibility because AF variance removes information about specific clusters of peak events, including the semantic meanings of the words spoken, and instead gauges their variance in cluster and duration sizes. Therefore, variance in clustering or hierarchical temporal structure might be comparable in an individual's speech across different languages, even though different units of speech are produced.

We tested within-speaker matching by running a LMER to statistically predict each participant's AF slopes from the English and Spanish conditions with their AF slopes in the Mixed condition. Half of the participants spoke English and the other half spoke Spanish in the Mixed condition. Each participant therefore provided one data point speaking the same language (English and Spanish were merged for this analysis), and one data point speaking different languages. AF slopes were correlated within individual speakers across different conversations, $B = 0.77$, $t(37.96) = 15.63$, $p < .001$, displaying within-speaker matching. The predictor slope did not reliably interact with language condition or timescale (all $p > .05$), suggesting within-speaker matching was not affected by the language spoken or timescale of interest, the latter being unlike our findings across speakers. In sum, these results indicate that speakers exhibit patterns of nested peak clustering across all measured timescales that are consistent with themselves across languages and conversations.

We next tested whether complexity matching varied as a function of language experience or background. Speakers varied in how they used the ratings scale, and the ratings were mostly subjective. Therefore, we instead focused on a simple binary categorization: If both members of a dyad listed Spanish as a native language and English

as a secondary language, the dyad was categorized as being Spanish primary (13 dyads), and otherwise English primary (15 dyads). Similarly to the above models, we used another LMER to predict the speaker's AF slope with their own AF slope, and included the fixed interaction between the binary variable of experience with the predictor AF slope. For the Spanish and Mixed conditions, the degree of complexity matching was not reliably affected by language experience, B = 0.10, $t(17.08) = 0.65$, $p = .53$, suggesting language fluency did not vary enough in our sample of participants to affect complexity matching. This likely reflects that our participants had similar language backgrounds overall, in that they were native Californians from families with largely Mexican heritage who spoke a Californian dialect of Spanish and used it on a regular basis.

4.4.3. Jensen-Shannon Divergence and Lexical Matching

We tested for matching in lemma usage using a three-way mixed design ANOVA, with language condition (English, Spanish, or Mixed) and JSD type (original or surrogate) as independent within-subjects factors, language experience as an independent between-subjects factor, JSD value as the dependent variable, and dyad as the random effect. A significant main effect of JSD type was found, $F(1,150) = 11.76$, $p < .001$, $MSE = 0.02$, indicating differences between the original and surrogate JSD values. Figure 20 shows that the original JSD values were less divergent than surrogates, meaning there was an overall effect of lexical matching for the original pairs, but not the surrogate pairs. This effect cannot be attributed to using words that are common to a given topic of conversation because the JSD surrogates were drawn from the same conversational topic as their corresponding originals, where there would presumably be differences in the conversations even if the topic was similar, like discussing different television shows.

There was also a main effect of language condition, $F(2,150) = 81.42$, $p < .001$, $MSE = 0.11$, reflecting that JSD values were most divergent (or least convergent) in the Mixed condition, followed by Spanish, and then English. This effect may be due to differences in the lemma dictionaries used, or inevitable issues with translation. The differences between original and surrogate values were importantly not reliably different for the Mixed versus pure language conditions.

We used the same ANOVA to test whether lexical matching (i.e., the difference between original and surrogate JSD values) interacted with or varied as a function of language condition (English, Spanish, or Mixed), and found no reliable difference, $F(2,150) = 0.38$, $p = .68$, $MSE < 0.001$, further confirming matching was observed during the Mixed and was not significantly different from the other two conditions.

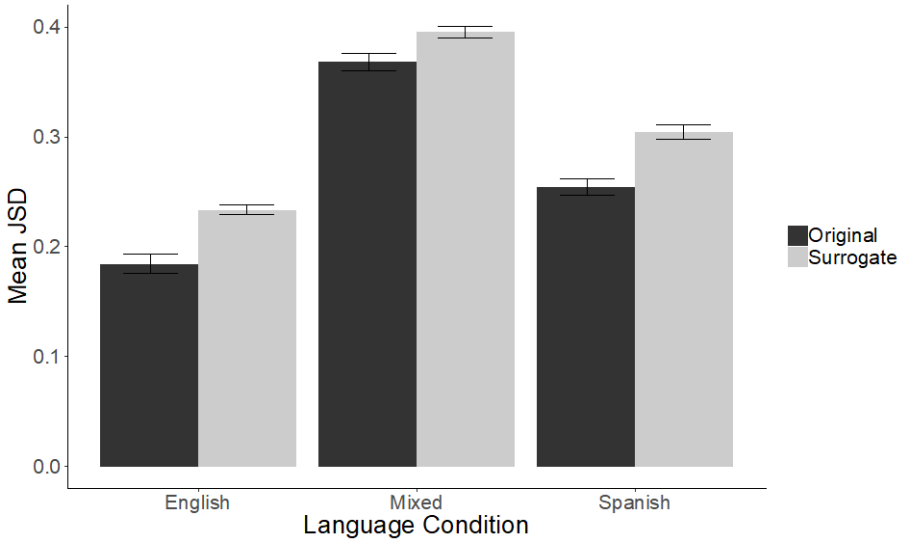

Figure 20. Mean Jensen-Shannon Divergence values (with standard error bars) for original versus surrogate pairings by language condition.

We also tested for effects of conversation order and topic on lexical matching using two additional three-way ANOVAs. To test for order effects, the trial number, language experience, and JSD type were set as independent variables, score as the dependent variable, and dyad as the random factor. The same ANOVA was used to test for conversational topic, where topic simply replaced trial number as an independent variable. Neither trial nor topic interacted with JSD type, all $p > .05$. In summary, like complexity matching, lexical matching appears to be robust to both intra- and inter-language interactions, and unaffected by variations in language experience in our participant sample.

Finally, we tested whether JSD score varied as a function of language experience and JSD type (original participant or surrogate). We again focused on a simple binary categorization: If both members of a pair listed Spanish as a native language and English as a secondary language, the pair was categorized as Spanish primary (13 pairs), and otherwise English primary (15 pairs). A three-way ANOVA tested the interaction between language experience (Spanish native or not Spanish native), JSD type (original or surrogate), and language condition (Spanish or Mixed). No interaction was found between these three conditions, nor was there a main effect of language experience, all $p > .05$. We therefore found no evidence of an effect of language experience, consistent with the similar lack of effect for complexity matching (see above), which may have been due to homogeneity of language fluency in our participant sample.

4.4.4. Relationship Between Complexity and Lexical Matching

To summarize, our results so far indicate that both complexity matching and lexical matching occur in inter-language Spanish-English conversations (i.e., the Mixed condition), with no reliable difference in the magnitude of matching compared with purely English or purely Spanish conversations. We last examined whether complexity

and lexical matching have a common basis by correlating their magnitudes. JSD difference values (original minus surrogate) are a direct measure of convergence in word usage for each given conversation, but we have so far only measured complexity matching at the aggregate level through AF slope correlations.

To provide a per-trial measure of complexity matching for each conversation, we computed the absolute differences of AF slopes in the longer timescales produced by each pair, which is inversely related to JSD difference scores. We ran a LMER with complexity matching as the predicted variable, the negative of the JSD difference scores (to undo the inverse relationship) as the predictor variable, and dyad as the random effect. Lexical matching was found to predict complexity matching and vice versa, B = .74, $t(82.0) = 3.71$, $p < .001$.

We next tested if the relationship between complexity and lexical matching was mediated by language condition by adding the interaction between JSD matching and language condition as a fixed effect into the above model. A marginal interaction was found for the Spanish and English conditions only, B = 0.40, $t(78) = 2.06$, $p = .04$. Upon further investigation, we found that the correlation between lexical matching and complexity matching was slightly stronger for only English, B = -0.95, $t(26) = -9.15$, $p < .001$, than only Spanish, B = -1.36, $t(26) = -7.75$, $p < .001$. The reason for this marginal effect is unclear and its unexpectedness warrants further investigation.

The observed relationship between complexity and lexical matching does not appear to be directly causal because word durations are mostly shorter than the second or greater timescales of complexity matching, and because surface forms of words are not directly matched in the Mixed condition. Therefore, overlap in the sounds of words was not the cause of complexity matching or vice versa. Instead, it appears that convergence may have an underlying basis that gives rise to both complexity and lexical matching.

4.5. Discussion

Bilingualism appears to be as natural and common to language development as monolingualism, suggesting that language processes have evolved to generally handle both linguistic conditions. This perspective led us to ask whether common principles and mechanisms apply to both cases. In the present study, we examined whether complexity matching—or the convergence of hierarchical temporal structure in speech sounds—occurs in monolingual or mixed bilingual conversations. We also tested if convergence occurs in the distributions of word usage. Our results provided clear evidence for both types of convergence in Spanish and English, as well as across the two languages. No effects of topic or order of the conversations were found, together demonstrating the robust and general nature of convergence in conversation.

4.5.1. Summary and Interpretation of Findings

Our primary finding was that both lexical and complexity matching were found across languages, and these effects were not modulated by either conversational order or topic. We also found that both forms of matching were correlated, suggesting they stem from common processes of convergence. For complexity matching, convergence was

only observed in the long timescales, but not the short timescales. This finding is consistent with prior studies indicating that prosodic and discourse processes may be more variable, and therefore more malleable to convergence. Since the longer timescales were in the range of hundreds of milliseconds to tens of seconds, whereas the shorter timescales ranged from tens to hundreds of milliseconds, AF slopes may be too coarse to detect any fine-grained effects of convergence between speakers like the voice onset times used in Balukas and Koops (2015). However, within-speaker complexity matching (or matching with one's self using different or the same languages) was found at both the long and short timescales within and across languages. This demonstrates how hierarchical temporal structure is reflective of one's style of speech rather than specific words, since different lexicons were used across languages, and complexity matching captures this convergence. In other words, we found participants matched with their own style of speech at both short and long timescales, but matching across speakers was only found at the long timescales, and we speculate that this difference is due to features of speech being more easily adjustable and more easily observed at the long timescales, at least for AF analysis.

The lack of differences observed for matching across languages is consistent with the cross-cultural phenomenon of lingua receptiva, where one bilingual speaker uses one preferred language and the other speaker uses the other available language (ten Thije, 2013). This suggests the processes underlying convergence appear to generalize over languages despite lexicon and structural differences. Lingua receptiva situations and code-switching while speaking are together relatively common experiences for bilingual speakers, where both languages are used interchangeably, and were likely also common experiences for our sample of Californian Spanish-English bilingual participants.

This observed equivalence of speech convergence can be explained in terms of interactive alignment theory, which proposes that interacting levels of representations converge when pairs speak with one another (Garrod & Pickering, 2009; Pickering & Garrod, 2004b; Trofimovich, 2016). This hypothesis seems to predict less matching across languages since some representations may not be directly aligned, and somewhat consistently with this, we found the timescales of analysis were affected differently. For example, prosodic style including voice quality and perhaps speaking rate may be similar across languages, which would allow complexity matching to be unaffected by using different languages. On the other hand, the use of different lexicons means that some aspects of speech, like phonological representations, cannot always become aligned during inter-language interactions. Since we found convergence across languages, our results seem to suggest that convergence is either a process independent of direct word alignment, or perhaps proficient speakers have learned how to share language processes and representations across languages (Guo & Peng, 2006; Kantola & van Gompel, 2011; Kroll, Bobb, & Hoshino, 2014; Marian & Spivey, 2003).

4.5.2. Future Directions and Limitations

Beyond direct lexical alignment, it is possible for language use to converge in the probability distributions of word usage and other levels of representation. Specifically, Zipf's law refers to the phenomenon of ranking word usage—theoretically in either

written or spoken text—and the frequencies of use producing a power law distribution. If we assume that speakers broadly produce such power law distributions (Zipf, 1935), as found in prior studies across different languages (e.g., Peterson, Tenenbaum, Havlin, Stanley, & Perc, 2012), then we may expect the shape of the power law, or its exponent, produced by one speaker to bend towards their conversational partner's and vice versa. However, since our trials were too short with too few words spoken, we could not test this hypothesis in the present study. With more data collected through longer interactions, it should be possible to analyze lexical matching using Zipf's law.

It would also be interesting to test both lexical and complexity matching in bilingual interactions between pairs of languages that vary in their phonological, grammatical, and lexical similarity. While we did not find any effects of language experience on complexity or lexical matching, because our participants were drawn from a fairly homogenous population of Spanish-speaking Californians with family roots primarily in Mexico, future studies should aim to analyze convergence in a wider range of language fluencies and backgrounds of speakers. With the results from Chapter 2 in mind, where flatter AF slopes corresponded to better speech production, we are curious if flatter AF slopes also correspond to more proficient speech.

One limitation of using AF analysis is that approximately four to five minutes of speech are required in order to accurately measure hierarchical temporal structure. Therefore, while we found no conversational order effects on either type of matching, we are unsure of exactly when complexity matching began within the first conversation. JSD likewise requires an entire word distribution over the course of each conversation, meaning it is also unclear when lexical matching began within the first conversation. Future studies should aim to pair these measures with more fine-grained measures of temporal analysis so that the time course of convergence can be better understood.

4.5.3. Conclusion

The present study explored speech convergence using two measures, complexity matching and lexical matching, during inter- and intra-language conversations. Our findings demonstrate that these forms of matching are robust to different language interactions. These measures of convergence appear to reflect basic principles of social interaction and shared processes of inter- and intra-language interactions.

Chapter 5

General Conclusions

5.1. Overview of Findings

Prior studies have established that prosody provides significant contributions to communication, and that forms of speech convergence including prosodic convergence often occur between speakers. However, few studies overall have explored the correlations between prosody, prosodic convergence, and various forms of performance, despite the potential for meaningful relationships. Furthermore, few studies have investigated how convergence is affected by different languages spoken. I therefore aimed to use this dissertation to analyze how prosody and prosodic convergence either vary by or are correlated with early language interactions (Chapter 2), task performance (Chapter 3), and bilingual conversations (Chapter 4).

To provide a summary of the primary results, in the second chapter, only the adult's hierarchical temporal structure was correlated with infant lexical ability, where less hierarchical temporal structure (visualized as flatter AF slopes) related to stronger lexicons. This conceptually replicates the findings from Abney et al. (2017), who found more advanced speech-related vocalizations were correlated with flatter AF slopes, as compared to non-speech-related vocalizations. Though not reliably so, the correlation between the infant's hierarchical temporal structure and their own vocabulary score trended in the same way for the two older age groups (9 and 18 months), but not the younger age groups (3 and 6 months), so that older infants displayed flatter AF slopes. In contrast, neither volubility, volubility matching, nor complexity matching were statistically predictive of vocabulary strength.

In Chapter 3, individual hierarchical temporal structure and speaking rate were correlated, but this effect was mediated by pair sex, where it was only observed for male-male pairs, but not female-female or mixed-sex pairs. Speaking rate was also moderated by epoch. Somewhat like Fusaroli et al. (2013), who found a correlation between complexity matching and their measure of performance in the second session, but not the first session, our correlation between score and speaking rate only became significant by the second epoch. These findings together demonstrate that it may take time for such a relationship to develop. We also found complexity matching between speakers, broadly replicating prior studies of convergence (Abney et al., 2014). Some speaking rate matching was found, but only between male-male pairs and not female-female or mixed-sex pairs. This broadly replicates and extends from the prior findings by Manson et al. (2013) who generally found speaking rate convergence, but also found differences between pair sexes, where male-male pairs had more laughter convergence than female-female pairs. In contrast to the findings with individual hierarchical temporal structure and speaking rate, neither form of matching was significantly related to score, but the relationships were descriptively in contrast, suggesting they reflect different forms of convergence despite both being prosodic measures of convergence.

Finally, in Chapter 4 I expanded on the current literature surrounding interactive alignment and bilingualism by analyzing convergence in inter- and intra- language

conditions. Hierarchical temporal structure was found whether participants were speaking English, Spanish, or speaking either language while their partner spoke the other (the Mixed condition). Both complexity matching and lexical matching were also observed regardless of the language spoken. This is particularly meaningful because the linguistic units themselves could not have been aligned during the Mixed condition, implying convergence must have been occurring at other levels, including the semantic level of representation observed via lexical matching. No effects of conversational topic were observed whether analyzing prosody or either form of convergence. However, complexity matching was only found at the longer timescales, and not at the shorter ones across speakers, but was found at both timescales within speakers when using different languages. This suggests hierarchical temporal structure is reflective of prosodic style of speech within speakers, but that complexity matching (at least as currently measured) may not be sensitive to convergence across speakers at the smaller units analyzed.

Across these chapters, several contributions to the existing literature were made. First, I found hierarchical temporal structure to be robust across conversations regardless of the conversational topic, pair sex, epoch, or language spoken. Building off of prior studies (Falk & Kello, 2017; Kello et al., 2017; Ramirez-Aristizabal et al., 2018), hierarchical temporal structure was again shown to be reflective of one's prosodic style of speech because complexity matching at the short timescales was found only within speakers. In other words, speakers matched with themselves, but not with others, at this level of speech.

Chapters 2 and 3 together demonstrated that greater hierarchical temporal structure is correlated with more prosodically exaggerated speech. That is, Chapter 2 showed greater hierarchical temporal structure when either addressing or speaking near less lexically advanced infants (the speaker directionality was not guaranteed since the recordings were created over the course of several hours as participants went about their day normally), and Chapter 3 showed greater hierarchical temporal structure was used by higher performing speakers in male-male pairs only. In more colloquial terms, this suggests more prosodic exaggeration is used when addressing young infants, and when excited about performing well on a collaborative task, at least for male speakers working with other males.

Hierarchical temporal structures were found to converge—or exhibit complexity matching—at the longer timescales across adult speakers, regardless of the language condition, pair sex, or conversational topic. Notably, all of the topics analyzed (find the differences, movies, music, and television) were relatively friendly, whereas argumentative conversations have been shown to not display complexity matching (Abney et al., 2014). Extending from adult conversations, infant-adult interactions were also shown to not display either complexity matching or volubility matching, which was surprising by the time infants reached 18 months. However, this null effect may have been due to the naturalistic character of the recordings. This leaves the question of what age interactive alignment between speakers begins to occur open for future studies to test.

While hierarchical temporal structure was at least somewhat predictive of both measures of "performance" (lexical ability or task performance), complexity matching was not related to either form. Furthermore, volubility matching was not related to lexical ability, nor was task performance related to speaking rate matching. Therefore, against

my expectations, convergence as a whole does not appear to be a particularly good measure of performance, though exceptions like co-laughter exist (Manson et al., 2013). Therefore, future studies may either find it useful to focus on individual measures of prosody in relation to performance, or to explore different types of tasks and measures of performance to continue to determine why convergence seems to be less related than the individual measures.

5.2. Future Directions and Limitations

This dissertation leaves several research questions open for future experiments. As mentioned above, the age at which interactive alignment begins to occur between children and caregivers (or others) is still unclear because we did not observe either complexity or volubility matching between them. However, this null effect may have been due to the fact that infants and caregivers may not have been addressing one another, despite both vocalizing near one another within the same period of time. To clarify the reason behind the lack of an effect, future studies may to some extent benefit from using more contrived tasks like Newman et al. (2016), simply so that researchers are certain of the directionality of the speech used by adults. Also briefly stated, it may be beneficial for future studies to continue to explore if the lack of a relationship between convergence and performance are specific to only the tasks described here, or if it extends to other tasks, like the prisoner's dilemma (Manson et al., 2013).

Since less hierarchical temporal structure (or flatter AF slopes, and less prosodic exaggeration) is related to infants with greater vocabulary sizes, speech-related vocalizations (Abney et al., 2017), and more advanced adult-directed speech as compared to infant-directed speech (Falk & Kello, 2017), it would be interesting to test if more proficient bilingual speakers also display less hierarchical temporal structure. If so, this would further confirm that more monotone speech is used either by or directed toward more advanced speakers. Likewise, it would be interesting to test if any relationships between hierarchical temporal structure and performance vary by the language spoken or linguistic situation (like a lingua receptiva scenario).

One limitation of using AF analysis to measure hierarchical temporal structure is that it requires at least a few minutes of audio data in order to be analyze the longer timescales. In turn, I was unable to study the time course of matching on a fine-grained temporal scale, meaning how convergence unfolded during the first few minutes of the interactions. It would be beneficial for future studies to pair AF analysis with complementary analyses that are more temporally friendly. Another limitation of using AF analysis is that its relationships with measures of pitch—and to some extent turn-taking—are less understood, and future studies should explore if and how AF analysis is affected by these two measures.

A final limitation of this dissertation is that while we know hierarchical temporal structure measured through AF analysis is related to variances in speaking rate and amplitude, it is unclear how pitch does or does not influence this measure. Thus, it would also be useful for future studies to pair analyses of hierarchical temporal structure with the fundamental frequencies of the speakers to better inform our understanding of this measure.

5.3. Conclusion

Prosody is a useful communicative device for studying measures of performance, but has gone relatively unanalyzed in prior studies. Considering several features compose prosodic structure, and that many measures of individual and collaborative performance are possible, this area of research is still rich for future exploration. Likewise, though I did not observe any relationship between the analyzed forms of prosodic convergence and performance, prior research did observe such relationships, indicating this area of research is still broadly complex and sometimes inconsistent. Hierarchical temporal structure is of particular use when studying prosody and prosodic convergence because it is indifferent to diverse language conditions and linguistic ability, making it a useful measure for acoustic data that may otherwise be difficult to compare.

References

Abney, D. H., Kello, C. T., & Warlaumont, A. S. (2015). Production and Convergence of Multiscale Clustering in Speech. *Ecological Psychology*, *27*(3), 222–235.

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology*, *143*(6), 2304–2315.

Abney, D. H., Warlaumont, A. S., Oller, D. K., Wallot, S., & Kello, C. T. (2017). Multiple Coordination Patterns in Infant and Adult Vocalizations. *Infancy : The Official Journal of the International Society on Infant Studies*, *22*(4), 514–539. https://doi.org/10.1111/infa.12165

Ambrose, S. E., VanDam, M., & Moeller, M. P. (2014). Linguistic input, electronic media, and communication outcomes of toddlers with hearing loss. *Ear and Hearing*, *35*(2), 139–147. https://doi.org/10.1097/AUD.0b013e3182a76768

Anders, J., Hovy, D., & Søgaard, A. (2015). Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*.

Anderson, A., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., … Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 1–21. https://doi.org/10.1177/002383099103400404

Anderson, A., Garrod, S. C., & Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *The Quarterly Journal of Experimental Psychology Section A*, *35*(3), 427–440. https://doi.org/10.1080/14640748308402480

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally Interacting Minds. *Science*, *329*(5995), 1081–1085. https://doi.org/10.1126/science.1185718

Bahtina-Jantsikene, D., & Backus, A. (2016). Limited common ground, unlimited communicative success. *Philologia Estonica Tallinnensis*, (1), 17–36.

Bahtina, D., ten Thije, J. D., & Wijnen, F. (2013). Combining cognitive and interactive approaches to lingua receptiva. *International Journal of Multilingualism*, *10*(2), 159–180. https://doi.org/10.1080/14790718.2013.789521

Balukas, C., & Koops, C. (2015). Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingualism*, *19*(4), 423–443. https://doi.org/10.1177/1367006913516035

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, *12*(2), 120–138. https://doi.org/10.1007/BF00986930

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250. https://doi.org/10.1016/j.tics.2012.03.001

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication*. https://doi.org/10.1016/0271-5309(88)90016-X

Blumenfeld, H. K., & Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes*, *22*(5), 633–660. https://doi.org/10.1080/01690960601000746

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. https://doi.org/10.1016/0010-0285(86)90004-6

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, *23*(2), 119–147. https://doi.org/10.1080/01638537709544986

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147. https://doi.org/10.1177/00238309010440020101

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*(6), 489–506.

Branigan, H. P., Tosi, A., & Gillespie-Smith, K. (2016). Spontaneous lexical alignment in children with an autistic spectrum disorder and their typically developing peers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000272

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Brennan, S. E., & Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, *1*(2), 274–291. https://doi.org/10.1111/j.1756-8765.2009.01019.x

Brennan, S. E., Kuhlen, A. K., & Charoy, J. (2018). Discourse and dialogue. In J. T. Wixted & S. L. Thompson-Schill (Eds.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Language and Thought* (Vol. 3). Wiley.

Brentari, D., & Crossley, L. (2002). Prosody on the hands and face. *Sign Language & Linguistics*, *5*(2), 105–130. https://doi.org/10.1075/sll.5.2.03bre

Brescoll, V. L. (2011). Who Takes the Floor and Why: Gender, Power, and Volubility in Organizations. *Administrative Science Quarterly*, *56*(4), 622–641. https://doi.org/10.1177/0001839212439994

Buchsbaum, B. R., Gregory, H., & Colin, H. (2001). Role of left posterior superior

temporal gyrus in phonological processing for speech perception and production. *Cognitive Science*, *25*(5), 663–678. https://doi.org/10.1207/s15516709cog2505_2

Buder, E. H., Warlaumont, A. S., & Oller, D. K. (2013). An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective. In B. Peter & A. A. N. MacLeod (Eds.), *Comprehensive Perspectives on Child Speech Development and Disorders: Pathways from Linguistic Theory to Clinical Practice* (pp. 103–134). Nova Science Publishers, Inc.

Buller, D. B., LePoire, B. A., Aune, R. K., & Eloy, S. V. (1992). Social perceptions as mediators of the effect of speech rate similarity on compliance. *Human Communication Research*, *19*, 286–311.

Byrne, D., Dillon, H., & Tran, K. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America*, *96*(4), 2108–2120. https://doi.org/10.1121/1.410152

Camp, B. W., Burgess, D., Morgan, L. J., & Zerbe, G. (1987). A Longitudinal Study of Infant Vocalization in the First Year. *Journal of Pediatric Psychology*, *12*(3), 321–331. https://doi.org/10.1093/jpepsy/12.3.321

Carmody, P. C., Mateo, J. C., Bowers, D., & McCloskey, M. J. (2017). Linguistic Coordination as an Unobtrusive, Dynamic Indicator of Rapport, Prosocial Team Processes, and Performance in Team Communication. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 140–144. https://doi.org/10.1177/1541931213601518

Cirelli, L. K., Einarson, K. M., & Trainor, L. J. (2014). Interpersonal synchrony increases prosocial behavior in infants. *Developmental Science*, *17*(6), 1003–1011. https://doi.org/10.1111/desc.12193

Cirelli, L. K., Wan, S. J., & Trainor, L. J. (2014). Fourteen-month-old infants use interpersonal synchrony as a cue to direct helpfulness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130400. https://doi.org/10.1098/rstb.2013.0400

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, L. B., M. John, S. Teasley, & D. (Eds.), *Perspectives on Socially Shared Cognition*. American Psychological Association.

Coco, M. I., Dale, R., & Keller, F. (2018). Performance in a Collaborative Search Task: The Role of Feedback and Alignment. *Topics in Cognitive Science*, *10*(1), 55–79. https://doi.org/10.1111/tops.12300

Cohen Priva, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, *141*(5), 2989–2996. https://doi.org/10.1121/1.4982199

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant-directed Speech in the First Month after Birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.1111/j.1467-8624.1990.tb02885.x

Costa, A., Pickering, M., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes*, *23*(4), 528–556.

Culpeper, J. (2011). It's not what you said, it's how you said it!": Prosody and impoliteness. In S. Mills (Ed.), *Discursive Approaches to Politeness* (pp. 57–83). Berlin: De Gruyter Mouton.

Cummins, F. (2018). *The ground from which we speak: Joint speech and the collective subject*. Cambridge Scholars Publishing.

Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern-day wild child*. https://doi.org/10.1016/C2013-0-07305-7

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

D'Imperio, M., Gorka Elordieta, S. F., Prieto, P., & Vigário, M. (2005). Intonational phrasing in Romance: the role of syntactic and prosodic structure. In S. Frota, M. Vigário, & F. João (Eds.), *Prosodies (With Special reference to Iberian Languages)* (pp. 59–98). Berlin: Mouton de Gruyter.

Dachkovsky, S., & Sandler, W. (2009). Visual Intonation in the Prosody of a Sign Language. *Language and Speech*, *52*(2–3), 287–314. https://doi.org/10.1177/0023830909103175

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. https://doi.org/10.1121/1.1613311

De Looze, C., Oertel, C., Rauzy, S., & Campbell, N. (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *International Conference on Phonetic Sciences* (pp. 1294–1297).

De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*. https://doi.org/10.1016/j.specom.2013.10.002

Dellwo, V., & Wagner, P. (2003). Relationships between rhythm and speech rate. In *15th International Congress of the Phonetic Sciences* (pp. 471–474). Barcelona, Spain.

Erekson, J. A. (2010). Prosody and Interpretation. *Reading Horizons: A Journal of Literacy and Language Arts*, *50*(2).

Falk, S., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80–86. https://doi.org/10.1016/j.cognition.2017.02.017

Fasolo, M., Majorano, M., & D'Odorico, L. (2008). Babbling and first words in children with slow expressive development. *Clinical Linguistics & Phonetics*, *22*(2), 83–94. https://doi.org/10.1080/02699200701600015

Fay, N., Ellison, T. M., Tylén, K., Fusaroli, R., Walker, B., & Garrod, S. (2018). Applying the cultural ratchet to a social artefact: The cumulative cultural evolution

of a language game. *Evolution and Human Behavior*, *39*(3), 300–309. https://doi.org/10.1016/J.EVOLHUMBEHAV.2018.02.002

Feinstein, A., Brown, R., & Ron, M. (1994). Effects of practice of serial tests of attention in healthy subjects. *Journal of Clinical and Experimental Neuropsychology*, *16*(3), 436–447. https://doi.org/10.1080/01688639408402654

Feldman, R. (2006). From biological rhythms to social rhythms: Physiological precursors of mother-infant synchrony. *Developmental Psychology*. https://doi.org/10.1037/0012-1649.42.1.175

Feldman, R. (2007). Parent–infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. *Journal of Child Psychology and Psychiatry*, *48*(3–4), 329–354. https://doi.org/10.1111/j.1469-7610.2006.01701.x

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98–116. https://doi.org/10.1037/0012-1649.42.1.98

Fitzpatrick, P., Frazier, J. A., Cochran, D. M., Mitchell, T., Coleman, C., & Schmidt, R. C. (2016). Impairments of Social Motor Synchrony Evident in Autism Spectrum Disorder. *Frontiers in Psychology*. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2016.01323

Fleischer, Z., Pickering, M. J., & McLean, J. F. (2012). Shared information structure: Evidence from cross-linguistic priming. *Bilingualism: Language and Cognition*, *15*(3), 568–579.

Francis, A. L., & Nusbaum, H. C. (1996). Paying attention to speaking rate. *Proceedings of the 4th International Conference on Spoken Language*, *3*, 1537–1540.

Francis, M. E., & Pennebaker, J. W. (1993). *LIWC: Linguistic Inquiry and Word Count*. Dallas, TX: Southern Methodist University.

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*(6), 244–249. https://doi.org/10.1016/J.TICS.2006.04.002

Fricke, M., & Kootstra, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, *91*, 181–201. https://doi.org/10.1016/j.jml.2016.04.003

Fusaroli, R., Abney, D., Bahrami, B., Kello, C., & Tylén, K. (2013). Conversation, coupling and complexity: Matching scaling laws predict performance in a joint decision task.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, *23*(8), 931–939. https://doi.org/10.1177/0956797612436816

Fusaroli, R., & Tylén, K. (2012). Carving language for social coordination: A dynamical

approach. *Interaction Studies*, *13*(1), 103–124. https://doi.org/10.1075/is.13.1.07fus

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–171. https://doi.org/10.1111/cogs.12251

Gallois, C., & Giles, H. (2015). Communication Accommodation Theory. *The International Encyclopedia of Language and Social Interaction*. https://doi.org/10.1002/9781118611463.wbielsi066

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292–304. https://doi.org/10.1111/j.1756-8765.2009.01020.x

Giles, H. (1973). Communicative effectiveness as a function of accented speech. *Speech Monographs*, *40*(4), 330–331. https://doi.org/10.1080/03637757309375813

Giles, H., & Baker, S. C. (2008). Communication Accommodation Theory. *The International Encyclopedia of Communication*. https://doi.org/10.1002/9781405186407.wbiecc067

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) Talk to Me: The Social Context of Infant-Directed Speech and Its Effects on Early Language Acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344.

Gorisch, J., Wells, B., & Brown, G. J. (2012). Pitch Contour Matching and Interactional Alignment across Turns: An Acoustic Investigation. *Language and Speech*, *55*(1), 57–76. https://doi.org/10.1177/0023830911428874

Graf Estes, K., & Hurley, K. (2013). Infant-Directed Prosody Helps Infants Map Sounds to Meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Greiser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features. *Developmental Psychology*, *24*(1), 14–20.

Gries, S. T. H., & Koostra, G. J. (2017). Structural priming within and across languages: A corpus-based perspective. *Bilingualism: Language and Cognition*, *20*(2), 235–250. https://doi.org/10.1017/S1366728916001085

Grosjean, F. (2010). *Bilingual: life and reality*. Harvard University Press.

Grünloh, T., Elena, L., & Michael, T. (2011). German children use prosody to identify participant roles in transitive sentences. *Cognitive Linguistics*, *22*(2), 393–419. https://doi.org/10.1515/cogl.2011.015

Guo, T., & Peng, D. (2006). Event-related potential evidence for parallel activation of two languages in bilingual speech production. *NeuroReport*, *17*(17), 1757–1760.

Haake, M., Hansson, K., Gulz, A., Schötz, S., & Sahlén, B. (2014). The slower the better? Does the speaker's speech rate influence children's performance on a language comprehension test? *International Journal of Speech-Language Pathology*, *16*(2), 181–190. https://doi.org/10.3109/17549507.2013.845690

Hancock, A. B., & Rubin, B. A. (2014). Influence of Communication Partner's Gender on Language. *Journal of Language and Social Psychology*, *34*(1), 46–64. https://doi.org/10.1177/0261927X14533197

Hardy, S. M., Messenger, K., & Maylor, E. A. (2017). Aging and syntactic representations: Evidence of preserved syntactic priming and lexical boost. *Psychology and Aging*, *32*(6), 588–596. https://doi.org/10.1037/pag0000180

Harris, J. D., & Nelson, D. (1993). Glottal pulse alignment in voiced speech for pitch determination. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 2, pp. 519–522 vol.2). https://doi.org/10.1109/ICASSP.1993.319357

Harrison, D. A., Mohammed, S., McGrath, J. E., Florey, A. T., & Vanderstoep, S. W. (2003). Time matters in team performance: Effects of member familiarity, entrainment, and task discontinuity on speed and quality. *Personnel Psychology*, *56*(3), 633–669. https://doi.org/10.1111/j.1744-6570.2003.tb00753.x

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, *15*(6), 409–414.

Healey, P. G., Purver, M., & Howes, C. (2014). Divergence in dialogue. *Public Library of Science: One*, *9*(6). https://doi.org/10.1371/journal.pone.0098598

Heilmann, J., Ellis, W. S., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in Identifying Language Abilities of Late-Talking and Typically Developing Toddlers. *American Journal of Speech-Language Pathology*, *14*(1), 40–51. https://doi.org/10.1044/1058-0360(2005/006)

Hilton, N. H., Schüppert, A., & Gooskens, C. (2011). Syllable reduction and articulation rates in Danish, Norwegian and Swedish. *Nordic Journal of Linguistics*, *34*(2), 215–237. https://doi.org/10.1017/S0332586511000175

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., … Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

Holmes, J. (1995). *Women, Men and Politeness*. London: Routledge. https://doi.org/10.4324/9781315845722

Hopkins, Z., Yuill, N., & Keller, B. (2016). Children with autism align syntax in natural conversation. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716414000599

House, J. (2007). The role of prosody in constraining context selection: a procedural approach. *Nouveaux Cahiers de Linguistique Française*, *28*, 369–383.

Howard, I. S., & Messum, P. (2014). Learning to Pronounce First Words in Three Languages: An Investigation of Caregiver and Infant Behavior Using a Computational Model of an Infant. *PLOS ONE*, *9*(10). Retrieved from https://doi.org/10.1371/journal.pone.0110334

Ireland, M. E., & Henderson, M. D. (2014). Language Style Matching, Engagement, and Impasse in Negotiations. *Negotiation and Conflict Management Research*, *7*(1), 1–16. https://doi.org/10.1111/ncmr.12025

Ito, J., & Mester, A. (2013). Prosodic subcategories in Japanese. *Lingua*, *124*, 20–40. https://doi.org/10.1016/J.LINGUA.2012.08.016

Iyer, S. N., Denson, H., Lazar, N., & Oller, D. K. (2016). Volubility of the human infant: Effects of parental interaction (or lack of it). *Clinical Linguistics & Phonetics*, *30*(6), 470–488. https://doi.org/10.3109/02699206.2016.1147082

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., & Jasnow, M. D. (2001). Rhythms of Dialogue in Infancy: Coordinated Timing in Development. *Monographs of the Society for Research in Child Development*, *66*(2), 1–132. Retrieved from http://www.jstor.org/stable/3181589

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011

Jun, S. (2003). The effect of phrase length and speech rate on prosodic phrasing. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 483–486).

Kantola, L., & van Gompel, R. P. G. (2011). Between- and within-language priming is the same: Evidence for shared bilingual syntactic representations. *Memory & Cognition*, *39*(2), 276–290.

Kello, C. T., Dalla Bella, S., Médé, B., & Balasubramaniam, R. (2017). Hierarchical temporal structure in music, speech and animal vocalizations: jazz is like a conversation, humpbacks sing like hermit thrushes. *Journal of The Royal Society Interface*, *14*(135). https://doi.org/10.1098/rsif.2017.0231

Kim, M., Horton, W. S., & Bradlow, A. R. (2011a). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 125–156.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011b). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 125–156.

Kim, S. K., & Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *The Journal of the Acoustical Society of America*, *142*(1), EL49–EL55. https://doi.org/10.1121/1.4991328

Ko, E.-S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. *Journal of Child Language*, *43*(2), 284–309. https://doi.org/10.1017/S0305000915000203

Koolagudi, S. G., & Krothapalli, R. S. (2011). Two stage emotion recognition based on

speaking rate. *International Journal of Speech Technology*, *14*(1), 35–48. https://doi.org/10.1007/s10772-010-9085-x

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., … Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In D. Burnham (Ed.), *International Speech Communications Association* (pp. 1692–1695). Brisbane, Australia: Proceedings of Interspeech 2008.

Kroll, J. F., Bobb, S. C., & Hoshino, N. (2014). Two Languages in Mind: Bilingualism as a Tool to Investigate Language, Cognition, and the Brain. *Current Directions in Psychological Science*, *23*(3), 159–163. https://doi.org/10.1177/0963721414528511

Kroll, J. F., Dussias, P. E., Bogulski, C. A., & Kroff, J. R. V. (2012). Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. *Psychology of Learning and Motivation*, *56*, 229–262. https://doi.org/10.1016/B978-0-12-394393-4.00007-8

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V, Ryskina, V. L., … Lacerda, F. (1997). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakin, J. L., & Chartrand, T. L. (2003). Using Nonconscious Behavioral Mimicry to Create Affiliation and Rapport. *Psychological Science*, *14*(4), 334–339. https://doi.org/10.1111/1467-9280.14481

Leaper, C., & Ayres, M. M. (2007). A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. *Personality and Social Psychology Review*, *11*(4), 328–363. https://doi.org/10.1177/1088868307302221

Leong, V., Kalashnikova, M., Burnham, D., & Goswami, U. (2014). Infant-Directed Speech Enhances Temporal Rhythmic Structure in the Envelope. In *INTERSPEECH-2014* (pp. 2563–2567). Singapore.

Levitan, R., Benus, S., Gravano, A., & Hirschberg, J. (2015). Entrainment and Turn-Taking in Human-Human Dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *INTERSPEECH-2011* (pp. 3081–3084). Florence, Italy.

Lopez, L. D., Walle, E. A., Pretzer, G. M., & Warlaumont, A. S. (2020). Adult responses to infant prelinguistic vocalizations are associated with infant vocabulary: A home observation study. *PLOS ONE*, *15*(11). Retrieved from https://doi.org/10.1371/journal.pone.0242232

Lowen, S. B., & Teich, M. C. (1996). The periodogram and Allan variance reveal fractal exponents greater than unity in auditory-nerve spike trains. *The Journal of the Acoustical Society of America*, *99*(6), 3585–3591.

Luyster, R., Lopez, K., & Lord, C. (2007). Characterizing communicative development in children referred for Autism Spectrum Disorders using the MacArthur-Bates Communicative Development Inventory (CDI). *Journal of Child Language*, *34*(3), 623–654. https://doi.org/10.1017/S0305000907008094

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An Item Response Theory–Based, Computerized Adaptive Testing Version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI:WS). *Journal of Speech, Language, and Hearing Research*, *59*(2), 281–289. https://doi.org/10.1044/2015_JSLHR-L-15-0202

Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, *34*(6), 419–426. https://doi.org/10.1016/j.evolhumbehav.2013.08.001

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16. https://doi.org/10.1111/j.1467-7687.2008.00671.x

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*(2), 97–115. https://doi.org/10.1017/S1366728903001068

Mariooryad, S., Kannan, A., Hakkani-Tür, D., & Shriberg, E. (2014). Automatic characterization of speaking styles in educational videos. In *International Conference on Acoustics, Speech and Signal Processing* (pp. 4848–4852). Florence, Italy: IEEE.

Marmelat, V., & Delignières, D. (2012). Strong anticipation: Complexity matching in interpersonal coordination. *Experimental Brain Research*, *222*(1–2), 137–148.

Matarazzo, J. D., Wiens, A. N., Saslow, G., Dunham, R. M., & Voas, R. B. (1964). Speech Durations of Astronaut and Ground Communicator. *Science*, *143*(3602), 148–150. https://doi.org/10.1126/science.143.3602.148

Matsuno, E., & Budge, S. L. (2017). Non-binary/Genderqueer Identities: a Critical Review of the Literature. *Current Sexual Health Reports*, *9*(3), 116–120. https://doi.org/10.1007/s11930-017-0111-8

McDaniel, V. F. (2020). *Acoustic interactions: Pitch Coordination during parent-infant interaction*. The University of Memphis.

McGregor, K. K., Eden, N., Arbisi-Kelm, T., & Oleson, J. (2020). The Fast-Mapping Abilities of Adults With Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *63*(9), 3117–3129. https://doi.org/10.1044/2020_JSLHR-19-00418

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant

directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378. https://doi.org/10.1016/j.cognition.2013.07.015

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science*, *317*(5834), 82. https://doi.org/10.1126/science.1139940

Mulac, A. (1989). Men's and women's talk in same-gender and mixed-gender dyads: Power or polemic? *Journal of Language and Social Psychology*, *8*(3–4), 249–270.

Mushin, I., Stirling, L., Fletcher, J., & Wales, R. (2003). Discourse Structure, Grounding, and Prosody in Task-Oriented Dialogue. *Discourse Processes*, *35*(1), 1–31. https://doi.org/10.1207/S15326950DP3501_1

Natale, M. (1975). Social Desirability as Related to Convergence of Temporal Speech Patterns. *Perceptual and Motor Skills*, *40*(3), 827–830. https://doi.org/10.2466/pms.1975.40.3.827

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers* (pp. 169–172). Association for Computational Linguistics.

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Ni Eochaidh, C. (2010). *The role of conceptual and word form representations in lexical alignment: Evidence from bilingual dialogue*. The University of Edinburgh.

Niederhoffer, K. G., & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360. https://doi.org/10.1177/026192702237953

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. https://doi.org/10.1016/j.wocn.2010.12.007

Oller, D. K., Eilers, R. E., & Steffens, M. L. (1994). Speech-like vocalizations in infancy: an evaluation of potential risk factors. *Journal of Child Language*, *21*(1), 33–58.

Overby, M., Belardi, K., & Schreiber, J. (2020). A retrospective video analysis of canonical babbling and volubility in infants later diagnosed with childhood apraxia of speech. *Clinical Linguistics & Phonetics*, *34*(7), 634–651. https://doi.org/10.1080/02699206.2019.1683231

Palomares, N. A. (2009). Women Are Sort of More Tentative Than Men, Aren't They?: How Men and Women Use Tentative Language Differently, Similarly, and Counterstereotypically as a Function of Gender Salience. *Communication Research*, *36*(4), 538–560. https://doi.org/10.1177/0093650209333034

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in*

Psychology, *4*.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197. https://doi.org/10.1016/j.wocn.2011.10.001

Pardo, J. S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). Influence of Role-Switching on Phonetic Convergence in Conversation. *Discourse Processes*, *50*(4), 276–300. https://doi.org/10.1080/0163853X.2013.778168

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254–2264. https://doi.org/10.3758/BF03196699

Pardo, J., Urmanache, A., Gash, H., Wiener, J., Mason, N., Wilman, S., … Decker, A. (2019). The Montclair map task: Balance, efficacy, and efficiency in conversational interaction. *Language and Speech*, 1–21. https://doi.org/10.1177/0023830918775435

Park, S. M., & Sarkar, M. (2007). Parents' attitudes toward heritage language maintenance for their children and their efforts to help their children maintain the heritage language: A case study of Korean-Canadian immigrants. *Language, Culture and Curriculum*, *20*(3), 223–235. https://doi.org/10.2167/lcc337.0

Patel, R., Niziolek, C., Reilly, K., & Guenther, F. H. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech, Language, and Hearing Research : JSLHR*, *54*(4), 1051–1059. https://doi.org/10.1044/1092-4388(2010/10-0162)

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal Patterns in Infants with Autism Spectrum Disorder: Canonical Babbling Status and Vocalization Frequency. *Journal of Autism and Developmental Disorders*, *44*(10), 2413–2428. https://doi.org/10.1007/s10803-014-2047-4

Perani, D., Abutalebi, J., Paulesu, E., Brambati, S., Scifo, P., Cappa, S. F., & Fazio, F. (2003). The role of age of acquisition and language usage in early, high-proficient bilinguals: An fMRI study during verbal fluency. *Human Brain Mapping*, *19*(3), 170–182. https://doi.org/10.1002/hbm.10110

Peterson, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, *2*, 943.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pickering, M. J., & Garrod, S. (2004a). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, *27*(2), 212–225. https://doi.org/10.1017/S0140525X04450055

Pickering, M. J., & Garrod, S. (2004b). Toward a mechanistic psychology of dialogue. *Behavioral & Brain Sciences*, *27*(2), 169–226.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2), 203–228. https://doi.org/10.1007/s11168-006-9004-0

Pretzer, G. M. (2019). *Multi-Domain Synchrony Within Vocal Development*. University of California, Merced.

Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. C. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 340–344. https://doi.org/10.1016/j.evolhumbehav.2007.05.002

Quigley, J., McNally, S., & Lawson, S. (2016). Prosodic Patterns in Interaction of Low-Risk and at-Risk-of-Autism Spectrum Disorders Infants and Their Mothers at 12 and 18 Months. *Language Learning and Development*, *12*(3), 295–310. https://doi.org/10.1080/15475441.2015.1075405

Ramirez-Aristizabal, A. G., Médé, B., & Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals*, *111*, 175–179. https://doi.org/10.1016/j.chaos.2018.04.021

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. https://doi.org/10.1111/desc.12172

Ray, M. L., & Webb, E. J. (1966). Speech Duration Effects in the Kennedy News Conferences. *Science*, *153*(3738), 899 LP – 901. https://doi.org/10.1126/science.153.3738.899

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. https://doi.org/10.1016/J.JML.2014.05.008

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, Gaze Coordination, and Beliefs About Visual Context. *Cognitive Science*, *33*(8), 1468–1482. https://doi.org/10.1111/j.1551-6709.2009.01057.x

Ritwika, V. P. S., Pretzer, G. M., Mendoza, S., Shedd, C., Kello, C. T., Gopinathan, A., & Warlaumont, A. S. (2020). Exploratory dynamics of vocal foraging during infant-caregiver communication. *Scientific Reports*, *10*(1), 10469. https://doi.org/10.1038/s41598-020-66778-0

Robinson, S. L., Sterling, H. E., Skinner, C. H., & Robinson, D. H. (1997). Effects of lecture rate on students' comprehension and ratings of topic importance. *Contemporary Educational Psychology*, *22*(2), 260–267.

Roe, K. V. (1975). Amount of Infant Vocalization as a Function of Age: Some Cognitive Implications. *Child Development*, *46*(4), 936–941. https://doi.org/10.2307/1128400

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Romaine, S. (2012). *The bilingual and multilingual community*. (T. K. Bhatia & W. C. Ritchie, Eds.). Malden, MA: Blackwell Publishing. https://doi.org/10.1002/9781118332382.ch18

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. https://doi.org/10.1006/jpho.1997.0051

Sawyer, T., Sierocka-Castaneda, A., Chan, D., Berg, B., Lustik, M., & Thompson, M. (2011). Deliberate Practice Using Simulation Improves Neonatal Resuscitation Performance. *Simulation in Healthcare*, *6*(6).

Schmidt, R., & Richardson, M. (2008). Dynamics of interpersonal coordination. *Coordination: Neural, Behavioral and Social Dynamics*, 281–308.

Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2020). Complexity Matching and Lexical Matching in Monolingual and Bilingual Conversations. *Bilingualism: Language and Cognition*, *23*(4), 845–857. https://doi.org/10.1017/S1366728919000774

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. https://doi.org/10.1016/0010-0285(89)90008-X

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. J. (2018). Infant–Mother Acoustic–Prosodic Alignment and Developmental Risk. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1369–1380. https://doi.org/10.1044/2018_JSLHR-S-17-0287

Selting, M. (2010). Prosody in interaction: State of the art. In D. Barth-Weingarten, E. Reber, & M. Selting (Eds.), *Prosody in Interaction* (pp. 3–40). John Benjamins Publishing Company. https://doi.org/10.1075/sidag.23

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326–332.

Skarakis-Doyle, E., Campbell, W., & Dempsey, L. (2009). Identification of Children With Language Impairment: Investigating the Classification Accuracy of the MacArthur–Bates Communicative Development Inventories, Level III. *American Journal of Speech-Language Pathology*, *18*(3), 277–288. https://doi.org/10.1044/1058-0360(2009/08-0035)

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the*

Acoustical Society of America, *118*(5), 3177–3186. https://doi.org/10.1121/1.2047107

Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *The Journal of the Acoustical Society of America*, *128*(1), 389–400. https://doi.org/10.1121/1.3419786

Sosa, A. V. (2016). Association of the Type of Toy Used During Play With the Quantity and Quality of Parent-Infant Communication. *JAMA Pediatrics*, *170*(2), 132–137. https://doi.org/10.1001/jamapediatrics.2015.3753

Speer, S. R., & Ito, K. (2009). Prosody in First Language Acquisition – Acquiring Intonation as a Tool to Organize Information in Conversation. *Language and Linguistics Compass*, *3*(1), 90–110. https://doi.org/10.1111/j.1749-818X.2008.00103.x

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Language Does Matter: But There is More to Language Than Vocabulary and Directed Speech. *Child Development*, *90*(3), 993–997. https://doi.org/10.1111/cdev.13125

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 127–132). Austin, TX: Cognitive Science Society.

Stoel-Gammon, C. (1991). Normal and disordered phonology in two-year-olds. *Topics in Language Disorders*, *11*(4), 21–32. https://doi.org/10.1097/00011363-199111040-00005

Street, R. L. J. (1984). Speech Convergence and Speech Evaluation in Fact-Finding Interviews. *Human Communication Research*, *11*(2), 139–169. https://doi.org/10.1111/j.1468-2958.1984.tb00043.x

Sultana, N., Wong, L. L. N., & Purdy, S. C. (2020). Natural Language Input: Maternal Education, Socioeconomic Deprivation, and Language Outcomes in Typically Developing Children. *Language, Speech, and Hearing Services in Schools*, *51*(4), 1049–1070. https://doi.org/10.1044/2020_LSHSS-19-00095

Tausczik, Y. R., & Pennebaker, J. W. (2009). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

ten Thije, J D, Gooskens, C., Daems, F., Cornips, L., & Smits, M. (2017). Lingua receptiva: Position paper on the European commission's skills agenda. *European Journal of Applied Linguistics*, *5*(1), 141–146.

ten Thije, Jan D. (2013). Lingua receptiva (LaRa). *International Journal of Multilingualism*, *10*(2), 137–139. https://doi.org/10.1080/14790718.2013.789519

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*.

Tobin, S. J., Nam, H., & Fowler, C. A. (2017). Phonetic drift in Spanish-English bilinguals: Experiment and a self-organizing model. *Journal of Phonetics*, *65*, 45–59. https://doi.org/10.1016/j.wocn.2017.05.006

Toribio, A. J. (2004). Convergence as an optimization strategy in bilingual speech: Evidence from code-switching. *Bilingualism: Language and Cognition*, *7*(2), 165–173. https://doi.org/10.1017/S1366728904001476

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is Infant-Directed Speech Prosody a Result of the Vocal Expression of Emotion? *Psychological Science*, *11*(3), 188–195. https://doi.org/10.1111/1467-9280.00240

Trofimovich, P. (2016). Interactive alignment: A teaching-friendly view of second language pronunciation learning. *Language Teaching*, *49*(3), 411–422. https://doi.org/10.1017/S0261444813000360

Ullmann, S. (1974). *Words and Their Meanings*. Canberra: Australian National University Press.

Vacharkulksemsuk, T., & Fredrickson, B. L. (2012). Strangers in sync: Achieving embodied rapport through shared movements. *Journal of Experimental Social Psychology*, *48*(1), 399–402. https://doi.org/10.1016/j.jesp.2011.07.015

Valdesolo, P., Ouyang, J., & DeSteno, D. (2010). The rhythm of joint action: Synchrony promotes cooperative ability. *Journal of Experimental Social Psychology*, *46*(4), 693–695. https://doi.org/10.1016/j.jesp.2010.03.004

van Baaren, R. B., Holland, R. W., Steenaert, B., & van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, *39*(4), 393–398. https://doi.org/10.1016/S0022-1031(03)00014-3

Vogel, C., Lopes, M. R., & Esposito, A. (2017). Gender differences in the language of the Map Task dialogues. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (pp. 151–156). https://doi.org/10.1109/CogInfoCom.2017.8268233

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A Social Feedback Loop for Speech Development and Its Reduction in Autism. *Psychological Science*, *25*(7), 1314–1324. https://doi.org/10.1177/0956797614531023

Watt, N., Wetherby, A., & Shumway, S. (2006). Prelinguistic Predictors of Language Outcome at 3 Years of Age. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1224–1237. https://doi.org/10.1044/1092-4388(2006/088)

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2018). Primed From the Start: Syntactic Priming During the First Days of Language Learning. *Language Learning*. https://doi.org/10.1111/lang.12327

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. In *Eleventh Annual Conference of the International Speech Communication Association* (pp. 1485–1488). Makuhari, Japan.

Wesnes, K., & Pincock, C. (2002). Practice effects on cognitive tasks: a major problem? *The Lancet Neurology*, *1*(8), 473.

West, B. J., Geneston, E. L., & Grigolini, P. (2008). Maximizing information exchange between complex networks. *Physics Reports*, *468*(1–3), 1–99.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://doi.org/ggplot2.tidyverse.org

Xia, Z., Levitan, R., & Hirschberg, J. (2014). Prosodic entrainment in Mandarin and English: A cross-linguistic comparison. In *Proceedings of the 7th International Conference on Speech Prosody* (pp. 65–69).

Xu, Y., & Reitter, D. (2016). Convergence of Syntactic Complexity in Conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 443–448). Berlin, Germany: Association for Computational Linguistics.

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., … Narayanan, S. (2004). An acoustic study of emotions expressed in speech. In *International Conference on Spoken Language Processing*. Jeju Island, Korea.

Yilmaz, G. (2015). What You Do and How You Speak Matter: Behavioral and Linguistic Determinants of Performance in Virtual Teams. *Journal of Language and Social Psychology*, *35*(1), 76–97. https://doi.org/10.1177/0261927X15575772

Yirmiya, N., Gamliel, I., Pilowsky, T., Feldman, R., Baron-Cohen, S., & Sigman, M. (2006). The development of siblings of children with autism at 4 and 14 months: social engagement, communication, and cognition. *Journal of Child Psychology and Psychiatry*, *47*(5), 511–523. https://doi.org/10.1111/j.1469-7610.2005.01528.x

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an Integrated Understanding of Speaking Rate in Conversation. In *Ninth International Conference on Spoken Language Processing*.

Zangl, R., & Mills, D. L. (2007). Increased Brain Activity to Infant-Directed Speech in 6- and 13-Month-Old Infants. *Infancy*, *11*(1), 31–62. https://doi.org/10.1207/s15327078in1101_2

Zhu, X., & Penn, G. (2006). Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*. Pittsburgh.

Zipf, G. K. (1935). *The psycho-biology of language*. Oxford, England: Houghton, Mifflin.

Appendix A: Pre-Experiment Language History Questionnaire

1. SONA ID:

2. Gender:

3. Age:

4. Do you have any visual and/or hearing problems? If yes, what are they?

5. What is your native country/ies?

6. What is your native langue(s)?

7. What language is spoken in your household?

8. At what age(s) did you start to learn each language, and for how many years?

9. What would you consider to be your primary second language?

10. What language are you most comfortable using on a daily basis?

11. On a scale of one to ten, with ten being the highest level of confidence, please mark your proficiency in the following areas:

a. English reading
1 2 3 4 5 6 7 8 9 10
b. English spelling
1 2 3 4 5 6 7 8 9 10
c. English writing
1 2 3 4 5 6 7 8 9 10
d. English speaking
1 2 3 4 5 6 7 8 9 10
e. English speech comprehension
1 2 3 4 5 6 7 8 9 10

12. On a scale of one to ten, with ten being the highest level of confidence, please mark your proficiency in the following areas:

a. Spanish reading
1 2 3 4 5 6 7 8 9 10
b. Spanish spelling
1 2 3 4 5 6 7 8 9 10
c. Spanish writing
1 2 3 4 5 6 7 8 9 10
d. Spanish speaking

1 2 3 4 5 6 7 8 9 10

e. Spanish speech comprehension

1 2 3 4 5 6 7 8 9 10

13. Estimate, in terms of percentages, how often you use your dominant language and other languages **per week** (in all weekly activities combined, circle which range best applies):

Dominant language:	0%	0-25%	50-75%	75-100%
Second language:	0%	0-25%	50-75%	75-100%

Appendix B: Post-Experiment Questionnaire

1. Have you ever met your partner before today? If so, are you just acquaintances, or friends?

2. On a scale of 1 to 5, how easy was the conversation in which you both spoken English, with 5 being the easiest?

1 2 3 4 5

3. On a scale of 1 to 5, how easy was the conversation in which you both spoke Spanish, with 5 being the easiest?

1 2 3 4 5

4. On a scale of 1 to 5, how easy was the conversation in which you spoke two different languages, with 5 being the easiest?

1 2 3 4 5