Title
Multimodal Dynamics of Extended Communication

Author
Alviar Guzman, Maria Camila

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, MERCED

Multimodal Dynamics of Extended Communication

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

in

Cognitive and Information Sciences

by

Maria Camila Alviar Guzman

Committee in charge:

Professor Christopher Kello, Chair
Professor Rick Dale
Professor Michael Spivey

2021

The dissertation of Maria Camila Alviar Guzman is approved, and it is acceptable in quality and form for publication in microfilm and electronically:

Professor Christopher Kello, Chair

Professor Rick Dale

Professor Michael Spivey

University of California, Merced
2021

To my mother, Alicia,
who always loved and pursued knowledge
and who even in the most insurmountable of distances
kept giving me the tools to succeed on this challenging journey

Table of Contents

List of Figures

List of Tables

Acknowledgements

Just as raising a child takes a community, preparing a scientist takes one too. I am grateful to all my colleagues, friends, and family for their continuous guidance, love, and emotional support. You all got me through this journey.

First, I would like to thank my committee members --Chris, Rick, and Spivey-- for their constant support, enthusiasm, and guidance that helped me grow as a scientist. Chris and Rick, thanks for bringing excitement back into hopeless times, and charting a path forward through the murky waters of naturalistic datasets. Spivey, thanks for your constant excitement for my quirky ideas. Thank you all for being fantastic mentors even through the times in which it was hard to mentor me!

I would also like to thank Alexia Galati, for her continuous support and mentorship. Thank you for always offering an empathic ear and thoughtful advice both on the science and the challenges of academic life.

I am also grateful to Javier Corredor and Florencia Reali, my college mentors, who gave me the chance to fall in love with cognitive science, and invited me to participate of a community where talking science was considered cool.

I would like to thank UC Merced, and the CIS department as a whole for providing me with the forum and the generous funding to think and talk about the mind in its many facets. It was fun to be part of such an interdisciplinary and supportive community.

Special thanks go also to my wonderful research assistants, Akeiylah, Alex, Nancy, and Aurora for the countless and painstaking hours of data collection and coding that made this dissertation possible. Thank you for all your work, and for letting me mentor you and share my excitement for science with you!

Thanks to my mom, my dad, my sister, and my brother for loving me and encouraging me so much as to give me the confidence to undertake this challenge. Thank you for constantly supporting me, cheering me, and also reminding me to set the necessary boundaries to get to the other side with my sanity and my health.

Thanks to my all my friends, old and new, for keeping me grounded, offering a commiserate ear, and providing good company and great excuses to get away from the science. You all are the best outcomes of my many years of schooling!

Last, but not least, I would like to thank Dan, my loving partner, for being an insightful interlocutor to my ideas, loving me through a pandemic, and hugging my anxiety away as many times as I needed it. You made this journey lighter and much happier.

The text of Chapter 2 of this dissertation is a reprint (with permission of John Wiley and Sons, Inc.) of the article “Complex Communication Dynamics: Exploring the Structure of an Academic Talk” as it appeared in *Cognitive Science*, which was co-authored by Alexia Galati and Rick Dale. The text of Chapter 3 of this dissertation is a reprint (with permission of Taylor and Francis, Ltd.) of the article “Multimodal Coordination of Sound and Movement in Music and Speech” as it appeared in *Discourse Processes*, which was co-authored by Rick Dale, Akeiylah Dewitt, and Chris Kello. The text of Chapter 4 will be submitted for publication with co-authors Rick Dale, Nancy Rodas de Leon, and Chris Kello. The text of Chapter 5 is under review and was co-authored with Chris Kello and Rick Dale.

Curriculum Vitae

Camila Alviar

Cognitive and Information Sciences
University of California, Merced
Email: malviarguzman@ucmerced.edu
Phone: (209) 777-7657

Education

2016 - 2021 **Ph.D.,** Cognitive and Information Sciences
University of California, Merced
Committee: Chris Kello, Ph.D.
Rick Dale, Ph.D.
Michael Spivey, Ph.D

2015 **B.S.**, Psychology
Graduated with honors
Universidad Nacional de Colombia

Additional Training

2019 *Self-Organization for Behavioral Scientists Workshop*. Center for the Ecological Study of Perception and Action, University of Connecticut, Storrs, August 3 – 7, 2019.

2017 *Nonlinear Methods for Psychological Science*. American Psychological Association Advanced Training Institute. University of Cincinnati, June 19 – 23, 2017.

Fellowships, Scholarships, and Awards

2021 Graduate Dean's Dissertation Fellowship. University of California, Merced
A semester of full tuition, stipend, and summer funding. $ 27400

2016 - 2017 Graduate Dean's Recruitment Fellowship. University of California, Merced A year of full tuition, stipend, and summer funding. $ 41100

2010 - 2015 Tuition waiver for excellent academic performance. Universidad Nacional de Colombia. 10 semesters of tuition waiver. $2000/semester

Publications

Articles, Proceedings, and Chapters

Zhao, Z., Tang, H., **Alviar, C.,** Kello, C., Zhang, X., Hu, X., Qu, X., Lu, J. (under review). Excessive and less complex body movement in children with autism during face-to-face conversation: An objective approach to behavioral quantification. *Molecular Autism.*

Alviar, C., Kello, C., & Dale, R. (under review). Multimodal coordination and pragmatic modes in conversation. *Language Sciences.*

Alviar, C., Dale, R., Dewitt, A., & Kello, C. (2020). Multimodal coordination of sound and movement in solo music and speech. *Discourse Processes, 57*(8), 682 – 702.

Alviar, C., Dale, R., Galati, A. (2019). Complex Communication Dynamics: Exploring the structure of an academic talk. *Cognitive Science, 43*(3), e12718.

Reali, F., Lleras, M., & **Alviar, C.** (2019). Asymmetrical time and space interference in Tau and Kappa effects. *Cogent Psychology*, *6*(1), 1568069.

Dale, R., Galati, A., **Alviar, C.**, Kallens, P. A. C., Ramirez-Aristizabal, A., Tabatabaeian, M., & Vinson, D. W. (2018). Interacting timescales in perspective taking. *Frontiers in Psychology*.

Alviar, C., Dale, R., Kello, C. (2018). The fractal structure of extended communicative performance. *Proceedings of the 2018 Cognitive Science Society Conference.* Retrieved from https://mindmodeling.org/cogsci2018/papers/0253/index.html

Corredor, J., Cure, S., & **Alviar, M. C.** (2018). La memoria mediada: la psicología de la memoria histórica [The mediated memory: The psychology of historical memory]. In G. Gutierrez (Ed.), *Integración teórica en las ciencias del comportamiento.*

Lleras, M., Realli, F., **Alviar, C.,** & Bermúdez, M. P. (2014). The metaphors we speak with affect the way we think about time and space. *Proceedings of the 2014 Cognitive Science Society Conference,* 1258-1263. Retrieved from https://mindmodeling.org/cogsci2014/papers/222/

Corredor, J., Pico, G., Castro, C., Hernández, M., Arias, C., **Alviar, C.**, Rojas, L., Silva, J., & Niño, J. (2013). La frontera digital: Efectos psicológicos de Internet en la región de la Orinoquía. [The digital frontier: Psychological effect of the Internet in the Colombian Eastern Plains]. *Revista Iberoamericana de Psicología: Ciencia y Tecnología, 6*(2), 55-67.

Manuscripts in Preparation

Alviar, C., Dale, R., Kello, C. (in preparation). Can you hear me now? Interpersonal and multimodal coordination in Zoom conversations.

Kello, C., Turner, M., **Alviar, C.**, Bhat, H. (in preparation). A probabilistic model of hierarchical temporal structure.

Galati, A., Alviar, C., Dale, R., Moreno, C. (in preparation). Task constraints on interpersonal coordination: Effects of task goals on alignment in eye-movements and speech.

Invited Talks

Alviar, C. (2020). *Multimodal coordination of sound and movement in solo music and speech*. Brownbag series. Department of Psychological Science, University of North Carolina at Charlotte, NC.

Conference Presentations and Posters

Asterisk (*) denotes presenter

Galati, A.*, **Alviar. C.**, Coco, M. & Dale, R. (2021, June). Task goals constrain interpersonal coordination: Evidence from the alignment of eye-movements. Oral presentation at the

Boorom, O.*, Muñoz, V., **Alviar, C.,** Kello, C., & Lense, M. (2021, June). Hierarchical Acoustic Structure During Parent-Child Interactions of Toddlers with Typical Development and Autism Spectrum Disorder. Oral presentation at the Rhythm Production and Perception Workshop held online from June 22 – 25.

Boorom, O.*, Muñoz, V., **Alviar, C.,** Kello, C., & Lense, M. (2021). Hierarchical Acoustic Structure During Parent-Child Interactions of Toddlers with Typical Development and Autism Spectrum Disorder. Abstract submitted to the International Society for Autism Research Annual Meeting.

Alviar, C.*, Dale, R., Dewitt, A., & Kello, C. (2020). Multimodal coordination of sound and movement in solo music and speech. Talk accepted at the Expression, Language and Music Conference, Storrs, Connecticut. (The conference has been postponed until 2022 due to covid-19).

Alviar, C.*, Dale, R., Dewitt, A., & Kello, C. (2019, August). Multimodal coordination of sound and movement in solo music and speech. Oral presentation at the Guy Van Orden Cognition and Dynamics Workshop, Storrs, Connecticut.

Alviar, C.*, Dale, R., Dewitt, A., & Kello, C. (2019, July). Multimodal coordination of sound and movement in solo music and speech. Poster presented at the International Conference for Perception and Action, Groningen, The Netherlands.

Alviar, C.*, Dale, R., Kello, C. (2018, July). The fractal structure of extended communicative performance. *Proceedings of the 2018 Cognitive Science Society Conference*. Retrieved from https://mindmodeling.org/cogsci2018/papers/0253/index.html

Alviar, C.*, Dale. R., & Galati, A. (2017, November). Complex communication dynamics: Exploring the structure of an academic talk. Poster presented at the Annual Meeting of the Society for Computers in Psychology, Vancouver, Canada.

Alviar, C.*, Dale, R., & Galati, A. (2017, June). Complex communication dynamics: Exploring the structure of an academic talk. Poster presented at the APA Advanced Training Institute in Nonlinear Methods for Psychological Science, University of Cincinnati, Cincinnati.

Lleras, M., **Alviar, C.**, & Reali, F.* (2016, October). Asymmetrical time and space interference in Tau and Kappa effects. Oral presentation at the 9th Embodied and Situated Language Processing Conference, Pucon, Chile.

Alviar, C.*, Corredor, J., & Reali, F. (2016. August). The words of emotion: Word-frequency effect on emotional perception. Poster presented at the American Psychological Association Annual Conference, Denver.

Corredor, J. A.*, Castro, C., Jimenez, T., Gonzalez, D., Castro-García, D., **Alviar, C.** (2016, August). The learning and understanding of historical memory in Colombia: Lessons for peace education. Poster presented at the American Psychological Association Annual Conference, Denver.

Bonilla-Hernández, C.*, Castro-García, D., **Alviar, C.,** González-Campos, D. A., & Corredor, J. (2015, July). Educación para el posconflicto en Colombia: memoria histórica y pedagogías alternativas. Un estudio cuantitativo. [Education for Colombian post-conflict: Historical memory and alternative pedagogies. A quantitative study]. Paper presented at the biannual meeting of the Interamerican Society of Psychology, Lima.

Lleras, M.*, Realli, F., **Alviar, C.,** & Bermúdez, M. P. (2014, July). The metaphors we speak with affect the way we think about time and space. Paper presented at the annual meeting of the Cognitive Science Society, Quebec.

Teaching Experience

Instructor of Record

Department of Cognitive and Information Sciences, University of California, Merced

Research Methods for Cognitive Science (Online course)	Summer 2020

Teaching Assistantships

Department of Cognitive and Information Sciences, University of California, Merced

Introduction to Ethics	Fall 2020
Mind, Brain, & Computation	Fall 2019

Introduction to Linguistics	Spring 2019, Fall 2017
Introduction to Cognitive Science	Fall 2018, Spring 2018, Spring 2020

Guest Lectures

Summer 2019 Social Cognition: “Language as Social Action”, UC Merced.

Summer 2018 Social Psychology: “Conformity”, UC Merced.

Research Assistantships

2013 - 2016 *Editorial Assistant*. Colombian Journal of Psychology [Revista Colombiana de Psicología]. Universidad Nacional de Colombia.

2013 - 2016 *Research Assistant*. Language and Cognition Research Group. Universidad de Los Andes.

2012 - 2016 *Research Assistant*. Education, Applied Cognition and Media Lab. Universidad Nacional de Colombia.

Mentoring

Research Assistants Supervised and provided training in experimental design, data collection, data analyses, academic writing, and research ethics to several undergraduate Research assistants at UC Merced.

Aurora Vargas Contreras	2020 - 2021
Nancy Rodas de Leon Now a doctoral student at the University of California, Merced	2020 – 2021
Alejandra Santoyo Now a doctoral student at the University of California, Merced	2017 – 2018
Akeiylah Dewitt Now a doctoral student at the University of Washington	2017 – 2018

Service

Organization of Local Events

2020 Co-organizer of the Graduate Student Visitation Weekend for the Cognitive and Information Sciences program at UC Merced

Leadership in Professional Organizations

2019 - 2020 Social and Community Officer Cognitive & Information Sciences Graduate Student Group

2019 - 2020 Women's tea co-organizer

2019 Undergraduate Mentor for the W-STEM organization at UC Merced

Past and Current Memberships

2019 – Present	Ecological Psychology Association
2014 - Present	Cognitive Science Society
2016 - 2018	American Psychological Association

Abstract

Language is a multimodal performance of remarkable coordination. Previous research has found this coordination responds to word-level variables as well as sentence-level variables. Coordination at the longer time scales of the discourse-level, however, is less studied. Bridging the rapidly changing multimodal behaviors of language to its diverse discursive contexts and pragmatic intentions is fundamental for our understanding of language use in social contexts. This dissertation takes initial steps in this direction by studying the dynamic organization and coordination of body movement and prosody over the extended time scales of diverse performances.

Chapter 2 explores the dynamics and multimodal patterns that speakers produce in the context of an academic talk. We analyze the organization and coordination of body movement, prosody, and PowerPoint slide transitions. Results show weak regularities in the coordination and organization of the modalities as a result of the shared discursive context, but also highlight the role of individual constraints in shaping the multimodal behaviors of speakers.

Chapter 3 contrasts the multimodal multiscale coordination of the movements and sounds of solo music performances and speech monologues. Results evidence different coordination patterns depending on performance goals, with higher local sound-movement synchrony and stronger multiscale coordination for speech compared to music. Coordination also varies across the discursive contexts analyzed, but not across instruments of interpretation.

Chapter 4 studies the effects that the limitations of videoconferencing have on interpersonal and multimodal coordination. The continuous perturbations introduced by videoconferencing reduce interpersonal coordination during remote conversations in ways consistent with the reduction of signal quality as compared to in-person interactions. Multimodal coordination, which is not mediated by the Zoom's audiovisual signals, is maintained, while speech convergence is reduced, and movement convergence is disrupted.

Chapter 5 outlines a proposal to connect the multimodal coordination patterns of language use to the pragmatic goals of social interactions. I argue that the context sensitivity and rapid adaptability of multimodal coordination is consistent with the characteristics of synergies. I propose the multimodal patterns of language result from metastable multimodal synergies that simultaneously provide stability for pragmatic goals while also dynamically adapting to ever-changing constraints and goals of conversations.

This dissertation, *Multimodal Dynamics of Extended Communication,* is submitted by Maria Camila Alviar Guzman 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

Using language involves a lot more than just using words (Clark, 1994). When people engage in a conversation or monologue, they also move their hands and their faces (McNeil, 1985). They draw things in the air, they manipulate invisible and visible objects (Clark, 2003). They shift their posture. They change the inflections in their voice and their volume. They pause solemnly or speed through their message (Cole, 2015). They look to their interlocutors' eyes, to their own hands, to the things around them (Bavelas & Chovil, 2018; Kendon, 1967). They laugh, smile, nod (Louwerse, Dale, Bard, & Jeuniaux, 2012). Years of research have established that all of these signals contribute to the message content and help with comprehension and delivery. The studies have also shown that, at least at the word and sentence level, the signals are orchestrated with remarkable coordination and precise timing (Özyürek, 2014 for a review of gesture research; Cole, 2015 for a review of prosody research). It is still an open question however, whether the different modalities are also organized and coordinated at the longer time scales of the discourse level. That is, over the entirety of a conversational topic or the full length of a lecture. Answering this question is fundamental to the development of models that address language in its face-to-face and situated nature (Holler & Levinson, 2019), and that bridge the low-level behaviors that make up language to the joint actions that language makes possible (Clark, 1994; Dale, 2015). In this dissertation, I aim to make progress on this question by exploring the organization and coordination of sounds and movements over time for monologues and conversations with varied goals. In this chapter, I first review the literature on gesture and speech coordination at the word and sentence level. Then, I present evidence suggesting that discourse level variables change how speakers use gestures and prosody. After, I elaborate on the theoretical reasons we should expect multimodal coordination patterns to vary over extended periods of time. To close, I present an outline of the remaining chapters of this dissertation, addressing their goals and main conclusions.

1.1. Coordination of Movements and Sounds at the Word and Sentence Level

The movements and sounds speakers perform during language use are importantly linked to each other. Perhaps unsurprisingly, the physical characteristics and movements of the vocal tract are directly related to changes in voice quality (Zhang, 2016), and the area of the opening of the mouth directly relates to changes in the amplitude of the speech wave (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). More surprisingly, gestures, in all their variations, are also deeply intertwined with speech in both meaning and timing (McNeill, 1985). Iconic gestures, which represent characteristics of their referent, tend to occur in close alignment to the words they represent (Özyürek, Kita, Allen, Furman, & Brown, 2005). In fact, the time windows in which iconic gestures are processed as part of the spoken message are very tight, allowing less than a 360-millisecond delay between the gesture and the referent word for clear speech conditions (Habets, Kita, Shao, Özyürek, & Hagoort, 2011), and up to 1-second delay for degraded speech conditions (Obermeier, Dolk, & Gunter, 2012). Listeners expect iconic gestures to be relevant to the content of speech becoming slower and less accurate to process mismatched word-gesture pairs (e.g., seeing "chop" and hearing "twist"; Kelly, Özyürek, Maris, 2010). Iconic gestures also respond tightly to the syntactic constructions used in speech. Speakers of languages such as English that combine manner and path in their syntax (e.g., "he rolled down the hill") also tend to produce gestures that combine manner and path, while speakers of languages such as Turkish that express them separately (e.g., "he descended as he rolled"), tend to depict just one of the dimensions (Özyürek et al., 2005). Importantly, requiring speakers of English to produce one syntactic form or the other, also results in iconic gestures that follow the syntax used in speech (Kita et al., 2007).

Pointing gestures are usually aligned with the prosodic characteristics of speech. The apex of the gesture usually occurs a few milliseconds after the onset of the deictic expression, and its precise timing is modulated by the distance that the hand has to travel to the referent (Levelt, Richardson, & La Heij, 1985). The timing of the gesture is also modulated by the position of the accented syllable with respect to the broader prosodic boundaries of words and sentences (Esteve-Gilbert & Prieto, 2013), and its execution follows the prosodic characteristics of speech, lengthening at prosodic boundaries as speech does (Krivokapić, Tiede, & Tyrone, 2017). The relationship between pointing gestures and speech starts early in development (Esteve-Gilbert & Prieto, 2014) and is robust to perturbations. When a pointing gesture is slowed down, speech lengthens, and when speech is disrupted, gestures are delayed as well, spontaneously adapting to maintain alignment between modalities (Chu & Hagoort, 2014).

Beat gestures, the repetitive rhythmic movements of the hands, are generally aligned with the peak amplitudes of the speech signal (Pouw, Harrison, & Dixon, 2019). High movement of the upper limbs is correlated with heightened F0 and increased sound amplitude at the time of the beat (Pouw, Jonge-Hoekstra, Harrison, Paxton, & Dixon, 2021). The link between beat gestures and the acoustics of the signal is actually strong enough to allow a listener to synchronize to the movements of the speakers using only the auditory information (Pouw, Paxton, Harrison, & Dixon, 2020). ERP studies have revealed that beat gestures also play a role in sentence parsing. They direct attention to the important words in the sentence (Biau & Soto-Faraco, 2013) by resetting the phase of theta band oscillations in the auditory cortex to make them synchronous with word onset (Biau, Torralba, Fuentemilla, de Diego Balaguer, & Soto-Faraco, 2015). Beat gestures also facilitate syntactic interpretations associated with the less common subject-object-verb order in German (Holle et al., 2012). And, conversely, they increase processing costs when aligned with the non-focused words in contrastive sentences (Dimitrova, Chu, Wang, Özyürek, & Hagoort, 2016).

The timing between iconic and beat gestures with respect to speech is robust to perturbation but flexible to context. Participants that heard their own speech with a 150 ms live delay as they were narrating a story, showed less variable timing between gesture peak velocity and peak pitch, suggesting stronger coordination between speech and gesture under delayed auditory feedback. Additionally, participants also showed delayed gesture peaks that shifted towards the delayed speech signal, suggesting gesture was also flexibly coordinated with the perceived speech. These results suggest pitch and gesture are dynamically coupled and their specific timing can adapt to the demands of context while maintaining local coordination (Pouw & Dixon, 2019).

1.2 Coordination of Movements and Sounds at the Discourse Level

While the evidence for the coordination of gesture and speech at the local scales of words and sentences is mounting, it is still an open question how much the coordination of gesture and prosody extends to the longer time scales of the discourse level. Are there patterns of multimodal coordination that systematically unfold over the extended time frame of slow changing discourse variables, such as the speaker's communicative goals, the conversational topics, the physical contexts of language use, the specific characteristics and responses of the audience? And do these patterns look different for specific instances of these variables? Answering these questions is key to build a bridge between the local and rapidly changing behaviors that make up language and the slower and more content driven processes of the joint actions that such language makes possible (cf. Dale, 2015). In other words, pursuing these questions will advance our knowledge on how movements and sounds turn into the explanations, requests, jokes, commands, etc. that are the bread and butter of everyday social interactions, and will make our models more relevant to the complex and situated nature of language use (Holler & Levinson, 2019).

Two complementary argumentative strands forecast the outlined quest to be a productive pursuit: one empirical and one theoretical. With respect to the former, empirical work has found differences for the kinematics of gestures, as well as, for the prosodic patterns of speech in relation to discourse level variables such as the speakers' goals and the audience characteristics, among others. With respect to the latter, theoretical proposals on the coordination of complex intelligent behavior (e.g., Kelso, 2009), language being one example of it, argue that such behavior emerges from the low-level non-linear interactions of the simple units that make up a complex system. In the phenomenon of interest here, such low-level units could correspond to the sounds, movements, gaze patterns, and other behaviors of language use. We expand on each of these two strands of argumentation next.

1.2.1 The Empirical Argument: Prosody and Gesture at the Discourse Level

A fair number of studies have examined gesture and prosody in relation to discourse level variables. Differences in their use have been found for variables such as the pragmatic intentions of the language users, the characteristics of the audience, the nature of task at hand, the channels that are available for information exchange with the interlocutors, and the contextual characteristics of the communicative situation, among others.

1.2.1.1 Prosodic Patterns

The prosodic inflections in speech change depending on the audience. A familiar example is infant-directed speech (IDS). When talking to their babies, parents use higher and more variable pitch, shorter sentences, and longer pauses (Fernald et al., 1989). The prosody of IDS is also more hierarchically organized than that of adult-directed speech, highlighting the nesting of information across the time scales of language (Falk & Kello, 2017). In adult speech, the prosodic events of monologues are less clustered than those of dialogues at the longer time scales, suggesting the lack of an interactive audience changes the way speakers structure speech (Kello at al., 2017). Prosodic contours also change depending on the pragmatic intentions of the speakers. For instance, listeners can identify basic speaker intentions such as prohibition, approval, comfort, and attention (particularly for IDS) in languages they do not speak based only on the prosodic features of the utterances (Bryant & Barret, 2007). Changes in prosody are also key in the production of irony and sarcasm. Although there is no common ironic prosodic pattern across speakers (Bryant & Fox Tree, 2005), the production of irony does involve a shift from the prosodic patterns a speaker uses for literal utterances in the same conversation (Bryant, 2010). Speech that is first spoken and later read exhibits differences in prosodic contours as well (Hirshberg, 2020). Prosody also codes for the information status of current words with respect to previous words, the illocutionary force of an utterance, and the speaker agreement with the ideas expressed by their interlocutor (see Cole, 2015 for a review). For instance, interlocutors having argumentative as opposed to friendly conversations, show less similarity in the temporal clustering of their speech (Abney, Paxton, Dale, & Kello, 2014).

1.2.1.2 Gesture Kinematics

The frequency and characteristics of gesture vary depending on discourse level variables. The presence of an interlocutor, for instance, makes gestures more abundant (Bavelas, Gerwing, Sutton, & Prevost, 2008; Holler, Turner, & Varcianna, 2013). Gestures are bigger and less redundant with the contents of speech when interlocutors are co-present and can see each other, as opposed to when they talk on the phone (Bavelas et al., 2008). Importantly, the intention to communicate is key for gesture kinematics. Participants that believe their partner to be learning the specifics of their movements, perform movements that are larger, faster, and more punctuated than participants that believe the partner is just observing the general organization of the task they are performing (Trujillo, Simanova, Bekkering, & Özÿurek, 2018). The familiarity of the addressee with the topic of the conversation affects the quality of the speaker's gestures as well. When speakers retell a story to an addressee that has already heard it, their gestures are smaller and less precise than when they tell the story to a new addressee (Galati & Brennan, 2014). Speakers also tend to encode information about size in both their speech and gesture when their addressee is unfamiliar with the object being discussed, while encoding it only in speech when the object is familiar (Holler & Stevens, 2007).

Gestures are sensitive to the contextual variables of the conversation. In noisy conditions, speakers segment their gestures more as to include more chunks of information in them (Trujillo, Özyürek, Holler, & Drijvers, 2020). Speakers also gesture across different axes depending on the location of their addressee, as to correctly represent the spatial relationships implied in speech with relation to the shared space (Özyürek, 2002). And they gesture more in the periphery when they have attentive addressees vs. distracted (Kuhlen, Galati, & Brennan, 2012). Speakers also produce more detailed gestures when they predict an utterance to be ambiguous for the listeners (Holler & Beattie, 2003). Changes in gesture also follow from the conversational goals of the interlocutors. Speakers produce more iconic gestures and move their hands and head faster during tasks that require them to demonstrate actions than during tasks that require them to just talk about their preferences between options (Danner, Barbosa, & Goldstein, 2018). Speakers that lie show more coordination of head movements with their interlocutors, especially when they have to pretend to disagree with their opinions (Duran & Fusaroli, 2017). Similarly, interlocutors show less movement coordination with each other during argumentative conversations on topics they actually disagree on (Paxton & Dale, 2013a). Lastly, gestures also play a role in the regulation of the interaction. The amount of movement in the eyebrows and the lips differs for felicitous and infelicitous exchanges of the floor (Danner, Krivokapic, & Byrd, 2019). And after listener feedback of miscommunication, speakers' gestures get bigger and more detailed as to help make them more communicative (Holler & Wilkin, 2011).

1.2.2. The Theoretical Argument: Coordination Dynamics of Complex Behaviors

As I briefly mentioned above, elements from the theory of coordination dynamics also encourage the search for coordination patterns at the discourse level. The theory of coordination dynamics of complex behavior proposes complex phenomena such as language, but also schools of fish or oscillations of neural ensembles, to be emergent from the non-linear interactions between low-level parts of the system with much simpler behavioral capabilities (Kelso, 2009, but see Dale, Fusaroli, Duran, & Richardson, 2012, and Rączaszek-Leonardi, & Kelso, 2008 for proposals specific to language). From the point of view of this theory, the interactions of the parts produce structures of higher complexity than the simple sum of the parts that produced them. Critical for our argument, the simple units are proposed to organize into their loosely coupled and temporarily stable groups of interaction, known as synergies or coordinative structures, in ways that directly respond to the goals of the system (Kelso, 2012). In the words of the phenomenon of interest here, the theory of coordination dynamics would expect the discursive structures of language to directly emerge from the temporary synergies of the lower-level behaviors such as the movements, the sounds, the gaze patterns, etc. that a speaker and a dyad perform during communication (Dale, 2015). And, since the synergies are by definition sensitive to the goals of the behavior, it would predict the signals to be assembled differently and vary hand in hand with the communicative goals of the speaker and/or the dyad.

The behavior of the system at the global level is also expected to be robust to perturbations that are not directly related to the goals, nor change the constraints of the behavior in important ways. The complex global coordination patterns of the elements, once formed, would constrain the behavior of the lower, faster changing elements to keep the slow changing and more complex emergent pattern stable for longer periods of time. Within a synergy, functional stability is the goal, and therefore, fluctuations within it or perturbations to its elements give rise to compensatory reorganization of the low-level elements participating in the synergy such that the function of the coordinative structure can be preserved, and behavior can be kept stable (Kelso, 2012). For our purposes here, this characteristic of synergies would mean that stability of coordination for the extended periods of time of discourse variables is possible and even expected. In line with this prediction, for example, hearing yourself with a delay while you narrate an event results in reduced variability in the timing between the peaks of the gesture and the pitch, promoting stronger coordination between the two modalities to achieve stability at the longer time scales of the task despite the perturbation (Pouw et al., 2019).

1.3 Outline of the Dissertation

Both the empirical evidence and the theory of coordination dynamics suggest that multimodal coordination patterns that are stable over longer time scales and that change in response to slow changing variables at the discourse level, should exist. However, the empirical evidence for these patterns is still limited. The existing studies on changes to modality organization as well as multimodal coordination at the discourse level, have mostly looked at local changes, and not at the extended dynamics of those changes' unfoldment. This dissertation explores the temporal and multiscale dynamics of speech and movement signals over extended communicative performances to determine whether the extended coordination of speech and gesture responds to discourse level variables and shows synergy like properties. The exploration of these extended multiscale patterns, and their connection to the broad communicative intentions of the speakers moves forward the study of multimodal situated language use.

The body of work presented in the following chapters aims to take the first steps to study multimodal coordination dynamics at the discourse level, particularly by focusing on the coordination of body movements and speech prosody. Across all the chapters we work with naturalistic datasets obtained from audio and video of naturally occurring communication in the world and in the lab. The focus on naturalistic datasets, although methodologically challenging, allows us to ecologically study language use as it occurs in common discourse contexts. We explore the influence of discourse level variables such as goals, discourse types, and task constraints on the coordination of sound and movements over extended periods of time. We make use of automated techniques such as frame-differencing and the Hilbert transform to extract the time series of movements and audio from the videos directly. And use both factorial analysis (Dale, 2015), cross-correlation, and spectral matching, a method similar to complexity matching (Marmelat & Delignieres, 2007) that we developed, to quantify multimodal synergies.

In Chapter 2, we explore the multimodal patterns of speakers delivering academic talks. We measure and analyze the temporal unfoldment of the speaker's speech rate, amplitude, pitch, body movement, and slide changes to see whether engaging in the type of communication required by scientific presentations results in similar multimodal patterns across speakers over time. We also analyze the coordination of those modalities across the total duration of the presentation to test whether the signals show stable patterns over extended periods of time indicative of stable multimodal synergies. We find some weak regularities in the ways speakers organize and coordinate their prosody, movements, and slides over the duration of the talk. This suggest that despite some influence of the global constraints, the constraints at shorter time scales within the main discursive context might be playing a bigger role in the multimodal behavior of the speakers. We also find high influence of individual differences, which points to the high importance of personal multimodal strategies over a single shared strategy stemming from the shared discursive context.

In Chapter 3, we compare the coordination of sounds and movements in monologues and solo music performances. Our goal is to see whether having an explicit communicative goal influences the coordination of the modalities in complex multimodal performances. We select and compare different discourse types within the speech group and different instruments within the classical music group to further investigate the influence of discursive contexts and manner of execution on the multiscale and local coordination of the modalities. Speech shows higher local synchrony between sounds and movements and stronger multiscale coordination than music, suggesting the overall goals of the performance have an effect on multimodal coordination. The analysis of the individual groups within the categories also shows differences in the multiscale coordination associated to changes in discursive contexts but not to the modes of execution. Of special interest are the cases of a capella singing, which shows multimodal patterns more similar to the music categories; and jazz improvisation, which shows magnitudes of coordination more like those of speech performances. Differences associated to variability in the goals of the performance as opposed to variability in their mode of execution are in line with predictions from the coordination dynamics theory related to the functional sensitivity of synergies.

In Chapter 4, we explore dyadic coordination of sounds and movements during informal conversations carried remotely via videoconferencing. This study aims to investigate interpersonal and multimodal coordination in the context of an extended perturbation: communication in Zoom. Zoom reduces the visual access of the participants to each other's movements to the torso, it interferes with eye-contact, and makes turn taking less natural because of the noise cancelling algorithm, introducing many perturbations into the natural coordination dynamics of the conversation. Within the limitations of the COVID-19 pandemic, we compare the results of our study to the findings of a similar in-person study to assess the effects of the technology in the coordination of modalities between and within interlocutors. The results indicate that the limitations of Zoom disrupt coordination more for movement than for sound, mirroring the level to which the two signals depart from their normal quality in in-person settings. In line with this, multimodal coordination of the sounds and movements of the participants, which isn't mediated nor affected by the quality of the Zoom signals, is also preserved. The comparison to the in-person dataset suggests that despite being less evident, Zoom's audio signals are perturbed enough that speech coordination is only achieved on the lowest frequencies. Due to the lack of an appropriate comparison, the conclusions from this study are only preliminary and more work is necessary to confirm them. However, these results raise interesting questions around the necessary conditions for the development of synergies and their adjustment to perturbation, and are generally in line with the predictions of the coordination dynamics theory: the global coordination is maintained in sound despite the perturbations introduced by Zoom, and perhaps even increased to make up by the lack of coordination in movement.

In Chapters 2 – 4, we establish the existence of multimodal patterns that vary over extended periods of time and as a result of discourse level variables in ways consistent with the predictions of the theory of coordination in complex systems. In Chapter 5, we sketch a proposal that more directly connects multimodal synergies to the pragmatics of language, and we outline some outstanding questions that can serve as future steps for this program of research. We argue that the adaptability to functional goals and the sensitivity to contextual constraints of extended multimodal synergies makes them of special pragmatic value for language interactions. We propose that transient multimodal synergies within an individual and a dyad might give rise to changes in the pragmatic goals as the conversation unfolds and the interactional constraints change. The specific multimodal synergies established during the conversation might even be used as a tool for direct pragmatic “inference” given their direct significance to the members of the dyad as participants of those synergies. Next steps in this line of research, will involve generating and testing empirical hypothesis in line with this proposal. Some initial ideas for future studies are outlined.

Chapter 2

Complex Communication Dynamics: Exploring the Structure of an Academic Talk

2.1. Prologue

In this chapter, I explore a natural dataset of academic lectures to determine how communication modalities are organized and coordinated over time in virtue of a shared discourse context: the presentation of complex information. Using automated and semi-automated techniques, I extracted and analyzed, from the videos of 30 speakers, measures capturing the dynamics of their body movement, their slide change rate, and various aspects of their speech (speech rate, articulation rate, fundamental frequency, and intensity). There were consistent but statistically subtle patterns in the use of speech rate, articulation rate, intensity, and body motion across the presentation. Principal component analysis also revealed subtle patterns of system-like covariation among modalities. These findings, although tentative, do provide initial evidence that the cognitive system is integrating body, slides and speech in a coordinated manner at the longer time scales of discourse during natural language use. The limited effect of time and the predominance of individual differences between speakers, however, indicate that individual constraints at shorter time scales than the ones measured, might take precedence to explain the multimodal dynamics of delivering a scientific talk.

2.2. Introduction

Human communication is highly multimodal. In fact, some researchers have described it as intrinsically multimodal (e.g., Enfield, 2013; McNeill, 1992). When talking to each other, humans don’t only use words, but they also move hands and bodies, they vary pitch, they make longer or shorter pauses, they may rush or slow down, they may whisper or raise the voice, and they may even signal to objects in the environment when they become relevant to the conversation. All these different signals have been shown in previous research to carry important information that contributes to meaning in both production and comprehension (Vigliocco, Pernis, & Vinson, 2014). Indeed, in natural communication, the cognitive system coordinates all these behaviors simultaneously.

Yet, the way in which humans control and coordinate multimodal signals during natural language use remains an underexplored puzzle. The present work takes a step towards filling important gaps in research on multimodality. First, as our review in the next section suggests, most research on multimodality focuses on the use of only a few modalities at a time, often in the context of very specific and controlled stimuli or tasks. Second, much of the research to date focuses on how language users vary a given modality when packaging information at the word or sentence levels. This scope overlooks the fact that natural human communication takes place in discourse contexts that demand extended, complex performances, beyond the sentential level. Third, most research on multimodality examines the coordination of different signals either by applying qualitative methodologies or by using quantitative methodologies on highly controlled tasks.

In this work, we address these issues in the context of a specific complex performance: the delivery of an academic talk. We obtain automated and semi-automated multimodal measurements from naturalistic video data and apply quantitative methods to explore the patterns of dynamic organization and coordination among different modalities, over time. During lecturing, as in any other communicative task, speakers have to coordinate simultaneously different sources of information to successfully transmit their message. Speech, gestures, gaze, lecture slides, and so on, all convey information and need to be coordinated in a way that supports both clarity and engagement, and accommodates the memory and attentional constraints of the speaker and the audience at the same time (e.g., see Abrahams, 2016 for review of a "big data" approach to behaviors supporting successful presentation).

Our goal is to make an initial foray into the manner in which various multimodal signals change over the course of a complex performance, such as giving a talk. There is considerable relevant research that precedes our present undertaking, although much of it has focused on a single modality, or on how different modalities predict talk quality. In the next section, we review prior research on how different modalities—specifically, body movement, voice, and the use of visuals—are deployed during the course of a talk (section 2.2.1). Then, we present some theoretical frameworks that accommodate the coordination of multimodal signals in extended communication, including situated cognition, the "extended mind" view, and dynamical systems (section 2.2.2). Finally, we describe our methods in more detail and state our exploratory predictions (section 2.2.3).

2.2.1. Adapting various modalities during a complex performance

2.2.1.1. Body movement

In the study of discourse and pragmatics, the way language users move their bodies has been thought to be central to their extended natural performance (Goodwin, 2000; Selting, 2010). Gestures, for example, are tightly intertwined with speech and complement spoken information in a non-redundant way (McNeill, 2008). In many approaches—including cognitive linguistics, conversation analysis, and gesture psycholinguistics—speech-accompanying gestures have been studied as conceptually and structurally critical for discourse management (for review and analysis see, Wehling, 2018). Gestures tend to occur in temporal alignment with pitch accents and help group related intonational phrases in bigger semantic units (McClave, 1994; Valbonesi et al., 2002; Yasinnik, Renwick, & Shattuck-Hufnagel, 2004). Postural changes, for their part, accompany shifts in conversational topics, as well as in conversational turns (Cassell, Nakano, Bickmore, Sidner, & Rich, 2001).

Some studies have examined body movement specifically in the context of giving a talk. Although several nuanced patterns are reported, there isn't yet a consensus about how body movement is implicated in such an extended performance. For example, Batrinca, Stratou, Shapiro, Morency and Scherer (2013) found that short presentations were judged by experts to be most effective when speakers gestured and paced more. In political speeches, motion energy was also shown to correlate positively with persuasiveness and overall effectiveness (Scherer, Layher, Kane, Neumann, & Campbell, 2012), and the amount and type of movement was found to predict perception of diverse personality traits, such as agreeableness and extraversion (Koppensteiner & Grammer, 2010).

Much of the work on extended communication rarely examines *how* speakers organize their body over time across their complex performance, whether there are certain inter-participant patterns, and how much individual variability there is. We address these issues in the present work.

2.2.1.2. Voice

Speakers use various acoustic parameters to structure information during speech. Acoustic measures have been shown to relate to the information focus of the sentence. For example, in a series of studies manipulating informational focus in sentences (e.g., subject, verb, or object), Breen, Fedorenko, Wagner, and Gibson (2010) found that speakers naturally employ higher intensity, longer duration, and higher mean F0, to highlight the focal information. Intonational patterns also seem to narrow the search space of possible interpretations of an utterance (Wilson & Warton, 2006). They help listeners predict later elements of the sentence, and are specially used by speakers when an ambiguous interpretation of a sentence is possible (Snedeker & Trueswell, 2003). Prosodic boundaries signal higher meaning groupings across sentences that help listeners understand hierarchical dependencies in the information (Tseng, Pin, Lee, Wang, & Chen, 2005), and reliably convey features of the syntactic structure of the message (Schafer, Speer, Warren, & White, 2000). Additionally, speech rate has been also found to vary depending on the complexity of the linguistic structures being used to convey a message (Cohen Priva, 2017), decreasing when less frequent words and structures are used and increasing for simpler linguistic constructions.

In much of this work, the acoustic feature of interest lies at the word or sentence levels. These are important levels of analysis, of course. In order to build cognitive theory, it is essential to understand the factors that guide the modulation of these signals at the word or sentential levels (e.g., frequency, contextual predictability of lexical information, etc.), as well as the interplay of these factors. For example, Breen and colleagues (2010) showed that, additional factors such as the speaker's awareness of prosodic ambiguity in a given information structure, influences the use of prosodic features, with speakers producing different patterns for contrastively (vs. non-contrastively) focused elements.

In the present work, we wish to scale up the study of vocal features to broader units of analysis in extended communication. In our domain of interest (i.e., giving a talk) vocal features are in fact the signal most commonly examined. Various aggregate acoustic features such as F0 range and variability, speech rate, and disfluencies were found to be related to teacher effectiveness in short lectures (Schmidt, Andrews, & McCutcheon, 1998). Pitch variation and speech rate were also shown to relate to perceptions of liveliness during presentations (Hincks, 2005). In analysis of extended classroom discourse, with a focus on initiation-response-feedback exchanges, Hellermann (2003) found that different prosodic packaging was used for different types of feedback.

Some researchers have also analyzed speech patterns in other domains of natural language performance, such as political speeches or news stories. Work by Hirschberg and colleagues has shown several correlates between acoustic variables and higher-level aspects of speeches, such as discourse structure (Grosz & Hirschberg, 1992; Hirschberg & Pierrehumbert, 1986) and a speaker's charisma (Rosenberg & Hirschberg, 2005). The standard deviation of acoustic measures, for example, predicts higher charisma in political speeches (Rosenberg & Hirschberg, 2005; see also: Shim et al., 2015). Such cues may also serve as strong markers of genre or interactive style, along with other cues such as lexical features (Hirschberg, 2000; Jurafsky, Ranganath, & McFarland, 2009).

2.2.1.3. Visuals

PowerPoint slides have become the norm in scientific presentations. Despite this, few studies have analyzed the *dynamics* of talk delivery using slides. In some domains, such as educational research and communication, the use of slides and visuals has been a critical subject of study (for some review see: Levasseur & Kanan Sawyer, 2006; Vyas & Sharma, 2014). Such research shows that fewer words on slides improve presentation effectiveness (Chen, Leong, Feng, & Lee, 2014), and that the dynamic elements of such visuals are an attractive feature (Moulton, Türkay, & Kosslyn,, 2017).

The manner in which these slides are delivered, however, is rarely a focal point of analysis. Pacing and relative rhythm with vocal and bodily modalities is unexplored. Some survey-based methods asking presenters what they do in their talks suggests considerable individual variability (Brock & Joglekar, 2011).

2.2.1.4. Multimodal coordination of signals

It is important to note again that considerable prior research supports our position that these various signals would be integrated during extended communication. In the case of co-speech gesture, evidence for this integration is plentiful (McNeill, 2000; McNeill, 2008). Gesture has been shown to complement spoken information in speakers' narrations (Cassell & McNeill, 1991; Melinger & Levelt, 2004) and to aid listeners in comprehension (Cassell, McNeill, & McCullough, 1999).

In the context of teaching, Pozzer-Ardenghi and Rother (2004, 2007), using a qualitative approach, found that the gestures and body orientation of instructors during lecture conveyed meaning that amplified and supplemented what was being said by drawing attention to important details. This evidence was taken to be consistent with the view that words, gestures, and visual aids come together to form a holistic meaning unit during the communicative act. Recent quantitative evidence supports the notion that expressiveness during speech connects both prosody and body motion (Pouw & Dixon, 2018; Voigt, Podesva, & Jurafsky, 2014).

2.1.2 Theoretical accounts for multimodal coordination

In this section we review some theoretical accounts that we consider useful for characterizing the coordination of multimodal signals in extended discourse. As our review so far suggests, speakers coordinate an array of multimodal cues, including external tools, during lecturing and other discourse domains.

From a *situated cognition* perspective, the use of multimodal signals, slides included, can be beneficial for cognition. Using external representations when thinking facilitates complex thought by transforming mental computation into perception and action and leveraging the advantages of these two processes. External representations (the slides in this case) also serve as shareable objects of thought that offer a common and persistent referent for different people to attend to and reason with at the same time (Kirsh, 2010). The slides, then, can help the speaker and the audience reduce the cognitive load associated with conveying or understanding the message by offering a shared external representation that supports thinking.

From a more radical point of view, the slides can even be said to become a constitutive part of the speaker's and the audience's cognitive systems. Such view is known as the *extended mind hypothesis*. According to this theory, first outlined by A. Clark and Chalmers (1998), the cognitive system and its processes are not only "internal", but also involve processes that extend outside the head into artifacts and the environment. In this view, the tools we use while thinking function as cognitive aids that, by sharing some of the representational load of the task at hand, become active elements of the cognitive system. As such, these tools need to be included in any explanatory account of that system.

Classical work within this theoretical framework has offered extensive descriptive evidence from complex tasks such as operating airplanes (Hutchins, 1995) or navigating the open sea without instruments (Hutchins, 1983). Researchers in this tradition argue that it is necessary to attend to the informational capabilities of the socio-technical system as a whole (including person and tools), and not just to the cognitive abilities of the individual minds operating the machines.

In the case of interpersonal communication, H. Clark (2003, 2005) proposes that to understand meaning and communication we ought also to extend the window of analysis to include other material elements besides words. The gestures we make, the way we place ourselves and the objects in the space with respect to other elements in the situation, and the specific moment at which we do it, serve communicative functions of their own. Signaling techniques, such as pointing and placing, among others, facilitate coordination during interpersonal communication and ground the speakers' message to the material world. In line with this proposal, Hutchins and Palen (1997) showed that to understand how meaning is created and communicated within a system, it is necessary to attend to all the modalities involved. The authors found that aircrew members engaging in a problem-solving simulation task communicated complex messages by coordinating their words, their spatial orientation, and the organization of their gestures in relation to the tools in the environment. Each of the modalities—external tools included—participated in the creation of the message, which could only be fully understood when analyzing the different modalities in relation to one another.

Besides being multimodal, communication unfolds over time in an organized way. *Dynamical systems* approaches to the understanding of the mind offer novel methodologies to quantitatively characterize the dynamics of a system (i.e., the behavior of the system over time) and measure coordination between the elements that produce them (Richardson, Dale, & Marsh, 2014). These approaches conceptualize cognitive processes as emergent phenomena resulting from the multiplicative interaction of simpler interdependent elements and processes (McClelland et al., 2010). The cumulative effects of the interactions between elements creates patterns of organization that can be measured at different time scales of the system's behavior (Kello et al., 2010). Self-organizing systems also tend to exhibit "synergies" between the elements that comprise them, such that multiple parts of the system come to operate together as a functional whole, exhibiting patterns of coordination and compensation that respond closely to the environmental constraints and the task demands (Kelso, 2009; Riley, Richardson, Shockley, & Ramenzoni, 2011; Shockley, Richardson, & Dale, 2009). In dialogue, interactional routines like turn taking, for example, seem to emerge from the low-level information of multiple channels. The words and syntactic structures being used (de Ruiter, Mitterer, & Enfield, 2006), the prosodic boundaries of intonational phrases (Bögels & Torreira, 2015), postural changes (Cassell et al., 2001), and the patterns of gaze orientation (Rossano, 2013) all seem to play a role on the coordinated dance of turn taking. Speech rate also seems to help with turn-ending prediction, as it may mediate the synchronization of conversational pace between partners (Wilson & Wilson, 2005).

Examining the way behaviors are temporally organized, independently and in relation to each other, offers a window into the characteristics of the system that is producing them. Studying multimodal communication within this framework allows us to explore the structure and coordination patterns of the cognitive system during communication in general, and during the presentation of complex information, in particular. Additionally, this approach allows us to expand on the situated cognition literature by making the first steps toward quantitatively exploring the way in which an extended cognitive system coordinates its parts. If extended cognitive systems do in fact emerge from the interaction of a person and their tools, resulting in interdependent systems, we should find evidence of structure in those systems' behavior across different time scales, and patterns of coordination between their behaviors.

2.1.3. The Present Study

In this work, we use a natural dataset of academic lectures to conduct an exploratory analysis of the temporal structure of multimodal behaviors during this complex extended performance. Analytically, we use a combination of automated and semi-automated methods. As our review suggests, the extant literature typically examines the coordination of modalities either by applying qualitative methodologies (e.g., Hutchins & Palen, 1997) or by using quantitative methodologies on highly controlled tasks (e.g., Valbonesi et al., 2002). Some exceptions include work that has used motion tracking or computer vision object-tracking to capture various features of the speakers' body movement (Chen, Leong, Feng, & Lee, 2014), and work that has used automation to assess public speaking skills (Batrinca et al., 2013; Scherer et al., 2012) and to explore multimodal coordination in short sentences (Voigt et al., 2014). But even so, there remains a paucity of research on the structure of these signals as they vary in time.

Automated and semi-automated methods are useful tools for studying extended communication. Uncovering organizational patterns during communication requires large amounts of data that are costly to obtain using the traditional hand-coding methods common to psychology. The use of extensive videos and automated techniques for analyzing them, allows us to frequently sample behaviors of interest with reduced effort and minimal time costs, while still achieving sufficient data quality. Frame-differencing methods are a useful application of computer vision to measure the movement in a video by quantifying the amount of pixel change across consecutive video-frames (see Paxton & Dale, 2013b for a review). When applied to videos in which the observed movement is coming from an empirically relevant source, or in which this source can be easily isolated, these methods offer a good proxy for variables of psychological interest, such as body movement and coarse gestures (see Pouw, Trujillo, & Dixon, 2018 for a comparison with motion tracking methods). Similarly, software like Praat (Boersma & Weenink, 2010) offers a broad set of tools to analyze and easily quantify diverse acoustic properties of audio signals, such as the intensity, fundamental frequency, formants, and power spectrum. This allows rapid access to prosodic information that contains clues to interesting psychological components of discourse.

In the present work, we are interested in addressing two main questions: Firstly, how do body, slide transitions, and prosodic components of speech—such as pitch, rate, and amplitude—change during the development of a talk as time advances and constraints change? And secondly, how are the changes in these different modalities dynamically related to each other? Are there covariation patterns between them? We focus on a segmentation that is relatively coarse-grained—analyzing subsets of behavior between 4 and 8 minutes in duration. The purpose is to find if particular periods of time in the extended performance enjoy a kind of highlighting in the multimodal signals we analyze.

Motivated by our review above, there are a number of possible patterns which we might observe here. Though we take an exploratory approach, it is instructive to consider these possibilities. One possibility is that the modalities show a linear *increase* over time: For example, speakers could show more irregular movements as their social motivations change during the talk (Koppensteiner & Grammer, 2010), increasing their overall body movement. Similarly, speakers could increase their slide rate towards the end of their talks as they run out of time to cover the material. Conversely, a second possibility is that the modalities show a linear *decrease* over time: Speakers attenuate their gestures (Galati & Brennan, 2014; Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015) and their articulation (Galati & Brennan, 2010) as they build common ground with their audience. Such adaptation would be consistent with a *lowering* of voice signals and body movement across a presentation. Speech rate may also decrease over time as people get into more complex parts of their talks that require more complex linguistic constructions (Cohen Priva, 2017). A third possibility is that the modalities show a *U-shaped* pattern: If the middle of the talk carried most of the new information speakers might utilize increased vocal and bodily expressiveness (Galati & Brennan, 2010) to elaborate the novel concepts at this point of the talk, leading to an inverted U pattern. All these described patterns suggest that some kind of packaging may be dependent upon the *region of time* during an extended performance. An additional possibility, however, is that we fail to find any structure over time in these signals. Such a null effect might suggest that *local* effects of prosodic or bodily variation are sufficient to explain performance within this circumscribed discourse context. In that case, variation may be detectable only at these lower levels (e.g., when considering words or sentences).

Beyond considering each modality separately, we are also interested in the covariation of these multimodal signals to tap into their interdependence and synergies as components of the same system. We do not posit any hypotheses in advance, as there are several possibilities. For example, voice and body movement may pattern in a parallel, reinforcing manner (e.g., Voigt et al., 2014), or may instead pattern in a more complementary fashion. The use of slides may also bear different relationships with the other signals—for example, an increase in the use of slides may be associated with a decrease in some signal (e.g., reduced body movement) but increase in another (e.g., faster speech rate). We examine the various possibilities in the systematic covariation of these multimodal signals by assessing their "compressibility" through principal component analysis (PCA).

2.3 Method

2.3.1 Data

The video recordings of 30 lectures given in the Cognitive and Information Sciences seminar series at UC Merced were analyzed. This seminar series covers a broad range of topics in Cognitive Science, and recordings of its talks are publicly available in YouTube[1] and Vimeo[2]. Each talk is typically an hour long followed by a 30-minute (on average) Q&A session, which was not included in the recording. Each lecture was videotaped from a fixed point within the audience, with the camera set up on a tripod. The camera was usually let to run without any modifications to its focus or position for the length of the talk. This made the videos suitable for meaningful analysis with frame differencing techniques described in the next section.

The recordings included in this study were selected using the following criteria to permit automated video and audio analyses: (a) the camera was fixed in the same position and focus during the whole lecture, (b) the speaker and the slides were always visible in the video, (c) the heads of audience members did not occlude the speaker from the camera view at any point, (d) there was low overlap between the speaker and the slides (determined through visual inspection), and (e) the quality of the audio track was sufficient (i.e., low background noise, no echo) to perform automated analysis in the speech signals.

The resulting sample of 30 videos used in the study included talks by 8 females and 22 males, from various fields, including philosophy, psychology, neuroscience, linguistics, computer science, and math. The sample also involved participants from diverse locations within the US: Of the 30 speakers, 15 were affiliated with institutions in the west coast of the United States, and 15 with institutions in other parts of the country. The presentations included in the analysis were between 34.6 and 81.3 minutes long, with a mean duration of 56.45 minutes (*SD* = 10.21 minutes).

[1] www.youtube.com/channel/UCRcuWjRqxZ2RHvEdZGAliWw

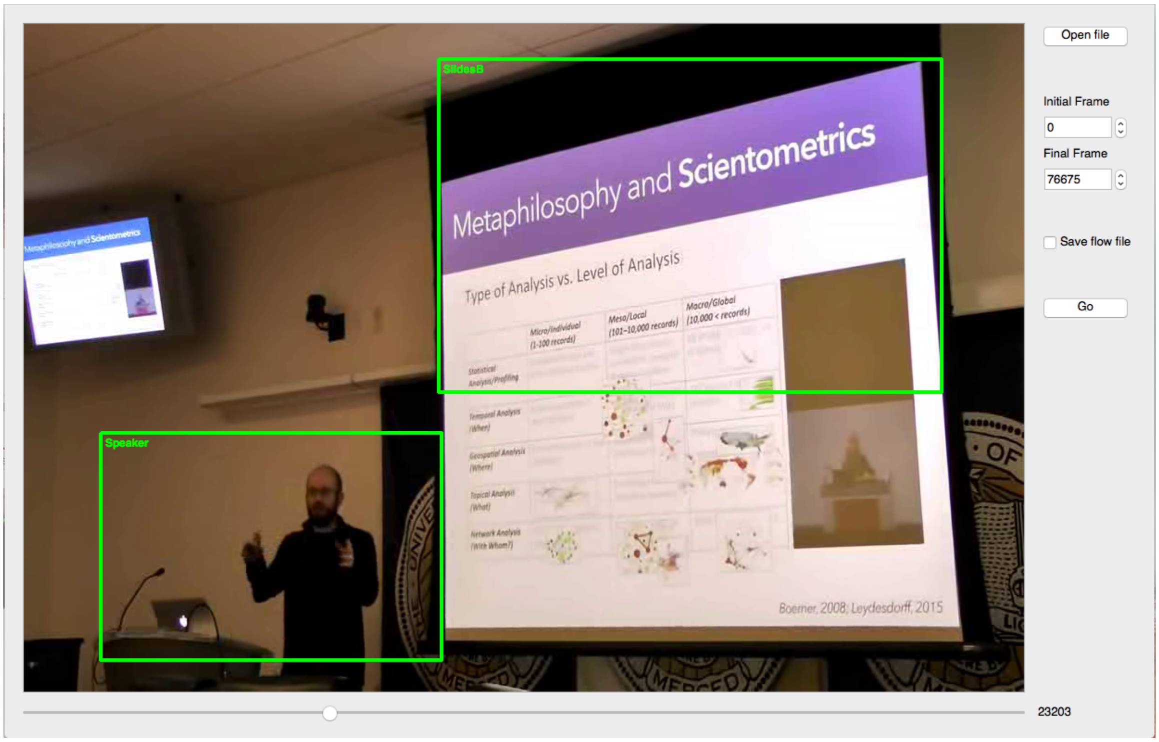
[2] https://vimeo.com/user8418321

2.3.2 Instruments and procedure

2.3.2.1. Video analysis

The videos were downloaded from YouTube and Vimeo in the best possible quality using the iSkysoft Video Downloader (i.e., 1080p for YouTube videos, and 360p for Vimeo videos). These were then converted to AVI format using the Any Video Converter software. The Optical Flow Analyzer (Barbosa, Yehia, & Vatikiotis-Bateson, 2008) was used to conduct the frame-by-frame analysis. This software compares the locations of the pixels in the current frame with the location of the pixels in the previous frame of the video. It calculates the magnitude and direction of the displacement of every pixel within some defined area of interest and, using the frame sampling rate, produces a velocity vector for each of the pixels in that area. The Optical Flow Analyzer then sums all the individual vectors coming from each pixel in the delimited area, and returns a vector reflecting the overall pixel change within each area of interest for every frame of the video (25 frames/s). Two main areas of interest were defined for each video (see Fig. 2.1): one demarcating the area in which the speaker was moving during most of the lecture; and the other one demarcating the area occupied by the slides. The areas of interest were placed in a way that avoided any overlap between them to make sure that any pixel change detected in any of them, presumably corresponded only to the movement of the object they were enclosing (i.e., either the speaker or the slides).

We faced two general issues in using the Flow Analyzer technique with these videos. First, occasionally, the speaker would step in front of their slides. We carefully selected videos which minimized this artifact, though there remained some of these instances in the videos. However, this occasional overlap was neither captured in the speaker's nor the slides' movement data: the areas of interest were placed so that they did not include regions where there was frequent overlap between the speaker's movement and the slides (See Fig. 2.1). Therefore, the area corresponding to the slides usually included only the top portion of the slides to avoid capturing any of the speaker's movement (e.g., when walking in front of the slides) in the slides' area of interest. This strategy allowed us to avoid erroneously registering the speaker's movement as slide movement and vice versa. However, it also resulted in data loss from the lower parts of the slides. This, as discussed below, makes our analysis less sensitive to slide animations (e.g., the use of animated bullet points) in comparison to slide transitions (i.e., changes from one slide to the next). Second, slide changes sometimes changed the illumination of the presenter's body, which could in turn briefly cause a change in pixels, leading to a punctate moment of spurious body movement. In order to ensure that the algorithm was producing coherent results, untainted by these artifacts, we carried out a validation analysis that is presented in a later section.


SlidesB
Speaker
Metaphilosophy and Scientometrics
Type of Analysis vs. Level of Analysis
Boerner, 2008; Leydesdorff, 2015
Open file
☐ Save flow file
Go
23203

Initial Frame 0

Final Frame 76675

Figure 2. 1. Screenshot of the Optical Flow Analyzer showing the placement of the two areas of interest in one of the videos. Note that the areas are placed in a way that avoids information from the slides being captured within the speaker's area, as well as movement from the speaker being captured within the slides' area.

For each video, we divided each time series obtained from the Optical Flow Analyzer into ten windows of equal size (each window representing 10% of the duration of the presentation in question—the duration of the window varied depending on the length of the talk) and got the aggregated measures of the speaker movement, the slide changes, and the prosodic components of the speech signal for each of the windows. This aggregation helps overcome the noise that may be present at finer-grained time scales: if subtle fluctuations in pixel change occurred from the slides, or from the speaker moving in front of the slides, these would only count as brief noise when averaging over several *minutes* of the presentation. Moreover, segmenting the signals into a number of windows as opposed to a number of minutes allowed us to align and compare pragmatically similar discourse segments of the talks (e.g., introduction to the topic, presentation of research findings, concluding remarks) regardless of the talk's exact duration. Two variables were obtained from the video analysis using R (R Core Team, 2016): body movement and slide change[3].

[3] The R scripts and datasets we used in this project, as well as additional figures are publicly available in GitHub: https://github.com/camialviar/ComplexCommunicationDynamics

Body movement. We obtained the mean pixel change within the speaker's area of interest for each window of each talk.

Slide change rate. As the slides change discretely, they produce big spikes of pixel change observed in the output from the Optical Flow Analyzer. We used an automated algorithm to find the spikes of pixel change in the slides signal. For this, we defined a threshold for each talk and identified the times in which big changes occurred. A minimum of 10 seconds between spikes was set to avoid counting multiple spikes coming from videos in the presentation, or other noise sources to be counted as slide changes. To determine the threshold, we plotted the standardized signal and determined visually the number of *SD*s required to minimize the amount of noise (i.e., constant small amounts of pixel change) and maximize the number of spikes (i.e., sporadic big amounts of pixel change) being captured. To validate the algorithm, we hand coded the slide changes (attending also to animations within the same slide) of 3 talks and compared these time series to the ones obtained using the algorithm. We calculated two common signal processing measures to evaluate the algorithm's performance: precision, which captures the ratio of correctly identified events vs. the total number of events found by the algorithm; and recall, which captures the ratio of correctly identified events vs. the total number of events actually present in the signal. When counting slide transitions (i.e., full slide changes) and slide animations (i.e., small changes within the same slide) as a slide change, the average precision and recall rates were 0.849 and 0.421, respectively. When considering only slide transitions as a slide change, the average precision and recall rates were 0.829 and 0.662, respectively, suggesting that the algorithm was especially successful at capturing full changes in the slides. For our analyses, we used the event count within each time window determined by the algorithm, and computed the rate of slide change per minute.

Validation of video analysis. To validate the data obtained with the Optical Flow Analyzer we selected and hand coded 200 5-second segments of 5 randomly selected videos (40 segments from each). The segments from each video were chosen using an algorithm in R that identified and selected 10 5-second high body movement windows (i.e., windows with high mean and low *SD* pixel change from the speaker area), 10 5-seconds low body movement windows (i.e., windows with low mean and low *SD* pixel change from the speaker area), 10 5-second high slide change windows (i.e., windows with high mean and high *SD* pixel change from the slides area) and 10 5-second low slide change windows (i.e., windows with low mean and low *SD* pixel change from the slides area).

All three co-authors coded each of these segments for amount of body movement (on a scale from 1-*low movement* to 7-*high movement*) and presence of slide changes (on a binary 1-*Change,* 0-*No change* scale). The Krippendorff's alpha (Hayes & Krippendorff, 2007; R package: Gamer, Lemon, & Fellows, 2012) indicates there was good interrater reliability for both the body movement judgements (alpha = .768) and the slide change judgements (alpha = .964). To dichotomize the body movement variable (making it comparable to the information we used to select the segments: high or low movement), we calculated the mean movement out of the three raters scores and performed a median split on the resulting values. Everything over and equal to the median movement rating was considered high movement, and everything below, as low movement. The comparison of the software's classification against the human raters' classification showed that: (a) 46 out of the 50 high body movement segments were judged to contain high speaker movement, (b) 47 out of the 50 low body movement segments were judged to contain low speaker movement, (c) 41 out of the 50 high slide change segments were judged to contain a slide change, and (d) 48 out of the 50 low slide change segments were judged to contain no slide changes.

2.3.2.2 Audio analysis

The speech signals were extracted from the videos, filtered, and segmented using Audacity. The noise reduction feature with the default settings was used to reduce the background noise in the sound waves, and the regular interval labeling feature was used to segment the files in 10 parts. Praat (Boersma & Weenink, 2010) was used to calculate the fundamental frequency, speech rate, articulation rate, and intensity of every speech excerpt.

Fundamental frequency. The fundamental frequency was computed using the autocorrelation method ("To pitch" feature) in Praat. The pitch floor and ceiling were set to 75Hz and 400Hz, respectively, for male speakers, and to 100Hz and 500Hz for female speakers, following the recommendations in the Praat manual. The male ceiling was set 100Hz higher than recommended because a good number of the male speakers used perceptually higher pitches when talking. The mean F0 for every time window was obtained. This measure was preferred over the range of F0 because the automated algorithm in Praat produces occasional spurious overestimations and underestimations of the pitch. While these occasional errors in the estimation have serious effects on the estimation of the range, they are less relevant for the calculation of the mean, making the latter a better measure.

Speech and articulation rates. The speech and articulation rates were calculated using de Jong and Wempe's (2009) Praat script. This script uses the peaks in the intensity contour to identify syllable nuclei. Speech rate is calculated as the number of syllables divided by the total duration of the signal; and articulation rate as the number of syllables divided by the duration of the voiced parts of the signal. The silence threshold was set to -26 dB, the minimum pause duration to 3 seconds, and the minimum dip between intensity peaks to 4 dB, given that we are working with noisy signals.

To validate the data, we randomly selected 100 30-second segments from 5 of the videos (20 segments each) and counted the number of syllables on each excerpt. The correlation between our syllable count and the algorithm's was 0.63.

Intensity. We used the "To intensity" feature of the Praat software to determine the intensity contours. The default settings were utilized. The mean intensity for each time window was calculated using the "energy" averaging method.

2.3.3 Data analysis

We conducted two separate analyses on these multimodal signals. First, we tested whether speakers consistently vary these behaviors across their talk. This first analysis tested each modality separately, independently of the others. We were interested in uncovering whether the cognitive system responds differently in various moments of the presentation as it advances. For example, do speakers begin with rapid speech and slow slides changes? And do they end on slower articulation and quick sequences of slide changes as time limitations become more pressing? Because we were interested in both linear and nonlinear patterns over time (see below), we built mixed effects models for each of the dependent variables (pitch, body movement, etc.) using time window (1-10) both as a linear and a quadratic predictor. To control for the effect of between subject variation in the observed trends we introduced speaker identity as a random effect. A maximal random effect structure, with a random intercept and slope for speaker identity, was specified in the model following the recommendations of Barr, Levy, Scheepers, and Tily (2013) and Mirman (2014). The models were built in R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015), and their p-values were calculated using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

The second analysis, as foreshadowed in the introduction, explored the covariation *among* these behavioral, multimodal signals in order to capture their interdependence. If the cognitive system is coordinating these signals together, as interdependent parts, we should be able to substantiate these interactions through a quantitative method. As a first step in this direction, we used principal component analysis (PCA), a dimensionality reduction technique, to explore if these multimodal measures are correlated and how "compressible" they are.

Though we used PCA, there are many dimensionality reduction techniques to choose from (Van Der Maaten, Postma, & Van Den Herik , 2009). Given the relatively sparse data here (300 rows) and our exploratory purpose, we chose PCA as it is among the simplest—it simply performs orthogonal rotation of our observed data. We used the prcomp (R Core Team, 2018) function in R, from the core 'stats' package. For our purposes, we took the unrotated output of the prcomp function. Importantly, rotation would have no effect on the observed compression of our variables, only their interpretation. We return to this issue in the General Discussion, where we discuss future directions for dimensionality reduction.

The prcomp function performs PCA using singular value decomposition (SVD). Variables were neither centered nor scaled beyond our own transformations (as described below). We interpret the variance accounted for by a given component as its corresponding eigenvalue.

Our primary goal is to determine whether there is a compression among the modalities that we measured. In this PCA approach here, this would appear as a smaller number of components accounting for a disproportionate percentage of the variance in these data. There are many approaches to determining this in PCA, such as a bend in a scree plot of eigenvalues or cumulative variance with a cut-off. Here we report eigenvalues above a value of 1.0, indicating that a component is accounting for more variance relative to the original standardized data. We also report percentage of variance accounted for, expecting that a disproportionate variance will be accounted for by fewer components than the dimensionality of the observed data.

2.4. Results

2.4.1. Patterns of variation of the modalities over time

The first goal of the present study was to explore and describe the ways in which the different modalities are organized and change as the presentation progresses and the constraints the speaker has to face change. Fig. 2.2 shows the average patterns of variation over time for each dependent measure. Most variables, with the exception of slide rate, exhibit a decreasing trend over time. In the case of body movement (Fig. 2.2, top left), speech rate (Fig. 2.2, middle left), and intensity (Fig. 2.2, bottom left), this decrement appears to be principally linear. However, in the last segment of the intensity time series (Fig. 2.2, bottom left) a sudden drop is visible. In the time series for articulation rate (Fig.2.2, middle right) an inverted U-shaped pattern is observed. These two characteristics of the observed patterns suggest that non-linear components could also be necessary to capture the distribution of the data.

Table 2.1 shows the raw and standardized coefficients for the mixed regression models of each dependent variable as a linear or quadratic function of time. We used linear mixed effects models with maximal nested random effects to help avoid false positives for each modality, by factoring in a more complex nested model of each individual speaker. Despite that, the fact that we tested multiple models—for the several modalities–does leave the possibility of Type I error lingering. The overall purpose of this paper was to explore these patterns in time, and so future work should follow up by expanding the available dataset in this domain and others, and perhaps expand model complexity (e.g., including more nonlinear terms, individual differences, etc.). This may help fine-tune which of our observed effects are real, and which effects we may have missed (Type II error).

The *t*-tests confirm our observations above, indicating a statistically significant linear effect of time on speech rate (p = .016) and intensity (p = .016), and a marginally significant effect (p = .054) on body movement. The negative sign of the slope for all variables suggests that the rate and intensity of speech, and the amount of movement decrease as the presentation advances (see Fig. 2.2: middle left, bottom left, and top left).

Significant effects for the quadratic component of the regression models were found for intensity (p = .023) and articulation rate (p = .032). In the case of intensity, this suggests that the decrease in the amplitude of the signal over time, observed in the bottom left panel of Fig. 2.2, is not completely linear: intensity decreases at a somewhat regular pace, and dramatically drops off at the last window. In the case of articulation rate, the significant quadratic term suggests that the number of syllables pronounced during actual voiced time, increases as speakers go into the middle section of the talk and decreases again as they move into the final segments (see Fig. 2.2, middle right).

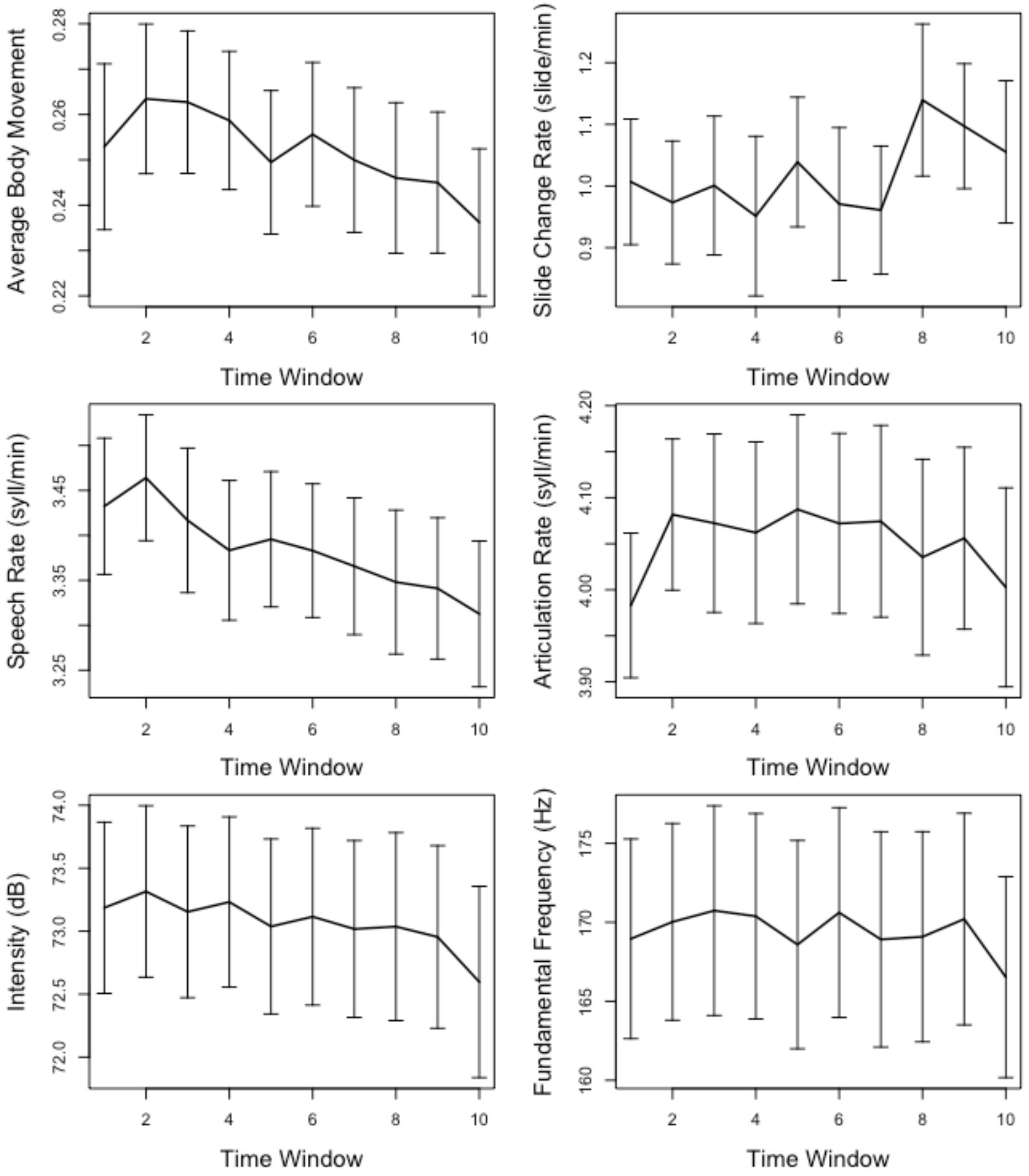


*Figure 2. 2***.** Overall patterns of variation over time for the different modalities of information: body movement (top left), slide change rate (top right), speech rate (middle left), articulation rate (middle right), intensity (bottom left), and fundamental frequency (bottom right). The time windows correspond to equally long segments of the talks, each comprising 10% of the duration of them. Error bars indicate SE of the mean and are corrected by within-subject measurements.

Table 2. 1.*Coefficients for the linear and quadratic mixed effects models.*

DV (units)	Predictor	*B*	ß	*SE*	*t*	*p*
Body movement (pixel change)	t	-0.0023	-0.026	0.013	-2.00	.054
	t^2	-0.00039	-0.0045	0.0031	-1.41	.17
Slide rate (slides/min)	t	0.012	0.019	0.017	1.13	.27
	t^2	0.0023	0.0037	0.0070	0.53	.59
Speech rate (syll/s total)	t	-0.014	-0.034	0.013	-2.57	**.016**
	t^2	-0.00017	-0.00042	0.0037	-0.11	.91
Articulation rate (syll/s voiced)	t	-0.0010	-0.0019	0.013	-0.14	.89
	t^2	-0.0036	-0.0069	0.0030	-2.26	**.032**
F0 (Hz)	t	-0.19	-0.0053	0.0067	-0.80	.43
	t^2	-0.080	-0.0023	0.0018	-1.24	.22
Intensity (dB)	t	-0.054	-0.014	0.0056	-2.55	**.016**
	t^2	-0.0087	-0.0023	0.00096	-2.40	**.023**

Note: The reported *SEs* correspond to the models using the standardized data. *p*-values < .05 are shown in bold.

The effects of time on the different variables are modest, as observed by the small beta coefficients in Table 2.1. This suggests that the specific patterns we have examined may be insufficient to fully describe the ways in which the different modalities change during these extended periods of time. More specifically, individual differences may play an important role in the observed trends. In fact, an initial exploration shows that different individuals do seem to be quite variable in their organization of the modalities. For example, speaker 6 and speaker 29, whose body movement and speech rate data are presented in Fig. 2.3. (left and right, respectively), show completely opposite patterns for both of these modalities. Speaker 6's body movement decreases and her speech rate increases as the lecture advances, whereas speaker 29's body movement increases and her speech rate decreases over time. The high variability of the individual speakers' patterns may be causing the random effect structure in the model to take up most of the variance, which in turn may be reducing the contribution of time as a fixed effect.

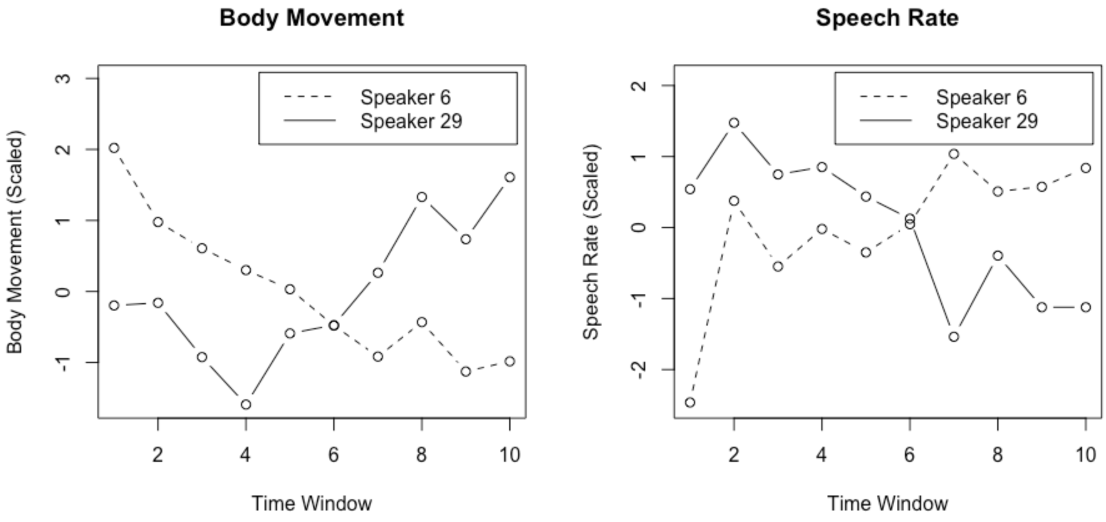


Figure 2. 3. Body movement (left) and speech rate (right) data for speaker 6 and speaker 29. The subjects show opposite patterns for each of the variables and exemplify the types of unique strategies that different speakers exhibit during their talks. Individual differences in the use of each modality might have a role in explaining the small effect of time in the mixed-effects models.

2.4.2. Covariation between modalities

The second goal of this study was to determine whether there were "system-like" properties in the speakers' multimodal behaviors—that is, to test if the different modalities varied in a coordinated way. We combined all the behavioral measurements, window-by-window, in one data matrix. Each of 300 rows in this matrix represented a speaker (N = 30) in a given 1/10th window, with each column reflecting one of the 6 modalities, as measured above. We used PCA to examine whether a small number of dimensions (components), discovered through the spectral decomposition of the data matrix, would adequately describe multimodal performance in a "compressible" manner. We applied PCA to two different covariation matrices: one standardizing the data across the whole sample, and the other standardizing the data within subject.

The standardization across the whole sample highlights the differences *between speakers* by scoring their performance on different variables relative to the group mean: thus, each speaker's score on a given variable would be high or low relative to the other speakers. This standardization procedure allowed us to ask questions about the covariation patterns in the distribution of "high" and "low" scores across the different variables. For example, do people who move more tend to have more slides and speak faster? Or, in contrast, do those who move more have fewer slides?

Table 2.2 shows the loadings of this PCA solution, as well as the eigenvalues for each of the factors. The first 3 components account for approximately 75% of the variance, have eigenvalues greater than 1 (i.e., accounting for more variance than any single original variable), and suggest the existence of patterns in the strategies that different speakers adopt when giving a talk. The first component accounts for 34.63% of the variance and involves a mixture of four variables: body movement, speech rate, articulation rate, and intensity. All these variables load positively on this component (they all have the same sign), suggesting that speakers who tend to speak faster also tend to speak louder and move more. The second component accounts for 20.94% of the variance and contains the variation of three variables: body movement, slide rate, and intensity. The signs of the loadings suggest that speakers who have more slide changes also tend to move more, and speak softer. The third component accounts for 17.33% of the variance and combines the variation of two variables: fundamental frequency and intensity. The signs of the loadings suggest that speakers who have a higher pitch tend to speak more softly.[4]

Table 2. 2**.** *PCA solution standardizing across the sample.*

DV	C1	C2	C3	C4	C5	C6
Body movement	**.22**	**.66**	-.01	.72	-.05	-.01
Slide rate	-.07	**.74**	.09	-.64	.15	.07
Speech rate	**.61**	-.08	-.09	-.12	-.15	.76
Articulation rate	**.57**	.02	-.03	-.23	-.50	-.60
F0	.06	.04	**-.95**	-.05	.28	-.12
Intensity	**.49**	**-.12**	**.28**	.01	.79	-.22
Eigenvalues	**2.07**	**1.25**	**1.03**	0.68	0.62	0.32
Percent σ^2	**34.6**	**20.9**	**17.3**	11.4	10.4	5.3

In a second exploration, we wished to investigate how modalities vary across the talk *irrespective* of the overall magnitude of the behavioral variables. In other words, if a speaker increases *her* body movement does she also increase *her* slide changes? To explore this, we conducted a PCA on *z*-scores computed *within* speakers. In this standardization procedure, we used each speaker's individual mean for each variable to standardize their score for each of the time windows of that variable. This way of standardizing highlights the differences *between time windows*, allowing us to ask questions about the way the different variables co-vary at different moments in time as they go above or below the speaker's mean.

Table 2.3 shows the loadings and eigenvalues of this PCA solution. Only the first component has an eigenvalue greater than 1, and it accounts for 29% of the variance. This solution is more difficult to interpret than the first PCA solution, as it shows less compressibility of the measures, as evidenced by the low percentage of variance accounted by the only component with an eigenvalue greater than 1. Nevertheless, we attempt to

[4] This last component could reflect sex differences in pitch. This was not a variable of interest to us, and so we do not test this in our dataset. However, it is also important to note that some males exhibited perceptually higher pitches, so the relationship suggested by component 3 between pitch and intensity might not be exclusively due to sex.

unpack what the first component's loadings suggest. The first component (29% of the variance) mainly includes variance from speech rate, intensity, articulation rate, and F0. This suggests that as speakers increase their speech rate, they also tend to increase their pitch and their loudness.

Table 2. 3**.** *PCA solution standardizing within subject*

DV	C1	C2	C3	C4	C5	C6
Body movement	**-0.10**	0.79	-0.22	-0.08	-0.54	0.10
Slide rate	-0.09	0.31	0.93	0.17	0.07	0.00
Speech rate	**-0.53**	-0.26	-0.04	0.51	-0.24	0.58
Articulation rate	**-0.50**	-0.33	0.16	-0.34	-0.50	-0.50
F0	**-0.46**	0.12	0.00	-0.66	0.43	0.39
Intensity	**-0.49**	0.28	-0.25	0.39	0.46	-0.50
Eigenvalues	**1.74**	.99	.89	.79	.70	.30
Percent σ^2	**29.0**	16.6	14.8	13.2	11.7	5.0

In general, the covariation patterns observed between the modalities suggest there could be system-like synergies in how different modalities vary over time. However, it is difficult to uncover their specific patterns of covariation, since some of the dependent variables load together in more than one component but exhibit contradictory relationships across components (see for example in Table 2.3 the loadings of body movement and articulation rate in C1 and C2).

Using orthogonal data rotation yielded only subtle features of compression. Probing further into the results of this PCA solution may require a larger sample size. Future explorations of the data may involve using more fine-grained time windows or including other relevant variables for the behavior under study (e.g., informational complexity of the information being conveyed). In addition, it may be useful to explore other dimensional reduction techniques (Van Der Maaten et al., 2009).

2.5. Discussion

This study explored how different modalities of communication are used by speakers during the delivery of an academic talk, and the relationships between those modalities. In regards to our first objective—to explore how different modalities are organized and change over the course of the lecture—our findings suggest some small but significant effects. Out of the 12 coefficients we tested (2 fixed effects per 6 modalities), 4 showed significant but small effects.[5] In line with a dynamical systems account, time

[5] We set alpha for *p* to 0.05. By chance, one would anticipate no more than about 1 of our coefficients to yield significance. Four significant coefficients suggest our results are not due to type I error, though it is important to note that the observed effects are small and future analyses, especially where models might become more complex, would benefit from a correction for alpha.

predicted significantly the speakers' speech rate, articulation rate, intensity, and marginally the speakers' body movement. As the presentation advanced, speakers reduced their volume, their speaking rate, and their movement. They also had higher articulation rates in the middle of the talk than at the beginning.

A simple and quite obvious explanation for this general decrease across modalities is a fatigue effect, given the speakers' sustained performance over the one-hour lecture. However, another, perhaps more interesting possibility is that the observed trends might be related to an increase in informational complexity as the talk advances. As more complex or less established information is reached (e.g., the results or conclusions of the presented research), it is reasonable to expect presenters to speak more slowly, make longer pauses, and decrease their volume as a result of uncertainty or difficulty in conveying the information. Consistent with this possibility, speech rate has been shown to depend on informational complexity, decreasing as complexity increases (Cohen Priva, 2017). Testing this possibility would require further exploration of the data. Future analyses of datasets of this kind may involve classifying or quantifying the slides' content in order to gauge the complexity of the information presented at each time window.

The magnitude of the effect sizes resulting from the regression analysis suggests that the dependency of the different modalities on time is limited. As we have mentioned before, this could reflect the importance of individual differences. It is possible that speakers vary widely in their use of the different modalities during the talk, such that the unique impact of time on each speaker (captured by the maximal random effect structure) accounts for most of the variance in the models, obfuscating the impact of time on individual modalities. Some initial explorations of the individual patterns of the speakers seem to support this possibility, as illustrated in Fig. 2.3.

We also explored model comparisons between models with different random effect structures to select those with the best fit, following a reviewer's suggestion. Ultimately, we decided to continue reporting the models with the maximal random effect structure here for two reasons. First, specifying the maximal random effect structure is a more conservative approach that reduces the possibility of type I error (Barr et al., 2013). Second, in light of the individual differences we've noted, we wished to control for the effects of individual subjects on the general trend (thus maintaining the maximal random effect structure). Model selection could be most valuable when having to specify more complex models, with additional parameters. In future work, such models could include the slide content or informational complexity, or the state of the other modalities at current or previous time steps. The best fitting of these models could provide insight about the specific variables that predict the organization and change of communication modalities during lectures. Doing model selection for the parameters of the random effect structure could also be a good way of testing for individual differences: the change in R-squared and in model fit indices between a maximal model and one without a random slope could illustrate the contribution of individual trends to the change of modalities over time.

We analyzed relatively large bin sizes across time, averaging measurements into 10 temporal bins. This methodological choice enabled us to stabilize behavioral measurements (e.g., reduce noise from small fluctuations in pixel change by averaging over several minutes), and to compare pragmatically similar discourse segments across speakers (e.g., the dip in many of the behavioral signals at the conclusion of the talk, as illustrated in Fig. 2.2, or a peak in slide change rate in window 8, top-right panel of Fig. 2.2[6]). However, our choice of relatively long time windows may have contributed to the modest effect sizes observed, since constraints at longer time scales (e.g., at the approximately 6-minute time windows we have used) may be less important than local constraints in guiding the organization of behavior during a talk. As we have discussed, it's possible that the way in which speakers recruit multiple modalities during a talk is significantly influenced by the complexity of the information presented and, specifically, by how concepts are represented in the slides (e.g., through text or figures). These factors (the complexity of the information, and the modality of information in the slides) vary locally, at shorter time scales than the one we used. Previous research has shown that factors pertaining to content can influence the coordination of multimodal behavior: for example, the use of images during classroom instruction is accompanied by gestures that help disambiguate them and aid integration between what is being said and what is being shown (Pozzer-Ardenghi & Roth, 2007). The interdependence of gesture, language, and external objects might be of such importance (Clark, 2005; Hutchins & Palen, 1997) that the best predictor of the state of each modality at any given time is the state of the other modalities, given the constraints of the specific message being conveyed and the objects available in the environment. Future work could delve deeper into behaviors at a smaller temporal grain size than the one we have chosen here, as it might be a more appropriate time scale to observe this behavior of the system.

Regarding our second objective—of examining whether modalities vary in a coordinated way, revealing system-like properties—results from both PCA solutions offer only suggestive glimpses of system-like covariation among modalities. These behavioral signals are, at least to a small extent, "moving together," suggesting that the cognitive system is partly controlling them as a coordinated unit. From a dynamic systems perspective, this would help understand the ease of multimodal coordination, as the degrees of freedom the cognitive system needs to control during the task are reduced (Kelso, 2009; Riley et al., 2011; Shockley et al., 2009). The first PCA revealed patterns in the general strategies that different speakers adopt when presenting complex information: it suggested that people who talk faster also tend to move more and talk more loudly. This is in line with some of the findings in the co-speech gesture literature indicating that gesturing facilitates fluency in speech, in particular when it involves spatial language (Morsella & Krauss, 2004; Rauscher, Krauss, & Chen, 1996). Similarly, people who have more slide changes in their presentations tend to move more. This makes sense given evidence that shows that when speakers change topics (Cassell et al., 2001) or integrate new ideas with prior discourse (Alibali et al., 2014; Enfield, 2009), they produce increased movement and gestures.

[6] We examined whether this peak reflects an increase in graph usage, following a reviewer's suggestion. We calculated the percentage of slide changes involving graphs for 5 speakers, for whom window 8 had a big peak. The results suggest that the peak was not driven by an increase in the percentage of graphs being used between window 7 (M=47.44%) and window 8 (M=49.31%). A detailed content analysis would be required to better clarify the variables driving the increase.

This pattern from the first PCA, in conjunction with the trends from the second PCA, has implications for the extended mind hypothesis (Clark & Chalmers, 1998). It suggests that lecture slides are integrated with the speaker's behavior and may be functioning as a part of the cognitive system itself. Studying the specific roles of the slides in the communicative process in more depth might help further clarify this possibility. Note that although multiple combinations of the modalities are possible speakers tend to adopt and maintain a stable strategy across their whole presentation[7]. This indicates that the system is organizing itself in a way that achieves stability at longer time scales. Speaker behavior enters a particular "region of activity space" (i.e., a specific subset of all the possible combinations of the modalities), and it remains stable in that region. This is consistent with the idea that individual differences are prominent, which was also suggested by our regression analyses. Although in the present work we cannot identify the causal locus of the stabilization of behavior, a variety of factors likely contribute, such as physical constraints of the environment, and pragmatic constraints of the topic or the discourse context (including audience feedback) contribute to that process.

The second PCA analysis also reveals some slight covariation across the different modalities as they change over time. Though our results are only suggestive, if speakers subtly adapt their modalities relative to one another during a talk, this would be characteristic of soft-assembled, context-dependent systems (Kelso, 2009, Richardson et al., 2014). More broadly, and perhaps more boldly, such a synergy would support the notion that different modalities weave into each other, and act as a functional group to give rise to meaning (Hutchins & Palen, 1997; Pozzer-Ardenghi & Roth, 2007). In general, the patterns from the second PCA are difficult to interpret, and some of its emerging patterns are puzzling. For instance, literature examining prosodic prominence (Cole, Mo, Hasegawa-Johnson, 2010) shows that speech rate decreases as F0 and intensity increase. This is in contrast with some of the patterns of the second PCA (i.e., positive covariation between all three), which suggest that there might be more than prosodic prominence to the speakers' multimodal performance. One possibility is that the observed relationships are purely physical and respond to the physiology of the vocal tract: a correlational study that asked people to change the vocal effort of their speech found that when people speak softer, their fundamental frequency, and speaking rate decrease as well (Black, 1961). An alternative possibility is that the covariation reflects emotional activation during public speaking. Though still debated, studies on the acoustic properties of emotion, have found that high emotional activation is usually accompanied by speech with higher fundamental frequency, higher intensity, and higher rate (see Scherer, 2003 for a review). Still another possibility is that the acoustic correlates of prosodic prominence that have been reported at the word and sentential level, do not hold when looking at the dynamics of longer speech excerpts: the temporal scale of the observations is too coarse to capture more fine-grained relationships between modalities.

[7] For the interested reader, Figures showcasing the clustering of the data when projected in the first two components of the first PCA solution are available in the GitHub repository for the project:
https://github.com/camialviar/ComplexCommunicationDynamics

It is possible that additional measures related to the informational flow of the talk or the quality of the performance are needed to clarify the patterns in the data. For instance, it might be the case that effective and ineffective presenters have completely different strategies of multimodal performance and that the mixture of ability levels in the sample is making it difficult to discern the specific patterns of covariation in the speakers' multimodal behavior. Previous research has shown that individual differences in verbal working memory (Gillespie, James, Federmeier, & Watson, 2014), spatial working memory (Chu, Meyer, Foulkes, & Kita, 2014), and phonemic fluency (Hostetter & Alibali, 2007), among others, affect the type and rate of co-speech gestures speakers make. Other work has suggested that the effectiveness of a speaker's message is associated with multimodal behavioral signatures. For instance, the proportion of grooming movements and gestures that a speaker makes modulates how effective gestures are in aiding message comprehension (Obermeier, Kelly, & Gunter, 2015). For future studies, it might be interesting to obtain performance measures like learning outcomes of the audience or overall enjoyment of the presentation to test if these are associated with distinctive combination patterns of the modalities. It may be the case, for example, that speakers that show more consistency in the patterns of multimodal behavior they use achieve higher comprehension of their message.

Finally, some general caution is required when interpreting the results presented here. Firstly, while natural datasets allow to study behavior as it happens in the real world and afford more ecologically valid inferences, they do so at the price of experimental control. Aspects of the video recording that were out of our control such as distance from the camera, general light level in the room, calibration of the microphone volume and the camera brightness at each talk could have affected the results presented here. We made a deliberate effort to reduce the sources of noise by carefully selecting the videos and preprocessing the data, but it is important to acknowledge this issue. Secondly, while automated methods constitute great alternatives to analyze bigger datasets, they have some limitations in their accuracy. The methods used in this paper showed good levels of accuracy when compared to the performance of human coders. However, they may have introduced some noise in the signals that could have made it more difficult to detect the actual coordination of modalities. Thirdly, we did not consider the content of these presentations, as it was outside the scope of the present analysis. Still, as we have acknowledged, the ebb and flow of slide changes, body, and voice may have been time-locked to certain aspects of the information presented in the slides or talk. Future research could use a co-registration of transcribed presentations, slides, and dynamics derived from automated methods in order to test this, although such corpora are still not widely available. Lastly, our dataset comprised talks given by mostly male experienced presenters, based in US institutions, and under minimal time constraints. The trends and patterns found in this study might (and most likely would) change when dealing with other levels of experience in presentations, more gender-balanced samples, more geographically and culturally diverse samples, and different time constraints. Exploring these variables and their effects on multimodal signal control is an interesting direction for future research.

2.6. Conclusion

The use of different modalities of communication during complex information presentation exhibits some weak but interesting regularities. System-like compression is observed in the way the different modalities are coordinated. Body, speech, and slides may begin to approximate an integrated unit during communicative performance across the presentation giving some support to the idea of an extended cognitive system. Further exploring individual differences and including measures of informational complexity could help further clarify the specific coordination synergies that speakers engage in during multimodal communication.

Chapter 3

Multimodal Coordination of Sound and Movement in Music and Speech

3.1. Prologue

In this chapter, I compare the multimodal coordination of movement and sounds for speech monologues and solo music performances to study the influence of explicit communicative goals on multimodal coordination. I also compare various discourse contexts within the speech group and various instruments within the classical music group to investigate whether multimodal coordination varies as a function of discourse type or physical means of execution. I examined similarities and differences in multimodal, multiscale coordination of speech and music using two complementary measures: I computed power spectra for envelopes of acoustic amplitudes and motion amplitudes, and correlated the spectral powers across modalities as a function of frequency to study coordination over several minutes at a time. I also correlated the smoothed envelopes of sounds and movements and examined peaks in their cross-correlation functions to study coordination over a few seconds a time. Speech performances yielded stronger, more reliable relationships between sound and movement compared with music for both measures.

Multimodal coordination within groups showed subtle differences for changes related to discourse type but not instrument performed. Interestingly, our two test groups also followed this pattern: a cappella singing, in which the voice is used as an instrument. patterned more with music, and improvisational jazz piano, which has been likened to a conversation, patterned more with speech. Taken together, results suggest that nested temporal structures in sound and movement are coordinated as a function of communicative aspects of performance and vary for different discursive contexts. The differences associated to variability in the discursive goals of the performance as opposed to variability in their mode of execution are in line with predictions from the coordination dynamics theory related to the functional sensitivity of synergies.

3.2. Introduction

Whenever we see someone speaking or playing an instrument, the sound we hear is experienced as integrated with the movement we see. In fact, movements of the lungs, vocal folds, tongue and lips, causally produce the speech we hear, and movements of the fingers, arms, lungs, torso, and even feet, causally produce the music we hear. These movements may seem somewhat isolated and subtle at times, even invisible to the perceiver, but the musculoskeletal system works in concert to produce many kinds of movements (Bernstein, 1967). Posture is dynamically adjusted, and poise maintained with visible body movements, even when targeted movements may be difficult to perceive. The overarching aim of our study is to investigate whether these visible movements are coordinated with sounds in the production of speech and music, and whether coordination depends on the intentional category of behavior, rather than its physical manifestation. To illustrate, talking and singing are two different categories produced with the same physical apparatus. Talking is primarily communicative in nature, whereas singing is more expressive and evokes feelings and connection with the music. We aim to learn about the roles of multimodal coordination in communicative and expressive performances.

There are many domain-specific aspects of sound versus movement, and speech versus music, but we can start to compare and relate them by first recognizing that both acoustic and motion signals vary in amplitude over time, for both speech and music. Moreover, the fluctuations in amplitude are far from random—sound and movement energies are clustered in time, sometimes periodically and sometimes aperiodically, but always with temporal patterning (Martin, 1972; Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013; Rohrmeie, Zuidema, Wiggins, Scharff, 2015). Specifically, amplitude envelopes have *multiscale structure*—acoustic energy comes in very brief clusters (phones and notes) that are grouped together to former larger clusters (syllables and motifs), which are themselves grouped to form even larger clusters, and so on across a wide range of time scales. Multiscale structure is also found in human movements (Delignières & Torre, 2009; Delignières, Deschamps, Legros, & Caillou, 2003; Delignières, Lemoine, & Torre, 2004; Hausdorff et al., 1996) and can be illustrated in the movements of speech and music. For example, multiple jaw oscillations (syllables) are nested within the movement of each breath group, and multiple guitar notes and chords may be nested within each fret change.

The commonality of multiscale structure makes it an apt basis for comparing and relating multimodal coordination in speech and music. In the present study, we quantified multiscale structure of sound and movements via temporal structure in the amplitude envelopes of corresponding acoustic and motion signals. Specifically, we measured nested clustering in the spectral power of amplitude envelopes, i.e. the *modulation spectrum*, and we compared clustering in audio and video recordings of speech and music performances. We also measured multimodal coordination more directly in the amplitude envelopes by measuring the peaks of their cross-correlation functions.

In general, we found reliable correlations in both measures of relationship between acoustic and motion signals, as expected based on previous research. However, we also observed differences in effects across different types of speech and music that provide evidence for tighter multimodal coordination during communicative performances. Next, we review the prior literatures on multimodal coordination in speech and music, starting with perception and followed by production. We then review the relevant prior studies of multiscale structure that, taken together with the literature on multimodal coordination, lead us to the present study.

3.2.1. The Coordination of Sound and Movement in Speech and Music

The integrated perception of sound and movement is supported by studies of both speech and music production and perception. In the case of speech production, for example, the area of the opening of the mouth at any given time has been shown to be robustly correlated with the amplitude of the speech wave 100 to 300 milliseconds later (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). More generally, recordings of tongue and vocal tract movements can be used to learn a causal (forward) model to produce corresponding speech sounds (Kello & Plaut, 2004).

There is also evidence that bodily movements causing little or no sound may nonetheless be coordinated with the production of speech sounds. For instance, amplitudes of movements like beat gestures of the arms have been shown to correlate with peak amplitudes in the speech signal (Pouw, Harrison, & Dixon, 2019). The temporal patterning of such amplitudes carries enough information for listeners to synchronize their own beat gestures to the speaker based on sound alone (Pouw, Paxton, Harrison, & Dixon, 2020). The coupling of speech and gesture is continuously ongoing, as evidenced by a study showing that disruption in one modality quickly spreads to delay movements in the other modality (Chu & Hagoort, 2014). In fact, recent evidence shows stronger coupling of speech and gesture (i.e., more synchronization) under a delayed auditory feedback perturbation, suggesting that speech-gesture synchrony might play a role in maintaining the stability of the speech production system (Pouw & Dixon, 2019). Articulatory movements and manual gestures also reflect the broad prosodic structure present in speech, lengthening under prosodic prominence and around prosodic boundaries (Krivokapić, Tiede, & Tyrone, 2017), and in some cases even extending over them to indicate larger prosodic groupings (Yasinnik, Renwick, Shattuck-Hufnagel, 2004).

Studies of musical performances have similarly revealed correlations between movements performed and sound produced. For example, the peak height of the pianists' fingers (Dalla Bella & Palmer, 2011), as well as the acceleration of the clarinetists' fingers (Palmer, Koopmans, Loehr, & Carter, 2009), both increase in response to faster tempi as way of preserving the temporal accuracy of the melody. Similarly, a study with saxophonists showed that the coordination between the tongue and the fingers was important for stability in tempo, and that different articulation techniques produced sound qualities distinct enough for expert saxophonists to correctly identify the technique used to produce the melody (Hofmann & Goebl, 2014). Moreover, ancillary gestures not directly involved in producing sound have been shown to correlate with the timing, timbre, and loudness of the music, especially in moments of harmonic transition (Teixeira, Loureiro, & Yehia, 2018).

The link between sound and movement in speech and music can also be seen in the effects of perceived movements on perceived sounds. For instance, evidence shows that body movement affects how infants (Phillips-Silver & Trainor, 2005) and adults (Phillips-Silver & Trainor, 2007) perceive ambiguous rhythm sequences by facilitating groupings in auditory stimuli that mirror rhythms in the corresponding movements. In the case of speech, the classic McGurk effect highlights the influence of the perceived movement of the mouth on the perceived phoneme being uttered (McGurk & MacDonald, 1976). Evidence from fMRI indicates that premotor and motor cortex, as well as the left cerebellum, are involved in beat perception by simulating periodic movements that predict upcoming beats (see Gordon, Cobb, & Balasubramaniam, 2018; Patel & Iversen, 2014). More generally, areas of the premotor cortex active during music production are also active during music perception, particularly in trained musicians (see Zatorre, Chen, & Penhume, 2007 for a review). An analogous effect has been found in the sensorimotor cortex during passive speech comprehension tasks (see Schomers & Pulvermüller, 2016 for a comprehensive review).

The literature reviewed thus far provides a wide range of evidence confirming and detailing the relationship between movement and sound for speech and music. However, the relationship is by no means fixed, as in gestures and other ancillary movements that may vary in their relationship to sound, and aspects of sound that have complex relationships to movements not measured or perceived (e.g. changes in loudness due to subtle changes in air pressure). The variable and complex relationship between movement and sound leads us to ask whether the relationship might vary as a function of the kind of speech being spoken or music being performed. Our approach to addressing this question is based on recent studies, reviewed next, that found large and consistent differences in the multiscale structure of sound in different kinds of speech and musical performances, measured directly and automatically in sound recordings.

Multiscale structure refers to the organization of behavior in time or space, specifically across temporal or spatial scales of observation. Behavior is said to have multiscale structure when structured, non-random variability is found in behavioral signals that are windowed across a wide range of temporal or spatial resolutions. With respect to structure across temporal scales, dozens of studies have found human behavior to exhibit a particular kind of multiscale structure (Kello et al, 2010), in which temporal clustering grows in proportion with timescale. This type of multiscale structure is generally referred in the literature as *1/f scaling* (e.g., Van Orden, Kloos, Wallot, 2011) because growth of temporal structure with timescale can be expressed as an inverse relationship between spectral power and frequency. We use the more general term *hierarchical temporal structure* here to refer to a general trend for temporal structure to grow with timescale, including deviations from this trend that may be informative about underlying processes.

Most relevant to the present study, Kello, Dalla Bella, Médé, and Balasubramaniam (2017) quantified the hierarchical temporal structure in speech and music recordings, as expressed in the amplitude envelopes of the acoustic waveforms. Specifically, they extracted above-threshold peak amplitude events, and quantified the degree of event clustering as a function of timescale. Short time scales measured small-scale clustering that roughly corresponded to individual phonemes and notes, whereas larger time scales measured larger-scale clustering that roughly corresponded to words and musical phrases, and so on. Allan Factor analysis (Allan, 1966) was used to quantify the pattern of clustering across time scales that corresponds to the degree and shape of hierarchical temporal structure for a given sound recording.

Kello et al. (2017) measured and compared Allan Factor functions for over 160 sound recordings, covering time scales from about 30 milliseconds to 30 seconds, and showed that different genres of music (popular, classical, jazz) and different types of speech (monologue, dialogue, synthesized) could be distinguished from each other on the basis of their Allan Factor functions. Nested clustering across time scales was found to increase with musical composition, speech interaction, and prosodic emphasis. These and other communicative and expressive differences in speech and music corresponded with consistent changes in the shapes of Allan Factor functions that reflected the underlying behavioral processes (see also Ro and Kwon, 2009).

It stands to reason that hierarchical temporal structure in the sounds of speech and music may have corresponding structure in the underlying movements that produce the sounds, and indeed there is evidence in support of this conjecture. Chandrasekaran et al. (2009) found hierarchical temporal structure in spectral analyses of articulator movements during speech that suggested the nested clustering in the movements of the lips mirrors the nested clustering in the resulting speech sounds. As for gestures during speech, Pouw and Dixon (2019) used cross-wavelet analysis to show that the amplitude envelope for speech sounds is coupled with the acceleration profile of hand movements across time scales. Finally, in a recent study, we used frame differencing analyses of video recordings to illustrate hierarchical temporal structure in the movement amplitudes of lecturers giving academic talks (Alviar, Dale, and Kello, 2018).

The present study builds upon prior work by analyzing structure in the sound and movement amplitude envelopes for a range of different kinds of speech and music recordings. We measure this structure in two complementary ways: one by correlating smoothed versions of the envelopes themselves, the other by correlating power estimates from modulation spectra of the envelopes. We computed the acoustic amplitude envelope directly from the acoustic waveform of each recording, whereas we used video frame differencing to compute amplitude envelopes from the video signal.

Prior studies showing different kinds of hierarchical temporal structure for different kinds of speech and music performances (Kello et al., 2017) lead us to expect variations in multimodal, multiscale coordination depending on the category of behavior. Without more specific predictions, we start by sampling from a range of different types of speech and musical performances that are readily available on YouTube, and analyzable in terms of recording conditions. We chose speech videos that varied primarily in their communicative and expressive properties, and music videos that varied primarily in their physical properties (type of instrument). We analyze whether our measures of coordination vary more as a function of functional properties in speech, or physical properties in music, and we then present more targeted analyses based on these results.

3.3. Method

3.3.1. Audio and Video Recordings

We used a convenience sample from YouTube as a source of natural data for this study. Recordings are publicly available, and contain time-locked video and audio from a wide range of people and backgrounds—different ages, different culture, different gender, and different parts of the world. Convenience samples can be biased relative to the population in question. In our case, biases might arise from YouTube's search algorithm criteria, or characteristics of people who have the means and interest to make and post videos, especially unproduced amateur videos. Despite these possible biases, the diversity of our sample makes it broadly generalizable and more representative of the general population than samples of college undergraduates by comparison.

The following criteria were used for selecting videos. First, we ensured that all recorded sound and movement (beyond low levels of background noise) was produced by only one performer per recording, which means the camera had to be still and shot from only one angle (hence most recordings were amateur and not edited or produced). We did not constrain our selection of recordings based on camera angle, but most were directly facing the person being recorded, except piano recordings which were mostly 90 degrees in profile, and most captured motion at least from the torso to the head. Performances were at least 3.5 minutes so that we could analyze time scales comparable to prior studies.

Applying these selection criteria to videos available on YouTube, we initially found 80 videos, half speech and half musical performances. We deemed this number of videos sufficient for our purposes given that Kello et al (2017) found significant differences in the shape of Allan Factor functions with a similar sample size per group. There were eight different categories of video, four for speech and four for music, for a total of 10 recordings per category. The four different categories of speech performances were chosen to sample a range of communicative styles and degrees of spontaneity: Spoken word poetry, scripted acting, unscripted spontaneous monologues, and teleprompter reading. The four different types of music recordings were all classical music performances chosen to sample a range of instruments i.e. physical means of sound production— flute, guitar, piano, or violin.

After analyzing results with these 80 recordings as presented below, we added two additional test categories, again with 10 videos each: *A cappella* singing was added to test the use of speech as an instrument, and improvisational jazz piano was added to test the use of an instrument in a more spontaneous, conversational style of performance (Kello et al., 2017). Recordings were 5.66 minutes long on average, and there were 57 male performers and 43 female performers distributed roughly evenly across the categories. Additional details about each recording are listed in the Appendix. We did not expect to get any systematic influences of other possible common characteristics across the groups, nor tested for them.

To facilitate comparisons across recordings and modality, all audio and video signals were resampled to the lowest common denominator of 30 Hz. This procedure constituted a massive downsampling of the audio signal, which was no less than 44.1 KHz as originally recorded, into the range of typical video sample rates, which was 20 to 30 frames per second in our sample (*M*=28.45, *SD*=2.49). Making the sample rate a constant of 30 Hz, with a minimum recording length of 3.5 minutes, resulted in all pre-processed signals having the same range of time scales available for analysis.

3.3.2. Sound and Movement Analyses

Figure 1 diagrams an outline of the basic steps of waveform analysis that we took. Processing began with converting audio and video signals into amplitude envelopes using the Hilbert envelope as a standard technique for audio signals (Falk & Kello, 2017). For video signals, we computed movement amplitudes using a simple frame differencing algorithm (all code available at https://github.com/camialviar/AVCoordMusicSpeech) that quantified the greyscale change (absolute differences) summed over pixels from frame to frame as a relatively instantaneous measure of overall movement amplitude (see Paxton & Dale, 2013b for a review of frame differencing techniques). Given that videos only captured movements of the performers, frame differencing combined both fine and coarse movements of the face and body. Frame differencing techniques have been found to produce comparable data to those obtained with motion tracking and computer vision techniques (Pouw, Trujillo, & Dixon, 2018).

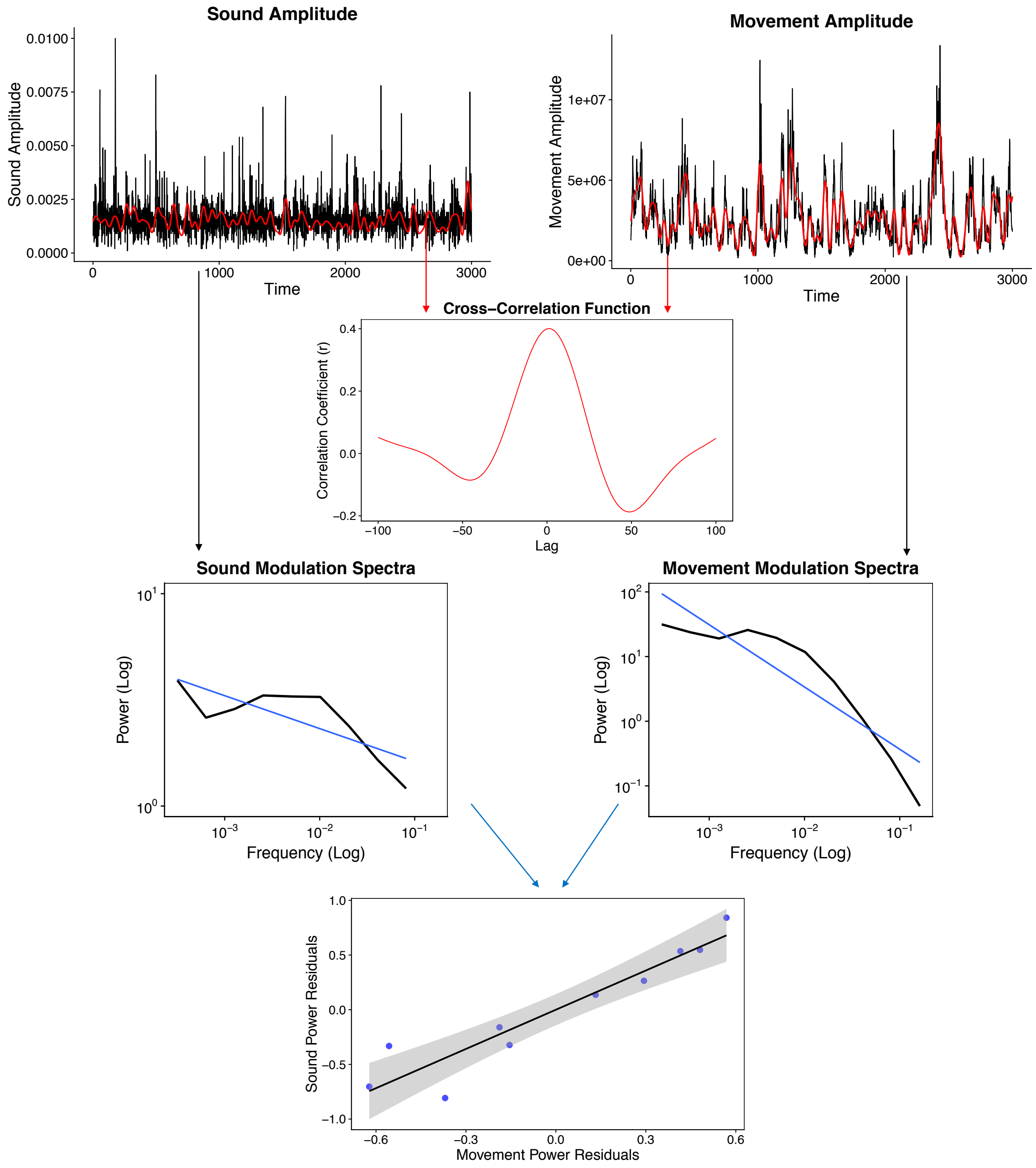


*Figure 3. 1***.** Illustration of the data analysis steps using a spontaneous speech recording as an example. Top: Downsampled sound and movement amplitude envelopes shown in black with the smoothed versions superimposed in red. Top-Middle: Cross-correlation function for the smoothed signals. Bottom-Middle: Modulation spectra of the unsmoothed amplitude signals shown in black, with the regression line for residualization superimposed in blue (y-axes have different ranges to better visual the similarity in spectral shapes). Bottom: Scatterplot of log-binned spectral power residuals for sound regressed onto the same for movement.

3.3.2.1. Cross-Correlation Peaks

One of our two measures of coordination was based on correlating the amplitude envelopes. To reduce the effects of high-frequency noise and idiosyncratic variations, the envelopes were first smoothed using a 6th order low-pass Butterworth filter with a cutoff frequency of 0.5 Hz (red line in top of Figure 1). We also trimmed the first ~12 seconds of data points to avoid the edge effects due to filtering. The cutoff frequency was selected empirically by inspecting the changes in the coefficient of variation of the Pearson correlation coefficients for different filter cutoffs. The 0.5 Hz cutoff was the frequency in which the decrease in the coefficient of variation stopped being linear, indicating a change in the relationship between the mean and the variability in the correlations. Cross-correlation functions were computed at lags up to ±100 data points (about 3 seconds, see top middle of Figure 1).

The cross-correlation functions can be compared directly to each other to test relative effects, but to test whether there is multimodal coordination beyond chance, we need a surrogate chance baseline. Within each category of ten recordings, each movement amplitude series was paired with the nine sound amplitude series from each of the other recordings in the same category. We selected the highest correlation coefficients in the cross-correlation functions and averaged them across the nine surrogate pairings to obtain a unique surrogate value for each original recording. Signals of different lengths were trimmed to match the shortest one automatically as part of the cross-correlation function in R.

3.3.2.2. Modulation Spectra

Our other measure of coordination was based on correlating power estimates of the modulation spectra of unfiltered sound and movement amplitude envelopes. We applied Fourier analysis to each amplitude time series to compute spectral power estimates over the range of available frequencies, i.e. a range of time scales, hence the multiscale nature of this analysis. The Fourier analysis decomposes a time series into sine waves at different frequencies where the amplitude of each sine wave estimates the power (amplitude squared) of each frequency in the signal. Each Fourier analysis was computed over a window of 3.5 minutes of amplitudes resampled at 30 Hz. For each recording longer than 3.5 minutes, the window was shifted forward in time, up to the recording length, and spectral power estimates were averaged over windows. Given the sample rate and set window size, the range of available frequencies for our spectral analyses was 15 Hz to 0.0048 Hz. This frequency range was divided into 10 logarithmically-spaced bins and spectral power estimates were averaged within each bin (see bottom middle of Figure 1). Logarithmic spacing meant that the lowest frequency bin contained just the one lowest frequency spectral power estimate, then the next bin averaged the next two power estimates, and so on, doubling the number of estimates averaged per bin from lowest to highest frequencies, i.e. 2^n spectral estimates per bin, with n equal to bin number 1 to 10, low to high frequency. Logarithmic binning evens out the amount of time series data contributing to each spectral power estimate (Thornton & Gilden, 2005), and it averages out idiosyncratic variance to enable comparison of spectra across modalities.

We measured the degree of spectral matching by regressing the log-binned power estimates for movement spectra onto those for sound spectra, after residualizing the effects of frequency (see regression line in bottom middle of Figure 1). Residualization was an important step because modulation spectra for speech and music are generally known to approximate a 1/f scaling relation (Voss & Clarke, 1973), as are movement time series (Alviar et at., 2018; Delignières et al., 2004; Hausdorff et al., 1996). Indeed we replicated this well-established effect (see the Spectral Matching Results section for more details). A similar technique termed *complexity matching* was introduced recently as a measure of coordination that assumes the linear multiscale relationship in which power increases with frequency. This assumption leads one to measure complexity matching in terms of correlation between the linear coefficients of log-log regression fits (Abney, Paxton, Dale, & Kello, 2014; Coey, Washburn, Hassebrock, & Richardson, 2016; Fine, Likens, Amazeen, & Amazeen, 2015; Marmelat & Delignières, 2012). Complexity matching was not appropriate for our purposes because we needed a measure that was sensitive to the specific bends and kinks of modulation spectra which were recently shown to reflect communicative and expressive aspects of speech and music (Kello et al., 2017). Therefore, we first removed the general trend of an inverse relationship between spectral power and frequency, and then submitted the residuals to mixed effects models, as presented below.

Residualization was necessary to observe a correlation between sound and movement that was not driven by the general $1/f^{\alpha}$ scaling relation between power and frequency, where $0 < \alpha < 2$. Residualization did not detract from the multiscale nature of the analysis because power estimates still spanned a wide range of frequencies. Correlating residuals served as a measure of the relationship between the deviations from a 1/f trend for sounds and movements produced by the same individual. Correlated residuals reflect similarities in the hierarchical temporal structure of sound and movement amplitudes.

3.4. Results

3.4.1. Cross-Correlation Analyses

Figure 2 shows cross-correlation functions in each of the four main speech categories and four main music categories, for each individual recording as well as the mean of each category. At a glance, the functions for speech generally have distinct peak correlations near a lag of zero, whereas functions for music have flatter profiles with more varied and less distinct peak lags. Figure 3 shows the mean peak lags and correlation coefficients for all categories. We tested for temporal coordination between sound and movement amplitudes by using peak correlation coefficients as dependent measures (Fisher Z transformed for normality) and comparing them with the mean peak coefficient for each recording's set of surrogate pairs. An analysis of variance was conducted with two independent variables, correlation type (original or surrogate) and performance type. Coefficients were stronger for original pairings (M=.25) compared with surrogates (M=.07), $F(1,64)=38.136$, $p<.001$, but there were no differences between original pairings and surrogates across groups, $F(7,64)=0.639$, $p>.7$. There was also no reliable difference between speech and music, *F*(1,78)=0.159, *p*>.6, nor between the 8 different performances, *F*(7,64)=0.595, *p*>.7. It appears that all recordings exhibited temporal coordination above chance at some lag, but with no reliable differences in the strength of correlation between different categories of music and speech.

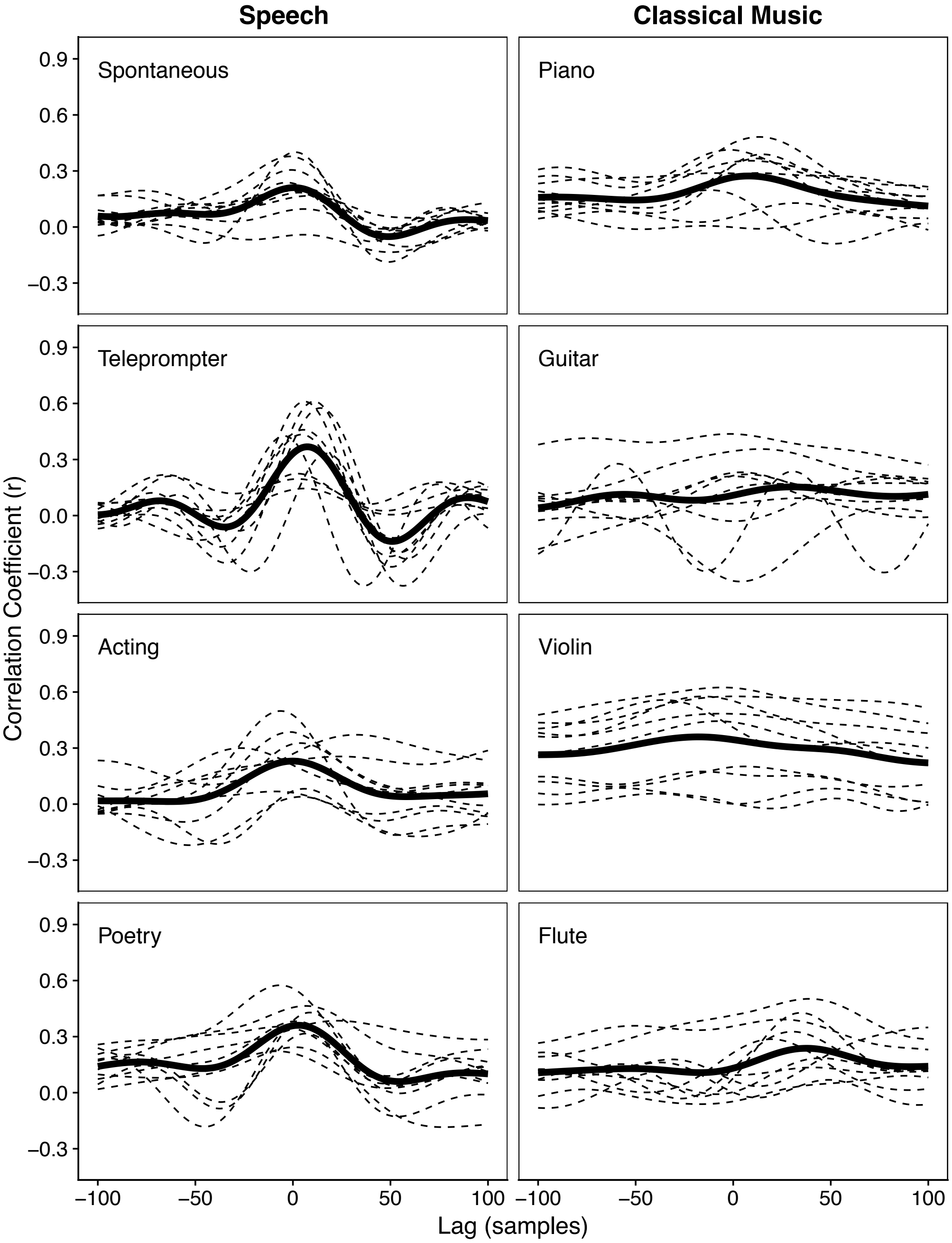


Figure 3. 2. Cross-correlation functions for each recording category. The dotted lines show functions for individual recordings, and solid lines show category averages.

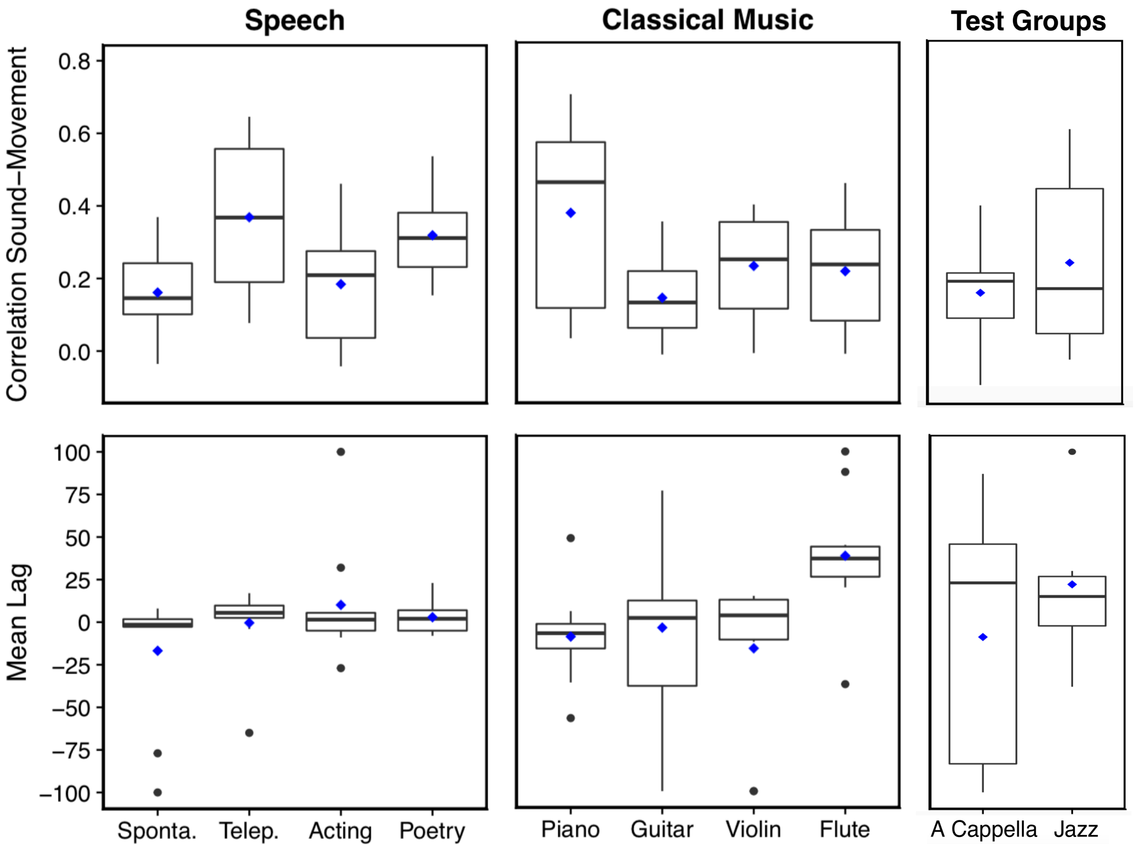


*Figure 3. 3***.** Box plots showing the distributions of peak correlation coefficients (top) and lags (bottom) for sound-movement cross-correlation functions, separated by recording category. The blue diamonds show the mean for each group. A cappella and improvisation jazz piano were analyzed separately from the main speech and music categories.

In contrast with correlation strengths, examination of peak lags in Figure 3 suggests a difference between categories: Lags appear relatively close to zero (*M*=2.75, 0.09 s) for the four pure speech categories, with relatively little variation around this central tendency, indicating synchronous coordination of sound and movement. Lags appear different for music, in that their variances are greater both within and across subcategories. We tested this apparent difference between speech and music using two analyses. First, we conducted an F-test of equality of variances and found that the variance among peak lags for speech recordings (S^2=867) was reliably less than that for music recordings (S^2=2134), *F*(39,39)=0.406, *p*<.01.

Second, we compared the cross-correlation function for each recording with the mean cross-correlation function of the recording's category, akin to a standardized score (e.g. a Z-score), but one that measures the consistency of an individual cross-correlation function with a mean cross-correlation function. If recordings have one or more peaks at consistent lags, then the category mean should preserve the peaks present in individual functions. By contrast, if peaks are inconsistent over recordings, then they should be averaged out in the category mean, and the mean will therefore covary less with the individual recordings. We tested these competing hypotheses in two steps: First, we correlated each individual cross-correlation function with its corresponding category mean to get a measure of consistency between each individual cross-correlation profile and the group's average profile. Second, we compared these measures of consistency across the different categories by running an ANOVA on the Fisher Z transformed correlation coefficients obtained in the previous step. The results showed that individual recordings were more correlated with their category means for speech (*M*=.752) compared with music (*M*=.506), *F*(1,78)=8.779, *p* < .01. This result supports the hypothesis that peak cross-correlation lags between sound and movement amplitudes were more consistently near zero across speech recordings compared with music recordings. Using correlation to create a kind of standardized score is unusual and may raise questions about its statistical validity, but results provide convergent evidence with the visible differences seen in cross-correlation plots, and with results from the more standard test of equality of variances.

Also, it is important to note that cross-correlation methods have been questioned recently because autocorrelation can interfere with the ability to determine causality of effects (Dean & Dunsmuir, 2016). In our study, we aggregated cross-correlations across many time series and examined only their peaks to assess relationships, which avoids recent criticisms (for instance, see Figure 4 in Dean & Dunsmuir, 2016). Future studies may investigate causal relationships in multimodal signals using alternative methods such as transfer entropy and Granger causality.

3.4.2. Spectral Matching Analyses

Figure 4 shows the mean modulation spectra for sound amplitudes and movement amplitudes in each recording category. All modulation spectra generally showed the expected inverse relationship between power and frequency in log-log coordinates, i.e. hierarchical temporal structure: Spontaneous speech α = -0.41, Teleprompter α = -0.37, Acting α = -0.59, Poetry α = -0.44, Piano α = -0.6, Guitar α = -0.5, Violin α = -0.54, Flute α = -0.71. We are unaware of prior results showing the same pattern in movement, but the ubiquity of hierarchical temporal structure in behavior provides an empirical basis for expecting it in movement as well, and indeed that is what we observed: Spontaneous speech α = -0.9, Teleprompter α = -0.56, Acting α = -0.85, Poetry α = -0.84, Piano α = -0.99, Guitar α = -0.9, Violin α = -0.8, Flute α = -0.87.

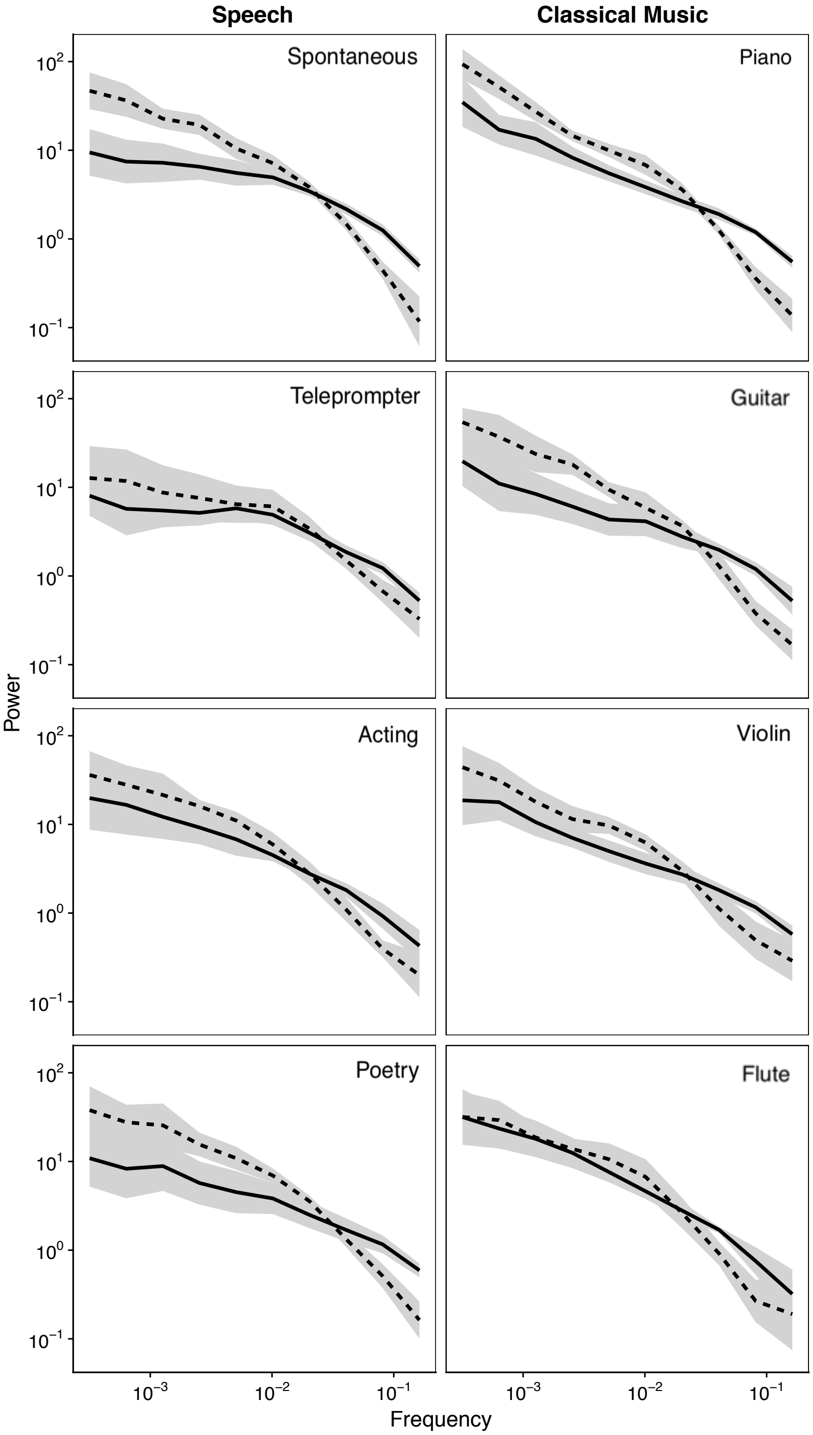


*Figure 3. 4***.** Mean modulation spectra for sound (solid line) and movement (dotted line) for each of the four main speech categories and music categories, in log-log coordinates. The shaded bands show 95% confidence intervals.

To test whether individual sound spectra were related to their corresponding movement spectra, we formulated mixed effects models to predict residual power estimates for sound (after removing the effect of frequency; see previous section) based on the same for movement. In one model we added the factor of main category (speech or music), and in another model we added subcategory as a factor (eight altogether, four speech and four music). We specified a maximal random effect structure following the recommendations of Barr, Levy, Scheepers, and Tily (2013), but the model failed to converge. We instead used a partial random effect structure including a random intercept and slope for movement residuals as a function of recording to account for multiple observations from each recording. The R notation for our model was as follows: *Residualized Sound Power ~ Residualized Movement Power*Group + (1 + Residualized Movement Power|Video ID)*, with Group being either the two main categories or the eight subcategories. Models were run using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) and the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) to calculate p-values.

We should note that, when mixed effects models fail to converge, random effects related to between-subjects factors can be removed with little consequence for the Type I error rate of the model (Barr, 2013). We kept the random intercept and the random slope for movement residuals (and dropped the random slope for group) since this was the within-subjects fixed factor in our model (i.e., involved repeated measures from the same recording). The random effects structure controls for variability that comes from individual differences: it adds it to the unexplained variability of the model, so that the standard errors correctly reflect the unexplained variance and make for a good baseline to judge the significance of the fixed effects (Mirman, 2014). The random intercept builds into the model the assumption that different individuals will have different mean deviations from the 1/f trend in their sound signals. The random slope for movement residuals builds into the model the assumption that the strength of the relationship between sound and movement residuals will be different for different individuals.

Figure 3.5 shows scatterplots of sound and movement power residuals based on the mixed effects model using subcategory as a factor. The scatterplots show generally positive trend lines for individual speech recordings, and more variable, inconsistent trend lines for music. A model with the main recording category (speech versus music) as a factor, and its interaction with movement power residuals (AIC=404.33), fits the data better than a model without main category (AIC=415.70; χ^2=15.372, p<.001). The interaction term was reliable for both speech (B=0.757, p<.001) and music (B=0.255, p<.01), but the relationship was significantly stronger for speech (p<.001). We repeated this analysis with subcategory as the fixed factor instead of main category, and this time the model had a marginally worse fit with the addition of subcategory, (AIC=418.48; χ^2=25.22, p<.05). We were interested in exploring finer-grained differences among subgroups, so we planned to inspect this model regardless of its fit compared to the simpler model that did not contain subcategory. Consistent with the model including main category (speech vs. music) as a factor, all interaction terms were reliable for speech subcategories, but none for classical music subcategories: Spontaneous speech (B=1.108, p<.001), teleprompter speech (B=0.832, p<.001), scripted acting (B=0.623, p<.001), poetry (B=0.447, p<.01). These differences between the groups in speech follow a pattern that is descriptively interesting, however, only statistically significant at the extremes (spontaneous speech vs. poetry:

p=.009). We should note too, that the null results for the music groups are most likely the result of a lack of power to detect the smaller effects, as suggested by the significant result for music.

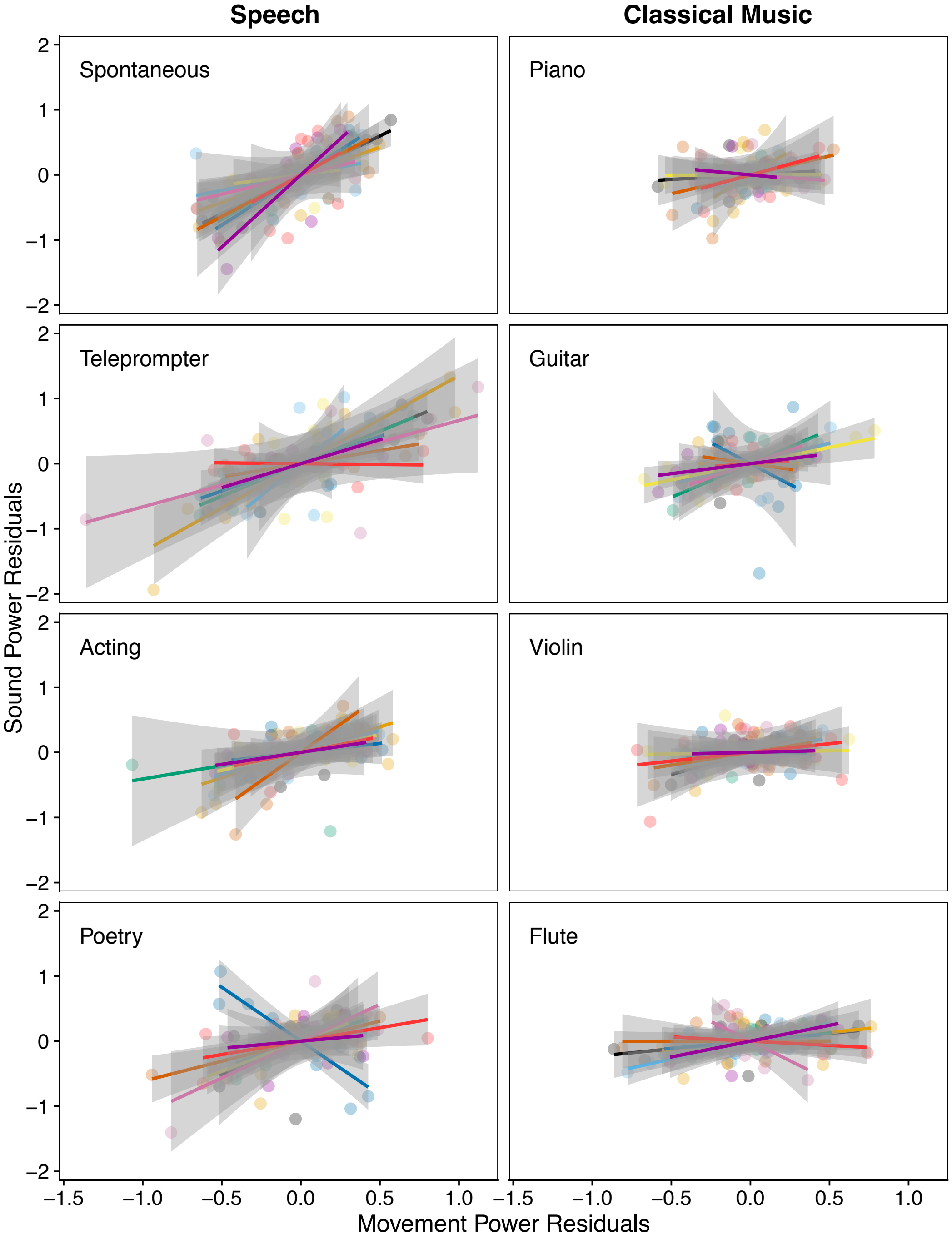


*Figure 3. 5***.** Movement power residuals plotted against sound power residuals based on the mixed effects model with recording subcategory as a fixed effect. Linear trend lines are shown for individual recordings in each subcategory.

3.4.3. Follow-up Test of Category Effects

Results so far provide evidence that multimodal, multiscale coordination between movement and sound is different for speech versus musical performances. Our four main types of speech recordings showed more consistent synchronization between sound and movement amplitudes (although correlation coefficients did not differ by category), and the speech categories showed more multiscale spectral matching compared with music categories. We also found that spectral matching varied within speech categories in a pattern related to the type of communication or performance. As we mentioned above, two caveats apply to this result: the differences between speech groups were only reliable at the extremes of the pattern, and the lack of differences among the music groups may be due to a lack of power.

Our category manipulation confounded speech and music with functional (communicative) differences versus physical (instrumental) differences. Given that this study is based on naturalistic observations of YouTube videos, we could not fully test these manipulations independently. However, we collected and analyzed data for two additional recording categories to provide an initial test of the effects observed thus far. One category was *a cappella* singing, which uses the vocal apparatus like speech (i.e. physical similarity), but in which the voice is used more like an instrument than a means of language communication. The other category was improvisational jazz piano, which is a musical genre that has been likened to spoken conversation (Kello et al., 2017; Sawyer, 2005).

Results for these two categories are shown in Figure 6. Singing patterned more with music than speech, in that peak lags of cross-correlation functions were much more variable (S^2=5194) than those for speech (S^2=867), F(39,9)=0.166, p<.001. Spectral matching also patterned with music in that the relationship between sound and movement spectra was weak for a cappella singing, albeit marginally reliable (B=0.404, p<.05). Jazz also patterned with music in terms of peak lags, but the category was unique in terms of spectral matching because power residuals for sound and movement amplitudes were *negatively* correlated (B=-0.53, p<.01). On the one hand, this negative correlation is like speech because it indicates multiscale, multimodal coordination, but on the other hand, the specific mode of coordination is different than observed for speech. While further investigation is needed to understand this unique result, the two additional test categories provide further evidence that coordination between sound and movement in speech and music depends on functional aspects of performance rather than the physical apparatus of sound production.

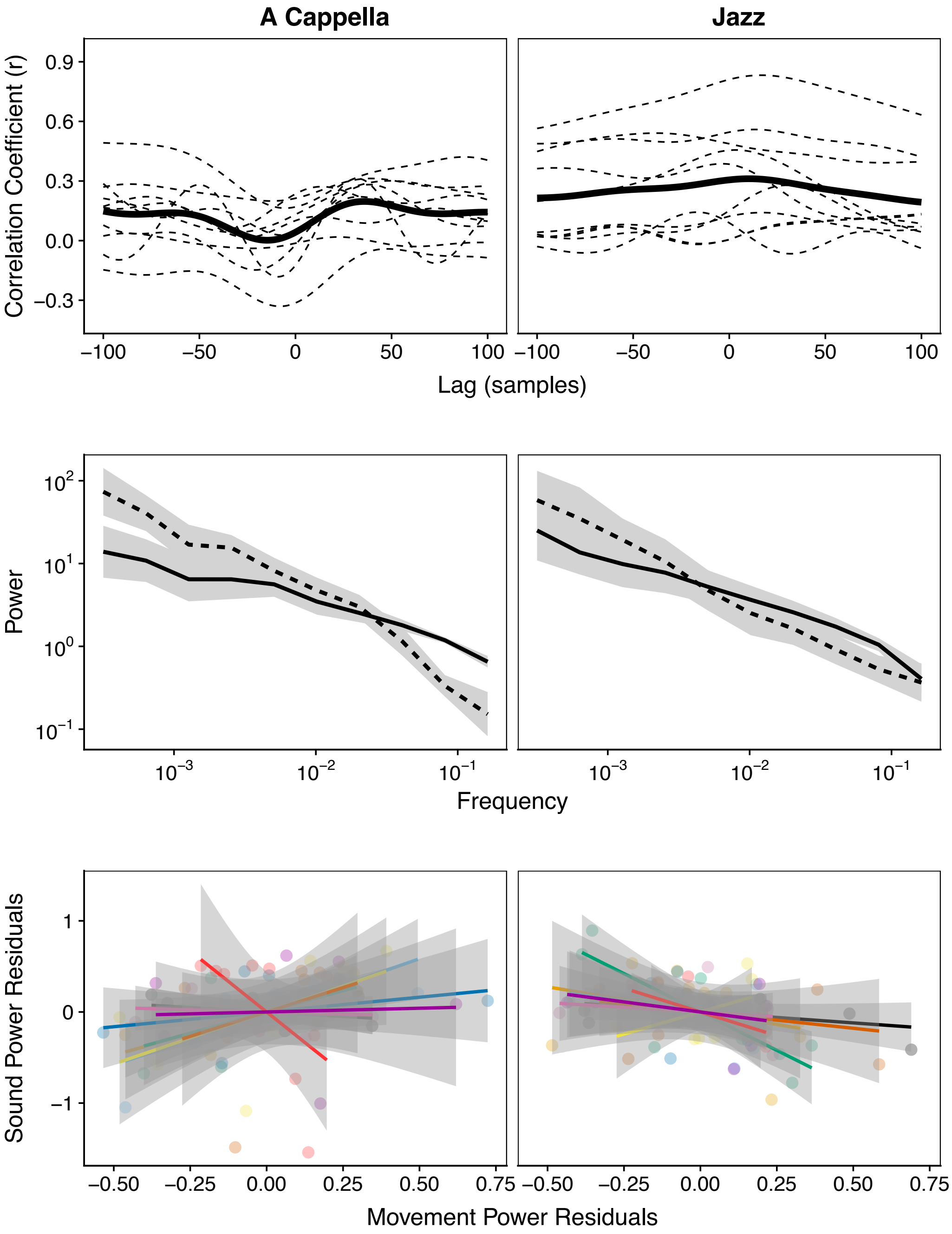


Figure 3. 6. Top: Cross-correlation functions for A Cappella and Improvisational Jazz Piano test categories. The dotted lines show functions for individual recordings, and solid lines show category averages. Middle: Mean modulation spectra for sound (solid line) and movement (dotted line), in log-log coordinates. Shaded bands show the 95% CI. Bottom: Movement power residuals plotted against sound power residuals based on the mixed effects model with recording subcategory as a fixed effect. Linear trend lines are shown for individual recordings in each subcategory.

3.5. Discussion

In this study, we explored the coordination between sounds and movements that produce or accompany a variety of speech and musical performances. Our main objective was to compare the coordination of movement and sound for categories of speech and music as a function of functional and physical properties of the performances. In general, we found reliable relationships between movement and sound for both speech and music, as was expected from previous research. We also found reliable differences in the multimodal coordination of speech compared with music, and these differences depended on communicative and expressive aspects of performance rather than the physical mechanism of sound production.

The test group of a cappella singing showed that multiscale coordination depends on the act of speech communication rather than use of the vocal tract per se, and the test group of jazz improvisation showed that coordination depends on the musical genre rather than use of a particular instrument such as the piano. Results also showed that our two measures of coordination are complementary, in that spectral matching showed greater multiscale coordination for speech compared with music, whereas the magnitude of cross-correlation showed coordination in both categories. Furthermore, the lags of peaks in cross-correlations highlighted the relative synchrony of sound-movement coordination in speech but not music, and both measures were necessary to show the unique pattern of coordination of jazz compared with speech and classical music.

One might wonder if the differences between speech and music may be attributable to the way movements are captured differently in audio or video signals. However, the nature of recordings was largely the same across subcategories of speech and of music, yet subcategory differences were found herein and also in Kello et al. (2017) for sound recordings using Allan Factor analysis. Also, across categories, there was no appreciable difference between videos of speech performances and singing, and yet their coordination patterns were distinct. In both categories, performers used their vocal tracts and were filmed from the same basic angle and distance.

We find a functional explanation to be more consistent with the results. In particular, speech performances served the purpose of language communication whereas musical performances did not. McNeill (1985) showed that gestures mirror and complement the meanings and discursive organization of speech communication, and listeners expect gesture to carry communicative information. Speakers gesture more when talking to an interlocutor (Bavelas, Gerwin, Sutton, & Prevost, 2008), they adapt their gestures to the needs of their audiences (Galati & Brennan, 2014; Özyürek, 2002), and they use gestures purposefully to disambiguate ambiguous sentences for the listener (Holler & Beattie, 2003). Speakers' gestures also carry relevant information about the referents in the narrative aiding discourse comprehension (Debrieslioska, Özyürek, Gullberg, Perniss, 2013), and help the listener gauge the speaker's confidence on the information being conveyed (Roseano, González, Borras-Comes, & Prieto, 2016). Listeners are slower and are less accurate to match meanings when presented with incongruent speech-gesture pairs (e.g., hearing "chop" and seeing a "twist" gesture) even when the task requires attention to just one of the modalities (Kelly, Özyürek, & Maris, 2009), and listeners show differences in event-related potentials (the N400 associated with semantic irregularities in sentences) for semantically congruent vs. incongruent speech-gesture pairs (Habets, Kita, Shao, Özyürek, & Hagoort, 2011). Listeners also take advantage of congruent gestures to facilitate processing of ambiguous speech, as evidenced in the reduction of the N400 for ambiguous sentences preceded by a disambiguating gesture (Holle & Gunter, 2007). This prior research suggests an intimate temporal relationship between speech and co-occurring signals, such as gestures. The relationship is profoundly functional in the sense that the success of communication depends on this dynamic relationship. Thus, evidence for greater multiscale coordination in speech may stem from the demands of speech communication.

Our functional interpretation leads to a general prediction that may be tested in future studies. First, an increase on communicative demands should increase the strength of the multiscale, multimodal coordination in speech. For instance, communication in more constraining or noisy conditions should lead to stronger spectral matching and greater synchronicity between sound and movement (cf., Boker, Rotondo, Xu, & King, 2002; Paxton & Dale, 2017). Similarly, the communication of more elaborate, complex messages should lead to the same pattern of effects. The same logic may be applied to jazz improvisation, which has been likened to a conversational interaction (Kello et al., 2017, Sawyer, 2005). Improvisation among two or more musicians should yield greater multiscale, multimodal coordination due to communication among the musicians compared with solo performances. Musical gestures have been shown to play a communicative role in a string quartet leading to improvement of their performance (Hospelhorn & Radinsky, 2017). And in the context of jazz, a previous study suggests that movement coordination between two jazz players engaged in improvisation follows complex and interesting coordination patterns of their heads and their arms at different time scales (Walton, Richardson, Langland-Hassan, & Chemero, 2015). The negative relationship of the spectral power in jazz suggests that decoupling movements and sounds might be a strategy to perform creatively in jazz, and perhaps also in other genres. Doing similar analyses for improvisation in other musical genres, and non-improvisational jazz will help get a clearer picture of multimodal coordination across musical performances.

Our functional interpretation is also supported by differences in magnitudes of spectral matching among the subcategories of speech performances. Beta coefficients patterned with the degree of message information carried by the movements. Movements during spontaneous speech appeared to convey the most information about the content of the message, followed by teleprompter speeches. By contrast, movements during monologue acting and spoken poetry appeared to be more associated with emotional, stylistic aspects of performance. Spoken poetry, for example, emphasized beat gestures and rhythmic movements that reflect the prosody of the poem more than its semantic content. These observations are admittedly subjective and require more systematic experimentation to test them.

Another possible factor that might affect multiscale, multimodal coordination is the degree of spontaneity in a performance. Previous studies have indicated that patterns of prosodic variation and gesture production for spoken poetry, for example, are more tightly scripted than they are for spontaneous conversation (see Novak, 2011 and Barney, 1999 for an analysis of the delivery of spoken word and Henning, 1955 for an analysis of speech delivery). By contrast, spontaneous speech and musical improvisation, for instance, may result in stronger synergies to reduce the additional degrees of freedom and ease coordination (see Kelso, 2009). Experiments that evoke different combinations of spontaneity and communication could help delineate and tease apart their influences on coordination. For instance, academic presentations may be more scripted compared with spontaneous retellings of cartoons (see Galati & Brennan, 2014).

In summary, our study shows how naturalistic recordings can provide a wealth of information about speech, music, and other human behaviors as they occur “in the wild”. We analyzed YouTube recordings and found evidence that the coordination of sound and movement is more consistent and synchronous in speech than in music, and this conclusion seems to depend on functional aspects of the performances and not the physical apparatus per se. The relevant functional factors may be related to communicative constraints, degree of spontaneity, and the general properties of language communication as a multiscale, multimodal performance (Dale & Kello, 2018). Future studies may test these hypotheses with more experimental control to minimize possible differences in recording conditions that might exist despite our careful selection criteria and controls.

Chapter 4

Can You Hear Me Now? Interpersonal Coordination during Remote Communication Using Zoom

4.1 Prologue

In this chapter I analyze interpersonal and multimodal coordination during remote communication over Zoom. The goal is to explore the effects of continuous perturbations on the interpersonal and multimodal coordination patterns that unfold during informal conversations. The theory of coordination dynamics predicts that interpersonal and multimodal synergies should be robust to perturbations that are not relevant to the goal of the system. During in-person conversations, partners produce verbal and non-verbal behaviors that tend to converge with each other. I study whether the technological limitations of Zoom hinder interpersonal convergence and multimodal coordination during remote interactions due to degraded transmission of sound and motion signals. To do this, I computed the power spectra of the sound and movement amplitudes for each member of a dyad, and correlated the logged deviations from the 1/f relationship between frequency and power for the sound and movement of conversational partners. The results show convergence only for sound but not for movement; and only for the lower frequencies of the sound power spectrum for interlocutors interacting over Zoom, as compared to convergence on the low and high frequencies for in-person interaction. Multimodal coordination is also observed for the remote conversations, which combined with the results of the in-person dataset, suggests that the mediation of the audiovisual signals through the technological limitations of Zoom disrupts coordination across interlocutors.

There are confounds in the datasets that make our conclusions tentative, however, in line with the predictions of coordination dynamics theory, the global coordination is maintained in sound despite the perturbations introduced by Zoom. The effect size in the remote conversations is higher compared to the in-person effect size. This pattern is not statistically significant but is intriguing. Could it be hinting to the reorganization of the dyadic synergy to maintain coordination under the new constraints of remote interaction, for which movement convergence is no longer achievable? Could the higher convergence in sound develop to compensate for the lack of coordination in the movement modality? Future research should more systematically study these hypotheses.

4.2. Introduction

Subtle coordination patterns emerge when humans interact with one another face-to-face. This behavioral convergence is robust and appears to promote prosocial behaviors between people (Chartrand & Lakin, 2013) and improve their task performance (Fusaroli et al., 2012). With the COVID-19 pandemic, videoconferencing suddenly became widespread, and the signals normally available during in-person communication were absent or substantially degraded. In online interaction, field of view is restricted, noise cancellation masks important sound signals, two-way eye contact is impossible, and bandwidth limits regularly disrupt signal transmission. As a result, conversations through videoconferencing feel stilted and anecdotally it seems more difficult to coordinate, but is behavioral convergence robust enough to occur nonetheless? In the present study, we used a recently developed method for analyzing *spectral matching* in sound and motion signals (Alviar, Dale, Dewitt, & Kello, 2020) to investigate whether the speech and movements of conversational partners are coordinated during Zoom interactions as they typically are in person.

During face-to-face conversation, interlocutors coordinate their behavior locally, mimicking each other's behaviors within short windows of time, and also globally, adapting behavioral "styles" to match that of their interlocutors. Locally, they use the same words their interlocutors are using (Clark & Brennan, 1991), they coordinate their smiles and nods, and mimic each other's movements (Louwerse, Dale, Bard, & Jeuniaux, 2012), they follow their interlocutors' gaze as they listen to them (Richardson & Dale, 2005), and they adjust the duration of their turn-taking pauses to their partner's (Cappella & Planalp, 1981). Globally, they match speech rates (Levitan & Hirsberg, 2011), and accents (Pardo, 2006). They also match the temporal structure of prosodic patterns (Abney, Paxton, Dale, & Kello, 2014), even when conversing in two different languages (Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2020), and exhibit similar temporal structures in their movements (Abney, Paxton, Dale, & Kello, 2021).

Coordination plays an important role in the affective and performance outcomes of interactions. For instance, increased mimicry results in smoother interactions and partners that are perceived as more likable (Chartrand & Bargh, 1999), whereas there is less convergence in movements and prosody in argumentative conversations compared with friendly ones (Paxton & Dale, 2013a; Abney et al., 2014). In therapy settings, movement, physiological, and linguistic coordination correlate with better therapeutic outcomes and rapport (Ramseyer & Tschacher, 2014; Wiltshire et al., 2020). Moreover, convergence in speech rate may increase the likelihood of cooperation in the prisoner's dilemma (Manson, Bryant, Gervais, & Kline, 2013).

With respect to performance outcomes, listeners that followed the speakers' gaze exhibited higher comprehension and performance in a memory task (Richardson & Dale, 2005). Interlocutors that reused task-related linguistic expressions, and converged on a stable set of them, performed better in a joint perceptual task (Fusaroli et al., 2012), and the same was found for reuse of lexical, prosodic, and pause patterns (Fusaroli & Tylén, 2016). Additionally, convergence in the temporal organization of speech acoustics (e.g., prosodic patterns) correlated with performance in a cooperative tower-building task (Abney et al., 2021).

The rise of videoconferencing has led researchers to study whether affective and performance outcomes are compromised as a result of remote interactions. Burgoon and collaborators (2002) manipulated co-presence (i.e., being physically in the same room), as well as the amount of multimodal information that remote conversational partners could access (i.e., only text, only audio, or both audio and video). Results showed that co-presence made the interaction feel easier and more engaging, and resulted in better social judgements between participants. However, remote interactions lead to better performance, presumably because participants were more task oriented during computer-mediated communication. In a different study that involved manipulating objects in the environment, co-presence and access to shared visual information was central to performance (Fussell, Kraut, & Siegel, 2000). In general, the affective outcomes of interactions seem to be more adversely affected in remote conversations (Walther, 2012). As a result, the same prosocial outcomes require more time to develop (Walther, 1995) and more creativity in text-related techniques (e.g., emojis) to bolster remote interactions (Walther, Loh, & Granka, 2005).

The adverse effects of remote interactions may be reflected in reduced convergence between interlocutors, but few studies have examined the effects of remote communication on behavioral convergence per se. Some evidence indicates that convergence can occur when interlocutors can hear but not see each other. For instance, Shockley et al. (2003) had participants interact to solve a collaborative task and found that dyads became more similar in their postural sway even when they could not see each other. In another study, participants speaking on the phone were able to synchronize their gait based on acoustic cues only (Murray-Smith, Ramsay, Garrod, Jackson, & Musizza, 2007). More direct measures of coordination have shown convergence in word choice during phone conversations (Mirzaiyan, Parvaresh, Hashemian, & Saeedi, 2010), and more formal turn-taking during remote interactions, regardless of whether visual information was available or not (Sellen, 1995). Studies so far suggest that convergence may occur in remote interactions, but results do not speak to the dynamics of interpersonal coordination that may be directly impacted by the technological limitations of platforms like Zoom. Does the degradation of audio and visual signals interfere with the prosocial convergence of speech sounds and bodily movements? Or is enough information preserved for perceptual and motor systems to support convergence nonetheless?

Videoconferencing platforms like Zoom allow us to address these questions directly in audio and video signals that are recorded separately for each participant. We can analyze the temporal structure of signals to see if they match across participants, and potentially across modalities as well, but we need a method suited to comparing the hierarchical temporal structure common to speech and movement (see Kello et al., 2017). We adopted a method of *spectral matching* (Alviar et al., 2020), akin to complexity matching (Marmelat & Delignières, 2012), to quantify the statistical similarity of the two signals across a range of nested time scales. Temporal structure at each timescale can be viewed as one or more rhythms within a frequency band, and spectral matching measures the degree to which the prominence (i.e., spectral power) of rhythms is correlated. Alviar et al. (2020) found spectral matching in the audio and video signals produced by *individual* speech and music performances: fluctuations in speech sound amplitudes were correlated with fluctuations in bodily movement amplitudes in recordings of individuals giving speeches and other solo performances. The present study uses the same method to test whether fluctuations in sound and motion are correlated across conversational partners on Zoom, and effect sizes are compared with analyses of a prior experiment in which conversations were recorded in person (Schneider et al., 2020).

4.3. Method

4.3.1. Participants

A sample of 20 dyads were analyzed in the present study. Undergraduate participants were recruited using the research participation system of the University of

California, Merced for course credit. Data for an additional 18 dyads were collected but excluded from analyses due to problems with the recording conditions. In particular, dyads were excluded based on the following criteria: (a) considerable video or/and audio distortion due to internet connection instability (more than 10% of the frames of the video frozen for any dyad), (b) use of virtual backgrounds that randomly occluded the participant, (c) repeated interruptions from other individuals in the participant's household, and (d) setups that resulted in constant illumination changes in the video (e.g., bright lights in the background). The sample size of 20 was chosen to be similar to prior in-person studies of convergence in conversations. Participants were at least 18 years old (mean = 20.5, 76% women), and to elicit natural conversations, they were asked to sign-up for the study with a friend. Prior to starting the experiment, partners were asked to rate the closeness of their relationship on a scale from 1 (*We are acquaintances*) to 10 (*We are like siblings*). The mean rating was 7.0, and 75% of partners reported being friends for longer than one year. The study was approved by the Institutional Review Board of the University of California, Merced. All participants signed an informed consent and were informed about their rights to withdraw from the study without repercussions.

4.3.2. Procedure

All experimental sessions were run via Zoom, and audio and video signals were collected for each participant using Zoom's native recording functions. Audio was recorded at 32 KHz and video at 25 Hz. Each participant was instructed to have their audio, video and microphone enabled and to set their device on a firm, stable surface in a room with good lighting and minimal background noise. The Zoom display was set so that participants were shown in two main windows side-by-side, and experimenter windows were minimized.

After joining the Zoom call, participants were instructed to have their chat enabled to receive prompts from the experimenter. After requiring verbal consent, the experimenter explained to the dyad that they would be given 5 conversational prompts to discuss for about 3-5 minutes each. To keep the conversations more naturalistic, the transition from one prompt to the next was driven by the participants themselves and not prompted by the experimenter. Participants were given two practice prompts that were timed by the experimenter to provide feedback on the expected duration of each conversation. During the practice, they were prompted to keep talking for at least three minutes, and were asked to transition to the next prompt after about five minutes. After answering any questions about the protocol, the experimenters turned off their microphones and videos and began the experiment by pasting the five experimental prompts into the chat, in random order. If participants finished in less than 15 minutes, the experimenter prompted them in the chat to continue talking, and then ended the experiment after 15 minutes of conversation. Experimental sessions lasted 19 minutes 40 seconds (*SD*=7 min, 52 sec) on average, with the conversation for each prompt averaging 3 minutes 49 seconds (*SD*=2min, 9 sec).

The practice prompts asked participants to talk about what they did yesterday or will do tomorrow. For the experimental prompts, half of the prompts were about typical topics of conversation for undergraduate students, i.e. their favorite music, television shows, and movies (following Schneider et al., 2020). The other half were atypical topics that were challenging to address directly because they all involved discussing favorite things about generally disliked situations, i.e. going to the dentist, having the flu symptoms, being broke, sitting in traffic, and using a public restroom. The two kinds of prompts are part of a broader study to be analyzed and reported elsewhere. Each dyad received five prompts chosen at random such that there were three of one kind and two of the other.

4.3.3. In-Person Conversations

COVID-19 prevented us from collecting data in-person, but Schneider et al. (2020) recorded similar conversations for 28 dyads before the pandemic using the same typical prompts as in the present study, but without the atypical prompts. Each dyad engaged in three five-minute conversations, one per topic with order randomized. The in-person experiment was different from the Zoom experiment in three respects: (a) in-person participants spoke both English and Spanish, including a mixed condition in which one partner spoke English and the other Spanish; (b) in-person partners were not acquainted with each other, whereas Zoom partners were friends, and (c) in-person recordings were audio only.

We used the Schneider et al. (2020) recordings as comparisons with conversations held via Zoom. Language had no effect on convergence in the original study (as measured by complexity matching, similar to the method used here), so we did not include language as a factor in the present analyses. The method of spectral matching that is explained below was applied equally to Zoom and in-person audio recordings, and results with speech sound signals were compared to provide an initial test of the effects of co-presence on spectral matching.

4.3.4. Data Preparation

We used frame differencing to extract movement automatically from the video recordings (see Figure 1, top). In particular, overall amount of movement from frame to frame was quantified by the amount of pixel change between consecutive frames, thereby creating a time series of movement amplitudes in the form of summed pixel changes (all code available in Github and see Paxton & Dale, 2013b for a review of frame differencing methods). Participants were the primary sources of movement in recordings, so frame differences are driven by face and bodily movements. The two are not distinguished using our method, but frame differencing has been shown to produce data comparable in quality to that obtained with other computer vision and motion tracking techniques (Pouw, Trujillo, & Dixon, 2020). We computed frame differences for the two Zoom windows separately to extract a time series of movement amplitudes for each participant.

The amplitude envelope of the audio recordings was computed for each participant using the Hilbert transform, and envelopes were downsampled to 25 Hz to make their sample rates equal to those of movement envelopes. Amplitude envelopes were downsampled and smoothed by taking the mean over adjacent windows of 40 msec in duration (see Figure 1, top). No other filtering or pre-processing algorithms were applied.

4.3.5. Sound and Movement Analyses

4.3.5.1. Modulation Spectra

We applied spectral analysis to the movement and sound amplitude of each participant to get a measure of the signals' organization across time scales (see Figure 1, middle). In particular, we used the fast Fourier transform to decompose a time series into sine waves of different frequencies (i.e., a range of time scales) and amplitudes, where the squares of amplitudes estimate the spectral power for each frequency. We divided each conversation into three sections of equal length to increase statistical power and account for the effects of time in the development of the conversation (i.e., beginning, middle, end, Alviar, Dale, & Kello, 2018). Prior studies using spectral matching and related methods have used four to five minute intervals to capture approximately three orders of magnitude in frequency range, from fluctuations on the order of tens of milliseconds to tens of seconds (Abney et al., 2014; Alviar et al., 2020; Schneider et al., 2021). To compare our results with prior studies, we ran analyses over four minutes of recording at a time. Given that each of the three conversation sections was at least five minutes long, we analyzed multiple four-minute intervals and averaged the resulting spectra per conversation section (i.e., beginning, middle, end) for each dyad (the last two intervals were overlapping when the total recording time was not a multiple of four minutes).

Given a sample rate of 25 Hz and four-minute analysis intervals, the available frequencies ranged from 12.5 Hz to 0.0083 Hz. This range was logarithmically binned so that the amount of data represented by each power estimate was even across frequencies (Thorton & Gilden, 2005), and corresponded to exactly half the length of each larger bin. In other words, each bin averaged 2^n spectral power estimates with *n* being the bin number from 1 to 10, lowest to highest frequency. The modulation spectra for most participants showed flattening in the highest frequency bin that includes effects of recording noise, so this bin was removed from the spectral matching analyses described next.

4.3.5.2. Spectral Matching

To measure coordination in the temporal organization of speech and movement between interlocutors and across modalities, we adopted a procedure similar to Alviar et al. (2020). We first removed the general $1/f$ relationship between spectral power and frequency by subtracting the average logged power spectrum for each modality from each participant's individual logged power spectrum (See Figure 1, middle). Subtracting the mean spectrum allowed us to avoid spurious correlations due to the common $1/f$ relationship by measuring spectral matching in terms of correlated deviations from the common power law between partners. Specifically, we regressed deviations for one partner onto those of the other using mixed effects models (see Figure 1, bottom).

Convergence in the amplitude envelopes of speech sounds may vary as a function of frequency. Fine-grained aspects of speech and movement may be less controllable and hence adaptable compared with slower, more deliberate time scales. Schneider et al. (2020) divided time scales in two halves and found that complexity matching was more prevalent in the lower frequencies, as anticipated. We did the same by splitting the spectral frequency bins into high (i.e., 4 highest: 6.25Hz – 0.78Hz) versus low (i.e., 5 lowest: 0.39Hz – 0.025Hz) frequency ranges and then we included the high/low variable as a fixed effect (see Figure 1, middle and bottom). The random effect structure for each model was composed of random slopes for the dyad, to control for dyadic differences in matching; as well as, random slopes for the conversation sections to control for the variation in the temporal unfolding of the conversation in matching from beginning, middle, and end. More complex models with random intercepts and random slopes failed to converge. We kept the random slopes only model because the random slopes directly control for dyadic differences relevant to matching (i.e., differences in matching strength), whereas random intercepts account for differences that are not likely to affect matching (i.e., differences in the size of the individual deviations from the power law trend). The models were run in R (R team, 2018) using the *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *lmetest* (Kuznetsova, Brockhoff, & Christensen, 2017) packages.

To establish the significance of each relationship, we compared the base model (intercept as the only fixed effect) with the one that included the other participant residuals (our fixed effect of interest), and then we compared this last one with a model that also included the interaction with spectral frequency (i.e., high vs. low). If significantly different, we used the betas as a more direct measure of the strength of the relationship for higher and lower frequencies. We created separate models to assess spectral matching in: (a) speech, (b) movement, (c) multimodal coordination (each participant's speech with their own movements), and (d) cross-modal coordination (one participant's sound spectra with the movement spectra of their partner). We calculated the effect size (*d*) of the relationships as a proportion between the beta of the fixed factor and the square root of residual variance (i.e., portions explained by random effects as well as error; see Westfall, Kenny, & Judd, 2014).

The average percentage of frozen frames per participant amounted only 2.4% of the total frames of the recordings. Thus, frozen frames are unlikely to have an effect on the spectral functions. However, to test whether any observed correlations in spectral powers might be attributed to correlated disruptions in signal transmission, we calculated their percentage in each section of the conversation and added the percentage as an additional fixed factor to the models. None of the models with percentages of frozen frames as a fixed factor were better fits compared with models without this fixed factor. The consistent lack of effect indicates that the upcoming statistical tests of spectral matching were not affected by correlations in Zoom disruptions.

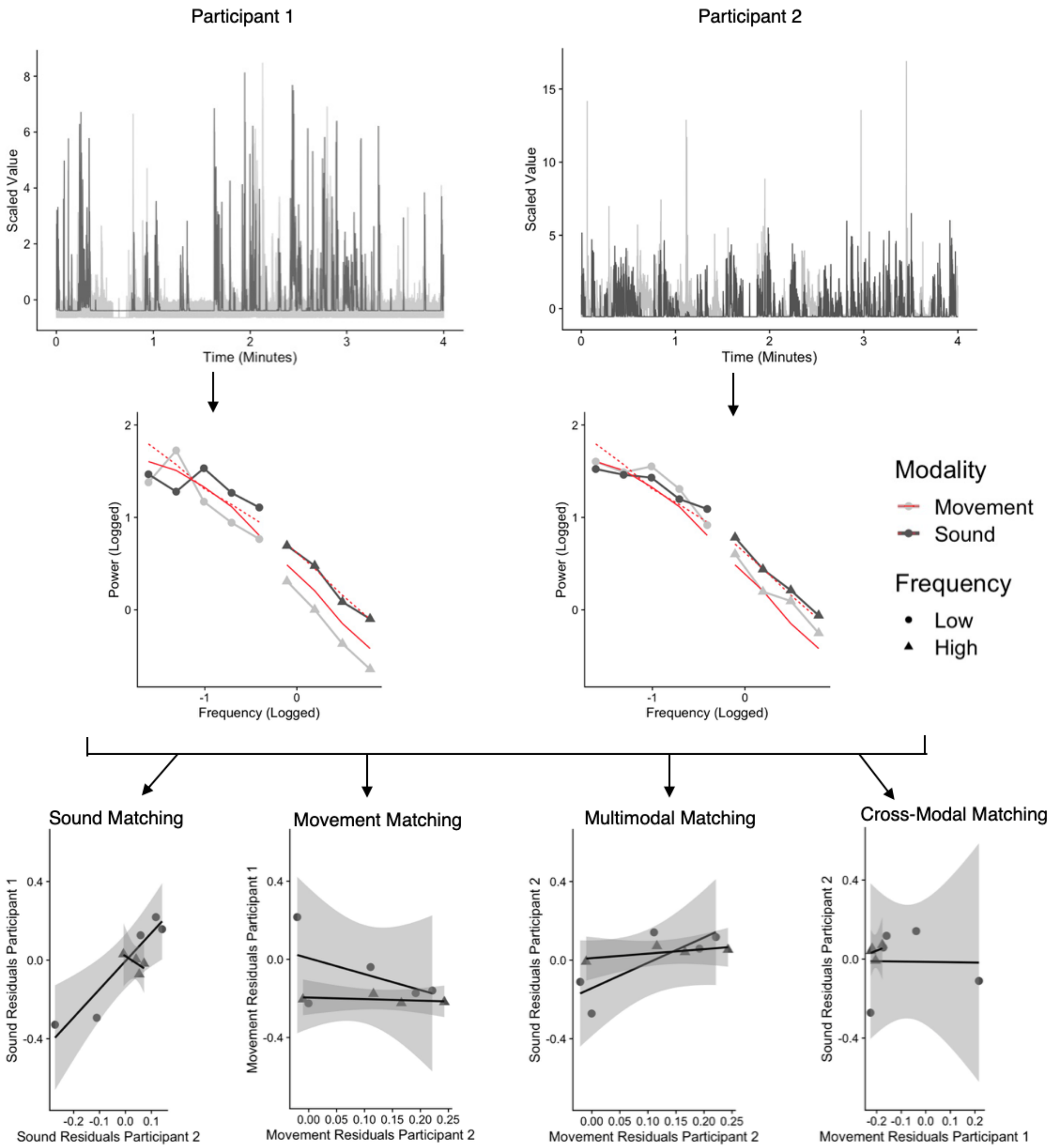


*Figure 4. 1***.** Illustration of the data analyses steps for one of the dyads. Top: Scaled sound and movement amplitudes for each participant in the dyad. Middle: High (triangles) and low (circles) frequency halves of the sound and movement modulation spectra for each participant, with the average spectrum for residualization shown in red (dotted line for the sound average, continuous line for the movement average). Bottom: Scatter plots of the (a) sound power residuals of both interlocutors (left), (b) movement power residuals of both interlocutors (middle-left), (c) speech vs movement power residuals of each interlocutor (middle-right), (d) speech power residuals of one interlocutor vs. movement power residuals of the other interlocutor (right).

4.4. Results

4.4.1. Partner Matching in Speech Sound Signals

Residuals of the log spectral power estimates for speech sounds produced by one partner were regressed onto those produced by the other, and then the interaction with frequency was added. The model with partner residuals and the interaction with frequency (*AIC*=-649.90) was a better fit compared with the model that only included the partner residuals (*AIC*=-640.84, c^2=11.05, *p*<.001). Both of these models were a better fit than the intercept only model (*AIC*=-636.29, c^2=6.55, *p*=.010). Results showed that the degree of speech sound convergence over Zoom was modulated by the timescale of the behavior—spectral residuals for one partner significantly predicted those of the other for the lower frequencies (*B*=0.42, *t*=4.25, *p*=.016, *d*=1.47), but not for the higher frequencies (*B*=0.04, *t*=0.36, *p*=0.721. See Figure 2, left).

We ran the same analysis on the in-person conversation data from Schneider et al. (2021). For in-person analyses, adding the partner's residuals (*AIC*=-489.80) to the base model improved fit (*AIC*=-483.97, c^2=7.83, *p*=.005), whereas the addition of frequency did not result in further improvement (*AIC*=-487.9, c^2=0.09, *p*=.754). Results showed reliable convergence again, as expected given results from the original study, but unlike the Zoom results, spectral residuals of one in-person partner significantly predicted those of the other (*B*=0.23, *t*=2.94, *p*=.006, *d*=0.57) for both low and high frequency ranges (See Figure 2, right). In the original study using complexity matching (i.e., correlating the slopes of log-log regression lines fit to spectra), Schneider et al. only found matching at the lower frequencies. Therefore, it appears that the additional statistical power and precision of spectral matching revealed an effect in the higher frequencies that went undetected when using complexity matching. Complexity matching is coarser because it uses only one regression line slope per participant and frequency range, whereas spectral matching uses the more fine-grained measurements provided by each spectral power estimate.

Results thus far are tentative because we cannot determine why we observed spectral matching in person but not on Zoom. The prior experiment by Schneider et al. (2020) has differences in participants, protocol, and sets of conversation prompts, and these could be factors driving the difference. That said, the evidence is consistent with the hypothesis that the signal limitations on Zoom disrupted convergence in the fine-grained aspects of speech sounds.

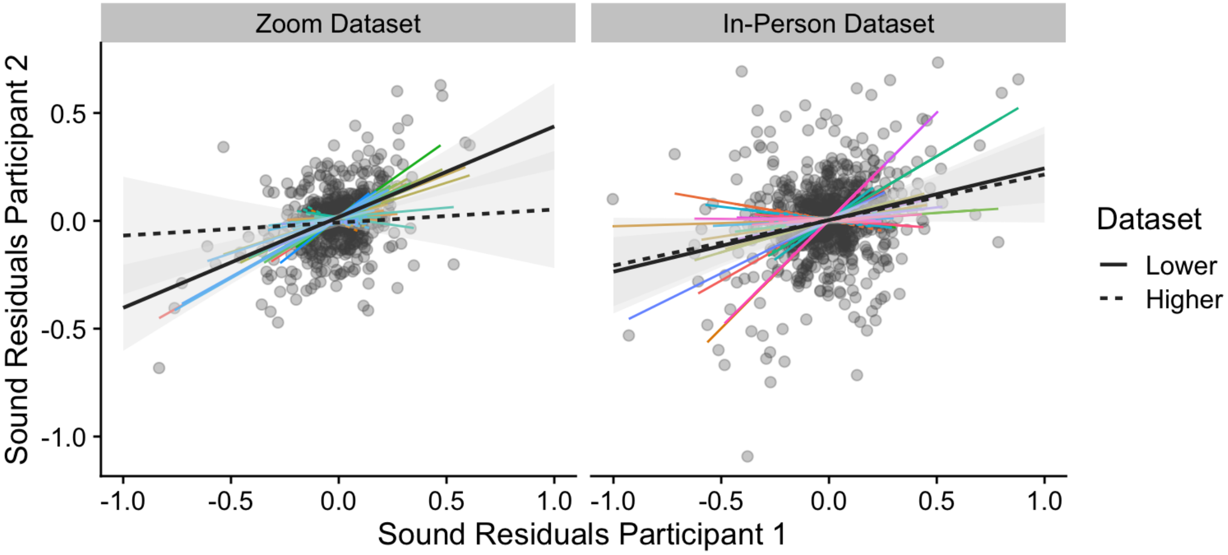


*Figure 4. 2***.** Relationship between the sound power residuals for members of a dyad for the lower and higher frequencies of the power spectra for both the Zoom (left) and in-person (right) datasets. The colored lines show the fit for each individual dyad, and the black lines shows the general prediction of the mixed-effects model for low (solid) and high (dotted) frequencies after controlling for the dyad and the conversation section. The shaded region represents the 95CI around the model predictions.

4.4.2. Partner Matching in Movement Signals

We ran the same analyses as in the previous section, but for the spectra of movement amplitude envelopes instead of those for sound amplitudes. Results indicated that neither the addition of spectral residuals (*AIC*=-109.31, c^2=0.36, *p*=.548), nor the addition of their interaction with frequency (*AIC*=-111.89, c^2=4.94, *p*=.084), produced a better fitting model as compared to the base model (*AIC*=-110.94). Therefore, we did not find reliable evidence for spectral matching in movement signals (see Figure 3) as we did for sound signals (Figure 2). We did not have directly comparable in-person data, but Abney et al. (2021) found complexity matching in movement amplitudes derived from frame differencing. Their experimental conditions were too different from ours to draw any direct comparisons. We can only conjecture that the limits of Zoom video signals were responsible for the lack of convergence. The following section provides an additional result in support of this conjecture.

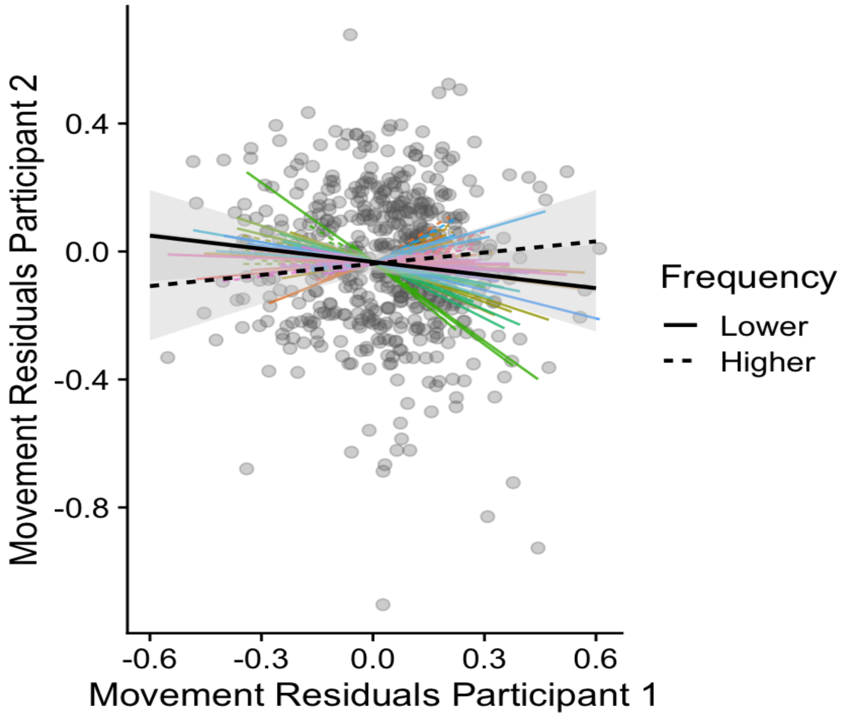


*Figure 4. 3***.** Relationship between the movement power residuals for members of a dyad for the lower and higher frequencies of the power spectra. The colored lines show the fit for each individual dyad, and the black lines shows the general prediction of the mixed-effects model for low (solid) and high (dotted) frequencies after controlling for the dyad and the conversation section. The shaded regions represent the 95% confidence interval around the model predictions.

4.4.3. Individual Matching of Sound and Movement Signals

One possible reason for the lack of spectral matching in movement amplitudes compared with sound amplitudes is that video signals are more degraded than audio signals relative to in-person interactions. Some of this degradation is inherent to video recording, like limits on camera field of view, and some may arise from lower quality signals due to bandwidth relative to audio signals. If lower quality of the video signal is the primary factor in eliminating reliable matching, then it should be eliminated for any test of matching using the video signals. We can run another test of matching by regressing spectral residuals for movement amplitudes onto residuals for sound amplitudes produced *by the same individual* in each conversation. If one individual's speech sounds and movements are coordinated when talking, then signal quality issues due to bandwidth should not affect this kind of multimodal coordination because it is not mediated by audio or video signals—it is instead mediated more directly through the body and nervous system. Indeed, Alviar et al. (2020) found spectral matching in the sound and motion signals from video recordings of individuals talking or playing music under various conditions.

We regressed the spectral residuals for sound signals onto those of movement signals for the same participant, and replaced the dyad random slope in the model with the random slope for participant to control for individual differences in multimodal matching. The model that included the movement residuals as predictors of the speech residuals produced a better fit (*AIC*=-1028.9) than the base model (*AIC*=-1023.5, c^2=7.41, *p*=.006), and the addition of the interaction with frequency did not reliably improve the fit (*AIC*=-1028.1, c^2=1.24, *p*=.265). This result replicates the results of Alviar et al. (2020) and extends them into Zoom recordings. It also provides evidence that the signal quality of Zoom video recordings was not so bad as to eradicate any matching with movement signals. Instead, the lack of partner matching in Zoom movement signals suggests that the limits of Zoom video signals disrupted the convergence of movements between conversational partners.

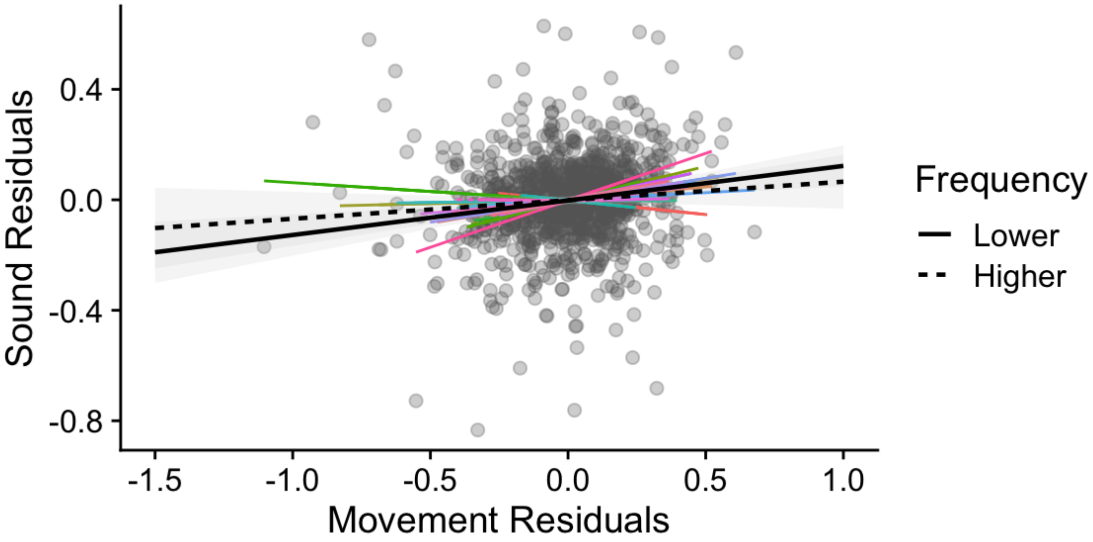


*Figure 4. 4***.** Relationship between the movement power residuals and the sound power residuals of each participant for the lower and higher frequencies of the power spectra. The colored lines show the fit for each individual participant, and the black lines shows the general prediction of the mixed-effects model for low (solid) and high (dotted) frequencies after controlling for the participant and the conversation section. The shaded region represents the 95CI around the model predictions.

4.4.4. Partner Matching of Sound Crossed with Movement

Results so far cast doubt on the possibility of cross-modal matching between partners given that no reliable partner matching was found for movement amplitudes. However, people may have learned to adapt to the limits on Zoom audio transmission by engaging in cross-modal coordination. In particular, noise cancellation on Zoom may limit coordination via sound alone, so partners may learn to use both sound and movement signals for backchanneling and other means of conversational coordination that may result in spectral matching. We regressed the spectral residuals for sound amplitudes for each individual onto those for the partner's movement amplitudes, and vice versa, and then we added the interaction with frequency. The model predicting the movement residuals from sound residuals was not a better fit (*AIC*=-1019.1, c^2=0.92, *p*=.337) than the base model (*AIC*=-1020.2), and the addition of frequency did not improve fit (*AIC*=-1017.8, c^2=0.73, *p*=.392). The lack of cross-modal matching (Figure 5) may be attributed to the same factors as those underlying the lack of partner matching in movement signals, or there may be other factors at play, including limits to cross-modal matching that may apply to in-person conversations as well. Further experiments are needed to adjudicate.

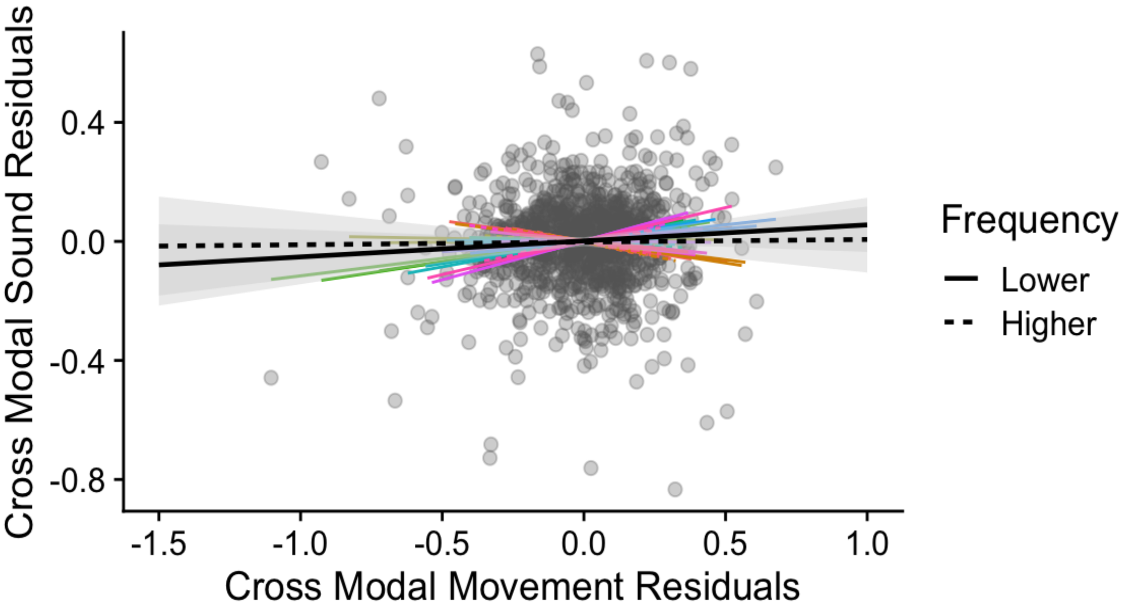


*Figure 4. 5***.** Relationship between the movement power residuals of one interlocutor and the sound power residuals of the other for the lower and higher frequencies of the power spectra. The colored lines show the fit for each dyad, and the black lines shows the general prediction of the mixed-effects model for low (solid) and high (dotted) frequencies after controlling for the dyad and the conversation section. The shaded region represents the 95CI around the model predictions.

4.5. Discussion

The present study investigated whether the limitations of videoconferencing disrupted the behavioral convergence common to in-person conversations (e.g., Abney et al., 2014; Schneider et al., 2020; Shockley et al., 2003). Results suggest that communicating via Zoom does affect interpersonal convergence, but it does so differentially for speech and movement. While speech convergence is still present at the longer time scales of behavior, movement convergence is absent both at long and short time scales.

These results are consistent with the intuition that visual information is most impacted by mediating interactions through teleconferencing. In Zoom, interlocutors have access to similar auditory information, while visual information is confined to a small window of lower resolution that may omit gestures and details in the periphery. The lack of movement convergence, however, is at odds with previous findings of movement coordination in person (Abney et al., 2021) and in the absence of visual access to interlocutors (Shockley et al., 2003; Murray-Smith et al. 2007). This may be partly explained by methodological differences across the studies, and as with every null effect, it is also possible that statistical power is an issue, in particular when dealing with noisy signals. However, the finding of multimodal matching using the same video signals as for the movement matching, suggests the quality of these signals was sufficient to detect an effect, and thus, the lack of matching might result from the limitations of Zoom. A possibility to explore in future studies is that mediating interaction via software might be disrupting coordination beyond simply disrupting visibility. For example, having access to our own image during the interaction might be a distraction that actively interferes with interpersonal coordination

The comparison of speech coordination between Zoom and in-person datasets suggests that even though our audio experience in Zoom is similar to in-person contexts, it still doesn't support the same amount of coordination compared with in-person conversations. While interlocutors on Zoom still exhibit matching in their speech at the lower frequencies, they do not match at the higher frequencies, as is the case for in-person conversations. It is possible that the limitations of videoconferencing alter the quality of the auditory signal enough that it disrupts convergence at the finer time scales. An important but unlikely alternative to this interpretation stems from an additional difference between datasets: our participants were friends whereas Schneider et al.'s (2020) were strangers. Crucially, however, if the friendship status of the participants were driving the results, the pairs of friends in our study would be expected to converge more strongly than the strangers, and not the other way around (Fujinara, Kimura, & Daibo, 2020; Latif et al., 2014). Our results show speech convergence for friends at lower but not higher frequencies, making it more likely that the difference comes from interacting over Zoom. Future studies should further explore the differences between remote and in-person communication as a function of familiarity.

In line with Alviar and colleagues (2020), multimodal matching within individuals was present in the dyadic interactions as well. Interestingly, participants showed multimodal coordination between their movements and sounds both at the low and high frequencies. The presence of intrapersonal coordination in audio and video signals suggests that the relative lack of interpersonal convergence was due to limitations of videoconferencing, as opposed to the other factors considered. When there is no mediation of signals via Zoom, as with intrapersonal coordination and in-person conversations, the temporal organization of sound and movement converges for both finer and coarser time scales. The observed differences suggest that the loss of quality and detail in the signals due to the technological limitations of videoconferencing is likely responsible for the reduction in convergence during remote interactions.

Recordings of naturalistic conversations via Zoom yielded signals imbued with the challenges that Zoom poses to convergence. However, it also meant reduced experimental control and increased signal noise, which led to more data exclusion than is usual in studies of conversations. Dyads were discarded prior to running spectral analyses, and the remaining dyads show diverse trends in their individual spectral matching (see the colored lines in the Figures). Therefore, the high exclusion rate should not interfere with interpretation of the results. With regards to limitations of generalization, it is worth mentioning that we intentionally recruited friends, and it is likely that strangers would show different patterns of coordination (cf. Fujinara, Kimura, & Daibo, 2020; Latif et al., 2014). The role of familiarity has been shown to impact remote text-based conversations (Riordan et al., 2016). These prior findings suggest that a theoretically relevant test of teleconferencing would use familiarity as a key factor. If video teleconferencing mediates important psychological and relational factors, then similar results between friends and strangers should be obtained in that context. Future research would benefit from exploring the factor of familiarity and its relation to spectral matching in online settings.

In conclusion, disruptions to interaction dynamics during videoconferencing limit the natural convergence of behavioral rhythms and may be responsible for the stilted feel of Zoom conversations. More research on the effects of different technologies on convergence and prosocial outcomes might guide technological improvements and contribute to understanding the dynamics of interpersonal processes.

Chapter 5

Multimodal Coordination and Pragmatic Modes in Conversation

5.1. Prologue

In the previous chapters I have discussed preliminary evidence for the multimodal coordination of movements and sounds over the extended time scales of the discourse level. I have also found differences in the multimodal coordination patterns across discourse contexts that show synergy like properties in the coordination of speech and movements: patterns that change for different discourse contexts and goals, and that persist and perhaps adapt to the perturbations of online interactions. In this chapter, I outline a theoretical proposal that connects the changes in the multimodal coordination patterns of verbal and non-verbal signals to the pragmatic goals or modes that interlocutors engage on during language use. I argue that pragmatic modes or goals might be realized as metastable multimodal synergies that dynamically reassemble at the individual and dyadic level as the goals and the interactional constraints of the conversation change. I propose these multimodal synergies might even act as pragmatic affordances that are directly perceived by the interlocutors that have jointly created them given a set of shared constraints. I outline an approach to study the multimodal coordinative structures that emerge as different pragmatic modes change over the course of a conversation, as an important next step to advance our understanding of human communication and related applications.

5.1. Introduction

Human communication is intrinsically multimodal. Across cultures and in very many contexts, one observes language as mostly a face-to-face endeavor, in which at least two people are involved. Not only words are exchanged, but also looks, tones, gestures, and objects are weaved together smoothly to form meanings (Clark, 1994; Dale, Fusaroli, Duran, Richardson, 2013; Enfield, 2013; Mondada, 2016). One would also observe that most instances of language use involve a joint purpose. Language helps to coordinate interlocutors to perform a specific task, sometimes telling a story to a crowd, sometimes fighting or flirting, or simply playing and laughing without an apparent goal. In all such instances of interaction, one sees a process of coordination among interlocutors, even in one person's imagination when humans speak to themselves (Clark, 1994). In addition to coordination, and perhaps no less important, is the multiscale nature of these complex performances (Dale & Kello, 2018). Phonemes constitute syllables, syllables words, words utterances, and utterances conversational turns and topics. Gesture and prosody also respond to this structure, mirroring it to help group information, aid turn taking, and accomplish pragmatic goals.

Language is a multimodal performance that happens over time both within individuals and among them. People follow each other's gaze (Jermann & Nüssli, 2012; Nomikou et al., 2016; Richardson & Dale, 2005), subtly mimic each other's movements and tones (Levitan & Hirschberg, 2011; Louwerse, Dale, Bard, & Jeuniaux, 2012), coordinate words over time (Brennan & Clark, 1996; Fusaroli et al., 2014), continuously exchange the floor with each other with little delay (Stivers et al., 2009), adapt their words and constructions to each other's needs (Braningan, Pickering, & Cleland, 2000; Clark, 1994), and in general jointly construct every aspect of a conversation into a smooth flow and without much conscious effort (Dale et al., 2013; Garrod & Pickering, 2004). Such coordination is astonishing, even puzzling, given all the modalities and behaviors available to convey any single message (cf. Buder et al., 2010; Oller et al., 2013). The multitude of coordinative potentials can be regarded as the degrees of freedom available to interlocutors, who must select and constrain them on the fly while communicating, all while abiding the importance of timing and rhythm in language use.

Despite these observations, the scientific study of language has focused principally on single or small numbers of behaviors, void of context, and individual in nature, focusing in the "linguistic modality" and the cognitive processes that allow individuals to individually produce and understand words (Garrod & Pickering, 2004; Linell, 2004; Perniss, 2018; Rączaszek-Leonardi, 2014; Tanenhaus & Brown-Schmidt, 2008). This focus has given us a vast and valuable amount of knowledge on the mechanisms of some aspects of the workings of language like grammar and phonology, but has also left us with an incomplete view that is far from actual language use in everyday situations. Over the past 20 years multimodal research has started to gain traction in the language sciences, and although still mostly bimodal instead of multimodal and individual instead of dyadic, findings have consistently shown the importance of the other modalities and joint actions for both discourse regulation and meaning making (Clark, 1994; Stivers & Sidnell, 2005; Streeck & Jordan, 2009).

An important next step towards understanding the intrinsic multimodality of language is to integrate it into a broader theoretical framework. What are the underlying structures of multimodal "synergies" across scales, and what theoretical framework may fruitfully account for them? Importantly, what new empirical tests and investigations might this framework lead to? The goal of the present paper is to consider these questions as they relate to the formation and change of pragmatic goals in conversation. We argue that a theoretical framework rooted in ecological psychology has great promise for understanding the kind of flexible structure we see in interaction. In an ecological framework, we argue that pragmatic "modes" should be understood as metastable multimodal coordinative structures. We enter and depart these structures fluidly while communicating: hopping from a greeting, to joke-telling, compassionate disclosures, heated disputes, and so on. These modes are transient but statistically identifiable configurations in which an individual language user becomes integrated with their conversational partners and topics. They form the kind of "flow" through which loosely coupled members of a conversation proceed. We discuss some of the benefits and implications of such an approach. This approach establishes linkages to many core ecological concepts, such as specifying affordances in the social realm, and the natural statistical structure that configures our many modalities (cf. Chemero, 2009).

Before we explicate our approach to multimodal coordination and discourse, we first sketch a history of three primary and competing models for language use. This review will serve as a contrast to the proposal here, and an illustration of key ingredients required for a theoretical approach to be rich enough to capture multimodal coordination.

5.2. Three Theoretical Traditions

5.2.1. Classical Symbolic Approach: Language for Externalizing Representations

Language as a multimodal object of study has only recently become a core agenda of the language sciences. For most of the early decades of modern cognitive science, approaches inspired by classical symbolic computing dominated the scientific study of language (Chomsky, 1957; Fodor, 1983). Researchers under this tradition focused their efforts on understanding the generativity of language and its importance for thought, while leaving aside its communicative function (Bechtel, 2001). Grammar was (and remains) at the center of such language studies. Grammatical structure is a powerful conceptual tool for describing the organization of thought and meaning comprehension (Rączaszek-Leonardi, 2014). From this point of view the many nonverbal signals that accompany language, though they may be critical for communication, are not as important and are rarely brought to bear on linguistic theory (Perniss, 2018, Rączaszek-Leonardi, 2014). Language was also primarily approached scientifically as an *individual* endeavor, with a primary function of organizing and making public the private representations inside one's mind through words (Rączaszek-Leonardi, 2014; Spivey & Richardson, 2008; Trueswell & Tananhaus, 2005).

In an early challenge to this approach, McNeill argued for the "verbality" of gesture, claiming that it was as instrumental as words in language production and comprehension (McNeill, 1985). From McNeill's account, gesture stems from the same mental representation that gives rise to linguistic material, in some cases expressing other aspects of the mental model that couldn't be as easily encoded into words. In this view, gesture reflects discursive aspects of speech in close synchrony to prosodic and linguistic components. Language is still about making external the contents of the individual minds, but the study of gesture became a central and necessary part of understanding language use more comprehensively. The inclusion of gestures as important aspects of language interactions has led to progress on understanding the timing and mechanisms that make possible the gesture-speech synchronization that we observe in language use (e.g., Kita & Özyürek, 2003; Hostetter & Alibali, 2008).

The classic symbolic and modular approach has been extended to models of the coordination of prosodic contours with spoken words. In such models, decisions about prosodic contours are made early on in the message planning process based on the speaker's competence with the pragmatic and prosodic rules of her language, as well as features of the mental model being put into words (Arnold & Watson, 2015). Models of dialog interpretation and production in pragmatics, inspired by speech-act theory (Searle, 1969), also involve a representational and modular approach. These models propose that the form of an utterance is decided by a speaker via the rules of pragmatics and conversation in their language, and the meaning is inferred by a listener via the sequential and systematic application of the shared logical rules that constitute the pragmatic knowledge of the language (Jurafsky, 2004). For the models reviewed in this section, multimodality is studied as a means for expressing different aspects of internal representations. The multimodal aspects of discourse serve to express linguistic content, pragmatic intentional structure, and participate (when they do) as epiphenomena to that underlying symbolic and intentional structure.

5.2.2. An Alternative Approach: Coordinating Representations

A quite different understanding of language use emerged after these early models. The emergence of this new approach is often illustrated with Clark (1994), who reinterpreted several findings in the 1980s and early 1990s. From his perspective, language participates in *joint* action, and its main purpose is to facilitate coordination between individuals to do things in the world together. In this view, it is not about the expression of internal representations, but about the development of *shared* representations that facilitate understanding and joint action. As a mainly social activity, language needs to be studied both at the level of the individual action and at the level of the joint product the individual actions make possible. At the level of the individual actions, interlocutors have a diverse set of communicative signals at their disposal: gestures, words, prosody, gaze, body orientation, interactions with objects in the shared environment, and even fillers play a role in communication.

At the level of shared actions, the dyad has an expanding common ground as a result of their communicative exchanges that can be observed in their increasing linguistic coordination (Clark, 1994). Language is played at different levels that build up in hierarchy to organize the scales of the joint action (e.g., paying attention to each other, sending signals, decoding meaning, taking on common goals, etc.). These distinct (but interacting) "layers" depend on the topic and protagonists of the conversation (e.g., a greeting in the corridor, a story about something or someone that's not present, a joke that assumes impossible worlds). They are also organized functionally according to interactive structure, depending on the relevance of the signal for the content of the conversation or the regulation of the interaction (e.g., backchannel and primary channel).

Clark's theoretical perspective and related work at that time has inspired an alternative tradition in the language sciences. Topics like audience design and the factors that modulate it, development of common ground over time and its effects on joint action, the role of listener's feedback in the development of the interaction, and the role of less obvious signals (like fillers, and placings) have been an important part of research in this tradition. Multimodality from this theoretical perspective is at the service of increasing common ground, and as such the goal has been to understand how people use diverse communicative signals to coordinate relevant representations for joint action. Although Clark does accept that the products of the interactions might be an emergent phenomenon, most of the research in the field has focused on the *intentional* use of signals to accomplish specific communicative movements, and has put a lot of emphasis in the coordination of intentions and representations (cf. Tollefsen & Dale, 2012).

In a conceptually related tradition contemporary with this one, conversation analysts have complemented Clark's view by emphasizing the systemic nature of language use in context (Goodwin & Heritage, 1990). Here multimodality is still about creating shared representations and doing things together, but with special attention to the coupling between interlocutors and their use of semiotic resources. Research from this perspective carefully analyzes many of the semiotic movements interlocutors make over time, to identify all the subtle ways of creating meaning together (see Mondada, 2016; Perniss, 2018 for recent examples of work from this approach). Some models inspired by this tradition express dialog acts as Bayesian inference on sets of surface cues that are present in the utterances, such as the specific words being used, the prosodic elements (F0 in particular), and the probability of observing different sequences of conversational moves (Jurafsky, 2004). Importantly, both of these approaches highlight the intentional use of the signals to achieve specific goals in the creation of meaning.

5.2.3. Beneath Intention: An Ecological Approach Rooted in Dynamics

So far, we have sketched a theoretical transition from classical approaches that emphasize grammatical processes exclusively in individuals, to alternative approaches that highlight the critical importance of coordination. Both, however, assume that underneath the language user is a sophisticated intentional system that maps meanings neatly onto behaviors.

Now we turn to the ecological approach that began this paper. This alternative theoretical perspective for understanding language claims that the sort of coordination patterns that we observe in conversation are not simply the result of intentional choices of the speakers. Instead, these patterns can emerge from low-level interactions and are modulated by multiple situational constraints operating at a lot of different scales (Shockley et al., 2009; Wallot & Van Orden, 2011). These constraints include the limitations and capabilities of the individual brain and body (Dale, Kello, & Schoenemann, 2016; Shockley et al., 2009), what is possible under a given situation and goals of interaction (Chemero, 2009; Gibbs & Van Orden, 2012), the local culture and conventions of language use (Rączaszek-Leonardi & Kelso, 2008), the emergent behavioral patterns at the level of the dyad (Shockley et al., 2009), and more. Language use is seen as a *self-organized* phenomenon in which these constraints operate dynamically, and often unconsciously, to shape human interaction. These constraints convey the cognitive system from states of uncertainty and possibility, to the linguistic and pragmatic choices in conversation, and back again, iteratively (Gibbs & Van Orden, 2012; Newtson, 1993; Wallot & Van Orden, 2011).

In this tradition, linguistic elements such as words or grammatical frames are treated as another source of constraint, albeit more abstract than bodily and environmental constraints. Linguistic units act as constraints by virtue of their repeated pattern of use through learning, and by sustaining predictable effects over the dynamics of a system (cf. Elman, 2001). In that way, they are brought into the interaction to help regulate and constrain the dynamics of the other variables of the system during conversation in specific ways (Rączaszek-Leonardi & Kelso, 2008). Multimodality thus has to do with the configuration of language as an integrative process of brains, inside bodies that move, embedded within contexts, and that pursue particular goals in coordination with other brains and bodies around them. While such a perspective may seem overly broad, one could argue that this wide relevance of language, from brains to bodies and external reference, is a hallmark feature demanding closer theoretical attention (Anderson, 2014; Dale et al., 2016). The question of multimodal coordination here has to do with successful action in such complex situations. Specifically, it has to do with mechanisms for the regulation of degrees of freedom of these acting bodies in a way that facilitates cognitive control. Consider the questions that may underlie one cognitive system that is communicating with another: “When should I take the floor? What’s the best way to describe this? Should I gesture this way or that? How long should I look at their eyes, how long away? Was that the right tone?” The stultifying effect of such conscious thinking is not often faced by our cognitive systems; they somehow, and rather fluidly in most contexts, simply do it all.

An analogy may be fruitfully drawn with theories of motor control (Fusaroli et al., 2014). Most movements are highly complex, even seemingly simple movements. Multiple muscles are marshalled in the service of the movement, sometimes with complex timings (d'Avella et al., 2003). A dominant approach in motor control is that specific muscles are not controlled *separately* but rather in *combination*. These combinations are called “synergies” (Bernstein, 1967) or “coordinative structures” (Tuller & Turvey, 1982; Shockley et al., 2009). Multimodal coordination has been proposed to self-organize too into goal-dependent coordinative structures, both at the level of the individual and at the level of the dyad or group (Dale et al., 2013; Fusaroli et al., 2014). The research in multimodality from this theoretical line is still at early stages, but gaining traction. It has principally focused on studying the emergence of the coordinative structures within and between individuals, identifying the mechanisms and constraints that modulate their emergence, configurations, and changes over time, and quantifying the effects of coordination in performance (e.g., Riley et al., 2011).

This ecological approach to language highlights the idea of language as joint action that Clark (1994) helped propel. Yet this approach helps to integrate a few clear characteristics of language, bringing them to the forefront of its scientific study. These characteristics are the *complex and dynamic* nature of language. By recognizing the very complexity of language in context, and that its use offers a moving target among interlocutors, it alters our sense of what must underlie it. It highlights how language can be emergent and not necessarily representationally or intentionally driven (at least, not all the time). Importantly, the perspective explains how we can sustain smooth dynamics amidst such multimodal complexity: Each interlocutor in an interaction is shaped by varied constraints not just “inside” them, but also from the environment, which includes other interlocutors in these interactions.

This section contrasted the proposed ecological approach with two prominent theoretical traditions in language and communication. It will be useful now to offer a well-developed example of the kind of multimodal dynamics we are describing. In the next section, we offer the example of how gesture and speech are richly integrated, constrained by each other and by the individuals and contexts underlying an interaction. Then, we expand our example to involve other modalities and more varied pragmatic contexts to further illustrate the complexity of the coordination endeavor that conversation entails.

5.3. Illustration: Audiovisual Coordination in Gesture

In the prior sections, we presented a kind of historical progress from classical theories, to more recent dynamical approaches to language. This dynamical approach takes language as a kind of management of complexity through constraint: Multiple processes, within and between people, are operating in parallel, and help shape the way people communicate. This process can be so multidimensional, with so many potential degrees of freedom, that the constraints are likely as important as traditional concepts of conscious control.

Are there specific examples of this kind of synergizing force among behaviors and constraints? In this section, we give some of these considerations more direct empirical substance. We begin with gesture. Perhaps the most common strand of research in the multimodality of language has been devoted to the coordination between gesture and speech. Several facts emerge from the review of the literature in this area: the importance of synchrony, the semantic value of gestures, and the sensitivity of the relationship to diverse functional constraints.

5.3.1. Synchrony of Gesture and Speech

The first robust fact about the relationship between speech and gesture has to do with their remarkable synchronization. At a very basic level, the movements of the vocal apparatus that produce speech, like the area of the opening of the mouth, directly relate to the amplitude of the speech signal (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). At a more general level, iconic gestures tend to occur at the times that are relevant for the content of the uttered sentence (Özyürek, Kita, Allen, Furman, & Brown, 2005), and beat gestures tend to align with the peak amplitudes of the speech signal (Pouw, Harrison, & Dixon, 2019). Importantly, such synchronization seems to be necessary for the successful integration of the two signals.

At the level of production, desynchronizing speech and gesture has consequences for fluency. Disrupting gesture in virtual reality by delaying the movement of the pointing marker or freezing it during a pointing task delays speech onset as well, regardless of the moment of disruption in the speech production process (Chu & Hagoort, 2014). Similarly, disrupting speech by changing the color of the object that needs to be named delays the onset and maximum extension of gesture as well (Chu & Hagoort, 2014), suggesting the coupling of the two modalities is bidirectional. A protocol of delayed auditory feedback in which participants listen to their own speech with a 150 ms live delay, shows that gestures are modified to still be synchronous with the original — and now disfluent — speech signal despite the perturbation. Importantly, gesture is not only coupled with the speech, but is also slightly shifted towards the delayed signal, showing that gesture also gets entrained to the perceptual feedback as it is happening (Pouw, & Dixon, 2019).

At the level of perception, gestures are integrated with the content of speech more easily if they occur within a tight window of the speech onset. An ERP study of matching and mismatching gesture-speech pairs with different synchronization conditions found an increase in the N400 for mismatched pairs only for the 0 and 160 milliseconds speech delay condition, but not for the 360 milliseconds delay. Such lack of effect in the longer delay condition suggests that gestures were not integrated with the relevant word, thus avoiding the increased load to process the mismatch (Habets, Kita, Shao, Özyürek, & Hagoort, 2011). Importantly, integration of gesture and speech is sensitive to communicative constraints and will take place even under longer asynchronies (1000 ms) when presented in degraded speech conditions or to listeners with hearing impairments. Both are situations in which gesture becomes a central cue for message comprehension (Obermeier, Dolk, & Gunter, 2012).

All the subtleties mentioned up to here in speech-gesture synchrony are important because they suggest the relationship between the two modalities is not fixed and impenetrable, but it is sensitive and adaptable to environmental constraints. Such a flexible and dynamic compensation of the two systems is incompatible with all modular theories of multimodal language. In such models, the modules are supposed to be opaque and informationally encapsulated. While it is true that the modules are allowed to interact by sharing the products of their internal processes, the speed of this linear exchange wouldn't account for the kind of rapid adaptation to online perturbations that are reported in the studies above. Concepts derived from the theoretical domain of complexity and self-organization, on the other hand, can help explain the sensitivity to constraints. In particular, the surprisingly rapid shifts and reorganizations of the dynamics that allow for function while maintaining behavioral stability may reflect different parts of the system coming together for their coordinated performance (cf. Wallot & Van Orden, 2012).

5.3.2. The Semantic Role of Gestures

A second fact that has become evident about the gesture-speech relationship (for both iconic and beat gestures) is their complementarity and mutual influence with the semantic interpretation of a message. Listeners expect gesture to inform the content of speech and be congruent with it. In fact, when presented with non-matching speech-gesture pairs (e.g., hearing "chop" and seeing "twist") participants become slower and less accurate to recognize the pair as matching an action prime (Kelly, Özyürek, Maris, 2010). The same effect is observed in the ERP signals, in which incongruent speech-gesture pairs produce an N400 effect suggesting higher semantic processing load (Habets et al., 2010). Iconic gestures appear to recruit the same brain areas involved in speech: the left inferior frontal gyrus, the superior temporal sulcus (bilaterally), the superior temporal gyrus (bilaterally), and the medial temporal gyrus are all active when processing only speech and also speech accompanied by gesture (Özyürek, 2014; Willems, Özyürek, & Hagoort, 2006). Their involvement in processing are differential however, with the LIFG, STS and STG being more active for non-matching gestures (Willems et al., 2006), the MTG for expected gesture-speech pairings (Özyürek, 2014), the left STS more active for gesture-speech signals vs. nonsense hand movements, and the right planum temporal more active for all multimodal cases (Hubbard, Wilson, Callan, & Dapretto, 2009).

Beat gestures also play a role in the semantic processing of speech by facilitating interpretations in line with less preferred syntactic constructions. For instance, the P600, previously linked to syntactic reinterpretations, is eliminated when the subject of an utterance with the less common OSV order is marked with a beat gesture. Importantly such an effect was only observed for beat gestures, but not for prosodic markings, or a visual depiction of the beat gesture path using a moving ball. Such specificity points to the importance of the communicative intention of the cue as well as its saliency in the multimodal signal (Holle et al., 2012). One possible mechanism for the influence of beat gestures in language processing can be found in their power to modulate neural oscillations. Beat gestures have been shown to reset the phase of the theta band in the auditory cortex to make it synchronous with the word onset, presumably providing an advantage for word processing. They also seem to affect alpha oscillations in the inverse way, making them less phase locked with the word, possibly further modifying the attention mechanisms for word processing (Biau, Torralba, Fuentemilla, de Diego Balaguer, & Soto-Faraco, 2015).

5.3.3. Sensitivity to Functional Constraints

A third fact of the gesture-speech relationship is its sensitivity to diverse functional constraints that modulate the gestures produced. The preferred syntactic encoding of the language being spoken seems to be one of such functional constraints. Turkish and English speakers, asked to describe the movement of cartoon figures in an animation, showed radical differences in their gesturing patterns that were in line with their language's preferred way of linguistically encoding the movement. English speakers that can include both the path and the manner of movement jointly in the verb of the utterance, tended to produce compound gestures showing both the path and manner of movement at the same time, whereas Turkish speakers that are forced to include path and manner in their sentences using separate words, tended to produce gestures that depicted just one of the two characteristics but not both (Özyürek et al., 2005).

A second functional factor that is of key importance for gesture production is the degree to which the production situation involves the intention to communicate. This factor can be said to be encompassed in a larger debate in the field about the primary function of gestures. Some scholars argue that gestures are produced to help the speaker retrieve words more easily, and thereby support fluency (see Krauss, Chen, & Gottesman, 2000). While others argue that gestures are produced to help the addressee understand a message, and thereby facilitate communication. Empirical evidence suggests gestures play both roles, but they seem to be especially tuned to the presence and needs of the addressees, or in other words to the general communicative character of language. In the general sense of the debate, the quality of the movements and their relationship to the sounds seem to be different depending on the communicative constraints of the situation. For instance, gestures and actions produced under a communicative condition in which the partner is believed to be required to learn the movements performed by the participant are larger, faster, and more punctuated, than gestures produced in a non-communicative condition in which the partner is believed to be learning the general logistics of the experiment (Trujillo, Simanova, Bekkering, & Ozÿurek, 2018). Also, the dynamic structure in the movement amplitude and the sound amplitude of a performance is more tightly correlated in situations in which the movements are more informative for the message (e.g., spontaneous speech) than in situations in which the movements are responding more strongly to stylistic (e.g., poetry) or non-communicative constraints (e.g., classical music; Alviar, Dale, Dewitt, Kello, 2020).

In the more specific sense of the debate, speakers modify their gesture rate and quality as a function of their addressees. For example, keeping lexical difficulty constant in a tip-of-the-tongue task while varying the degree of social interaction, results in higher numbers of iconic gestures when participants interact with a partner, either face to face or through a screen, than in solo conditions (Holler, Turner, & Varcianna, 2013). Relatedly, participants adapt their gestures to their addressee's location and knowledge. Regarding the former, in a cartoon retelling paradigm, speakers performed gestures accompanying prepositions like *into* or *out of* along the lateral axis or the sagittal axis depending on the location of the addressees, and therefore, the location of the shared space with them (Özyürek, 2002). Regarding the latter, speakers will produce bigger and more precise gestures when retelling a cartoon to a new addressee, even when retelling the story for the third time, while making their movements smaller and sloppier for people that have heard the story before (Galati & Brennan, 2014). They will also predict possible ambiguity for the listener when referring to homonyms and produce more iconic gestures to help disambiguate the meaning, producing more detailed gestures the less disambiguation is included in the verbal modality (Holler & Beattie, 2003). Speakers will also perform bigger more detailed gestures after an addressee's feedback suggests a problem in understanding, making their gestures more communicative (Holler & Wilkin, 2011).

The responsiveness of this gesture-speech synergy to such a diversity of factors like the syntactic structure of language, the location and knowledge of the addressees, the ambiguities in the information, and the communicative nature of a situation, is a feature that is hard to explain in modular models. All this information would have to be specified in advance in the form of conditions that mandate the selection of specific collections of rules inside a contextual module that would later serve as input for the operation of the other modules to make their outputs sensitive to the context (cf. Onnis & Spivey, 2012). This may be possible cognitively, but it would not be a very parsimonious model (Casasanto & Lupyan, 2015). On the other hand, the dynamical proposal includes at its core the idea of constraints, both internal and external, individual and collective, that give rise to different emergent coordinative structures and shape the state space of the system differently depending on their specific combination over time. From this theoretical point of view, speaker and addressee could also be understood as another emergent system with coordinative structures of its own (Dale & Spivey, 2019; Schmidt et al., 1990; Shockley et al., 2009). This system encompasses the two individuals, and its reorganization happens not for the speaker or for the listener uniquely, but for the two-person system's functional stability and coordinated response to changing constraints.

5.4. Multimodal Coordination in Broader Contexts of Language Usage

Most of the research reviewed to this point covers only a few of the pragmatic goals humans pursue with conversation. In the day-to-day use of language, people do more than describe scenes to one another and say ambiguous sentences out loud. People use language to flirt, to make each other laugh, to teach, to be ironic, to make fun of themselves and others, to meet new people, to be rude, to disagree. And they do it using a wide variety of discursive tools: they make requests, they reenact what they saw, they demonstrate what they mean, they indicate understanding, they nod and smile to acknowledge their interlocutors, they give orders, they let one another know they are not done speaking but are thinking what to say next. All these instances of language use involve a complicated multimodal dance of words, gestures, tones, and gaze that regulate the interaction.

Much of the research history of pragmatics involved understanding how speakers choose to use one word over another. Even in contemporary theories of pragmatics, word choice serves as a focal point of inquiry (see Gibbs and Van Orden, 2012). However, language is not only words. Speakers also get to “choose” the prosodic inflections they give to the message, the gestural accompaniments, the facial expressions they perform, and even the coherence among all of these to give the listener additional cues about their intentions. In the following paragraphs, we review research on the multimodal performance of pragmatics. These insights open additional avenues of exploration and point to understudied variables of language use in natural conversation.

Gesture in general does not only have a semiotic function, but also a pragmatic one that reflects the kind of discourse movements and social goals that speakers wish to achieve in an interaction (Kendon, 2017). Gestures in their pragmatic relationship to speech can be (a) operational, reflecting some of the general content of the utterance (e.g., expressing negation), (b) modal, indicating a frame of interpretation for what’s said (e.g., gestural quotations, epistemological stance), (c) performative, showing emphasis or indicating the speech act (e.g., pointing to the salt while asking for it), (d) parsing, marking punctuation and relationships between parts of the speech content more noticeable (Kendon, 2017). Pointing gestures, for example, have been shown to be at the center of successful recognition of indirect requests (Kelly, Bar, Church, & Lynch, 1999). However, they not only serve to index, but are also interpreted by viewers as affirming and giving emphasis to what is being said, while brushes are conversely interpreted as establishing a negative attitude towards the information uttered (Freigang & Kopp, 2015). Exaggerated gestures, by their part, seem to be a key component of humorous self-mockery distinguishing it from the self-deprecating type, and inviting others to exaggerate and laugh with oneself (Yu, 2013).

Beyond hand gestures, facial gestures seem to be very common and instrumental in the regulation of an interaction, though less studied. Gestures like the *thinking face* help speakers communicate they are having problems finding the words but are not ready to give up the floor to their interlocutors, nor require action from them. *Smiles* have a variety of functions as a gesture too, like indicating humor, comprehension, awareness of reference to obvious things for the interlocutor, awareness of mistakes in the conversation, etc. Lastly, the *pointing with face and eyes to the hands* gesture helps highlight important hand gestures to the addressee indicating is time to pay attention to what the hands of the speaker are going to show next (Bavelas & Chovil, 2018).

Prosody, by its part, is also central to the use of language for different conversational goals. Sarcasm, and verbal irony in general, have been repeatedly shown to be enacted through the use of prosodic features, although not specific and unique to irony, nor generalizable across speakers (Bryant & Fox Tree, 2005), but markedly shifted in comparison to the surrounding non-ironic utterances in the conversation (Bryant, 2010). Intentions like prohibition, approval, attention, and comfort are also robustly recognizable in an unfamiliar language for the listeners via the prosodic characteristics of the speech signal, particularly when infant directed (Bryant & Barrett, 2007). Differences between spontaneous and then read speech in a variety of English corpora involving diverse task-directed conversations also point to prosodic variations between these two types of speech (see Hirshberg, 2000 for a review). Hirschberg (2000) suggests it is possible that global features vary with broad differences like speech type, while local features vary as a function of subtler semantic differences.

Gesture and prosody rarely, if ever, play their roles by themselves in the conversation. In humorous self-mockery for example, exaggerated gestures come hand in hand with exaggerated prosody (Yu, 2013), and in re-enactments, they come hand in hand with specific gaze and word patterns. When engaging in the reenactment of a scene, speakers look away from the interlocutors (Sidnell, 2006) and exaggerate their gestures and their prosody to different extents depending on the "quotative expression" (e.g., like, says, go), or lack thereof, used to introduce the reenactment episode (Blackwell, Perlman, & Fox Tree, 2015). With facial gestures, *thinking face* gestures for example, are also usually accompanied by expressions like "uh" and "um" that are also indicative of speech difficulties of specific lengths (Clark & Fox Tree, 2002), and also involve redirecting gaze away from the conversation (Bavelas & Chovil, 2018). In dry verbal irony, having access to only prosody or only words is not sufficient to recognize irony, and their combination is needed (Bryant & Fox Tree, 2005). And in automatic speech act recognition, accuracy gets better when not only the words and the sequences of utterance types are included, but also the prosodic information (Stolcke et al., 1998). All these examples point to the same general intuition: language use is always multimodal, and different pragmatic goals and discourse tools are constructed using different combinations of modalities that seem flexible, and highly sensitive to one another, the interlocutor, and the goals of the conversation.

While face-to-face language is always performed with the entire body and in coordination with the environment, scientific research on language, however, rarely includes more than two modalities at a time. This is likely to simplify the measurement problem by focusing on particular modalities of interest. It begs the question though: are there multimodal coordinative structures that mediate the comprehension of semantic and pragmatic intent during conversation, and could be used to, for example, better classify speech acts and pragmatic situations? It is very likely that the specific combinations of movements, pitch, speech rate, facial expressions, words etc., are dependent on pragmatic intent. For instance, a diverse combination of prosodic features, discursive moves, and topics of conversation as extracted from words, contribute to the accuracy of the automatic detection of flirting, friendliness, assertiveness, and awkwardness in a speed dating audio corpus (Ranganath, Jurafsky, & McFarland, 2013). Although not all features are relevant for the categorization of all groups, and words seemed to carry most of the weight, the diversity of features that people vary consistently when flirting, just trying to be friendly, or having an awkward interaction is impressive, and goes along with the idea of subtly diverse multimodal coordinative structures for different interaction purposes. Similarly, Pentland (2008) and collaborators (e.g., Kim et al., 2012) have founded an entire research literature on automatically extracting clusters of diverse features that, when integrated, predict deception, outcomes of negotiation, teamwork collaboration, and more.

Earlier in this paper, we defined the concept of "coordinative structure" as describing a loosely assembled configuration of various states of individuals and groups acting together. These coordinative structures may emerge from a collection of constraints at various levels, including the goals of conversation partners. Pragmatic goals, for example, organize speakers into temporary configurations like telling jokes, giving directions , showing compassion, and so on. The behaviors that manifest under these coordinative structures adapt to *together*, in parallel, much as the research on gesture and speech shows a kind of rich mutual adaptation. If different coordinative structures are at the base of using language for different pragmatic goals, studying the types of coordinative structures that form under different pragmatic goals in a conversation and how they change is a natural next step for the study of pragmatics, and multimodal coordination in general. We refer to these quasi-stable configurations that dynamically emerge and transition across each other as *pragmatic modes.* It might be that pragmatic modes are realized as changes between dynamic regimes and metastable states of the cognitive system (see Dale, 2015, for an example of analyzing discursive modes along this lines).

Another insight from multimodal research in pragmatics, but also multimodal research in general, is that interlocutors can be highly responsive to one another. Listeners laugh and join in when presented with the exaggerated gesture and prosody of humorous self-mockery, but rarely do it when speakers engage in the self-deprecating version (Yu, 2013). Listeners stop all other activities and attentively look to speakers when they engage in reenactments, playing along when they are used as characters of the reenacted situation (Sidnell, 2006). They politely wait and listen while speakers perform a thinking face and search for the next word, participating in the search only when directly or indirectly asked to do so (Bavelas & Chovil, 2018). When allowed to interact normally, interlocutors subtly coordinate their movements with each other during knock-knock jokes at all the time scales important for the interaction (Schmidt, Nie, Franco, & Richardson, 2014). They also coordinate their facial gestures, their gestures, their language use, and even their laughs while describing a route to a listener, and they may become even more synchronized the more difficult the task gets (Louwerse, Dale, Bard, & Jeuniaux, 2012). Even at a neural level, interlocutors show synchronization in their neural oscillations while engaged in conversation above and beyond the synchronization that would be expected from entrainment to the same speech signal. Synchronization in the alpha oscillations, related to attentional mechanisms, in conjunction with synchronization in beta oscillations, related to motor movements, are proposed to reflect the emergence of an interactive prediction process that encompasses the dyadic system (Perez, Carreiras, & Duñabeitia, 2017; cf. Hasson et al., 2012).

The notion of a pragmatic mode as a coordinative structure of the multi-person system, i.e. a temporary set of states or conditions on its way to the next mode, implies that multimodal coordination is labile under different situations. Some qualitative conversation analysis research offers illustration of this. In one example, Jensen (2018) offers a qualitative analysis of spontaneous humor in which he analyzes the coordinated verbal and non-verbal exchanges between interlocutors and shows how humor is jointly constructed by multimodal signaling that communicates the existence of humorous affordances and values in the situation. It is possible that this new mode of interacting emerges as people shift to playfulness, and humor looks different at the level of the multi-person system than the mode of interaction they were pursuing before. It is possible that this change is not obvious at the individual level (e.g., prosody: Flamson, Bryant, & Barret, 2011), but may instead be observable in the *dyadic* system. It might be the case that not only are individuals travelling across metastable states when changing pragmatic modes, but that dyadic systems also show a reorganization of coordinative structures for different conversational goals and modes. Having quantitative measures and analyses of the dyad as the basic unit is another natural step towards a more dynamic study of pragmatics that follows the intuitions of complex dynamic systems theory applied to language (Dale & Spivey, 2019).

5.5. Summary, Conclusion and Future Directions

We began this paper by putting the multimodal complexity of natural language performance into stark relief with prior theoretical traditions. We argued that an ecological approach that integrates an individual language user with her conversation partners and their mutual environment is critical to understanding *how* such multimodal complexity is possible. Using speech-gesture coordination as an illustration, we showed that these modalities subtly constrain each other, involving constraints from both listeners and speakers, and can be influenced by the context and goals of an interaction. We then expanded our argument to describe how this specific instance of gesture and speech extends to the cognitive system more generally. We proposed pragmatic modes as an organizational feature of language use that emerge as a result of multimodal coordinative structures and associated goals of interaction partners. These coordinative structures are temporarily stable ("metastable") and transition from one to the next during interaction. This theoretical proposal explicitly links the level of pragmatic goals with the "nexus" of multimodal signals that manifest these goals, and transition them to new ones as conversations unfold.

Our approach is explicitly aligned with ecological psychology. Consider, for example, some key concepts of ecological psychology (Chemero, 2009; Lobo et al., 2018). For each, we can articulate how the present proposal for understanding multimodal dynamics in conversation relates to them. The *perception-action loop* suggests that production and comprehension among interlocutors are an active process, in which language comprehension is active, even anticipatory (Pickering & Garrod, 2014), and makes use of varied nonverbal structures to shape and guide the speaker and the listener (Clark, 1996). This active exchange among interlocutors represents a kind of coupling that echoes the concept of *organism-environment couplings*, but here we have cognitive agents who are mutual members of each other's environments, thus forming a kind of duality (cf. Dale & Spivey, 2019). The *sine qua non*, to many, of ecological psychology are the related concepts of *information pickup and affordances*. Here we find a potential bridge with pragmatic modes, but one that is subtler. Pragmatic modes are neither fixed nor invariant. Interlocutors may create new coordinative structures which are unique to their own specific moment in space and time. These new coordinative structures are understood by them uniquely as active members of the system that created them.

The influence of multimodal structure may relate in some ways to the kind of specification that the visual array supplies to organisms that engage their environment. Instead of visual arrays, we have multimodal arrays that are in flux, constraining each person in the dialogue and only temporarily establishing conditions for dialogic structure. Consider some recognizable examples: looking down at a watch, turning up one side of one's mouth, furrowing the brow, a long pause while looking at the ground, etc. An intriguing hypothesis is that members of successful conversational systems may *directly perceive* the significance of these moments *as* members of these conversational systems. The "mesh" of multimodal structure may be a kind of visual (haptic, auditory, etc.) array that forms the Gibsonian-like structure of conversational synergies.

There are obvious limitations to our proposal here. More work is needed to develop the concept of a pragmatic mode to the point of generating testable, informative hypotheses. Nevertheless, we would argue that the examples of gesture and speech serve as very strong illustrations of these concepts, even if "pragmatic mode" is a relatively new frame for them. A proposal of this kind would be successful to the extent that it recommends new empirical research. Here we conclude by offering two such examples, one more theoretical, and the other more applied.

First, consider a potential theoretical test of these ideas in the real-time dynamics of conversations and their disruption. One prediction that follows from the idea of pragmatic modes has to do with overhearers or interlocutors that join a conversation at a later point, after its coordinative structures have been established by other conversation partners. As external to the system, the new interlocutors should show a different dynamic regime than the other two, and they should show variability while they reorganize themselves to participate in the existing system. The dyadic system should also show signatures of reorganization as members adapt to the new interlocutor and create coordinative structures that support the new constraints that come with it. Articulating these dynamics across multiple modalities under these contexts would serve two exciting theoretical purposes. First, articulated dynamics would specify the manner in which varied modalities and behavioral patterns are assembled uniquely, and how they flexibly adapt. And second, they would shed more light on the nature of the underlying control processes of language and communication themselves.

An intriguing field in which one could apply our approach is the creation of virtual agents with human-like communicative behaviors. Building realistic looking avatars has proven to be a challenge despite all our knowledge on audiovisual coordination in conversation, in part because most of our research, in its bimodal nature, has only bimodal insights to offer for the creation of a multimodal avatar. Engineering naturalistic agents will require an understanding of the coordination of gestures, words, prosody, facial expressions, gaze direction and even changes in posture and body sway. Ignoring any of the pieces or their timing with respect to the interlocutor's behavior produces virtual agents that are perceived as non-realistic with limited ability to establish rapport with human interlocutors (Bergmann, Kahl, & Kopp, 2013; Bergmann & Kopp, 2009; Gratch et al., 2006; Gratch, Wang, Gerten, Fast, & Duffy, 2007; Huang, Morency, & Gratch, 2011). In general, the work in virtual conversational agents suggests that, to be more natural and human-like in their behaviors, virtual agents benefit from architectures that work less as a modular system with rigid rules and symbols, and more as a complex embodied system. Seeking to replicate the sort of multimodal dynamics in artificial systems will allow researchers to better understand the viability of the constraints that guide the human case, too.

Human language is a multimodal phenomenon that has at its core the coordination of dyadic and group actions in complex environments. In this paper, we have argued that there is much evidence that the relationship among signals like gestures, prosody, words, and gaze is not a simple and rigid one as would be predicted from the modular models of multimodal language. Speakers adapt their multimodal behaviors to their interlocutors needs, the changing constraints of the environment, the goal and constraints of the tasks, and reorganize their behavior to keep the system stable when faced with perturbations that jeopardize the successful communication of the message. Such adaptability of behavior seems to be more in line with the ideas of loosely coupled coordinative structures that self-organize and emerge to accomplish successful communication in the midst of an ever changing set of constraints. The research on multimodal coordination in different pragmatic settings is scarce, mostly bimodal, and mostly individual instead of dyadic, but it suggest similar adaptation and flexibility to the research on more generic instances of language use. Speakers also use their modalities differently depending on the pragmatic goals pursued and discursive moves used in the conversation. We propose pragmatic goals and modes might be carried out by putting together different coordinative structures within the individual system and the dyadic system, and can therefore, be understood as metastable states of the language system with differences in their dynamic regimes. Advancing a project like this is in our opinion central to advancing the understanding of language use beyond the lab and controlled tasks, beyond pencil and paper tasks, and towards its occurrence in the wild with all its nuances, all its beauty, and all its power.

6. Concluding Thoughts

Years of studies have now established that gesture and speech are organized and coordinated at the word and sentence level (see Özyürek, 2014 for a review of gesture research; Cole, 2015 for a review of prosody research). In this dissertation, I aimed to determine whether such coordination also extended to the discourse level by studying: (a) whether body movement and speech prosody were stably organized and coordinated over the extended time scales of a full conversation or monologue; and (b) whether such coordination and organization varied in response to changes in discourse level variables, such as goals and communicative context.

The results provide initial evidence for modality organization and coordination that expands over several minutes at a time. Although subtle, speakers communicating complex information during academic presentations, show patterns in their prosody, their movements, and their PowerPoint slides' transitions, that follow common shapes over the course of their talk (Chapter 2). Also, the multiscale organization of speech and movement during monologues and dialogues is coordinated over several minutes of observation at a time (Chapter 3 and 4), and for both the fast and slow fluctuations of the sounds and movements of language (Chapter 4).

The results also provide initial evidence for differences in extended multimodal coordination patterns that are related to changes in discourse goals and contexts. The multiscale structure of sounds and movements is more tightly coordinated for language, where speakers have an explicit and specific communicative goal, than for music, where performers' goals are more aesthetic and less communicatively specific (Chapter 3). Multimodal multiscale coordination may also change between monological and dialogical contexts of language use. The effect sizes for multimodal coordination in Chapter 3 and Chapter 4, suggest stronger coordination of the movements and sounds of an individual in the context of monologues as compared to dialogues. Future studies should investigate this descriptive observation more directly and systematically.

Lastly, some of the patterns observed in the coordination of prosodic patterns and body movement may offer initial support to the proposal outlined in Chapter 5, as they are reminiscent of the characteristics that are expected of synergies. The results of Chapter 2, for example, allude to the reduction of degrees of freedom that synergies mediate in a complex system (Kelso, 2009). The prosody, body movements, and PowerPoint slides' transitions speakers produce during a scientific talk can be compressed into fewer components that explain more variance than each of the original signals alone, suggesting at least some of the modalities are being coordinated as one integrated stable unit. The results of Chapter 3, by their part, are consistent with the functional sensitivity and robustness to perturbation that is attributed to synergies (see Kelso, 2009). For instance, performances with radically different goals —such as talking vs. playing a musical part, a capella singing vs communicating, or jazz improvising vs. performing classical music— exhibit different levels of multimodal multiscale coordination, whereas performances with similar goals but differences in their execution —such as classical pieces performed with different instruments— result in similar levels of multimodal multiscale coordination. And in this same spirit, the reduced multimodal coordination in the dialogues of Chapter 4 compared to the monologues of Chapter 3, might correspond to a new stable configuration of the multimodal synergy that reflects precedence of dyadic goals over individual goals in the context of interpersonal coordination for conversation. All these observations although promising and theoretically interesting, are still admittedly mostly conjectures that are in need of more evidence for their support.

The proposal of pragmatics as driven by the unfoldment of metastable multimodal synergies offers some other predictions derived from the characteristics of synergies. For instance, talking about multimodal *synergies* as opposed to just different *coordination patterns*, suggest we should be able to find evidence of compensatory element reorganization in the face of perturbation (Kelso, 2009). Some of the patterns of Chapter 4 might point in this direction for the interpersonal coordination of multimodal signals. For instance, stronger speech convergence at the lower frequencies for online compared to in-person conversations, could be the result of compensatory synergy reorganization to keep dyadic coordination stable when movement convergence is disrupted. A proper in-person control condition and direct manipulations of the size and timing of the perturbations (e.g., more detailed vs. blurrier video signals in Zoom) could better test this possibility in a naturalistic dyadic context. In a monological experimental context, the use of robotic exoskeletons, for example, could prove useful to determine whether perturbing the movements of speakers as they tell a story alters the structure or organization of their prosodic patterns as well.

Another prediction of the account presented in Chapter 5 relates to the dynamic adaptations of the multimodal synergies as associated with the changes in pragmatic goals that speakers and dyads are dynamically pursuing. Just as perturbations should be dissipated by the synergies, changes in the pragmatic goals that speakers are pursuing — such as momentarily telling a joke, or getting serious in an otherwise lighthearted conversation to express hurt— should be preceded by the emergence of new and statistically differentiable multimodal synergies. And, following yet another prediction, the *transitions* to new transiently stable states, should exhibit a critical slowing down (i.e., a slower decay of autocorrelation) and an enhancement of variability at the local level (Kelso, 2009, 2012). If two pragmatic goals in a conversation are indeed realized as two different metastable states of the system as I argue, then we should be able to see increased autocorrelation and increase local variability just before the system is qualitatively perceived to be in a new pragmatic "mode".

The approach taken here has a couple of limitations. In the first place, working with naturalistic datasets affords inferences that are ecologically valid and relevant. However, it also means reduced control over the variables affecting the phenomena of interest. This, in conjunction with the use of automated methods to extract the data, results in lower signal to noise ratios that increase the likelihood of Type II errors and make the patterns observed less interpretable, as in Chapter 2. In this same line, naturalistic datasets are full of uncontrolled variability in the environmental constraints that is likely to affect the linguistic behaviors of speakers. In fact, from the perspective of coordination dynamics, those constraints will differentially shape the state space of a system to afford stable states of behavior that are appropriate to the situation (Kelso, 2012). Not being able to control or properly quantify the constraints that were in play in any particular performance, makes it harder to obtain clear pictures of the shared multimodal patterns and interpret their differences. This problem could be reduced in the future by: (a) controlling for the situational characteristics in which the behavior takes place (naturalistic behavior but in the control of the lab), or alternatively, (b) coding for some of the characteristics believed to be most important to the behavior and studying shared patterns and individual differences with respect to those constraints. These options would likely bring some clarity to the patterns reported in Chapter 2, and perhaps further refine the results of the other chapters. More generally, these options would also likely help trace the observed statistical patterns back to the psychological variables that are more meaningful and familiar to scholars from more traditional approaches to language research. Lastly, the methods I used here prioritized the coarse measurements of the temporal dynamics over finer ones. This permitted smoother measurements that were necessary given the noisier quality of the data obtained with automated methods. However, it also limited the possible inferences to the coarser attributes of the discursive contexts, leaving out of reach more fine-grained explorations of the dynamics of the performances as they unfolded.

Establishing a bridge between the multimodal coordination and organization patterns, and the variety of interpersonal goals that language fulfills is an important step for our understanding of language use in situated social contexts (cf. Dale, 2015). This endeavor promises to be a challenging and productive life-long research program for which this dissertation provides only the first steps. The use of methods that allow for higher temporal resolution such as wavelet analysis, recurrence quantification analysis, or even windowed spectral matching will facilitate the exploration of the kinds of rapid adaptations in dynamic regimes that are hypothesized in Chapter 5. A bigger challenge perhaps, will be in designing the tasks that elicit and perturb different pragmatic modes in naturalistic ways in the lab, or in finding existing audiovisual datasets that capture different pragmatic phenomena in natural conversation with sufficient quality for its study using automated methods.

References

Abrahams, M. (2016, April). A big data approach to public speaking. *Insights.* Retrieved from https://www.gsb.stanford.edu/insights/big-data-approach-public-speaking

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

Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2021). Cooperation in sound and motion: Complexity matching in collaborative interaction. *Journal of Experimental Psychology: General*. Advance online publication.

Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and Instruction*, *32*(1), 65-100.

Allan, D. W. (1966). Statistics of atomic frequency standards. *Proceedings of the IEEE*, *54*(2), 221-230.

Alviar, C., Dale, R., Dewitt, A., & Kello, C. (2020). Multimodal Coordination of Sound and Movement in Music and Speech. *Discourse Processes*, *57*(8), 682-702.

Alviar, C., Dale, R., & Kello, C. (2018, July). The fractal structure of extended communicative performance. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1292 - 1297). Austin, TX: Cognitive Science Society.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. MIT Press.

Arnold, J. E., & Watson, D. G. (2015). Synthesizing meaning and processing approaches to prosody: Performance matters. *Language, Cognition and Neuroscience, 30*(1-2), 88-102.

Barbosa, A. V., Yehia, H. C., & Vatikiotis-Bateson, E. (2008). Linguistically valid movement behavior measured non-invasively. In *AVSP* (pp. 173-177).

Barney, T. (1999). Readers as text processors and performers: A new formula for poetic intonation. *Discourse Processes*, *28*(2), 155-167.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). Cicero-towards a multimodal virtual audience platform for public speaking training. In *International Workshop on Intelligent Virtual Agents* (pp. 116–128). Springer.

Bavelas, J., & Chovil, N. (2018). Some pragmatic functions of conversational facial gestures. *Gesture, 17*(1), 98-127.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495-520. doi: 10.1016/j.jml.2007.02.004

Bechtel, W. (2001). Linking cognition and brain: The cognitive neuroscience of language. *Philosophy and the neurosciences: A reader*. Oxford: Basil Blackwell.

Bergmann, K., Kahl, S., & Kopp, S. (2013, August). Modeling the semantic coordination of speech and gesture under cognitive and linguistic constraints. In *International Workshop on Intelligent Virtual Agents* (pp. 203-216). Springer: Berlin, Heidelberg.

Bergmann, K., & Kopp, S. (2009, September). GNetIc–Using bayesian decision networks for iconic gesture generation. In *International workshop on intelligent virtual agents* (pp. 76-89). Springer: Berlin, Heidelberg.

Bernstein, N. A. (1967). *The coordination and regulation of movements.* Oxford, UK: Pergamon Press.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143-152.

Biau, E., Torralba, M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*, *68*, 76-85.

Black, J. W. (1961). Relationships among fundamental frequency, vocal sound pressure, and rate of speaking. *Language and Speech*, *4*(4), 196-199.

Blackwell, N. L., Perlman, M., & Tree, J. E. F. (2015). Quotation as a multimodal construction. *Journal of Pragmatics, 81*, 1-7.

Boersma, P. & Weenink, D. (2010). Praat: Doing phonetics by computer (Version 5.1.31) [Software]. Available from http://www.praat.org

Bögels, S. & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, *52*, 46-57.

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13-B25.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7-9), 1044-1098.

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.

Brock, S., & Joglekar, Y. (2011). Empowering PowerPoint: Slides and teaching effectiveness. *Interdisciplinary Journal of Information, Knowledge, and Management*, *6*(1), 85–94.

Bryant, G. A. (2010). Prosodic contrasts in ironic speech. *Discourse Processes, 47*(7), 545-566.

Bryant, G. A., & Barrett, H. C. (2007). Recognizing intentions in infant-directed speech: Evidence for universals. *Psychological Science, 18*(8), 746-751

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice?. *Language and Speech, 48*(3), 257-277.

Buder, E. H., Warlaumont, A. S., Oller, D. K., & Chorna, L. B. (2010). Dynamic indicators of mother-infant prosodic and illocutionary coordination. In Speech Prosody 2010-Fifth International Conference.

Burgoon, J. K., Bonito, J. A., Ramirez Jr, A., Dunbar, N. E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication*, *52*(3), 657-677.

Cappella, J. N., & Planalp, S. (1981). Talk and silence sequences in informal conversations III: Interspeaker influence. *Human Communication Research*, *7*(2), 117-132.

Casasanto, D., & Lupyan, G. (2015). All concepts are ad hoc concepts. *The conceptual mind: New directions in the study of concepts*, 543-566.

Cassell, J., & McNeill, D. (1991). Gesture and the poetics of prose. *Poetics Today*, *12*(3), 375–404.

Cassell, J., McNeill, D., & McCullough, K.-E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *7*(1), 1–34.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001, July). Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 114-123). Association for Computational Linguistics.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893.

Chartrand, T. L., & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology*, *64*, 285-308.

Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.

Chen, L., Leong, C. W., Feng, G., & Lee, C. M. (2014). Using multimodal cues to analyze mla'14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 45–52). ACM.

Chomsky, N. (1957). *Syntactic structures*. Mouton & Co.

Chu, M. & Hagoort, P. (2014). Synchronization of speech and gesture: Evidence for interaction in action. *Journal of Experimental Psychology: General*, *143*(4), 1726-1741. doi: 10.1037/a0036281

Clark, H. H. (1996). *Using language*. Cambridge University Press: Cambridge.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243-268). Hillsdale: Erlbaum.

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, *7*(4-5), 507-525.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, 7-19.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73-111.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694.

Coey, C. A., Washburn, A., Hassebrock, J., & Richardson, M. J. (2016). Complexity matching effects in bimanual and interpersonal syncopated finger tapping. *Neuroscience Letters*, *616*, 204-210. doi: 10.1016/j.neulet.2016.01.066

Cohen Priva, U. C. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition, 160*, 27-34.

Cole, J. (2015). Prosody in context: a review. *Language, Cognition and Neuroscience*, *30*(1-2), 1-31.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*(2), 425-452.

Dale, R. (2012). Integrating and extending the distributed approach in cognitive science. *Interaction Studies*, *13*, 125-137.

Dale, R. (2015) An integrative research strategy for exploring synergies in natural language performance. *Ecological Psychology, 27*(3), 190-201

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 59, pp. 43-95). Academic Press.

Dale, R. & Kello, C. T. (2018). "How do humans make sense?" Multiscale dynamics and emergent meaning. *New Ideas in Psychology*, *50*, 61-72. doi: 10.1016/j.newideapsych.2017.09.002

Dale, R., Kello, C. T., & Schoenemann, P. T. (2016). Seeking synthesis: The integrative problem in understanding language and its evolution. *Topics in Cognitive Science, 8*(2), 371-381.

Dale, R., & Spivey, M. J. (2019). Weaving oneself into others. *Eye-tracking in Interaction: Studies on the role of eye gaze in dialogue*, 10, 67.

Dalla Bella, S., & Palmer, C. (2011). Rate effects on timing, key velocity, and finger kinematics in piano performance. *PloS One*, *6*(6), e20518. doi: 10.1371/ journal.pone.0020518

d'Avella, A., Saltiel, P., & Bizzi, E. (2003). Combinations of muscle synergies in the construction of a natural motor behavior. *Nature Neuroscience*, *6*(3), 300-308.

De Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385-390.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515-535.

Dean, R. T., & Dunsmuir, W. T. (2016). Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models. *Behavior Research Methods*, *48*(2), 783-802.

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural viewpoint signals referent accessibility. *Discourse Processes*, *50*(7), 431-456. doi: 10.1080/0163853X.2013.824286

Delignières, D., Deschamps, T., Legros, A., & Caillou, N. (2003). A methodological note on nonlinear time series analysis: Is the open-and closed-loop model of Collins and De Luca (1993) a statistical artifact?. *Journal of Motor Behavior*, *35*(1), 86-96. doi: 10.1080/00222890309602124

Delignières, D., Lemoine, L., & Torre, K. (2004). Time intervals production in tapping and oscillatory motion. *Human Movement Science*, *23*(2), 87-103. doi: 10.1016/j.humov.2004.07.001

Delignières, D., & Torre, K. (2009). Fractal dynamics of human gait: a reassessment of the 1996 data of Hausdorff et al. *Journal of Applied Physiology*, *106*(4), 1272-1279. doi:10.1152/japplphysiol.90757.2008

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, *28*(9), 1255-1269.

Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics*, *25*(3), 359–410. doi:10.1515/cog-2014-0024

Duran, N. D., & Fusaroli, R. (2017). Conversing with a devil's advocate: Interpersonal coordination in deception and disagreement. *PloS One*, *12*(6), e0178140.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*(4), 547-582.

Enfield, N. J. (2009). *The anatomy of meaning: Speech, gesture, and composite utterances* (Vol. 8). Cambridge University Press.

Enfield, N. J. (2013). *Relationship thinking: Agency, enchrony, and human sociality*. Oxford University Press.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research: JSLHR*, *56*(3), 850–864.

Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, *57*, 301-316.

Falk, S., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80-86.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477-501.

Fine, J. M., Likens, A. D., Amazeen, E. L., & Amazeen, P. G. (2015). Emergent complexity matching in interpersonal coordination: Local dynamics and global variability. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 723. doi: 10.1037/xhp0000046

Flamson, T., Bryant, G. A., & Barrett, H. C. (2011). Prosody in spontaneous humor: Evidence for encryption. *Pragmatics & Cognition, 19*(2), 248-267.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Freigang, F., & Kopp, S. (2015). Analysing the modifying functions of gesture in multimodal utterances. In Proceedings of the 4th Conference on Gesture and Speech in Interaction (GESPIN).

Fujiwara, K., Kimura, M., & Daibo, I. (2020). Rhythmic features of movement synchrony for bonding individuals in dyadic interaction. *Journal of Nonverbal Behavior*, *44*(1), 173-193.

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.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147-157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145-171.

Fussell, S. R., Kraut, R. E., & Siegel, J. (2000, December). Coordination of communication: Effects of shared visual context on collaborative work. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 21-30).

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35-51.

Galati, A., & Brennan, S. E. (2014). Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451.

Gamer, M., Lemon, J., & Fellows, I. (2012). irr: Various coefficients of interrater reliability and agreement. R package version 0.84. https://CRAN.R-project.org/package=irr

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy?. *Trends in cognitive sciences*, *8*(1), 8-11.

Gibbs, R. W., & Van Orden, G. (2012). Pragmatic choice in conversation. *Topics in Cognitive Science, 4*(1), 7-20.

Gillespie, M., James, A. N., Federmeier, K. D., & Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, *132*(2), 174-180.

Goodwin, C. (2000). Action and Embodiment within Human Situated Interaction. *Journal of Pragmatics*, *32*, 1489–1522.

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual Review of Anthropology, 19*(1), 283-307.

Gordon, C. L., Cobb, P. R., & Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: An ALE meta-analysis of fMRI data. *PloS One*, *13*(11), e0207213. doi: 10.1371/journal.pone.0207213

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Second International Conference on Spoken Language Processing*.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L. P. (2006, August). Virtual rapport. In *International workshop on intelligent virtual agents* (pp. 14-27). Springer: Berlin, Heidelberg.

Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., & Badler, N. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent systems*, *17*(4), 54-63.

Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007, September). Creating rapport with virtual agents. In *International workshop on intelligent virtual agents* (pp. 125-138). Springer: Berlin, Heidelberg.

Habets, B., Kita, S., Shao, Z., Özyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845-1854. doi: 10.1162/jocn.2010.21462

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114-121.

Hausdorff, J. M., Purdon, P. L., Peng, C. K., Ladin, Z. V. I., Wei, J. Y., & Goldberger, A. L. (1996). Fractal dynamics of human gait: Stability of long-range correlations in stride interval fluctuations. *Journal of Applied Physiology*, *80*(5), 1448-1457.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89.

Hellermann, J. (2003). The interactive work of prosody in the IRF exchange: Teacher repetition in feedback moves. *Language in Society*, *32*(1), 79–104.

Henning, J. H. (1955). How to deliver a speech. *Communication Quarterly*, *3*(1), 3-21. doi: 10.1080/01463375509384871

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*(4), 575–591.

Hirschberg, J. (2000). A Corpus-Based Approach to the Study of Speaking Style. In *Prosody: Theory and Experiment* (pp. 335–350). Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9413-4_12

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th annual meeting on Association for Computational Linguistics* (pp. 136–144). Association for Computational Linguistics.

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79*, 1-17.

Hofmann, A. & Goebl, W. (2014). Production and perception of legato, portato, and staccato articulation in saxophone playing. *Frontiers in Psychology*, *5*, 690. doi: 10.3389/fpsyg.2014.00690

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, *19*(7), 1175-1192. doi: 10.1162/jocn.2007.19.7.1175

Holler, J. & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener?. *Gesture, 3*(2), 127-154. doi: 10.1075/gest.3.2.02hol

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639-652.

Holler, J., & Stevens, R. (2007). The effect of common ground on how speakers use gesture and speech to represent size information. *Journal of Language and Social Psychology*, *26*(1), 4-27.

Holler, J., Turner, K., & Varcianna, T. (2013). It's on the tip of my fingers: Co-speech gestures during lexical retrieval in different social contexts. *Language and Cognitive Processes*, *28*(10), 1509-1518.

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics, 43*(14), 3522-3536.

Hospelhorn, E., & Radinsky, J. (2017). Method for analyzing gestural communication in musical groups. *Discourse Processes*, *54*(7), 504-523. doi: 10.1080/0163853X.2015.1137183

Hostetter, A. B. & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture*, *7*(1), 73-95.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15*(3), 495-514.

Huang, L., Morency, L. P., & Gratch, J. (2011, September). Virtual Rapport 2.0. In *International workshop on intelligent virtual agents* (pp. 68-79). Springer: Berlin, Heidelberg.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping, 30*(3), 1028-1037.

Hutchins, E. (1983). Understanding Micronesian navigation. *Mental Models*, 191-225.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science*, *19*(3), 265-288.

Hutchins, E. & Palen, L. (1997). Constructing meaning from space, gesture, and speech. In L. B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, tools and reasoning* (pp. 23-40). Berlin Heidelberg: Springer.

Jensen, T. W. (2018). Humor as interactional affordances: An ecological perspective on humor in social interaction. *Psychology of Language and Communication*, 22, 238-259. doi: 10.2478/plc-2018-0010

Jermann, P., & Nüssli, M. A. (2012). Effects of sharing text selections on gaze cross-recurrence and interaction quality in a pair programming task. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 1125-1134).

Jurafsky, D. (2004). Pragmatics and computational linguistics. In L. Horn, & G. Ward (Eds), *Handbook of pragmatics* (pp. 578-604). Oxford: Blackwell.

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 638–646). Association for Computational Linguistics.

Kello, C. T., Brown, G. D., Ferrer-i-Cancho, R., Holden, J. G., Linkenkaer-Hansen, K., Rhodes, T., & Van Orden, G. C. (2010). Scaling laws in cognitive sciences. *Trends in Cognitive Sciences*, *14*(5), 223-232.

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), 20170231. doi: 10.1098/rsif.2017.0231

Kello, C. T., & Plaut, D. C. (2004). A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, *116*(4), 2354-2364. doi: 10.1121/1.1715112

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language, 40*(4), 577-592.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, *21*(2), 260-267. doi: 10.1177/0956797609357327

Kelso, J. S. (2009). Synergies: Atoms of brain and behavior. In D. Sternad (Ed.), *Progress in motor control* (pp. 83-91). US: Springer.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22-63.

Kendon, A. (2017). Pragmatic functions of gestures. *Gesture, 16*(2), 157-175.

Kim, T., McFee, E., Olguin, D. O., Waber, B., & Pentland, A. S. (2012). Sociometric badges: Using sensor technology to capture new forms of collaboration. *Journal of Organizational Behavior*, *33*(3), 412-427.

Kirsh, D. (2010). Thinking with external representations. *AI & Society*, *25*(4), 441-454.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language, 48*(1), 16-32.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, *22*(8), 1212-1236.

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*(38), 15443-15448.

Koppensteiner, M., & Grammer, K. (2010). Motion patterns in political speech and their influence on personality ratings. *Journal of Research in Personality*, *44*(3), 374–379.

Krauss, R., Chen, Y., & Gottesman, R. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and Gesture* (pp. 261-283). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620850.017

Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, *8*(1).

Kuhlen, A. K., Galati, A., & Brennan, S. E. (2012). Gesturing integrates top-down and bottom-up information: Joint effects of speakers' expectations and addressees' feedback. *Language and Cognition*, *4*(1), 17-41.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. doi:10.18637/jss.v082.i13

Latif, N., Barbosa, A. V., Vatiokiotis-Bateson, E., Castelhano, M. S., & Munhall, K. G. (2014). Movement coordination during conversation. *PLoS One*, *9*(8), e105036.

Levasseur, D. G. & Kanan Sawyer, J. (2006). Pedagogy meets PowerPoint: A research review of the effects of computer-generated slides in the classroom. *The Review of Communication*, *6*(1–2), 101–123.

Levelt, W. J., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, *24*(2), 133-164.

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In P. Cosi, R. De Mori, G. Di Fabbrizio, and R. Pieraccini (Eds.), *INTERSPEECH 2011 Twelfth Annual Conference of the International Speech Communication Association*. ISCA Archive: Florence, Italy.

Linell, P. (2004). *The written language bias in linguistics: Its nature, origins and transformations*. Routledge.

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, *9*, 2228.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science, 36*(8), 1404-1426.

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.

Marmelat, V. & Delignières, D. (2012). Strong anticipation: Complexity matching in interpersonal coordination. *Experimental Brain Research*, *222*(1-2), 137-148. doi: 10.1007/s00221-012-3202-9

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review,* *79*(6), 487–509. https://doi.org/10.1037/h0033467

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45-66.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348-356.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746. doi:10.1038/264746a0

McNeill, D. (1985). So you think gestures are nonverbal?. *Psychological Review, 92*(3), 350. doi: 10.1037/0033-295X.92.3.350

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

McNeill, D. (2000). *Language and gesture* (Vol. 2). Cambridge Univ Pr. Retrieved from http://books.google.com/books?hl=en&lr=&id=DRBcMQuSrf8C&oi=fnd&pg=PR9&dq=gesture+mcneill&ots=jBDQ3Cwqqt&sig=mGXmcSQeJbs2-goD-1SeSF0CpVU

McNeill, D. (2008). *Gesture and thought*. University of Chicago press.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, *4*(2), 119–141. https://doi.org/10.1075/gest.4.2.02mel

Mirman, D. (2014). *Growth curve analysis and visualization using R.* Boca Raton, FL: CRC Press.

Mirzaiyan, A., Parvaresh, V., Hashemian, M., & Saeedi, M. (2010). Convergence and divergence in telephone conversations: a case of Persian. *International Journal of Human and Social Sciences, 5*(3).

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics, 20,* 336–366. doi: 10.1111/josl.1_12177

Morsella, E. & Krauss, R. (2004). The role of gestures in spatial working memory and speech. *The American Journal of Psychology, 117*(3), 411-424. doi:10.2307/4149008

Moulton, S. T., Türkay, S., & Kosslyn, S. M. (2017). Does a presentation's medium affect its message? PowerPoint, Prezi, and oral presentations. *PLOS ONE*, *12*(7), e0178774. https://doi.org/10.1371/journal.pone.0178774

Murray-Smith, R., Ramsay, A., Garrod, S., Jackson, M., & Musizza, B. (2007, September). Gait alignment in mobile phone conversations. In *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services* (pp. 214-221).

Newtson, D. (1993). The dynamics of action and interaction. In *A Dynamic Systems Approach to Development: Applications*, eds L. B. Smith and E. Thelen (pp. 241-264). Cambridge, MA, US: The MIT Press).

Nomikou, I., Leonardi, G., Rohlfing, K. J., & Rączaszek-Leonardi, J. (2016). Constructing interaction: the development of gaze dynamics. *Infant and Child Development*, *25*(3), 277-295.

Novak, J. (2011). *Live poetry: An integrated approach to poetry in performance* (Vol. 153). Rodopi: Amsterdam.

Obermeier, C., Dolk, T., & Gunter, T. C. (2012). The benefit of gestures during communication: Evidence from hearing and hearing-impaired individuals. *Cortex*, *48*(7), 857-870.

Obermeier, C., Kelly, S. D., & Gunter, T. C. (2015). A speaker's gesture style can affect language comprehension: ERP evidence from gesture-speech integration. *Social Cognitive and Affective Neuroscience*, *10*(9), 1236-1243.

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, A. S., Chorna, L., & Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of language. *Proceedings of the National Academy of Sciences*, *110*(16), 6318-6323.

Onnis, L., & Spivey, M. J. (2012). Toward a new scientific visualization for the language sciences. *Information*, *3*(1), 124-150.

Özyürek, A. (2002). Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46*(4), 688-704. doi: 10.1006/jmla.2001.2826

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Phil. Trans. R. Soc. B, 369*(1651), 20130296.

Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures?: Insights from cross-linguistic variations and similarities. *Gesture, 5*(1), 219-240.

Palmer, C., Koopmans, E., Loehr, J. D., & Carter, C. (2009). Movement-related feedback and temporal accuracy in clarinet performance. *Music Perception: An Interdisciplinary Journal*, *26*(5), 439-449. doi: 10.1525/mp.2009.26.5.439

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

Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience, 8*, 57. doi: 10.3389/fnsys.2014.00057

Paxton, A., & Dale, R. (2013a). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology, 66*(11), 2092–2102. https://doi.org/10.1080/17470218.2013.853089

Paxton, A. & Dale, R. (2013b). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods*, *45*(2), 329-343.

Paxton, A. & Dale, R. (2017). Interpersonal movement synchrony responds to high-and low-level conversational constraints. *Frontiers in Psychology*, *8*, 1135.

Pentland, A. (2008). *Honest signals: how they shape our world*. MIT press.

Pérez, A., Carreiras, M., & Duñabeitia, J. A. (2017). Brain-to-brain entrainment: EEG interbrain synchronization while speaking and listening. *Scientific Reports*, *7*(1), 4190.

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology, 9*, 1109.

Phillips-Silver, J., & Trainor, L. J. (2005). Feeling the beat: Movement influences infant rhythm perception. *Science*, *308*(5727), 1430-1430. doi: 10.1126/science.1110922

Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition*, *105*(3), 533-546. doi: 10.1016/j.cognition.2006.11.006

Pickering, M. J., & Garrod, S. (2014). Self-, other-, and joint monitoring using forward models. *Frontiers in Human Neuroscience*, 8, 132.

Pouw, W., & Dixon, J. A. (2018). Quantifying gesture-speech synchrony: Exploratory data report and pre-registration. *Open Science Framework*. https://doi.org/None

Pouw, W., & Dixon, J. A. (2019). Entrainment and modulation of gesture–speech synchrony under delayed auditory feedback. *Cognitive Science*, *43*(3), e12721. doi: 10.1111/cogs.12721

Pouw, W., Harrison, S. J. & Dixon, J. A. (2019). Gesture-speech physics: The biomechanical basis of gesture-speech synchrony. *Journal of Experimental Psychology: General.* Preprint PDF *PsyArxiv* doi: 10.31234/osf.io/tgua4

Pouw, W., Jonge-Hoekstra, D., Harrison, S. J., Paxton, A., & Dixon, J. A. (2021). Gesture-speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, *1491*(1), 89-105.

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences*, *117*(21), 11364-11367.

Pouw, W., Trujillo, J., & Dixon, J. A. (2018). The quantification of gesture-speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*. Preprint PDF *PsyArxiv* doi: 10.31234/osf.io/jm3hk

Pozzer-Ardenghi, L. & Roth, W. M. (2004). Photographs in lectures: Gestures as meaning-making resources. *Linguistics and Education*, *15*(3), 275-293.

Pozzer-Ardenghi, L. & Roth, W. M. (2007). On performing concepts during science lectures. *Science Education*, *91*(1), 96-114.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rączaszek-Leonardi, J. (2014). Multiple systems and multiple time scales of language dynamics: Coping with complexity. *Cybernetics & Human Knowing*, *21*(1-2), 37-52.

Rączaszek-Leonardi, J., & Kelso, J. S. (2008). Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. *New Ideas in Psychology, 26*(2), 193-207.

Ramseyer, F., & Tschacher, W. (2014). Nonverbal synchrony of head-and body-movement in psychotherapy: Different signals have different associations with outcome. *Frontiers in Psychology, 5*, 979.

Ranganath, R., Jurafsky, D., & McFarland, D. A. (2013). Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language, 27*(1), 89-115.

Rauscher, F., Krauss, R., & Chen, Y. (1996). Gesture, Speech, and Lexical Access: The Role of Lexical Movements in Speech Production. *Psychological Science, 7*(4), 226-231.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060.

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). Complex dynamical systems in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 253-282). New York: Cambridge University Press.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science, 18*(5), 407-413.

Riley, M. A., Richardson, M., Shockley, K., & Ramenzoni, V. C. (2011). Interpersonal synergies. *Frontiers in Psychology*, *2*, 38.

Riordan, M. A., Kreuz, R. J., & Olney, A. M. (2014). Alignment is a function of conversational dynamics. *Journal of Language and Social Psychology, 33*(5), 465-481.

Ro, W. & Kwon, Y. (2009). 1/f noise analysis of songs in various genre of music. *Chaos, Solitons & Fractals*, *42*(4), 2305-2311. doi: 10.1016/j.chaos.2009.03.129

Rohrmeier, M., Zuidema, W., Wiggins, G. A., & Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664), 20140097.

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2016). Communicating epistemic stance: How speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Processes*, *53*(3), 135-174. doi: 10.1080/0163853X.2014.969137

Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. In *Ninth European Conference on Speech Communication and Technology*.

Rossano, F. (2013). Gaze in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 308-329). Malden, MA: Wiley-Blackwell.

Sawyer, R. K. (2005). Music and conversation. *Musical Communication*, *45*, 60.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169-182.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227-256.

Scherer, S., Layher, G., Kane, J., Neumann, H., & Campbell, N. (2012). An audiovisual political speech analysis incorporating eye-tracking and perception data. In *LREC* (pp. 1114–1120).

Schmidt, C. P., Andrews, M. L., & McCutcheon, J. W. (1998). An acoustical and perceptual analysis of the vocal behavior of classroom teachers. *Journal of Voice*, *12*(4), 434–443.

Schmidt, R. C., Carello, C., & Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 227.

Schmidt, R. C., Nie, L., Franco, A., & Richardson, M. J. (2014). Bodily synchronization underlying joke telling. *Frontiers in Human Neuroscience, 8*, 633.

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.

Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. *Frontiers in Human Neuroscience*, *10*, 435. doi: 10.3389/fnhum.2016.00435

Searle, J. R., & Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction*, *10*(4), 401-444.

Selting, M. (2010). Affectivity in conversational storytelling. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, *20*(2), 229–277.

Shim, H. S., Park, S., Chatterjee, M., Scherer, S., Sagae, K., & Morency, L.-P. (2015). Acoustic and para-verbal indicators of persuasiveness in social multimedia. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 2239–2243). IEEE.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305-319.

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.

Sidnell, J. (2006). Coordinating gesture, talk, and gaze in reenactments. *Research on Language and Social Interaction, 39*(4), 377-409.

Snedeker, J. & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103-130.

Spivey, M. J., & Richardson, D. C. (2008). Language embedded in the environment. *The Cambridge Handbook of Situated Cognition*. Cambridge University Press, Cambridge, UK, 383-400.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587-10592.

Stivers, T., & Sidnell, J. (2005). Introduction: multimodal interaction. *Semiotica, 2005*(156), 1-20.

Stolcke, A., Shriberg, E., Bates, R., Coccaro, N., Jurafsky, D., Martin, R., ... & Van Ess-Dykema, C. (1998, March). Dialog act modeling for conversational speech. In AAAI Spring Symposium on Applying Machine Learning to Discourse Processing (pp. 98-105).

Streeck, J., & Jordan, J. S. (2009). Communication as a dynamical self-sustaining system: the importance of time-scales and nested context. *Communication Theory*, *19*(4), 445-464.

Tanenhaus, M. K., & Brown-Schmidt, S. (2008). Language processing in the natural world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1105-1122.

Teixeira, E. C., Loureiro, M. A., & Yehia, H. C. (2018). Expressiveness in music from a multimodal perspective. *Music Perception: An Interdisciplinary Journal*, *36*(2), 201-216. doi: 10.1525/MP.2018.36.2.201

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, *12*(3), 409-441. doi: 10.3758/BF03193785

Tollefsen, D., & Dale, R. (2012). Naturalizing joint action: A process-based approach. *Philosophical Psychology*, *25*(3), 385-407.

Trueswell, J. C., & Tanenhaus, M. K. (Eds.). (2005). *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*. MIT Press.

Trujillo, J., Özyürek, A., Holler, J., & Drijvers, L. (2020). Evidence for a multimodal Lombard effect: Speakers modulate not only speech but also gesture to overcome noise. *PsyArxiv.*

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition, 180*, 38-51.

Tseng, C. Y., Pin, S. H., Lee, Y., Wang, H. M., & Chen, Y. C. (2005). Fluent speech prosody: Framework and modeling. *Speech Communication*, *46*(3), 284-309.

Tuller, B., & Turvey, M. (1982). The Bernstein perspective: ll. the concept of muscle linkage or coordinative structure. In J. A. S. Kelso (Ed.), *Human motor behavior: An introduction* (pp. 253-270). New York: Taylor & Francis.

Valbonesi, L., Ansari, R., McNeill, D., Quek, F., Duncan, S., McCullough, K. E., & Bryll, R. (2002, September). Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures. In *Signal Processing Conference, 2002 11th European*. IEEE.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331.

Van Orden, G. C., Kloos, H., & Wallot, S. (2011). Living in the pink: Intentionality, wellbeing, and complexity. In *Philosophy of complex systems* (pp. 629-672). North-Holland.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B, 369*(1651).

Voigt, R., Podesva, R. J., & Jurafsky, D. (2014). Speaker movement correlates with prosodic indicators of engagement. In *Speech Prosody* (Vol. 7).

Voss, R. F., & Clarke, J. (1978). "1/f noise" in music: Music from 1/f noise. *The Journal of the Acoustical Society of America*, *63*(1), 258-263. doi: 10.1121/1.381721

Vyas, P., & Sharma, S. (2014). A study on the efficacy of PowerPoint for writing instruction. *Int J Instr Technol Distance Learn*, *11*(8), 29–42.

Wallot, S., & Van Orden, G. (2011). Grounding language performance in the anticipatory dynamics of the body. *Ecological Psychology, 23*(3), 157-184.

Wallot, S., & Van Orden, G. (2012). Ultrafast cognition. *Journal of Consciousness Studies*, *19*(5-6), 141-160.

Walther, J. B. (1995). Relational aspects of computer-mediated communication: Experimental observations over time. *Organization Science*, *6*(2), 186-203.

Walther, J. B. (2012). Interaction through technological lenses: Computer-mediated communication and language. *Journal of Language and Social Psychology*, *31*(4), 397-414.

Walther, J. B., Loh, T., & Granka, L. (2005). Let Me Count the Ways: The Interchange of Verbal and Nonverbal Cues in Computer-Mediated and Face-to-Face Affinity. *Journal of Language and Social Psychology, 24*(1), 36–65.

Walton, A. E., Richardson, M. J., Langland-Hassan, P., & Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Frontiers in Psychology*, *6*, 313. doi: 10.3389/fpsyg.2015.00313

Wehling, E. (2018). Discourse management gestures. *Gesture*, *16*(2), 245–276. https://doi.org/10.1075/gest.16.2.04weh

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020.

Willems, R. M., Özyürek, A., & Hagoort, P. (2006). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex, 17*(10), 2322-2333.

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*(10), 1559-1579.

Wilson, M. & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957-968.

Wiltshire, T. J., Philipsen, J. S., Trasmundi, S. B., Jensen, T. W., & Steffensen, S. V. (2020). Interpersonal coordination dynamics in psychotherapy: a systematic review. *Cognitive Therapy and Research*, *44*(4), 752-773.

Yasinnik, Y., Renwick, M., & Shattuck-Hufnagel, S. (2004, June). The timing of speech-accompanying gestures with respect to prosody. In *Proceedings of the International Conference: From Sound to Sense*.

Yu, C. (2013). Two interactional functions of self-mockery in everyday English conversations: A multimodal analysis. *Journal of Pragmatics, 50*(1), 1-22.

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547. doi: 10.1038/nrn2152

Zhang, Z. (2016). Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, *140*(4), 2614-2635.