Modelling the Relationship Between Gesture Motion and Meaning

Carolyn Saund

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Neuroscience and Psychology
College of Medical, Veterinary, and Life Sciences
University of Glasgow

September 2022

Abstract

There are many ways to say "Hello," be it a wave, a nod, or a bow. We greet others not only with words, but also with our bodies. Embodied communication permeates our interactions. A fist bump, thumbs-up, or pat on the back can be even more meaningful than hearing "good job!" A friend crossing their arms with a scowl, turning away from you, or stiffening up can feel like a harsh rejection. Social communication is not exclusively linguistic, but is a multi-sensory affair. It's not that communication without these bodily cues is impossible, but it is impoverished. Embodiment is a fundamental human experience.

Expressing ourselves through our bodies provides a powerful channel through which we express a plethora of meta-social information. And integral to communication, expression, and social engagement is our utilization of conversational gesture. We use gestures to express extra-linguistic information, to emphasize our point, and to embody mental and linguistic metaphors that add depth and color to social interaction.

The gesture behaviour of virtual humans when compared to human-human conversation is limited, depending on the approach taken to automate performances of these characters. The generation of nonverbal behaviour for virtual humans can be approximately classified as either: 1) data-driven approaches that learn a mapping from aspects of the verbal channel, such as prosody, to gestures; or 2) rule bases approaches that are often tailored by designers for specific applications.

This thesis is an interdisciplinary exploration that bridges these two approaches, and brings data-driven analyses to observational gesture research. By marrying a rich history of gesture research in behavioral psychology with data-driven techniques, this body of work brings rigorous computational methods to gesture classification, analysis, and generation. It addresses how researchers can exploit computational methods to make virtual humans gesture with the same richness, complexity, and apparent effortlessness as you and I. Throughout this work the central focus is on metaphoric gestures. These gestures are capable of conveying rich, nuanced, multi-dimensional meaning, and raise several challenges in their generation, including establishing and interpreting a gesture's communicative meaning, and selecting a performance to convey it. As such, effectively utilizing these gestures remains an open challenge in virtual agent research. This thesis explores how metaphoric gestures are interpreted by an observer, how one can generate such rich gestures using a mapping between utterance meaning and gesture, as well as how one can use data driven techniques to explore the mapping between utterance and metaphoric gestures.

The thesis begins in Chapter 1 by outlining the interdisciplinary space of gesture research in psychology and generation in virtual agents. It then presents several studies that address presupposed assumptions raised about the need for rich, metaphoric gestures and the risk of false implicature when gestural meaning is ignored in gesture generation. In Chapter 2, two studies on metaphoric gestures that embody multiple metaphors argue three critical points that inform the rest of the thesis: that people form rich inferences from metaphoric gestures, these inferences are informed by cultural context and, more importantly, that any approach to analyzing the relation between utterance and metaphoric gesture needs to take into account that multiple metaphors may be conveyed by a single gesture. A third study presented in Chapter 3 highlights the risk of false implicature and discusses this in the context of current subjective evaluations of the qualitative influence of gesture on viewers.

Chapters 4 and 5 then present a data-driven analysis approach to recovering an interpretable explicit mapping from utterance to metaphor. The approach described in detail in Chapter 4 clusters gestural motion and relates those clusters to the semantic analysis of associated utterance. Then, Chapter 5 demonstrates how this approach can be used both as a framework for data-driven techniques in the study of gesture as well as form the basis of a gesture generation approach for virtual humans.

The framework used in the last two chapters ties together the main themes of this thesis: how we can use observational behavioral gesture research to inform data-driven analysis methods, how embodied metaphor relates to fine-grained gestural motion, and how to exploit this relationship to generate rich, communicatively nuanced gestures on virtual agents. While gestures show huge variation, the goal of this thesis is to start to characterize and codify that variation using modern data-driven techniques.

The final chapter of this thesis reflects on the many challenges and obstacles the field of gesture generation continues to face. The potential for applications of Virtual Agents to have broad impacts on our daily lives increases with the growing pervasiveness of digital interfaces, technical breakthroughs, and collaborative interdisciplinary research efforts. It concludes with an optimistic vision of applications for virtual agents with deep models of non-verbal social behaviour and their potential to encourage multi-disciplinary collaboration.

Contents

List of Tables

List of Figures

Acknowledgements

Life is truly a team effort. Without any of my role models below my life would be radically different. Although I cannot name the many outstanding mentors, allies, peers, colleagues, and friends I have had the pleasure of knowing and being inspired by, I deeply appreciate the complex interconnectedness that facilitates happiness, learning, and productivity. Any of my success is the success of my community, my achievements are the achievements of all who have supported me along the way.

Some in particular require special acknowledgement:

Agree, Connor, Rhyan, MJ, Josh, Joe, Ross, Graham, and Paul, who spent countless hours making me laugh and teaching me the fundamentals of both computer science and global citizenship.

Kat, Chris, Kevin, Daniel, Nick, and Sam, my erudite Cantabrigians with whom I can always wax philosophical. Our late-night salons have given me so much respect for alternative perspectives.

James, Christine, Kirsty, Laura, Kathrin, Carolyn, Mike, and Jack, who caught me when I landed in Glasgow, who always say "YES!" to life, and upon whose unwavering support I can always depend.

My utterly delightful lab group+, Mary, Nut, Tobi, and Chris, whose vast interests never fail to engage us in endless thought-provoking, idea-sparking conversation.

My goofy friend Alexander Lenail, who is relentlessly ambitious and insatiably curious.

My brilliant, dedicated friend Dr. Claire Duvallet, who embodies strength and justice.

My mother Laura Larkin, who personifies resolution, determination, and philanthropy, who is a true outstanding renaissance woman, and who I want to be exactly like when I grow up.

My brother and sister-in-law Dr. Brad Saund and Dr. Katie Saund, who are infinitely patient, compassionate, generous, empathetic, loving, and inspiring.

My father Dr. Eric Saund, from whom I get my passionate character and my deep curiosity about the mind, and who knew from the beginning that I am a scientist.

My loving, supportive, creative, patient partner, Alistair Blincow, who invariably reminds me that I can do anything, and who is always there for me when I inevitably try to do everything.

And finally, my advisor and mentor, the artist, the magician, the man, the myth, the legend, Professor Stacy Marsella. He has taught me uncountably many lessons about research, but moreover about how we each choose to live our own lives. I will be forever grateful for his guidance and, dare I say it, friendship (although he may be horrified to be associated with the likes of me).

Declaration

With the exception of Chapter 1, which contains introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated. A number of chapters have been conducted in conjunction with co-authors, which is described in the Contributions. In all cases, the author has held a leading role in the project and made the primary contribution to the work as presented here. This work has not been submitted in any previous application for a higher degree or professional qualification.

Signature:

Contributions

The great majority of original work presented here was conducted and written exclusively by myself. However, as scientific progress is an inherently collaborative effort, I here describe in detail the pieces of this thesis that are not explicitly my original work.

While the vast majority of previously published work was originally written by myself, I have re-written all sections of published work that were originally written by any of the co-authors listed below. In cases where peer-reviewed publications are included in this thesis, I have not altered the text of the original publication, except in cases where a co-author wrote the original text. Specifically, **I make the active choice to retain the pronoun "we," in reference to the original group of co-authors of the paper**. In all cases I am the lead author of the publication.

Marion Roth, Mathieu Chollet, and Stacy Marsella are co-authors on the study in Section 2.1. Marion Roth in particular helped to find and annotate two of the gestures used in the final experiment, and was instrumental in the decision of how to manipulate each gesture, and in the experimental design and interpretation of results. She also originally wrote some of the text of Section 2.1.2. Matheiu Chollet played a large role in ideating and helping to design the experiment, as well as editing our submission and final copy.

Discussions with Maki Rooskby were extremely helpful in the initial formulation of the experiment and gestures selected in Section 2.2. I deeply appreciate her cheerful and insightful advice on cultural differences and contextually aware translations.

Andrei Bîrlădeanu is a co-author of the work in Section 4.1, and wrote parts of Section 4.1.1.

In Section 4.2, Anna Weinstein wrote the initial code for processing and converting BVH files to more easily manipulated data frames and edited the final submission. Haley Matuszak provided the deep analysis of the example in Section 4.2.4 and, along with Professor Stacy Marsella, originally wrote the accompanying text found in Section 4.2.4. **They wrote the code for the semantic parser described in the same section, and which is used throughout experiments of the architecture.**

As the Principle Investigator on these projects, as well as my advisor and supervisor, Prof. Stacy Marsella played a guiding role throughout each section. He helped iterate through potential experimental designs, and pointed me towards resources to implement and recruit for each of the subjective studies presented here. He helped interpret results, organize, and outline the structure of each chapter, bring submissions to completion, and compile and edit the background presented in Chapter 1.

Publications

Parts of this thesis have been published as peer-reviewed conference proceedings and book chapters, and developed from work presented at previous venues.

Chapter 1

Saund, C. & Marsella, S. (2021). The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. September 2021 Pages 213–258. https://doi.org/10.1145/3477322.3477330

Chapter 2.

Saund, C., Roth, M., Chollet, M., & Marsella, S. (2019, September). Multiple metaphors in metaphoric gesturing. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019) (pp. 524-530). IEEE.

Saund, C., & Marsella, S. (2021, June). Interpretations of virtual agent performances of metaphoric gestures differ across cultures [conference presentation]. 5th Virtual Social Interaction Conference, Glasgow, UK.

Chapter 3.

Saund, C., & Marsella, S. (2021, December). The Importance of Qualitative Elements in Subjective Evaluation of Semantic Gestures. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1-8). IEEE.

Chapter 4.

Saund, C., Birladeanu A., and Marsella S. (2021, May). "CMCF: An architecture for realtime gesture generation by clustering gestures by motion and communicative function." In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (pp. 1136-1144) (AAMAS 2021)

Saund, C., Matuszak H., Weinstein A. & Marsella, S. (2022, December). Motion and Meaning: Data-Driven Analyses of The Relationship Between Gesture and Communicative Semantics. *accepted for publication in Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand. https://doi.org/10.1145/3527188.3561941*.

Chapter 1

The Importance of Gesture in Social Interaction

Gestures accompany our speech in ways that punctuate, augment, substitute, and sometimes even contradict verbal information. Co-speech gestures draw listeners' attention to specific phrases, indicate the speaker's feelings towards a subject, or even convey "off-the-record" information that is explicitly excluded from linguistic channels. The study of co-speech gesture stretches at least as far back as the work of Quintilian in 50 A.D., and draws from the disciplines of cognitive science, performance arts, politics, and, more recently, computer science and robotics. Gesture is a critical tool to enrich face-to-face communication, of which social artificial agents have yet to take full advantage.

This chapter demonstrates the far-reaching groundwork of gesture research that has been laid by behavioral observation and Psychological and Cognitive Sciences. Throughout, I discuss the importance of gesture in speech and conversation, historical axes for gesture classification, how it is that gestures actually come to carry meaning, and the cultural and contextual relevance in gesture production. I then connect this rich history with modern techniques and issues faced by the contemporary challenge of generating social gestures on virtual agents. I review several common approaches, modern implementations, and evaluation techniques, and discuss their merits and drawbacks.

The many lenses through which I approach gesture in this chapter is reflective of its complex, nuanced role in social communication. In this chapter, I show why it is important to understand the use of gesture in humans, and how researchers have approached this complexity. This interdisciplinary landscape provides the backdrop of the current work, which I outline at the end of this chapter.

Much of this chapter is unchanged from a Chapter published in The Handbook on Socially Interactive Agents, and can be cited as: **Saund, C.** & Marsella, S. (2021). The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. September 2021 Pages 213–258. https://doi.org/10.1145/3477322.3477330

1.1 What are gestures?

Gestures as I discuss them in this thesis are the spontaneous movements that accompany speech. Generally these are limited to hand and arm movements (McNeill, 1992) but can extend to the head, feet, or other body parts (Kendon, 2000). Users of sign language even use facial expressions in communicatively similar ways to how speakers use gestures in conversation. Our focus here, however is on hand and arm movements.

Specifically, this thesis focuses on gestures in conversation that usually accompany utterances, commonly referred to as co-speech gestures. This includes gestures that occur during speech in conversational or performative settings, such as interviews and monologues, with or without audiences. These can occur with or without conversational partners as well. Gestures serve a remarkably wide variety of communicative functions in conversation, including conveying information to observers as well as aiding in speech production and fluency for the speaker.

Importantly, the classifications provided here are by no means exhaustive. In this section, in addition to introducing one prevailing taxonomy (1.1.1), I discuss weaknesses and alternative proposals to classifying gestures using these dimensions (1.1.2), as well as many other factors which determine how researchers group gestures, both physically and functionally.

1.1.1 Classification dimensions

A common method of classifying co-speech gestures is by the five types or dimensions shown in Table 1.1. These correspond not only to differences in the motions used to realize the gesture, but more meaningfully to differences in the conversational contexts, their roles in speech production and communicative intentions of the speaker.

Gesture type	Co-speech necessary?	Viewer necessary?
Emblem	No	Sometimes
Beat	Yes	No
Iconic	Sometimes	No
Deictic	Sometimes	Sometimes
Metaphoric	Yes	No

Table 1.1: Table of gesture classical types and co-speech properties. In this case, "viewer necessary" refers to the necessity of the gesture to be seen for it to carry meaning. For example, if an emblematic gesture is not accompanied by speech, it must be seen by the viewer to carry meaning. Similarly, if a deictic gesture is not accompanied by speech, the viewer must see it in order for it to be understood. In both cases, if the gesture is not accompanied by co-speech, it must be seen by a viewer to have the intended communicative impact. Note, these gestures can still occur spontaneously even without a viewer present (i.e. over the phone).

Emblems are gestures which may essentially be thought of as replacements for spoken language. A prominent example is the "thumbs up" gesture which is common in several cultures, but often with strikingly different meaning. For example, if somebody asks a question, a "thumbs-up" response

in North America unambiguously means "yes," with or without verbal affirmation. Emblems carry equivalent meaning to their linguistic counterpart. Importantly, the interpretation of these gestures are culturally and linguistically dependent. For example, the "OK" symbol in western cultures, featuring a thumb and index finger pinch and three outstretched fingers, is a rude insult in Morocco.

Beat gestures, contrarily, are gestures that do not carry semantic content in their movements, but instead "reveal the speaker's conception of the narrative's discourse as a whole" (McNeill, 1992) by emphasizing specific words with small motions, often coinciding with the prosody of the spoken utterance. The movement of a beat gesture is short and quick, and often takes place only in the periphery of where the speaker uses other gestures (McNeill, 1992), and take generally similar form regardless of content of the co-utterance (Levy and McNeill, 1992). Beats may also aid in speech fluency by coinciding rhythmically to a spoken co-utterance, providing prosodic cues to word recall and comprehension (Hadar, 1989; Leonard and Cummins, 2011).

Iconic gestures are literal representations of real, physical counterparts. For example, if someone utters "we need a knife to cut the cake," they may produce a gesture with one flat palm held horizontally, and the other held vertically in a perpendicular "slicing" motion. In this instance, the hands are literally acting out the motion of a knife cutting something, with the hands embodying literal physical objects in the world. Similarly, an iconic gesture may be a mime of a literal motion. For example, if someone tells a story in which they were "running down the street," they may hold their arms to their sides and swing them up and down to emphasize, exaggerate, or depict their speed as though they were actually running.

Deictic gestures are pointing gestures which direct attention towards a referent in the environment. If you have an array of items on a table and tell someone to "pick up that one," the statement makes no sense without a verbal or gestural counterpart to identify the referent. Similarly, if somebody asks "which way did they go?" a person may simply point in lieu of providing a verbal response.

Metaphoric gestures "present an image of an abstract concept" (McNeill, 1992). For example, one may gesture in a bowl or container shape when describing "all of their ideas." Although the abstract notion of an "idea" can never be physically realized, the metaphoric gesture situates "ideas" in a metaphorical container that can be reliably referenced throughout the conversation by the speaker and viewers. Metaphoric gestures and the semantic concepts they communicate both individually and in conjunction with spoken language dominate the focus of my research presented in this thesis. Yet, it is essential to note that a single gesture plays many communicative roles simultaneously.

1.1.2 Multiple classifications

As McNeill (2006) has argued, these classifications are not strict types but rather dimensions which are overlapping and open to interpretation when considering the use of gestures in interactions. This refers to the notion that a particular gesture, within one particular context, may be interpreted to have different elements of the axes described above.

The same physical motion of a gesture may result in different interpretations depending on co-

(a) The beginning of the movement as she says "Anything at all."

(b) The second part of this gesture, creating the space where "anything" may metaphorically be.

Figure 1.1: The motion of the metaphoric gesture accompanying the phrase "Anything at all." This example originally appears in Lhommet and Marsella (2014a).

speech context. Consider the "slicing" motion described above. When applied to physical objects ("a knife to cut the cake"), this would be characterized as an iconic gesture. However, consider the same gesture if it accompanies the phrase "we need coordination to cut to the heart of the issue." In this instance, the *cutting* is metaphorical, as "issues" are not physical beings with literal "hearts." Similarly, "coordination" is not a physical object like a knife that can cut. However, the metaphor of "cutting to the heart of an issue" is grounded in physical space insofar as *cut* is a verb which describes a physical action. In the metaphoric condition, "coordination" may be represented metaphorically as a knife by the fingers falling into stiff, parallel lines. In this case the fingers may further be thought of as representing people falling into line. This motion thus illustrates two distinct utterances in which the same gesture occurs, one where the gesture is referring to actually cutting a physical object and one where the gesture is used metaphorically.

The use of a metaphor in speech is not necessary for the metaphor to be conveyed in the accompanying gestures. Figure 1.1 illustrates a metaphoric gesture accompanying the dialog "we can talk about anything at all". There is no metaphor used in the dialog while the gesture is based in metaphors whereby abstract things, such as topics of conversation, can be represented as physical objects and a set of these objects can be held in a physical container that is being depicted by the gesture. Despite this degree of independence between a metaphor's use in spoken language and accompanying gestures, the catalogue of metaphors used in speech provide a useful resource for researchers. Grady (1997) provides many such metaphors, for which gesture researchers commonly observe gestural counterparts [1]. These include *Similarity is Proximity* (e.g. "these fabrics aren't quite the same but they're *close*"), *Change is Motion* (e.g. "things have *shifted* since you were last here."), and *Moments in Time are Objects on a Path* (e.g. "Summer always *passes* too quickly"). These and many other metaphors often coincide with physical representations of these metaphoric actions (Lakoff and Johnson, 2008) represented gesturally. I frequently reference these metaphors throughout this thesis with gradually

[1] Grady does not propose or consider a framework for gesture analysis in this work cited. Instead, this work considers in depth the many ways in which metaphors permeate our speech, but does not explicitly discuss how we may use bodies to act out these metaphors as we say them.

deepening analysis.

The above are examples of how gestures may be used to emphasize or induce metaphors. Conversely, consider the straightforward presentation of two options "this or that," with the hands held flat, palm-up in front of the speaker. The speaker may say "this option," and beat with one hand, and then repeat the phrase "or this option," but move the other hand, clearly indicating that they are providing context for the different options. The indication is made by a beat motion, but also is a clarification of "which option," giving it attributes of a deictic gesture, referred to as an abstract deictic (McNeill, Cassell, and Levy, 1993). Additionally, the laying out of two different ideas in space is metaphoric as it relies on the metaphors of abstract concepts being physical objects and dissimilar concepts are far apart (Grady's *Categories/Sets are Bounded Spatial Regions*), thus incorporating yet another element of the dimensions described above into a single gestural motion.

Alternative classification schemes

In some modern works, gestures are often given multiple classifications, or the classification of gestures is skipped altogether, and gestures are judged solely by their communicative role or perceived intention. For example, Murphy (2003) proposes analysing gestures not by abstract representation, but instead by the production of those representations themselves. That is, gestures can be analysed exclusively by their body movements as opposed to attempting to interpret what those movements represent. He argues movement-based analysis is less prone to researcher bias and less likely to leave out body movements which do not fall neatly into the dimensions described above.

This is contrary to the idea proposed by Novack and Goldin-Meadow (2017). They suggest that iconic and deictic gestures are not simulations of actions they intend to portray, but instead consciously representational of abstract versions of those actions. This allows researchers to organize gestures according to their functional role in conversation. By focusing on gesture's function as opposed to its specific form, researchers can begin to focus on why a particular gesture occurs rather than how the intention maps to movement.

Although these alternative schemes are thought-provoking, their use is necessarily limited to specific applications. Gesture is a complex *physical* phenomenon generated from deep *cognitive* processes with acute *behavioral* consequences. It is only by considering the full picture of gesture production, from intention, to physical action, to function, that we can begin to create socially compelling gesture in artificial agents.

1.1.3 Timing

Another key axes of gesture which is tangential to the schemes described above is how they vary by the timing of their performance with a co-occurring utterance, ranging from nearly coinciding temporally with speech, to gestural performances many seconds in advance of or post-utterance (Calbris, 1995; Gibbs Jr, 2008; Nobe, 2000). Perception of appropriateness for different gestures with respect to

co-speech timing is not fixed (Leonard and Cummins, 2011) – a feature of conversation that adds complexity in generating gesture in VAs which I discuss in Section 1.6.1.

Growth Point Theory

The window of time for gestures to be relevant to corresponding speech is similarly fluid, and depends on context (Leonard and Cummins, 2011). Often, gestures anticipate the speech to which they correspond (McNeill, 1985; Nobe, 2000), indicating that cognitively, the meaning we attempt to convey is formulated and performed by the body before we are able to form (or at least utter) words for intended communication (Kendon, 2000). This similarly implies that the cognitive processes between communication intention and speech formulation are the same processes which initiate gesture production (Kendon, 2000). In non-verbal communication research this phenomenon of language and gesture generated from a common cognitive starting point and used in conjunction to convey a thought, with each channel able to convey related though non-redundant information is referred to as the *Growth Point Theory* (McNeill, 1985).

1.1.4 Gestural Phases and Units

There is a complex feature structure at the level of individual gestures. Many classifications include phases of gestural motion including the rest, preparation, stroke, holding, and relax phases, as well as the forms of motion, their locations, and changing hand shapes. However, people often gesture in an overall fluid performance involving gesture sequences (referred to by Kendon (2004) as *gesture units*) in such a way that not all phases may be present in every individual gesture. In sequences, co-articulations between gestures may eliminate the rest or relaxation phase of a gesture (McNeill, 1992).

Ideational Units

One way to refer to a sequence of related ideas that can span multiple gestures that is significant in the field of gesture research is as an ideational unit. This refers to a sequence of gestures that conveys multiple complex ideas over the course of their performance, during which time individual changes in the form of the gesture (i.e. the hand shape) correspond to a single idea to be conveyed. In other words, the changes in a gesture within an ideational unit convey meaning. Calbris (1990) argues that ideational units structure discourse and the kinesthetic segmentation of gestures. They impose requirements on gestural features both within and across ideational units in an overall performance.

Within an ideational unit some features such as hand shape, movement trajectory, or location in space, may be consistent across gestures while other features may, at times, serve key roles in distinguishing individual gestures from one another. This happens both physically and at the level of their meaning. For example, the hands may go into a rest position between gestures to indicate the end of an idea, a change of hand shape can serve to indicate the start of a new idea in the discourse (Calbris,

2011) or one gesture's location may serve to refer to a preceding gesture in an overall gestural scene where, for example, locations in gestural space take on specific meanings that may be referred to by subsequent gestures. The notion of continuity between gestures within an ideational unit underpins many aspects of the design of the frameworks I present in Chapter 4, and which I discuss in detail in Section 4.1.1.

1.2 Cultural Relevance

Another critical aspect to bear in mind when discussing gestures, especially in the context of artificial agents, is that nearly every aspect of gesturing is culturally dependent (Efron, 1941). Hand shapes (Calbris, 2011), gesture size and frequency (Kita, 2009), emblematic meaning (Calbris, 1990), and timing (Kita, 2009; Talmy, 1985) are a few examples of components of gesture which rely heavily on the native and contextual culture of the speaker. Some cultures use hardly any beat gestures, whereas some use them to punctuate almost every sentence (Levinson, 1996). As previously mentioned, emblems which are positive signals in one culture may be rude insults in another (Calbris, 2011).

Furthermore, different cultures' concepts of physical space and time inform their gestures as well (DiMaggio, 1997). In North American cultures, when talking about time individuals often gesture along a plane running horizontal to the speaker, with the left in the past and the right in the future. However, in French culture, time is often gestured as a plane running parallel to the speaker, as if the speaker is walking along the line of time with the future positioned in front and the past behind the back of the head (Calbris, 2011). But, in other cultures, the future may be referenced behind the speaker, with the past in front of the speakers eyes (Núñez and Sweetser, 2006). Contrast this yet again to Chinese culture, in which the vertical axis commonly applies in conceptualizing time where earlier times are viewed as "up" and later times as "down" (Radden, 2003). These different gestures show not only that cultural sensitivity must be taken into account for artificial agents when interpreting and performing gestures, but that the underlying conceptual representation of time may differ between cultures as well (Cienki and Koenig, 1998, referencing and extending the Sapir-Whorf hypothesis (Kay and Kempton, 1984)).

1.3 Gesture's role in conversation

The influence of gesture permeates social interaction. While we predominantly discuss gesture's role in human-human interaction, it is crucial to note that virtual agents elicit responses consistent to humans in many social contexts (Takeuchi and Naito, 1995; Poggi and Vincze, 2008; McCall, Bunyan, Bailenson, Blascovich, and Beall, 2009; Krämer, Kopp, Becker-Asano, and Sommer, 2013).

1.3.1 Dialog regulation

Gestures can help regulate conversation, for example by signaling the desire to hold onto, acquire, or hand over the dialog turn (Bavelas, 1994). Bergmann, Rieser, and Kopp (2011) explore a non-exhaustive list of the multitudinous ways gesture regulates dialog, which can be broadly broken into content-specific and content-agnostic behaviors. Content-specific gestures relate to the specific content being discussed, and include clarification requests, establishing confidence levels in mastery of the content of conversation, assessments of relevance, and indications and connections of topical information within the conversation. Content-agnostic gestures, however, have to do with the social rules of the conversation. These may include next-speaker selection, or handling of anti-social or non-canonical discourse behavior, such as interrupting.

1.3.2 Observer's internal beliefs

Gestures that accompany face-to-face spoken interaction convey a wide variety of information and stand in different relations to the verbal content that accompanies them. For the observer, gestures serve a wide variety of communication functions, such as commenting, requesting, protesting, directing attention, showing, and rejecting (Jokinen, Navarretta, and Paggio, 2008). In realizing these communicative functions, a gesture can provide information that embellishes, substitutes for, contradicts or is even independent of the information provided verbally (e.g., Ekman and Friesen, 1969b; Kendon, 2000).

As discussed above, gestures of course are physical actions, but these actions can convey both physical and abstract concepts. A sideways flip of the hand suggests discarding an object but can also be used to represent the rejection of an idea (Calbris, 2011). Similarly, gestures also serve a variety of rhetorical functions. Comparison and contrasts between abstract ideas can be emphasized by abstract deictic (pointing) gestures that point at the opposing ideas as if they each had a distinct physical locus in space (McNeill, 1992) (or beating in conjunction with "one option" or "another" such as described in Section 1.1.1). Similarly, beat gestures comprised of downward strokes in conjunction with speech are often used to emphasize the significance of a word or phrase in the speech or enumerate points (Crowder, 1996; Holle, Obermeier, Schmidt-Kassow, Friederici, Ward, and Gunter, 2012). This prosidic relationship between speech and motion is discussed in the context of gesture generation in Section 1.6.

Gestures are also used to reinforce and clarify their co-speech utterances. Jamalian and Tversky (2012a) show that different gestures in co-ordination with the same temporally ambiguous utterance ("the meeting was moved forward two days") successfully disambiguate temporal uncertainty. Similarly, gestures are able to allow observers to interpret statements as questions using the same audio (Kelly, Barr, Church, and Lynch, 1999), and to disambiguate linguistic homonyms (Holler and Beattie, 2003). It is precisely because gestures are used to clarify speech so often that some researchers suggest that gesture is the first tool humans use to disambiguate basic ideas and requests (Özçalışkan

and Goldin-Meadow, 2005). Further evidence suggests increased gesturing in this manner can lead to positive learning outcomes in teaching scenarios (Goldin-Meadow and Alibali, 2013).

Yet the impact of gesture is not always so explicit. For example, gestures are known to influence thought in the viewer. In the same publication, Jamalian and Tversky (2012a) showed that different types of metaphoric gestures change the way that individuals qualitatively describe certain systems and processes. Similarly, gestures have also been shown to influence memory recall in cases of eye-witness testimony (Gurney, Pine, and Wiseman, 2013), opening up discussion of gestures providing leading answers in a similar off-the-record manner.

1.3.3 Revealing the Speaker's mental states and traits

Gesture plays a critical role in human interaction, where it is not simply an addition to speech. Rather, it is an independent expression of thought that reveals underlying beliefs, intentions and processes of the speaker (Cienki and Koenig, 1998).

Gestures can present information about the speaker's state and views towards the subject of conversation. Pollick, Paterson, Bruderlin, and Sanford (2001) shows that viewers are able to read affect from arm motions alone, potentially giving the viewer valuable interpretable information about the gesturer's internal mental state. Seeing gestures used appropriately also bolster's viewers' impression of the speaker; Speakers who gesture in conversation are perceived as more composed, effective, persuasive, and competent than those who do not (Maricchiolo, Gnisci, Bonaiuto, and Ficca, 2009b). Furthermore, some work suggests that viewers tend to prefer virtual agents whose gestures mimic their own (Luo, Ng-Thow-Hing, and Neff, 2013).

A wide range of mental states and character traits can be conveyed gesturally. Placing hands on hips can display dominance or displeasure (Maestripieri, 2005), gestures performed with rapid acceleration can convey arousal or displeasure, a gesture with palm facing outward as if suggesting stop can convey displeasure at what a conversational partner is saying or doing (Schwartz, Tesser, and Powell, 1982).

Self-touching gestures or self-adaptors (Ekman and Friesen, 1969b), such as rubbing a forearm, are also believed to convey information about a person's mental state while also providing self-comfort. In particular, these behaviors can reveal negatively valenced emotional states such as anxiety, fear or guilt (Ekman and Friesen, 1969a). In the context of virtual agents, self-touching gestures have been shown to increase perceived warmth along with nervousness of the agent (Krahmer and Swerts, 2007).

Gestures may further be used to implicitly convey off-the-record information (Wolff, 2015). For example, a speaker may describe two people "getting together" with a co-speech gesture of either gently intertwining hands, or two fists clashing against one another. While the former may suggest harmony between individuals, forcing hands together at high velocity multiple times implies conflict and aggression (Morris, 2015). However, the speaker may specifically choose to convey this information outside of the speech channel. In doing so, the speaker both relays information in a fashion that is off-the-record, but still provides context of that information for the viewers. This is a crucial mech-

anism to convey sensitive information which could be exploited by virtual agents in high-intensity contexts, and also demonstrates the necessity of recognizing potentially severe consequences for false implicature.

1.3.4 Speaker impact

While gesture is an invaluable tool for face-to-face communication, it also acts as an aid for the speaker. Gestures occur regardless of whether a listener can actively view them. Individuals gesture at near the same rate when speaking to someone on the phone or in person (Iverson and Goldin-Meadow, 1998). Similarly, individuals gesture when they know that the viewer is blind (Iverson and Goldin-Meadow, 1997; Iverson and Goldin-Meadow, 1998). Fascinatingly, even congenitally blind individuals gesture at both sighted and other blind individuals (Iverson and Goldin-Meadow, 2001). This suggests that gesture plays an important role not only in social communication, but to aid in the speaker's own process of conveying information. One hypothesis for this is that using gesture helps lighten the cognitive load on the speaker (Goldin-Meadow, Nusbaum, Kelly, and Wagner, 2001).

While it is impossible to know the full extent of interaction between gesture and speech without understanding the underlying mechanism of going from thought to communication, we can observe ways in which communication is explicitly aided by gesture, or rather, hindered without gesture. Speakers speak less fluently when they lose the ability to gesture (Lickiss and Wellens, 1978). They also have more trouble recalling words when their hands are bound and they are unable to gesticulate during speech (Rauscher, Krauss, and Chen, 1996). This phenomenon points to deep relationships between physical embodiment, motion, and cognition, as discussed in the next section.

1.4 How gestures carry meaning

Although gestures perform a variety of functions in face-to-face interaction, often simultaneously, there is a limit to the complexity of information they can reliably convey. In this section, I discuss several traits and forms of gesture which have been shown to carry meaning to viewers. These traits can be combined freely, leading to the enormous physical variety we see in human gestures. As previously stated, although many physical cues can be grouped into gestures the focus here is explicitly on movements of the hands and arms.

There are many individual components of a gesture which may be responsible for viewer interpretation, and the information and capacity of each component varies by individual, and by culture. Broadly, when discussing co-speech gestures, we refer to the shape and trajectory of the hands, and all of the parameters which guide those components. Non-exhaustively, this includes velocity and amplitude of arm motions, orientation of the speaker towards the subject, the direction and symmetry of the hands, the paths of the wrists through space, and the timing of hand shape changes relative to conversational context.

These components and more are discussed at length by Calbris (2011), in which she discusses how parameters of these components (such as the plane of trajectory of the hands, or orientation of the hand relative to the arm) may augment or vary the communicative function of a gesture. Specifically, she uses gestural components specified in Calbris, Montredon, and Zaü (1986): Movement, localization, body part, orientation, and configuration in conjunction with hand shapes (such as those shown in Figure 1.2). Together, these components can be used as a framework to describe and analyze the shape and communicative function of conversational gestures. It is not only the components themselves, but moreover the dynamics (e.g. amplitude, speed and fluidity of movement) of these components that are integral in conveying these functions (Castellano, Villalba, and Camurri, 2007). In the same work, Calbris also explores how varying parameters of a gesture may result in multiple gestural representations of a single idea (the subject of Chapter 2), and how, because of the parameter space of gestures, one idea may be presented by many different conceivable gestures. The phenomenon one gesture conveying multiple potential concepts, *and* the ability of one concept to be feasibly conveyed by multiple gestures is often referred to as the many-to-many problem of gesture generation.

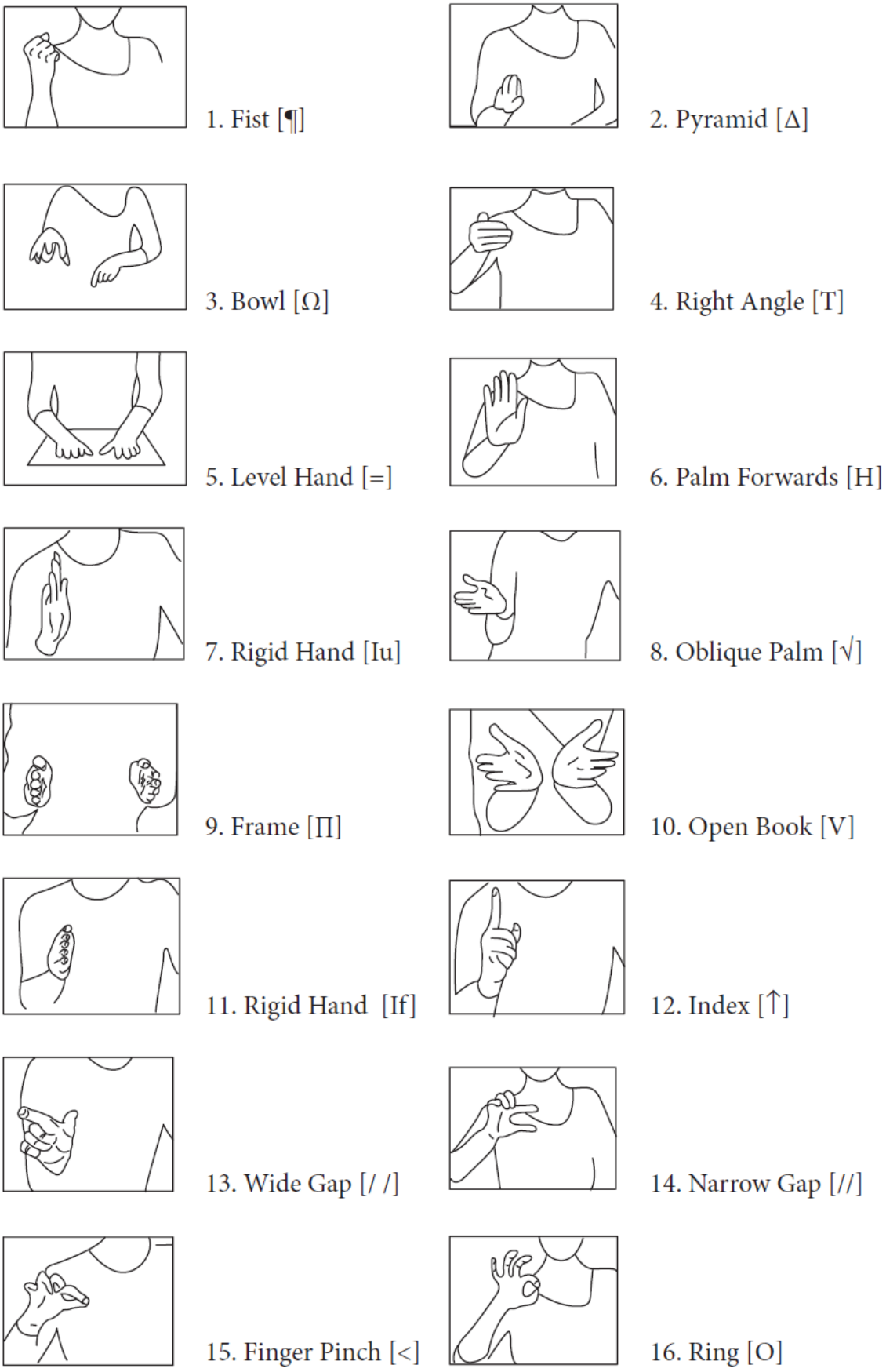


Figure 1.2: A selection of Calbris' handshapes provided in Calbris (2003).

1.5 Summary of Behavioral Gesture Research

Thus far this chapter has reviewed a history of gesture classification and observational research methods in psychology. It discussed the roles gesture plays in conversation for both the speaker and viewer, and how the various forms of gesture relate to their functional roles in communication.

This chapter also alluded to several elements of gesture that contribute to major challenges to gesturing in virtual agents. Growth Point Theory (Section 1.1.3) implies a human-like model of gestures requires an agent to use a deep model of communication and intentionality that generates language and gesture through separate but interwoven processes. The notion of Ideational Units (Section 1.1.4) suggests that agents' gestures must be fluid from one to the next, and any changes in gesture carry semantic or rhetorical significance. Moreover, the wide variety of impacts that gesture may have throughout conversation (Section 1.3) must be considered within a specific cultural context (Section 1.2). This presents considerable challenges to evaluating the efficacy of gesture generation models in virtual humans.

Having presented the history of gesture literature in behavioral psychology and demonstrated about gesture's importance in conversation, this chapter now shifts focus and examines current works implementing realistic, communicative, socially meaningful gesture generation in virtual agents. This section begins by discussing several major challenges in gesture generation in artificial agents, then presents broad approaches in current implementations that attempt to address these issues. It then critically appraises how these implementations are evaluated, as well as how gesture data is collected and analyzed within these frameworks. In doing this, it illustrates the difficulty of integrating the knowledge presented in the previous sections with modern data-driven gesture generation techniques.

1.6 Interdisciplinary Models and Approaches to Gesture Generation

While the importance of gesture in both the viewer and the speaker is clear, so too is the extent to which gesture is a complex, nuanced, and difficult task to perform. While spontaneous gesture is subjectively relatively effortless in most humans, it quickly presents many substantial challenges in machines. In social virtual agents, this difficulty can be broadly broken down into two tasks: selection and execution. This is not to downplay all the difficulty in collecting upstream knowledge on which to base selection, such as modeling or inferring intentions, leakage, dialog regulation, and predicting the effects of gesture performance. These phenomena represent substantial challenges in their own right, and have fields of research dedicated to them. For purposes of gesture generation, we will focus on approaches for these two sub-problems.

In this section, I focus on broad approaches to gesture generation, and their similarities and differences. While I provide contemporary examples of these various architectures, I do not deal with implementations of computational models or mechanical realization of gesture generation.

1.6.1 Challenges of gesture generation

The two challenges of selection and execution come with two important constraints which plague all aspects of intelligent social agent research: processing time and realization (animation or hardware). An acceptable pause between utterances is anywhere from 100–300 ms (Reidsma, Kok, Neiberg, Pammi, Straalen, Truong, and Welbergen, 2011), during which time an agent must gather or infer the relevant context, select a gesture given that context, plan, and perform the gesture in coordination with speech in order to appear natural. Similarly, choosing the contextually perfect gesture is useless if it cannot be performed on the required hardware. If choosing the optimal gesture would take 5 seconds, but a close-enough gesture only 0.05, that must be accounted for in the selection process.

In addition to these theoretical challenges, researchers also face the practical issue of how best to transcribe communicative functions using a common interface across different selection and execution implementations. The dominant framework for this is the SAIBA framework (Kopp, Krenn, Marsella, Marshall, Pelachaud, Pirker, Thórisson, and Vilhjálmsson, 2006) with stages that represent intent planning, behavior planning and behavior realization. SAIBA interfaces with two markup languages, Functional Markup Language (FML) and Behavior Markup Language (BML), to move between these stages. By beginning with intention of the agent, one can then derive the signals to produce. This decouples intention from implementations for different gesture generation mechanisms so they may be applied to different social agents, and forces architectures to drive gesture generation by intention and communicative function – an important theoretical parallel to Growth Point Theory (Section 1.1.3). Notably, this framework was explicitly developed with the goal of interdisciplinary collaboration in mind.

In reality, the major challenges of what motions to perform, how to communicate those motions,

and how to finally perform them must be considered in tandem throughout the gesture selection and performance process. Below, I dive deeper into the considerations of the process going from communicative intent to gesture performance.

Selection

Selecting a gesture (or class of gesture, such as a beat or deictic [Section 1.1.1]) comes with a range of considerations. Some driving factors may be the communicative intent of the speaker, from the motivation and sub-goal of a particular utterance, to any overarching goals dictated by the interaction. An agent must then incorporate relevant social context, such as the social status of the user, or the user's attentiveness to the conversation. This leads to considering the location of the conversation, both generally and to be aware of elements that may be constantly updating, such as people walking by. These factors drive the process of determining how to actually gesture, both with and without speech.

Selection must primarily be guided by the conversational goals of an agent. While gestures can be used to build rapport between agents and users (Wilson, Lee, Saechao, Hershenson, Scheutz, and Tickle-Degnen, 2017), this function may be considered unnecessary or even detrimental to an agent whose primary function is to direct or inform users efficiently. It is important that these dialog goals guide gesture selection, as random gesturing is not only confusing for the viewer and unnatural-looking (Lhommet and Marsella, 2014a), but can also lead to critical misunderstandings (Gurney, Pine, and Wiseman, 2013). The subject of false implicature is a main theme of Chapter 3 of this thesis and a driving force behind the design decisions made in the frameworks presented in Chapter 4.

As previously discussed, one role that gesture plays in human speech is to convey both explicit and implicit information to conversational partners in a contextually appropriate manner. Depending on the intended communicative function of the gesture, this context can be considered with great depth. One of the fundamental social skills for humans is the attribution of beliefs, goals, and desires to other people, otherwise known as Theory of Mind (Whiten and Byrne, 1988). In other words, an agent's concern with respect to gesture is not only "what does my gesture mean?" but "what does my gesture mean *to them*?" Özyürek (2002) provides evidence that speakers alter their gestures according to their mental models of their addressees. Scassellati (2002) provides an overview of how these challenges might be addressed in artificial agents, including implementations to find ways which can be used to predict internal state, and consequently, potential user responses.

Moreover, what may still be more relevant to an agent's gestures is it's own internal emotional state. Gesture can also be used to portray emotion in a way that is detectable by viewers (Pollick, Paterson, Bruderlin, and Sanford, 2001; Kipp and Martin, 2009) as discussed in Section 1.3.3. There is a considerable literature dedicated to computational models of emotion, with a summary found in Marsella, Gratch, Petta, et al. (2010). The breadth of this field in the context of gesture research suggests that an agent's own internal state may play a modulating role in gesture generation, with respect to both the type of gesture selected, as well as the way that gesture is performed. Research suggests

agents with understandable and consistent mental states and which act predictably are preferable for users (Mubin and Bartneck, 2015), making gesture a key potential avenue to facilitate positive social interaction.

Yet another consideration is when a gesture performance is appropriate by an agent. If given speech to perform, acoustic features such as emphasis and prosody can be key indicators of when a gesture performance may enhance or hinder communication (Krahmer and Swerts, 2007). Similarly, semantic information in speech may give clues as to when to gesture, or give parameter values to modulate gestures. For instance, it may be advantageous to refrain from gesturing, or use very low amplitude gestures, when discussing sensitive topics.

Execution

Equally important to the context and content an agent may access and express is the structure of potential gestures the agent can perform. Given the space of possible human gestures (e.g. the infinite planes on which hands can project and angles at which wrists can move, described in Section 1.4), they can be extremely challenging or impossible to replicate exactly, especially in physical robots with limited degrees of freedom, or even more challenging, robots with non-humanoid forms.

One area of concern in terms execution of a gesture is temporally aligning motion appropriately with co-speech utterances. Gestures seem to differ in terms of perceivers' sensitivity to their alignment with speech (Bergmann and Kopp, 2012), and there is a wide window in which gestures may be performed to be seen as aligned with linguistic meaning (see Section 1.1.3). Depending on agent implementation, coordination with other relevant body parts, such as the eyes, legs, and mouth, may present challenges for both dynamic animation and robotic movement. While virtual agents may have limited body points that can be controlled, a wide variety of tools from 3D modeling and animation tools (such as Maya by Autodesk, INC., 2019) to character animation engines (such as those implemented by Niewiadomski, Bevacqua, Mancini, and Pelachaud, 2009 or SmartBody by USC Institute for Creative Technologies, 2020) exist to both hand animate, use motion capture, or procedurally generate gestures on virtual agents.

Another challenge in gesture animation concerns the complex structure of gestures and the role of that structure in the performance of sequences of gestures (namely the phases described in Section 1.1.4). This includes the challenge of how to integrate individual gestures' features into fluid performances. To do so, virtual agent researchers have taken into account that human gesturing has a hierarchical structure that serves important demarcative, referential and expressive purposes. Xu, Pelachaud, and Marsella (2014) lay out an approach that uses this higher level of organization to realize gesture performances. Their approach determines when and which features are common versus which ones must be distinguishable and addresses issues concerning the physical coordination or coarticulation between gestures within gesture units, including determining whether individual gestures go into phases of relax, rests or holds. This work further draws on Calbris' concept of an ideational unit, described in Section 1.1.4.

Another challenge concerns the manipulation of the expressivity of gestures. For example, consider a gentle beat gesture that might convey a calm speaker emphasizing a point versus a strong beat gesture with larger, more accelerated motion that conveys a more agitated speaker strongly emphasizing a point. One approach to realizing such variation is to handcraft a suite of beat gestures, then blend them with other semantically-linked gestures. The technique of parameterized blending of animations, however, supports smooth variation between those extremes by controlling the amount of each gesture that is used in the blend so that the resulting gesture could vary the degree to which it emphasizes a point or conveys agitation. Blending presents challenges specifically to animators and graphic designers responsible for the presentation of gestures on virtual agents (Feng, Huang, Kallmann, and Shapiro, 2012).

Robots offer their own set of challenges. Often, robots have far fewer degrees of freedom than humans and virtual agents, with hard constraints on the extent and speed of motion. They are both different and severely limited compared to graphics based humanoid models. Specifically, robots suffer from physical limitations of their own hardware, with body parts being too heavy to move quickly without hurting themselves or others around them. Or, in order to alleviate danger to themselves or others, they may have a severely limited range of motion they can use to express gestures (Vasic and Billard, 2013; Hirth, Berns, and Mianowski, 2012).

1.7 Broad approaches in current implementations

Generating compelling gestures in socially intelligent agents is an open problem. Approaches to co-speech gesture generation can be characterized as existing on a continuum: rule-based vs. end-to-end machine learning techniques. One issue common to any approach that is often neglected in automated gesture generation is that of going from mental states to gestural performance. As noted, human gesturing is influenced by a wide variety on mental states, including communicative intentions within and across utterances, leakage or regulation of affective and cognitive states, traits and dialog management. The richness of human gesturing arises from this variety of mental state inputs.

However, the social agent field currently lacks a cognitive architecture of sufficient complexity to model such a variety of mental states, and has broadly moved away from holistic, all-encompassing behavioral architectures (with notable exceptions, e.g. Swartout, Gratch, Hill Jr, Hovy, Marsella, Rickel, and Traum (2006) and Kopp, Welbergen, Yaghoubzadeh, and Buschmeier (2014)). Consequently, the proxy input in gesture models is often reduced to the text and/or audio of the utterance which the agent is meant to perform, sometimes along with a limited communicative intent, for these elements are available to agents. This can limit an agent's gesture performance to what is available in these inputs. In other words, if the agent is not modeling emotion, social attitudes like skepticism, or what it wants to say on versus off the record, then its gestures cannot reflect this information. This is even true in the case of systems that use recorded voice, where potentially some of this information may be inferred from the audio, since the agent or agent designer must still be modeling such

information when selecting or recording the voice.

1.7.1 Rule-based models

One of the earliest non-verbal behavior generators is the Behavior Expression Animation Toolkit (BEAT) (Cassell, Vilhjálmsson, and Bickmore, 2004), which works by analyzing the relation between surface text and gestures. Text is parsed to attain information such as clauses, themes and rhemes, objects, and actions occurring in the discourse. This information is then used in conjunction with a knowledge base containing additional information about the world in which the discourse is taking place in order to map them onto a set of gestures.

Non-Verbal Behavior Generator (NVBG) (Lee and Marsella, 2006) extends the BEAT framework by making a clearer distinction between the communicative intent embedded in the surface text (e.g. affirmation, intensification, negation, etc.) and the realization of the gestures. This design allows NVBG to generate gestures that are rhetorically relevant even without a well-defined knowledge base.

Another approach which utilizes real-world utterance analysis is by Stone, DeCarlo, Oh, Rodriguez, Stere, Lees, and Bregler (2004). They proposed a framework to extract utterances and gesture motions from recorded human data then generate animations by synthesizing these utterances and motion segments. This framework includes an authoring mechanism to segment utterances and gesture motions then a selection mechanism to compose utterances and gestures. Similar to this, Neff, Kipp, Albrecht, and Seidel, 2008 created a comprehensive list of mappings between gestures types and related semantic tags to derive transmission probabilities of motion from sample data. This framework captures the details of human motion and preserves individual gesture style, which can then be generalized to generate gestures with varying forms of input.

This leads to a still more sophisticated method of generation, which is to combine this language-based method with making inferences from dialog about the mental state of the agent to determine which gesture to use. This approach may be effective without mapping to exact gestures. The outcome from different rules may, instead of prescribing an exact gesture, determine specific elements which should be present in a gesture (as seen in Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis (2005) and described in greater detail in Section 1.7.1). Additionally, various contextual information, such as speech prosody or detected listener attention, can determine other elements of gestural performance such as speed, co-speech timing, and amplitude.

This approach has been shown to be effective through multiple prominent examples in virtual agents. Using a combination of acoustic and linguistic elements, Cerebella (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013; Lhommet and Marsella, 2013) is a system currently in use in both virtual agent and social robotics applications. It dynamically generates gestures which appropriately correspond to speech both auditorily and semantically. Greta (Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005) is another example that typifies how high-level concepts can be used through external context to drive the motion of gestures of an agent. The architecture for these two systems, which provide excellent comparative examples of gesture generating architecture, are shown in Figure

1.3a and 1.3b.

Gesture catalogues vs. dynamic generation in rule-based approaches

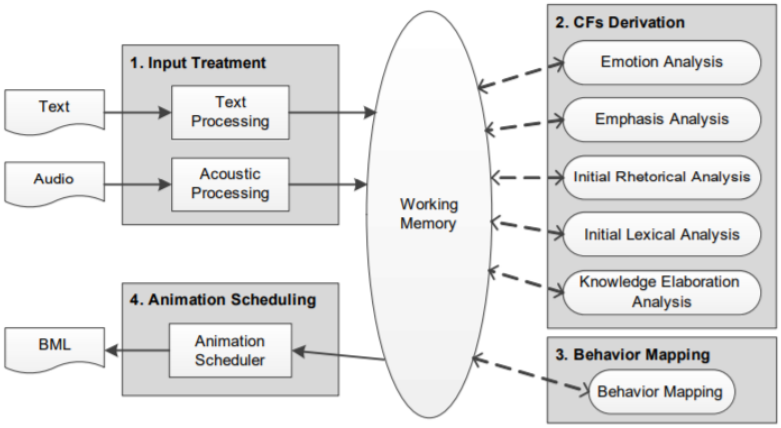
Broadly, rule-based gesture generation techniques can be implemented using either a set catalogue of gestures, or a set of parameters which drives dynamic generation of gestures on-the-fly.

Virtual agent designers and social roboticists often take the approach of using a fixed library of gestures. This is beneficial both because the agent designer may create gestures specific to the use case of the agent, either by having an animator create gestures using animation software or use motion capture of an actor. Another benefit is that by having pre-computed animations, the agent does not have to do extra work to actually compute the animation, but instead can act instantaneously in a motion that is guaranteed to satisfy the requirements of its software and hardware. However, while looking smooth and executing quickly are huge considerations in social agent research, this approach suffers from a lack of diversity in movements. By selecting only from a library of pre-animated gestures, agents risk looking particularly "artificial" by re-using gestures, by lacking a gesture for a particular social situation or by being unable to vary expressivity.

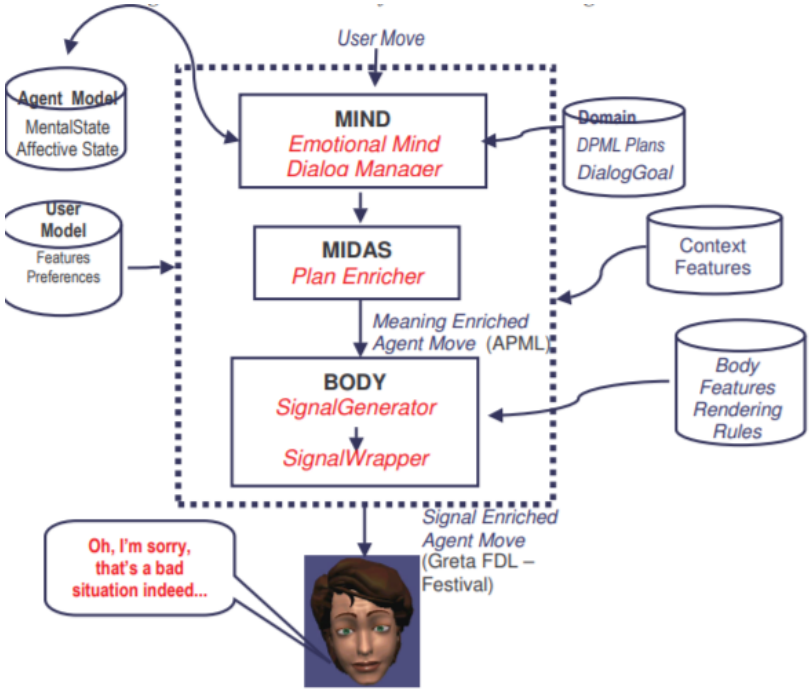
To address such limitations, research has explored parameterized gesture generation techniques as mentioned above that blend animations dynamically, providing a continuous range of variability between a mild beat gesture to a strong beat or small frame gesture or a large frame. This can also be done across multiple dimensions so that for example a beat may be varied both in intensity and direction. Another approach that I delve deeply into in Chapter 4 is parsing natural speech and motion into individual gestures, characterizing those gestures either physically or based on their co-speech semantics, and then utilizing this library of natural motion as a gesture catalogue during runtime.

Alternatively, an option of greater complexity is to allow agents to generate gestures entirely from a more complete parameterization of the motion such the hand shapes, the path the wrist takes, etc. This can be manifested in two ways: by generating gestures on-the-fly, or finding gestures from a library that satisfy any specified parameters, sometimes referred to as matching (Ferstl, Neff, and McDonnell, 2021b). These approaches must contain a model of how particular elements of the communicative context relate to gestural parameters, where the context might include, for example, whether the agent is trying to convey confusion, how agitated should the agent look and what hand shape and motion was used in the previous gesture. The alternative one might use is to simply have a table lookup approach, where the context select a set of pre-specified parameter values. For example, Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis (2005) uses context to derive hand-crafted parameters (such as amplitude, openness, etc.) which then select from a library of pre-created gestures. The use of pre-animated motions saves the calculation of motion planning during execution, while also supporting manipulation of the dynamics of those motions during execution to provide a level of novelty for the viewer.

Importantly, the resulting gesture from any method may still be adjusted through parameter manipulation. Gestures may be sped up, mirrored to adjust direction, or blended to create amplitudinal



(a) Cerebella Architecture



(b) Greta Architecture

Figure 1.3: The architectures of two generative gesture models.

"mild" or "extreme" versions of a gesture, all at run time.

1.7.2 Data-driven techniques

On the other end of the spectrum is completely text-agnostic end-to-end gesture production using deep learning. These models use large amounts of audio and video harvested from online sources like YouTube, and use video parsing tools such as OpenPose (Cao, Hidalgo Martinez, Simon, Wei, and Sheikh, 2019) to extract motion data to correlate audio to speaker movements. Using varying combinations of adversarial networks and regression, models are able to produce natural-looking gestures over a wide variety of speech-audio inputs (Ferstl, Neff, and McDonnell, 2020; Henter, Alexanderson, and Beskow, 2020). This approach undeniably leads to impressive results, particularly in the context of generating gestures based on an individual speaker (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019).

However, this approach lacks the sophistication of including multiple informative aspects of gesturing. By using audio input, these models are largely based exclusively on vocal cues like pitch and prosody. As a result, they fail to learn mappings between motion and semantic and rhetorical structure, and produce gestures that, while natural-looking, are less nuanced and complex than those we see in human performance (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Yoon, Wolfert, Kucherenko, Viegas, Nikolov, Tsakov, and Henter, 2022). While it has been argued that the middle layers of these networks can derive some of these aspects (Takeuchi, Hasegawa, Shirakawa, Kaneko, Sakuta, and Sumi, 2017) evaluations of gesture meaningfulness or semantic relatedness to co-utterances have not been done with end-to-end machine learning models based on audio.

Recently, end-to-end models have also been developed without audio, exclusively using co-utterance text of gestures (Yoon, Ko, Jang, Lee, Kim, and Lee, 2019) as well as models which incorporate vector mappings of words that can be included in generation mechanisms (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020). These have resulted in gestures which are judged as related to co-utterance, as well as life-like and likeable, although these judgements are somewhat questionable and are the subject of Chapter 3 of this thesis. Still, this work paves the way for promising avenues in the future of gesture generation, harnessing the power of both end-to-end machine learning models with speech qualities derived from both audio and textual cues.

Hybrid systems can offer the best of both worlds in terms of flexibility, novelty, and performance. From the examples above, I show how these two approaches exist on a continuum, although are not mutually exclusive. In the rule-based example, to recognize that a particular phrase has a negative intent necessarily requires some aspect of machine learning, as there is a robust body of literature on detecting affect in both written language (Pennebaker, Francis, and Booth, 2001; Hutto and Gilbert, 2014) and speech (Eyben, Wöllmer, and Schuller, 2009; Schuller, Batliner, Steidl, and Seppi, 2011). Similarly, we can detect transcripts from audio input and parse these using rhetorical and semantic cues through text parsers (e.g. Charniak, 2000; Joty, Carenini, and Ng, 2015; Pedersen, Patwardhan, Michelizzi, et al., 2004), many of which are used in the models above. These can be correlated with

gestures and may add crucial elements extra-auditory to deep learning models.

The Cerebella system realizes such a hybrid technique. It leverages information about the character's mental state and communicative intent to generate nonverbal behavior, when that information is modeled by the agent (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013; Lhommet, Xu, and Marsella, 2015). In addition, it relies on machine learning methods to also derive syntactic structure from the text and prosodic information from the spoken utterance. These sources of information are fed into a rule-based system and lexical database that perform additional lexical, pragmatic, metaphoric and rhetorical analyses of the agent's utterance text and audio to infer communicative functions that will drive the agent's nonverbal behavior.

This hybrid approach and the benefits offered by balancing trade-offs between dynamic generation and maintaining designer control over gestures inspires the core algorithms and implementations presented in Chapter 4.

1.8 Collecting data and evaluating generation algorithms

Researchers use a variety of techniques, tools, and analyses to study and understand naturally occurring gestures. Beyond simply classifying gestures (Section 1.1.1), it is necessary to elicit gestures in specific contexts in order to study how gestures influence conversation, and especially to form specific, testable hypotheses around gestures' role and influence in conversation. The ability to test hypotheses and therefore evaluate the performance of gesture generation algorithms is impeded by two major challenges: actually gathering natural gesture data, and then comparing gestures performed by generative algorithms to those performed by humans.

1.8.1 Gesture collection and analysis

Like many fields of behavioral psychology, researchers have used natural observation since the 70s and 80s. In the lab, classical techniques include solving spatial reasoning problems and game play (Alibali and GoldinMeadow, 1993), narrating videos (Kita and Özyürek, 2003), or telling written stories to conversational partners (Jacobs and Garnham, 2007). Historically, researchers have used more subjective techniques such as conversational scenarios (Ennis, McDonnell, and O'Sullivan, 2010) and questions, explicitly designed to elicit a variety of metaphoric gestures (Chu, Meyer, Foulkes, and Kita, 2014). More recently, some researchers have also used trained actors, either to perform their interpretation of an expression of an emotion, or to speak freely in a story-like, monologue fashion (Ferstl and McDonnell, 2018). Recently, current tools like YouTube with motion extraction software such as OpenPose (Cao, Hidalgo Martinez, Simon, Wei, and Sheikh, 2019) have provided troves of real-life examples of gestures by a huge variety of speakers in different contexts (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Yoon, Ko, Jang, Lee, Kim, and Lee, 2019).

A litany of tools is then used to dissect and analyze these gestures. Mainly from audio and video, a variety of annotation schemes have been developed for the purposes of segmenting and assigning

meaning to sections of gestures (Kipp, Neff, and Albrecht, 2007; Kipp, 2014; Chafai, Pelachaud, and Pelé, 2007). Such schemes are validated by determining internal consistency and inter-annotator agreement, thereby generating a reliable metric through which gesture elicitation techniques as well as gestures themselves can be compared along many axes.

Motion capture has also gained prominence in the gesture-capture space. Motion capture allows precise information on the spatial and temporal aspects of gesture, which can lead to powerful insights into how gesture correlates to speech and other elements of nonverbal behavior (Heloir, Neff, and Kipp, 2010). Importantly, this equipment also allows full-body capture, which is essential for producing natural and convincing gesture (Luo, Kipp, and Neff, 2009). However, motion capture equipment is expensive, can be cumbersome or distracting for participants, and still suffers from technical inaccuracies, particularly for capturing hands. And technological advances have allowed still other tools, such as gyroscopes, accelerometers, wiimote, and even VR controllers to sometimes be used to capture information about gestures (Corera and Krishnarajah, 2011).

Using these and other technologies, numerous data sets have gained popularity for use of studying, comparing, and animating gestures. This includes a wide range of visual technologies, from over 30 camera angles (Joo, Simon, Li, Liu, Tan, Gui, Banerjee, Godisart, Nabbe, Matthews, et al., 2017) to one central camera (Cooperrider, 2014), and from set gestures in tightly controlled staging conditions (Hwang, Kim, and Lee, 2006; Gunes and Piccardi, 2006) to spontaneous recordings collected completely outside laboratory settings (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Yoon, Ko, Jang, Lee, Kim, and Lee, 2019). Along with a growing interest in open science and data set production, new annotation tools such as the Visual Search Engine for Multimodal Communication Research (Turchyn, Moreno, Cánovas, Steen, Turner, Valenzuela, and Ray, 2018), which allows researchers to rapidly search data sets for specific types of motion, are becoming more sophisticated and widely used. Additionally, in recent years multiple meta-studies and evaluations on many non-verbal communication datasets have increased access and ease-of-use of many more gesture databases with varying additional annotated components (Ahuja, Lee, Ishii, and Morency, 2020; Lee, Deng, Ma, Shiratori, Srinivasa, and Sheikh, 2019; Ferstl and McDonnell, 2018; Metallinou, Yang, Lee, Busso, Carnicke, and Narayanan, 2016).

1.8.2 Evaluation

Evaluations of generative gesture models must be as application-driven as the selection and performance of the gestures themselves. And, while some metrics offer the comfort of traditional statistical analysis or straightforward interpretations, the right metrics to evaluate a model might be as difficult to determine as the gestures themselves.

Manipulating gesture can impact how viewers perceive an agent's personality traits (Neff, Wang, Abbott, and Walker, 2010) as well as common factors of interest such as trustworthiness, persuasiveness (Poggi and Pelachaud, 2008), affect (Pollick, Paterson, Bruderlin, and Sanford, 2001), and naturalness (Maatman, Gratch, and Marsella, 2005), often measured using self-reported subjective

measurement techniques. However, these factors are usually difficult to measure directly. Many individual gestures may be produced over the course of a relatively short utterance, leading to a litany of issues for how best to parse and recreate the timing of gestures (Wachsmuth and Kopp, 2001; Chiu and Marsella, 2014; Wilson, Bobick, and Cassell, 1996). This is even further complicated once a gesture has been selected for evaluation, because humans are notoriously bad at consciously discerning what does and does not look natural (Ren, Patrick, Efros, Hodgins, and Rehg, 2005), and accurately evaluating how they believe they are influenced by a behavior compared to how they actually respond (Lloyd, 1980).

For this reason, a variety of other metrics may be employed to measure the performance of generative models across axes of interest. Providing a forced-choice between the original input gesture and the model's output and comparing results versus a random production may be an alternative way to allow users to express preference for gesturing behavior (Lhommet and Marsella, 2013). For this evaluation method, a wide variety of comparisons may be employed to examine viewer exerience, depending on the goal of the study (Wolfert, Robinson, and Belpaeme, 2022). Mixed methods may also be used, for example giving users a chance to freely write an utterance that could accompany a gesture and perform a thematic analysis on the generated utterances. Minimally, this method can be used during pilot experiments to determine appropriate terminology for classic fixed-choice responses (Bryman, 2017).

Although it may seem intuitive that gestures should be evaluated by interpretability or clarity, this may not always be the case. For instance, an agent may actually intentionally perform a gesture which contradicts the utterance. The ultimate goal is to evaluate the gesture's consistency with the desired communicative function. That function, though, must be tailored to the particular context and uses for that social agent. And, crucially, the mapping between motion and meaning must be understood before it can be exploited.

As an alternative to subjective measurements one can evaluate gestures in terms of do they have the desired effect on behavior. For example, a range of experimental games have been used to explore the effect of an agent's nonverbal behavior on a human participant's behavior. For example, prisoners dilemma (De Melo, Zheng, and Gratch, 2009), the ultimatum game (Nishio, Ogawa, Kanakogi, Itakura, and Ishiguro, 2018), and the desert survival task (Khooshabeh, McCall, Gandhe, Gratch, and Blascovich, 2011).

When the physical motion properties of a gesture are available, as in the BVH (Biovision Hierarchy) file format used in motion capture and animation work, objective metrics concerning the physical properties can be used to evaluate gestures. Again, the challenge here becomes relating these properties to communicative functions and nonverbal behavior.

Tools to deploy evaluations are also advancing rapidly. Whereas previously researchers required individuals to make in-person evaluations of many gestures, crowdsourcing platforms such as Amazon's Mechanical Turk and Prolific now imbue the possibility of rapidly acquiring many "first-impression" measures on many different gestures. This has the added benefit of reducing the burden on viewers, as

well as reducing any fatigue effects of rating many different gestures. However, crowdsourcing platforms often offer varying quality in participant responses, and some demographic elements cannot be verified, making precise research challenging on this medium (Breazeal, DePalma, Orkin, Chernova, and Jung, 2013). Additionally, crowd-sourced participants may be non-naive "expert survey-takers," which can skew study results (Downs, Holbrook, Sheng, and Cranor, 2010). Study design elements such as verifying attentiveness, longitudinal studies, and mixed method qualitative analyses of free responses are able to overcome some of these challenges (Chandler, Mueller, and Paolacci, 2014; Rouse, 2015).

Ultimately, the evaluation of a model must be specific to both its implementation and application. I examine this issue in greater detail in a subjective experiment presented in Chapter 3.

1.9 Ongoing Challenges

The technology and tools for modeling and generating gestures continues to advance. Larger and larger data sets are being captured and new techniques are being used to process that data, further enabling machine learning approaches. These advances will provide new power to address challenges and opportunities. Here, I discuss some of these challenges.

1.9.1 Gestures and the context that informs their use

One of the key challenges we face in realizing gestures for social agents is the complex relationship between gestures and the context of the interaction and overall structure of the discourse. As has been pointed out repeatedly by gesture researchers (e.g. Kendon, 2000) and discussed in Section 1.3, gestures, specifically their communicative function, are not simply a vivid illustration of the dialog text. For example, pragmatics concerns the context in which the interaction occurs and the impact of that context on deixis, turn-taking, across utterance structure of the interaction, presuppositions and implicature. These factors have a profound effect on gesture use. An obvious example of this concerns deictic gestures. Utterances such as "You should talk to Michael," or "Leave by the door on the right," may or may not co-occur with a deictic gesture. Another example is the cross utterance use of gestural space, where one utterance can locate an abstract concept in gesture space and in a subsequent utterance, gestures can refer back to that location so as to refer to that original concept. Another example of the extra-utterance factors impacting gestures concerns how mental state leakage discussed above impacts gesture use and gesture performance. Further, the roles, cultures, and relational history of participants impact their gestures. Yet another example is when gestures are used to convey information off the record or even contradict the content of the utterance. Broadly, a gesture can be a distinct speech act from the speech act realized by the utterance.

These examples pose significant challenges to realizing rich gesturing in social agents, regardless of whether the approach is end-to-end machine learning, rule-based or some hybrid. Fundamentally,

capturing and exploiting context requires some approach to modeling or inferring this extra-utterance information.

In the case of end-to-end machine learning approaches that map an utterance to gesture, the external context of the utterance, the overall structure of the interaction, off-the record information to convey gesturally, and arguably even the the internal mental states and roles of the participants will not be apparent in the individual utterance text or prosody. This makes it unlikely that a mapping from utterance to gesture that takes into account just the utterance will capture the richness of human gestures. Even in the case of rule-based methods, algorithms must have some way of modeling cognitive and conversational information over the course of interaction in order to come close to the richness of human gesture.

1.9.2 Complex Gesturing

A related challenge concerns complex gesturing. As illustrated above and in Section 1.1.2, gesture categories are fluid, and a single gesture often combines elements of many different categories, which are related to elements of the interaction through multiple cues. This complexity is compounded by the fact that gestures can both stand alone individually as well as tie together pragmatic, semantic and rhetorical elements that span utterances. Moreover, a gesture can express multiple semantic or rhetorical concepts in a single motion. This complexity poses significant difficulty in mapping multiple communicative concepts to a variety of motions, and is explored subjectively in Chapter 2. My approach to overcoming this challenge is incorporated in the framework presented in Chapter 4.

In order to use these various sources of information to gesture effectively both for individual turns of dialog as well as coherently and naturally over an utterance and multiple dialog turns, researchers in gesture as well as conversational AI must come together to create a computationally organized model that tracks semantic, environmental, conversational, and spatial context for interactions. This underscores the tight relationship between gesture, speech, and the overarching interaction, and highlights how integrated gesture generation systems need to be with speech production and pragmatics in order for virtual agents to be as human-like as possible.

1.9.3 Role of Participants

A gesture model also needs to consider the participants themselves. In order to gesture appropriately the social agent should take into account their conversational partner. Humans tailor gestures to the individual to whom we are speaking (Marchena and Eigsti, 2014) which can have significant effects on how the speaker is perceived (Lee, Uhlemann, and Haase, 1985). This can include some basic automatic responses, like mirroring, but also encompasses extremely sophisticated complex modeling of the user's mental state. Adjusting gestures to be smaller or slower when discussing sensitive topics, taking into account the age of the listener or making large, pointed gestures to persuade a crowd are a few examples of acutely different circumstances during which the context must be detected, and the

implications analyzed, to adjust gesture parameters (Poggi and Vincze, 2008). Crucially, context must affect both the selection as well as production of gestures.

This raises the question of how an agent infers a conversational partners' reactions. Are they, for example, being persuaded or amused by the agent's use of expressive gestures? Clearly, an agent should select a gesture that is relevant and meaningful to its communicative function and consequently be able to infer whether that communicative function is being realized in the human partners in the interaction. Thus, gesture generation algorithms may also benefit from detecting user engagement and inferring mental state and feeding that information back into performances. Again, this requires an understanding of how the viewers own motions relate to *their* intended communicative functions. A mapping between motion and communicative meaning can be exploited by both selection and execution steps within gesture generation pipelines.

Inferring a viewer's mental state gives rise to a growing issue of concern in gesture research: cross-cultural interpretation. As the world becomes evermore interconnected and developers of social agents become increasingly interested in international markets, the importance of gesturing in a culturally sensitive way gains much greater importance. This includes not only the amount or style of gesture, but gets into deeper issues of conceptual organization and metaphorical hierarchies that exist in different cultures (such as the "Time is a Line" metaphor discussed in Section 1.2, and which I revisit in Section 5.2.2). This means that metaphoric gestures which convey a particular meaning in one culture may carry no or even an opposite meaning in another, which can result in critical misunderstandings between agents and users. The importance of cross-culture metaphoric interpretation is further explored in the experiment described in Section 2.2.

1.9.4 Ambiguity

Contrary to providing clarity in conversation, one might well argue that human-like or "natural" behaviors may cause ambiguity in the agent's message. Instead of an agent conveying agitation by the dynamics of their gestures maybe it is just as or even more effective to put a sign over agent saying it is agitated or altering the color of the agent. Specifically, the work shown in Section 2.1 suggests that when gestures are too complex in the sense of a single gesture conveying multiple pieces of information, they become less uniformly interpreted across subjects – muddling the message an agent may attempt to convey. As the ability to produce complex gestures increases, researchers will need to consider different ways to measure trade-offs in performance of generative models, from speed and complexity to optimizing for user understanding.

Finally, one question that still remains as an overarching guiding principle is just how human-like does the behavior of the agent have to be. If one ascribes to the Media Naturalness hypothesis (Kock, 2005), divergence from the naturalness of face-to-face interaction, broadly speaking but specifically here in terms of nonverbal behavior, can lead to an increase in cognitive effort, an increase in communication ambiguity, and a decrease in physiological arousal.

1.9.5 The Application

Unquestioningly, these trade-offs will be context- and application-dependent. In a social skills training application to train doctors to break bad news to patients (Kron, Fetters, Scerbo, White, Lypson, Padilla, Gliva-McConvey, Belfore II, West, Wallace, et al., 2017; Ochs, Montcheuil, Pergandi, Saubesty, Pelachaud, Mestre, and Blache, 2017), naturalness is a paramount consideration in part because people are trained to deal with ambiguities.

In contrast, a learning application for children that seeks to increase engagement as a child learns to count may forego any attempt at naturalness. Here there are opportunities to draw on a wide range of research. There is animal and human research on supernormal stimuli that can provoke primal responses in people (Barrett, 2010). The performance arts, specifically theatre and dance, can provide more stylized and less ambiguous means of conveying information (Smith and Cross, 2022). Notably, social agent researchers (Marsella, Carnicke, Gratch, Okhmatovskaia, and Rizzo, 2006; Neff and Fiume, 2008) have looked at Delsarte's work on gesture that heavily influenced early silent film acting as a means of gesture selection and performance, as well as Laban Movement Analysis to manipulate the animation of expressive gestures (Chi, Costa, Zhao, and Badler, 2000).

1.9.6 Evaluating Impact

One way to evaluate impact of gesture on behavior across large demographic populations is through increasingly popular crowdsourcing platforms (Breazeal, DePalma, Orkin, Chernova, and Jung, 2013; Morris, McDuff, and Calvo, 2014). In addition to evaluating a social agent's gesture performance, crowdsourcing opinions makes a combined approach to gesture generation possible: generative models which use crowd or expert input to create and refine generative models of dialog for a social agent (Feng, Sequeira, Carstensdottir, El-Nasr, and Marsella, 2018) could be extended to gesture. Research has begun using crowd feedback in model tuning to adjust gestures according to different social and conversational contexts. By using machine learning to uncover patterns in user preference and determine salient features in gesture motion, we may be able to increase model performance and produce gestures that are more contextually appropriate and complex than simply using top-down expert-driven rule-based techniques or end-to-end deep learning. While this is a relatively new technique in the field of gesture generation, finding ways to seamlessly incorporate human judgements into the generation process is a promising avenue for producing natural, meaningful, and relevant gestures in social artificial agents.

Methodologically, the most common tool to measure viewer impact is self-reported questionnaires, often focusing on a gesture's "humanlike-ness" and it's "appropriateness" for it's contextually presented co-speech (Wolfert, Robinson, and Belpaeme, 2022). While these metrics provide surface-level feedback that can be useful for researchers, these shallow metrics leave much open to interpretation. The challenge of how best to measure the impact of a gesture on a viewer, and the broad range of qualitative impacts a gesture can have on conversational interpretation, is the subject of Chapter 3.

1.10 The Current Work

This chapter discussed the many ways in which gesture enhances communication. Gesture acts as a guide for dialog, an influence on the observer, and a reflection of the speaker's internal beliefs. It briefly summarized a long history of gesture studies, including myriad ways to classify gesture by both motion and communicative function. It considered how these functions, combined with individual and cultural context, may reveal information about the speaker's attitudes and mental states, as well as more complex information about an individual's cognition.

Meanwhile, it also explored how digital technology is becoming more and more ubiquitous in social interaction, one aspect of this being the ongoing development of virtual agents and embodied characters. Embodied agents are increasingly called upon to perform face-to-face interactions. The ability of virtual agents to accurately perform embodied social non-verbal cues is crucial, as humans draw inferences about agents from even a lack of a nonverbal performance. Gesture is a crucial element of these embodied interfaces because of its many communicative and expressive roles.

Data-driven, deep learning approaches to gesture generation have relied more on computational advances and increasing availability of data than on lessons from psychological research on gesture. These approaches generally have nevertheless been effective at making an agent look lively, but often are not designed to convey rich meaning. Critically, these models are not interpretable or editable. Therefore, they do not give designers control over the implicit mapping from speech to the resulting form, except by selecting the dataset that the system learns on. Applying such models in virtual human applications runs several risks. Context, such as the cultural and task settings, influence what behavior is exhibited. Related to this, a poorly chosen or performed gesture can lead to unintentional false implicature, which can be especially critical in sensitive applications such as teaching, coaching, and therapeutics. This assumes two major points concerning the need for rich gestures and the risk of false implicature: People may infer complex meaning from a virtual human's gestures, and they may draw false implicatures from gesturing even when those gestures are not intended to convey specific meaning.

Contrary to deep learning gesture generation algorithms, rule based approaches are heavily informed from psychologically and ethnological studies. These models tend to be interpretable in terms of how gestures are selected and generated. This enables richness in the meaning conveyed as well as tailoring to the application context. However, these approaches tend to inherit the limitations of the studies on which they are based, specifically the limited coverage of the phenomena because research historically takes an observational as opposed to data-driven approach.

This background informs a main goal of this thesis: to build a data-driven analysis technique that is capable of harnessing insights from observational research. Throughout, this body of work demonstrates how metaphoric gesture generation presents a multitude of challenges for virtual agents, because of this type of gesture's many forms and functions.

These functions, specifically the relationship between gesture form and communicated metaphor, are the topic of Chapter 2 of this thesis. It breaks down the role of form in multi-metaphoric gestures

in communication in Section 2.1, then examines how this communication differs across cultures in Section 2.2. These studies argue that it is critical to understand gesture as a multi-faceted, multi-dimensional, nuanced phenomenon. A single gesture is capable of simultaneously delivering a broad variety of information, including potentially conflicting semantic concepts. Thus, any attempt to map the relationship between a gesture's motion and its communicative meaning must incorporate and be sensitive to this complexity.

Metaphoric gestures are also discussed the context of current implementations of gestures in virtual agents. There are many ways to realize compelling gestures in social agents, but these must be centered on the communicative function of the gesture. Using frameworks which abstract implementation from communicative function allows researchers to separate the problem of gesture selection and animation. Both machine learning and rule-based techniques offer promising solutions to these difficulties, but face similar challenges in terms of gesture collection and model evaluation. Such evaluation challenges inspire the experiment and discussion presented in Chapter 3. This Chapter emphasizes the danger of false implicature, and further reinforces the necessity of a comprehensive, interrogable mapping between motion and meaning. It secures the knowledge that the impact of gesture on viewers cannot be evaluated simplistically, nor divorced from linguistic context; Motion and language work in tandem to convey meaning.

Chapters 2 and 3 show that not only do viewers infer rich meaning from complex gesture, but furthermore that they readily infer meaning from gestures that are not intended or designed to convey semantic information. They establish that it is critical for generative gesture models to effectively navigate the mapping between motion, meaning, and interpretation.

Chapters 4 and 5 shift focus from the impact of gesture on viewers to incorporating psychological insights into data-driven mechanisms to produce and understand this mapping. Despite recent advancements, gesture generation still faces many challenges, such as generating conversationally (semantically) relevant movements, integrating complex or ambiguous gestures, and considering the role of the viewer when modulating gesture behavior. These must all be taken into consideration in order to achieve the greatest impact of gesture on an agent's audience. Balancing these trade-offs and understanding the mapping between gesture form, motion, meaning, and impact is imperative for gesture generation.

While recent times have shown growth in use of data in gesture generation, this abundance of data is largely used to create black-box machine learning algorithms which are neither transparent nor interrogable. In this thesis, I argue that researchers should use data to examine and better understand the relationship between communicative intention and gestural motion. Creating this mapping is the subject of the work presented in Chapter 4. This chapter presents a framework that is guided by psychological research and developed alongside modern computational techniques to create a comprehensive, human-readable mapping between motion and meaning that is, in particular, capable of considering simultaneously communicated concepts. Chapter 5 then demonstrates how this mapping can be applied to behavioral hypothesis testing. This Chapter then furthermore discusses the frame-

work's potential applications to evaluating generative models, facilitating subjective observational studies, and allowing interaction and virtual agent designers to explore potential situations in which false implicature may be a concern.

This hybrid psychological and data-driven approach allows us to explore the details of this mapping, with significant ramifications. While the framework I put forward is far from complete, it lays the groundwork for holistic understanding of how physical motion relates to communication in conversation. The many ongoing challenges faced by gesture generation researchers and potential future applications of this framework are discussed in Chapter 6.

Chapter 2

Exploring the Complexity of Metaphoric Gesture

As discussed in Section 1.3.3, gesture is not only a behavioral phenomenon, but a cognitive one (McNeill, 1992). Gestures act as windows into embodied cognition (Cienki and Koenig, 1998), especially as they physically illustrate both linguistic and contextually and culturally implied co-speech metaphors. This chapter presents two subjective studies, each consisting of two experiments, that probe deeply into the variety of interpretations of co-speech metaphoric gestures, both within individual viewers and across many individuals within and between cultures.

These studies explore gestures that can be grounded in complex meaning involving multiple metaphors (Lhommet and Marsella, 2013; Ravenet, Pelachaud, Clavel, and Marsella, 2018). Study 1 establishes that viewers interpret complex, multi-faceted messages through metaphoric gestures. Study 2 furthermore explores the extent to which these messages are culturally and contextually dependent. Together, they establish that viewers readily discern multiple metaphors from a single gesture. Therefore, any mapping from an utterance's meaning to an accompanying metaphoric gesture must take into account that multiple metaphors can be conveyed in a single motion.

These results have significant ramifications for how these behaviors are generated while raising concerns about potential false implicatures when a generation process is not tailored to the context of the interaction.

Section 2.1 is unchanged from published work that can be cited as: **Saund, C.**, Roth, M., Chollet, M., & Marsella, S. (2019, September). Multiple metaphors in metaphoric gesturing. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019) (pp. 524-530). IEEE.

Section 2.2 is based on work compiled and presented at the 5th Virtual Social Interaction conference and can be cited as: **Saund, C.**, & Marsella, S. (2021, June). Interpretations of virtual agent performances of metaphoric gestures differ across cultures [conference presentation]. 5th Virtual Social Interaction Conference, Glasgow, UK.

2.1 Study 1: Multiple Metaphors in Metaphoric Gesturing

The use of metaphoric gestures by speakers has long been known to influence thought in the viewer. What is less clear is the extent to which the expression of multiple metaphors in a single gesture reliably affects viewer interpretation. Additionally, gestures which express only one metaphor are not sufficient to explain the broad array of metaphoric gestures and metaphoric scenes that human speakers naturally produce. In this section we address three issues related to the implementation of metaphoric gestures in virtual humans. First, we break down naturally occurring examples of multiple-metaphor gestures, as well as metaphoric scenes created by gesture sequences. Then, we show the importance of capturing multiple metaphoric aspects of gesture with a behavioral experiment using crowdsourced judgements of videos of alterations of the naturally occurring gestures. Finally, we discuss the challenges for computationally modeling metaphoric gestures that are raised by our findings.

2.1.1 Introduction

Gesture plays a powerful role in human face-to-face interaction (McNeill, 1992; Kendon, 2004). This has led to research in the virtual human community both on how to realize gestures (Cassell, Pelachaud, Badler, Steedman, Achorn, Becket, Douville, Prevost, and Stone, 1994; Hartmann, Mancini, and Pelachaud, 2005; Bergmann and Kopp, 2009; Chiu and Marsella, 2014) and their impact on human-agent interaction (Poggi and Pelachaud, 2008).

Our interest here is specifically in a class of gestures called metaphoric gestures. Metaphoric gestures are described by McNeill (1992) as gestures which "present an image of an abstract concept,". For example, one may gesture in a bowl or container shape when describing "all of their ideas." Although the abstract notion of an "idea," can never be physically realized, the metaphoric gesture situates "ideas," in a metaphorical container that can now be spatially referenced by the speaker and viewers (Calbris, 2011). Metaphoric gestures have been shown to have an impact on human interaction, influencing the listener's perceptions about the speaker (Goldin-Meadow and Alibali, 2013; Maricchiolo, Gnisci, Bonaiuto, and Ficca, 2009b), increasing comprehension (Beaudoin-Ryan and Goldin-Meadow, 2014) and recall (Cook and Goldin-Meadow, 2006). Similarly, they directly influence the thought processes of the viewer (Tversky and Hard, 2009).

It has previously been hypothesized that metaphoric gestures provide a window into cognition by revealing how abstract reasoning is grounded in the representations and processes by which we perceive and take action in the physical world (e.g., Cienki and Koenig, 1998). The underlying content of metaphoric gestures often depict abstract concepts as physical objects with distinct properties, which can then have actions taken on them through subsequent gestures (Lhommet and Marsella, 2013). For example, one's gesture may form the shape of a physical object in order to identify an "idea," and later in the conversation mime throwing the object away to suggest rejecting the idea. In addition to depicting abstract concepts and taking actions on them, attributes of the abstract properties can be

suggested by a gesture that conveys physical properties. For example, the significance of that idea can be conveyed if the gesture simultaneously suggests the object representing the idea is heavy or large. Thus two metaphors, *Abstract Ideas are Physical Objects* and *Importance is Size*, are composed into one singular gesture.

It is with this type of metaphoric composition in mind that we approach metaphoric gesture as a way to examine embodied cognition. Specifically, we have been exploring two phenomena related to metaphoric composition in gestures:

1. Multiple metaphors can be creatively composed in the formation of a single gesture.

2. Sequences of gestures creatively compose metaphoric scenes involving multiple metaphors that relate the abstract concepts the speaker conveys in terms of physical properties of the scene, and gestures used to depict it.

An overarching goal of this research is to create computational models for virtual agents which dynamically produce natural, human-like gestures. Such models would not only have broad implications for embodied agents, but may bring us a better understanding of the underlying thought processes that are responsible for producing communicative performance. However, providing a virtual agent with a capacity to gesturally compose metaphors is only justified if speakers use such gestures, and observers are desirably impacted by them.

To that end, this work explores the communicative interpretability of individual metaphoric gestures that express multiple simultaneous metaphors. We break down metaphoric gesture sequences to demonstrate the rich metaphoric scenes that speakers naturally produce. Critically, we assess whether viewers infer multiple concepts when multiple metaphors are composed into a single gesture, and if there is a benefit to modeling them in the sense that they influence an observer's interpretation of the performance. Having illustrated the richness of this phenomenon and it's relevance to subjective interpretation, we discuss the challenges associated with realizing such performances in virtual humans.

2.1.2 Metaphoric Composition Examples

By using metaphoric gestures, we create representations of objects in physical space to convey abstract semantic meaning. The three examples below are in-depth analyses of gestures which employ multiple metaphors simultaneously to illustrate many shades of meaning, some of which are directly informed by speech, but which more often imply extra information not given by language. They were chosen for their clarity in simultaneously expressing multiple physical metaphors in a single gesture (Grady, 1997). We focus on three aspects of the objects created by metaphoric gestures:

1. The **properties** of object representations capture the attributes of the abstract concept conveyed.

2. The **locations** of object representations convey relationships between them, the attitudes of the speaker, and the compositions of sets.

3. The multiple **metaphors** which may be in play.

The metaphors we consider are informed by Grady (1997) and are of the form "*Semantic Meaning is Physical Characteristic*" (*e.g. Contrasting Ideas is Physical Separation*). Although Grady focuses on linguistic metaphors, gestures allow a broader range of semantic expression than speech alone. The semantic meaning that we associate with the physical characteristics of a gesture is largely interpretive, but is also informed by previous work (e.g., Lhommet and Marsella, 2016; Cienki and Koenig, 1998; Calbris, 2011; Kipp, Neff, and Albrecht, 2007).

We refer to these three gestures throughout this section as the "Entity," "Unpolite," and "Audience," examples respectively. Although we describe them below, they can be viewed from the link in the Supplementary Materials (Section 7.1.1; `https://osf.io/txv7g/?view_only=bccaefcd70f44e5e8d06f27accb1893a`).

Example 2.1 ("*Entity*"; "`kelly_entity.mp4`"): "It [HR dept.] was an entity completely controlled by the CEO." - The element "human resources" was established in the scene earlier on and is a reference to the human resources department. The speaker discusses harassment by the CEO, an issue that typically should be addressed by HR. Here (Fig. 2.1), the speaker picks up the element, representing HR in the physical space (*Abstract Idea is Concrete Object*), contains it in her hand, and pulls it outwards, away from her body. (*Contrasting Ideas is Physical Separation*). The bowl-shaped hand is positioned at the bottom of the element, with the palm up. The hand shape is tense and firm (*Certain is Firm*), as if the element was trapped between her fingers (inverse of *Accessible is Open*). In this example, we see how a single gesture employs four metaphors simultaneously. While she indicates the "control," the CEO has over the "entity," HR, she also removes the entity from herself by physically moving the metaphoric object representing HR away from her, and closing it off indicating it is beyond her reach. This not only matches the language which accompanies these gestures, but also effectively illustrates many of the subtleties in this sentiment.

Example 2.2 ("*Unpolite*"; "`colbert_unpolite.mp4`"): "That room gets very unpolite [sic]." With the palms facing inwards, the speaker moves both hands towards each other, interlocking his fingers (Fig. 2.2). He then moves them slightly apart and back together. The hands themselves represent the concrete and abstract elements he talks about: the people involved and the ideas being discussed (*Abstract Idea is Concrete Object*). It is conveyed that there is disagreement between opinions (*Conflict is Collision*) but the people involved are also working together (*Combining Idea is Physical Closeness*).

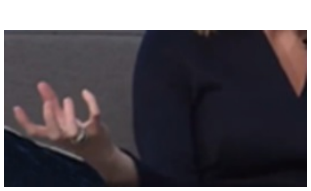


Figure 2.1: Firm, bowl-shaped hand in example 2.1

The speaker is discussing a room in which many writers must work together to produce a TV program in a very short amount of time. Although his language says the writers are "unpolite," (sic) to one another, the intertwining of the fingers also indicates a measure of closeness, cohesion, or unification. When we examine this gesture in behavioral experiments below, we see that although these two metaphors are seemingly in conflict, both messages are transmitted reliably to viewers.

Figure 2.2: Hands interlocking back and forth in example 2.2

Example 2.3 (“*Audience*”; “`megynkelly_available_audience_full_scene.mp4`”): A Metaphoric Scene - The following examples come from a large metaphoric scene in which the speaker simultaneously elaborates on her point and sets up a complex relationship between metaphoric objects using gestures (Kelly and Business Insider, 2017). We again see multiple metaphors combining in single gestures, but also observe an elaborate physical scene with metaphoric containers, subgroups with specific properties, actions taken on those subgroups, and multiple references to metaphoric objects according to their position in space (see an overview of the sequence in Fig. 2.3).

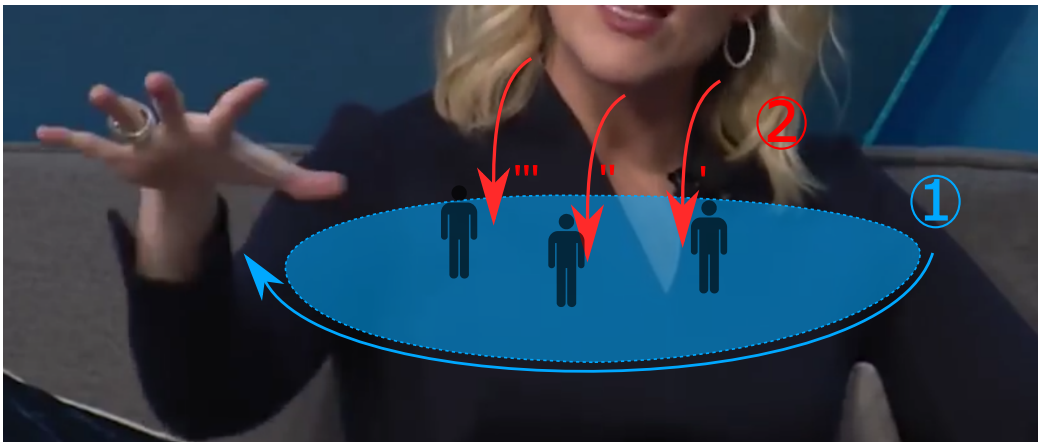


Figure 2.3: The entire motion of the scene she sets up.

Example 2.3.1: “Was there an available audience [. . .] of Republicans and Independents and Democrats” - The speaker establishes the set “available audience” (Fig. 2.3) by denoting the top of its shape in space with her flat open hand, palm facing downwards (*Sets are Containers*). The downward position of her palm indicates the “availability,” of the audience to hear her message (*Being in Control is Being Above*). The movement is large (*Quantity is Size*), and inclusive. The verbal context implies an intention of influencing the “audience,” and, interestingly, is elaborated on at the end of the phrase.

Subsequently, the speaker establishes herself as a point of reference (i.e. who is controlling the audience) by gesturing inwards. Then, she refers back to the already established set and creates three new reference points (Fig. 2.4; *Constituents are Contents, Elements have Location*). Pointing with a pyramid handshape to the three specific and clearly separate locations in space, the speaker indicates three distinct components (i.e. different political ideologies) of the already established set (*Contrasting*

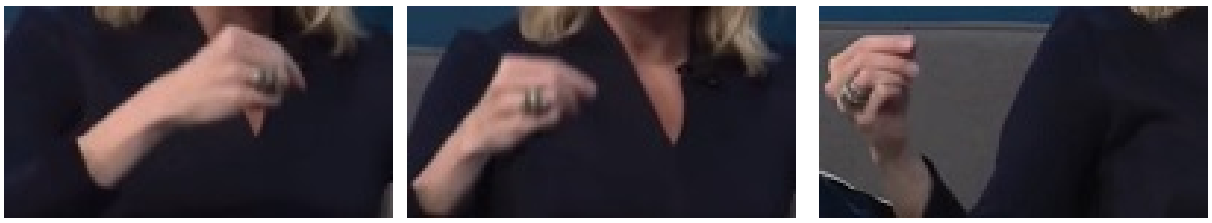


Figure 2.4: Three successive precise spatial references with pyramid handshape

Ideas is Physical Separation).

Example 2.3.2: "Who are open-minded" - With a flat, widely open hand, the speaker denotes the front of the previously established shape "audience", conveying its abstract properties by visualising them in physical space. The speaker shows that the audience's minds are open (*Abstract Concept is Concrete Object*), illustrating that new ideas can affect existing opinions (*Accessible is Open*).

Example 2.3.3: "Who are there for . . . persuading" - With the relaxed hand on the right side of the shape and the palm facing inwards (Fig. 2.5, left), the speaker grasps the different elements of the previously established set "audience" (*Sets are Containers*) selecting them collectively as one group (*Combining Ideas is Physical Closeness*). The initial open and inclusive movement progresses into a closed pyramid hand shape in front of the shape of reference, with the palm facing towards the speaker (Fig. 2.5, middle), suggesting ownership. All components of the "audience" are contained in the speaker's hand and are therefore in her control (inverse of *Accessible is Open*). The speaker then moves the pyramid-shaped hand in a half-circle outwards and away from her body, supporting the bottom of the shape with the palm upwards (Fig. 2.5, right). This open and forward motion illustrates the persuasion of the "audience" (*Progression is Forward*): the speaker is still in control of what the hand contains, but simultaneously conveys that her actions affect others' views (*Accessible is Open*).

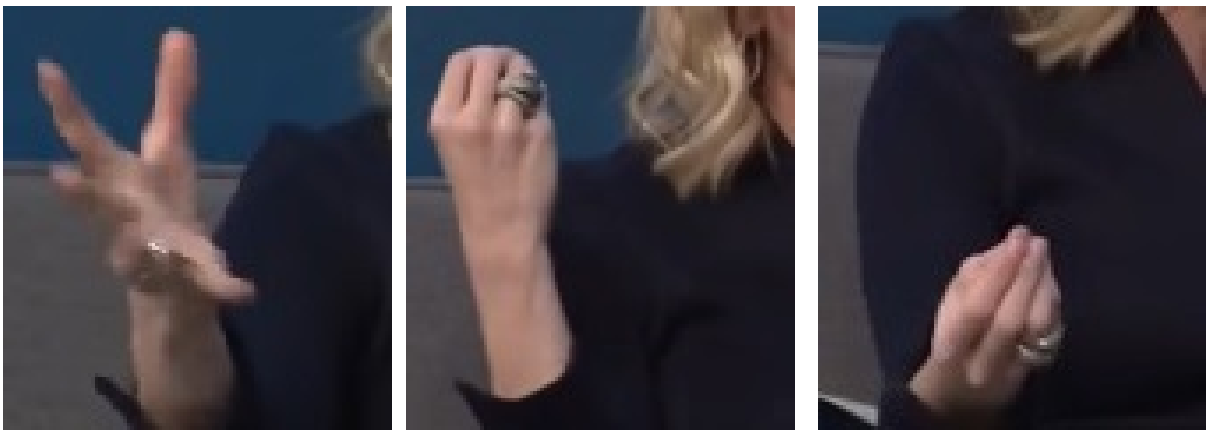


Figure 2.5: Speaker grasping, controlling and persuading the metaphoric audience established in example 2.3.1

2.1.3 Behavioral Experiments

Although it has previously been shown that metaphoric gestures influence conversational interpretation (Jamalian and Tversky, 2012b), it is less clear the extent to which multiple layered metaphors communicate nuanced information that can reliably be decoded. In this section, we describe two experiments where we explore this possibility. Given the extent to which we observe metaphors co-occur within a gesture, we expect complex metaphoric (multi-metaphoric) gestures to carry more information and be subject to a wider variety of interpretations than gestures which employ only one metaphor. However, we expect that some information from all of the metaphors can be decoded as interpretable input by the viewer. These experiments compare multi-metaphor gestures to physically closely related single-metaphor gestures (our gesture conditions) that we believe employ different metaphoric associations to understand whether multiple metaphors are interpretable in a single gesture. By choosing multi-metaphor gestures as base stimuli, we manipulate only a subset of metaphors to attempt to isolate the effect of that particular metaphor on viewer interpretation.

Although we present two related experiments, the results taken together paint a complex picture of cognitive metaphoric composition which carries nuanced information about how metaphors combine to influence viewer interpretation.

Method

Participants

All participants were recruited from Amazon Mechanical Turk platform from the United States and United Kingdom. 187 participants were used to gather pilot data for this experiment. 12 participant responses were removed for incomplete responses. 275 Participants were recruited for the first experiment, 371 for the second. Participants were paid £8 per hour, prorated for their individual time taken to complete the task. This number of participants allows a power of 0.83 and 0.85 for each study, respectively.

Pilot Procedure

Participants were given instructions to view one 1.5 second video clip of an actor performing a gesture without sound. Five gestures, three of which are also described in Section 2.1.2 (specifically, "Entity," in example 2.1, "Unpolite," in example 2.2 and the hand sweep of "Audience," example 2.3.1), were selected for their clarity in display of multiple simultaneous metaphors by four separate annotators. For each of the five gestures there was a performance of the original gesture, an alteration which emphasizes one metaphor, and another alteration which emphasizes another (Fig.2.6) (video of all gestures can be found in the link in Section 7.1.1). Participants randomly viewed one of 15 gestural performances, then were asked to fill out a free-response with the prompt of "She is talking about a group. What do you think she is saying?" These free responses were analyzed and grouped into response themes to derive statements for the subsequent experiments. Experiment 1 asked participants

to rank statements according to their agreement with the statement, while in Experiment 2 participants were asked to rate their agreement from 1-7 for each statement.

(a) Original (b) “Collide” (c) “Intertwine”

Figure 2.6: Screenshots of “Unpolite” gesture conditions.

2.1.4 Experiment 1

Experimental Procedure

From the pilot data, we identified 5 themes across all stimuli. These were used in combination with the metaphors we believed to be expressed to generate a total of 10 statements, with 2 corresponding to each of the 5 metaphors found throughout the gesture stimuli, all of which refer to group relationships (All stimuli can be found in the link in Section 7.1.1). Participants were asked to rank these statements according to how true they believed them to be (1-10), and then asked to fill out a free-response portion to check that they viewed the video. All statements can be found in Section 7.1.2.

Hypotheses

We predicted that our gesture alterations would cause subjects to rank the statements which most closely matched the metaphor(s) the alteration was meant to emphasize more highly, and that the original multi-metaphoric gesture would see statements representing all metaphors which appear ranked higher than distractor statements. Furthermore, we predicted that in more metaphorically ambiguous gestures, participants would rank seemingly contradictory statements that are associated with the active metaphors more similarly than in the alterations, which are designed to heavily emphasize one metaphor over another. One example of this may be ranking “There is tension in this group of people,” first, and “This group of people is working together,” second.

2.1.5 Experiment 2

Experimental Procedure

After preliminary analysis of the statement rankings above showed trends we expected, but non-significant results, we used the themes gathered from the pilot data in combination with the free-response and analysis of ranking data to generate 12 new statements to attempt to address gaps in

our statement-ranking for some gesture conditions (all statements for Experiment 2 can be found in Section 7.1.3). Participants were asked to rate their beliefs from 1-7, and again asked to fill out the free-response portion to check they attended to the video.

Hypotheses

We predicted that our alterations would push ratings to either higher or lower extremes when the statement closely matched the metaphor which the alteration emphasized. Specifically, we expected to find two things:

1. Increased scores for statements corresponding to the metaphors that are expressed in an alteration. Scores corresponding to distractor questions would not show significant differences across alterations of a single gesture. If all gesture conditions result in different interpretations for that statement, we expect to see 30-45 significant differences in total (3 conditions, 5 gestures, 2-3 statements per gesture).

2. If a metaphor is expressed in a gesture alteration, the original multi-metaphoric gesture would have similar scores for the same statement. If this is the case, we expect to see fewer significant differences between alterations, as the original multi-metaphor gesture would not differ significantly from one of the single-metaphor alterations. The alternative to this would be that the single metaphor conditions would show large concentrations in either high or low scores, but the combined metaphor would not, which would indicate the metaphoric message is lost in the combined gesture, and lead to significant differences in statement scores across all three gesture alterations.

We recognize that the relationship between metaphors and scores between gesture conditions is complex, and may depend on the degree to which the metaphors are conceptually related, and perhaps whether the physical attributes of each gesture relate to each other. If the metaphors are orthogonal in conceptual space, for example, then one would expect the difference between single-metaphor condition responses to be greater, but if they are conceptually associated this may co-influence their interpretation by the viewer, which could cause different response patterns than those hypothesized above.

2.1.6 Analysis and Results

For the first experiment, we calculated the difference between rankings for each statement for each condition to visualize how statement rankings are correlated with one another across conditions (Fig. 2.7, Top, complete results in Section 7.2). We analysed the impact of gesture condition on each statement ranking using a non-parametric 3x5 ordinal model as well as the impact of different gesture conditions using a Wilcoxon signed-rank test per statement. Although we observed trends which are visualized in Fig. 2.7, no statistically significant values were found.

For the second experiment, in order to interpret how gesture conditions affected participant scores, the density of score for each question across gesture conditions was visualized in a violin boxplot (selected graphs shown in Figure 2.7, Bottom). We used an ANOVA analysis followed by t-tests with Bonferroni correction (n=36 for 3 possible significant differences across 12 statements) to calculate significant differences in scores of statements meant to measure each metaphor across gesture conditions to understand which metaphor-gesture trends were significant. Significant values are reported in Table 2.1. This experiment finds 22 significant relationships, many for the condition and combinations we expect from the original gesture (explicit hypotheses found in Section 7.2). We see that out of a possible 180 combinations (3 conditions x 12 statements x 5 gestures), 22 represent significant differences in the scores between gestures for the same questions. One reason we believe more of our hypotheses were not confirmed is due to the metaphors each statement was intended to cover. Some metaphors had multiple statements that may apply (for example, *Conflict is Collision* could be described with either "Disagreement" or "Tension").

We see in Table 2.1 a variety of mixed results, with some questions resulting in significant differences between all three conditions, sometimes between only one condition and the original multi-metaphor gesture, and sometimes only between the two altered conditions. Note that only combinations for statements which were significant (with unadjusted p) are reported in the table.

Table 2.1: Significant Results of Experiment 2

statement	cond. 1	cond. 2	p	adjusted p	name
buttheads	collide	together	0.001	0.009	unpolite
buttheads	collide	original	0.001	0.012	unpolite
disagreement	collide	together	0.003	0.042	unpolite
disagreement	collide	original	0.001	0.089	unpolite
getalong	together	collide	0.03	0.391	unpolite
tension	collide	together	0.001	0.004	unpolite
tension	collide	original	0.000	0.001	unpolite
unified	original	collide	0.046	0.596	unpolite
buttheads	original	separated	0.009	0.121	entity
disagreement	original	separated	0.005	0.066	entity
disagreement	original	chest	0.009	0.128	entity
tension	original	separated	0.029	0.244	entity
buttheads	no_circle	original	0.022	0.311	audience
buttheads	palm_up	no_circle	0.028	0.296	audience
close	palm_up	no_circle	0.045	0.358	audience
goal	palm_up	no_circle	0.045	0.581	audience
tension	no_circle	original	0.003	0.047	audience
tension	no_circle	palm_up	0.013	0.165	audience
protected	original	frame	0.024	0.313	anything
protected	cover	frame	0.028	0.360	anything
worktogether	frame	cover	0.044	0.565	anything
worktogether	original	cover	0.017	0.224	anything

mean of condition 1 is larger than the mean of condition 2.

2.1.7 Discussion

Our results indicate multi-metaphor gestures lead to different interpretations from only atomic changes in gesture performance, but more investigation is required to understand the complexity of this phenomenon. Results are particularly clear in the case of the "Unpolite" gesture, which combines the

(a) Density of rankings by statement per condition

(b) Density of 1-7 scores by statement per condition

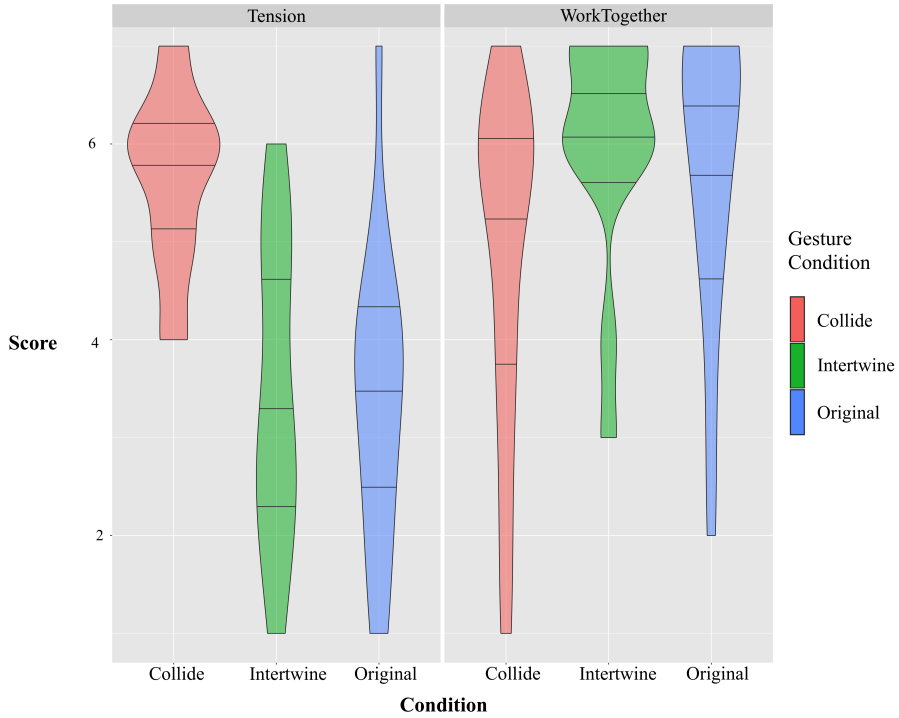


Figure 2.7: Statement Ranking distributions from Experiment 1 with ranking 1 being most applicable (top) and Score Distributions from Experiment 2 (bottom) for the "Unpolite"videos.

metaphors of *Combining Ideas is Physical Closeness* and *Conflict is Collision*. We see in the case of the original gesture (Fig.2.6a), statements about "Tension" and "Working Together" are ranked highly (Fig. 2.7, top) despite the fact that these two concepts are not obviously related. However with conditions which take away either of those metaphors' manifestations (intertwining fingers in Fig. 2.6c or physical collision in Fig.2.6b), we see statement rankings diverge. This supports our hypothesis that unrelated metaphors that are strongly interpreted in single-metaphor gestures are still recognizable when combined in multi-metaphor gestures.

In Experiment 2, we see the dominance of the "intertwine" motion pulling the 1-7 scores for the "Tension" and "Disagreement" questions towards the interpretation of *Combining Ideas is Physical Closeness* metaphor that is exhibited in the multi-metaphor condition (Table 2.1). The mean score for the original gesture is significantly lower than the "collide" condition for the tension question, but not significantly different from the "intertwine" condition for that question. However, the multi-metaphor gesture also shows more variation in scores than either single-metaphor condition. This suggests that these two metaphors are individually recoverable despite being combined into one gesture, but with some difficulty, further supporting our hypothesis in Experiment 1.

This also indicates that multiple nuanced thoughts can be simultaneously expressed and interpreted through exclusively nonverbal behavior, but suggests that one metaphor may show dominance over others in some cases, which supports Hypothesis 2 for this experiment. The implications of this finding for virtual humans is that models may want to creatively select, combine, and express multiple metaphors to perform in a single gesture, if the goal is to produce nonverbal behavior with the same degree of rich information as performed by humans. But, those gestures must take into account physical nuances which may cause them to be more heavily weighted than other co-occurring metaphors.

These effects naturally lead to three extensions of the current research. The first is to determine what physical attributes of a gesture specifically denote the presence of these metaphors. In this case, it appears the intertwining fingers carry the information of the group working together, whereas the collision between the hands indicates tension. However, in this case the hands themselves represent the group members. In the context of an overall metaphoric scene, how are physical properties of abstract objects related to metaphoric content, and how are these properties conveyed through physical gesture?

The second extension is around interpretation of metaphoric objects are used across time. How can metaphors combine across time in a gesture sequence to influence interpretation? How these metaphors interact temporally across objects in a scene may reveal clues about how humans use metaphors and narratives as a scaffold for thinking about abstract concepts.

Finally, are these metaphors universal? All of the above gestures were produced and decoded by speakers and viewers from the United States and United Kingdom, but if these gestures represent metaphors that are more broadly applicable to embodied cognition, then any viewer should have the same interpretation of these gestures. Uncovering who is able to decode these metaphors, and under what conditions, could lead to new discoveries about universal underpinnings of metaphoric gestures

in embodied cognition.

2.1.8 Ramifications for Computational Models

As we see in both experiments, multi-metaphoric gestures affect viewer interpretation of the message being conveyed. However, despite some success modeling single-metaphoric gestures, modeling such creative gestures presents considerable computational challenges both in terms of gesture animation and generation from underlying communicative intent.

The virtual human community has long been aware of gesture as an essential channel of communication, and there has been some progress on attempts at both inferring and realizing metaphoric gestures. Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro (2013) sought to infer possible mental states that underlies the spoken dialog's text and prosody, which allowed it to make inferences about metaphors not directly represented in the text, e.g., that political ideologies, collections of ideas, etc can be treated as physical objects with physical properties that convey abstract concepts. Lhommet and Marsella (2013) present a logical schema-based approach to inferring and conveying multiple metaphors across gestures, potentially for use in pre-defined ways within gestures. For example, if a set of abstract concepts is depicted as a container, additional gestures can be used to remove or add concepts to that set. Xu, Pelachaud, and Marsella (2014) demonstrate an approach to synthesizing the animation of sequences of metaphoric gestures within *ideational units* (Section 1.1.4). Ravenet, Pelachaud, Clavel, and Marsella (2018) extract image schemas (Cienki, 2005) from surface text to dynamically generate metaphoric gestures. However these approaches fall short of modeling a more open-ended creative process discussed below of multiple metaphors brought together and conveyed within and across gestures.

If one's goal in understanding these gestures is to eventually model and realize them in a virtual human capable of intelligent interaction, then the richness of these examples raise several daunting challenges. There is an inference task of knowing when the gestures should be used, in effect which metaphors need to be conveyed from moment to moment. In addition, there is the task of constructing the animation that conveys those metaphors.

With respect to inference, a traditional approach is to make inferences from the spoken dialog to generate nonverbal behavior (Cassell, Vilhjalmsso, and Bickmore, 2001; Lee and Marsella, 2006). However there is often not a simple inference from dialog to metaphoric gestures. In the "Audience" example (2.3), the speaker's sentence does not explicitly state that "democrats" should be separated physically from "republicans" and "independents", to indicate different political beliefs. Contrast that with a phrase like "put that idea aside" where the sentence clearly indicates a linguistic metaphor. When there is such a mapping, one can rely on mapping from metaphors in the text to metaphoric gestures, perhaps using recent advances in ML-based metaphor detection in language to drive the gesturing (Leong, Klebanov, and Shutova, 2018).

Constructing an animation of a single gesture from multiple metaphors raises additional challenges. A simple approach is handcrafting the physical properties of a compound metaphor ges-

ture using pre-specified parameters for a procedural animation system (e.g. Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005) or for a keyframe/motion capture animation system (e.g. Thiebaux, Marsella, Marshall, and Kallmann, 2008). However such approaches fail to capture what appears to be a product of a creative, embodied-cognition process where metaphors may be dynamically combined to form a gesture in line with those cognitive processes. Approaches that assume a fixed set of multi-metaphor gestures imply the underlying metaphors and associated cognitive processes are themselves constrained. Keeping in mind that such gestures are mapped many-to-many with linguistic expressions, the variety of subtle changes in gestural performance is still difficult to capture in a fixed-animation model. If instead we want to address this creative process in a more open-ended fashion, we may need a deeper model of the relation of a gesture's form and movement over the stroke of the gesture, and how the form and movement at different points in the gestural motion may convey different metaphors.

Furthermore, complex gestural scenes raise still other challenges. These sequences involve gestures unfolding over time that exploit space in order to convey abstract concepts. For example, the audience (a set), the speaker (physically present), and different subsets of political orientations (subsets of the audience) that are spatially separated from each other but within the region of the audience. Furthermore, this scene presents discontinuous elements in the utterance that are related to each other. The hand sweep (Fig. 2.4, left) appears to set up the audience location in gesture space, relates that set to the speaker (Fig. 2.4, middle) and then goes back to that audience space to make subsequent reference to, to essentially enumerate, the republicans, independents and democrats in that space (Fig. 2.4, right).

2.1.9 Conclusion of Study 1

In this section, we illustrated the rich structure of metaphoric gestures, both in terms of multiple metaphors conveyed in a single gesture, and sequential gestures setting up metaphoric scenes. Two experiments confirmed some of our hypotheses around how multiple metaphors are simultaneously conveyed through combining a gesture's physical properties, and produced avenues for future research. We discussed the challenges of implementing these multi-metaphor gestures with computational models for use in embodied agents. Ultimately, this raises additional questions around the creative process of going from abstract thought to communicative gesture performance.

2.2 Study 2: Cross-cultural interpretations of multi-metaphoric gestures

Metaphoric gestures influence viewer perception of information conveyed in social interaction. Some of this influence is attributed to a shared interpretation of physical metaphors underlying the gesture, grounded in embodied cognition. Although many gestures have been shown to be based in culture,

it is possible that due to the physical metaphor, this particular type of gesture is cross-culturally interpretable. In this section we explore the extent to which two cultures that are known to differ in their use of gesture are able to recover similar information in context-free metaphoric gestures. We use naturally-occurring complex metaphoric gestures with origins in either culture along with simplified manipulations, and crowdsource judgements to discover how semantic interpretations of these gestures vary between cultures. We find a complex pattern of differences and similarities between American and Japanese interpretations, and discuss how future experiments may further reveal what causes people to ascribe semantic interpretations to gesture characteristics across cultures.

2.2.1 Introduction

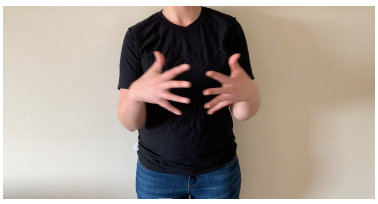
Most gestures have been shown to be largely non-culturally-transferable (Calbris, 1990; Kendon, 2004). For emblems, which are rote linguistic replacements for speech, this is intuitive; just as the same sound holds different meanings in different languages, so too do specific gestures mean different things across cultures. One example of this is the "thumbs-up," gesture, which in the western world is indicative of a job well done, but in many Middle Eastern and Asian cultures, is considered highly offensive. Aside from emblems, the amount of the different types of gesture varies significantly between cultures (Kita, 2009), and bi-cultural individuals commonly switch their gestural styles when switching languages (Cavicchio and Kita, 2013).

However, metaphoric gestures may still be culturally transferable in their interpretations. Gestural motion is abstract and reminiscent of the physical world, which all humans experience in largely the same way. Despite cultural differences, our shared physical embodiment may allow us to interpret physical metaphors cross-culturally, even though performances of those specific metaphoric gestures differ. The present question is: is there something inherent about the physical motion of a gesture that consistently evokes similar patterns of conceptual interpretation, even across cultures?

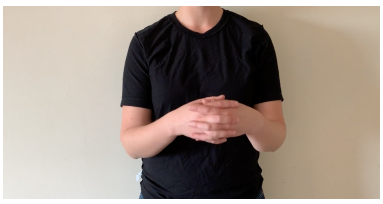
Here we present a preliminary study that examines this question of shared understanding through physical embodiment. We asked viewers of different cultures to interpret metaphoric gestures created by their own culture and by a culture with different gesture behavior. We use multi-metaphoric gestures, which seem to display two physical metaphors within one gesture, as well as versions of those gestures in which one element of that gesture has been changed to emphasize one metaphor over the other. In this case, we observe groups from the United States and Japan. Although the production of these physical metaphors may be different, there is reason to believe all participants will form similar interpretations of each gesture, as it is representative of an abstract notion that has a physical counterpart (Kita, 2009).

2.2.2 Experiment

While multi-metaphoric gestures are shown to result in complex interpretations, some information can be reliably conveyed to a single culture (Saund, Roth, Chollet, and Marsella, 2019) (Section 2.1



(a) Original Gesture (W1-0). The actor stands with hands raised to the elbow, palms inward and fingers spread, and forcefully moves the hands together and apart, interlacing the fingers when they come together, three times.



(b) Manipulation 1 (WG1-1), designed to increase "Unity." The actor stands with hands raised to the elbow, palms inward and fingers spread, and brings the hands together, interlocking the fingers, one time.

(c) Manipulation 2 (WG1-2), designed to increase "Conflict." The actor stands with hands raised to the elbow in closed fists, and forcefully moves the hands together and apart, maintaining a fist shape, three times.

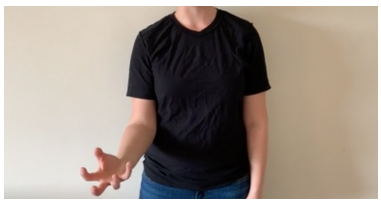
Figure 2.8: Screenshots from Western Gesture Set 1

in this thesis). What is less clear is the extent to which this information is interpretable to individuals from different cultures. In this section, we explore the possibility that cultures with different gesturing behavior could nevertheless interpret metaphoric gestures similarly to one another.

Although many gestures are non-culturally-transferable, it is theoretically plausible that metaphoric gestures will convey similar information to different cultures due to embodied cognition arguments. Embodied cognition is the idea that physical presence and grounding is integral to the way that humans organize and understand abstract information. This leads us to believe the shared experience of a physical body will lead individuals in different cultures to interpret metaphoric gestures in the same way for two reasons:

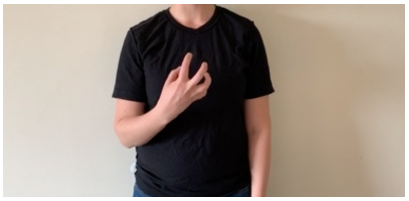
1. Abstract concepts can be represented in the mind, regardless of culture, by physical characteristics, and

2. Concepts that are represented by physical dynamics should be evoked using similar gestural motion dynamics for different groups of people.



(a) Original Gesture (W2-0). The actor stands with one hand raised to the center of the chest, with the hand in a rigid cupped position. They unbend the elbow, bringing the hand away from the chest, while maintaining the rigid cupped hand.

(b) Manipulation 1 (WG2-1), designed to increase "Openness." The actor stands one hand raised to the center of the chest, with a soft, open palm. They unbend the elbow, bringing the hand away from the gest, revealing an open palm in front of them.



(c) Manipulation 2 (WG2-2), designed to increase "Control." The actor stands with one hand raised to the center of the chest, with the hand in a rigid cupped position. They maintain this position with a soft beat at the time when the other gestures unbent the elbow.

Figure 2.9: Screenshots from Western Gesture Set 2

Method

To compare interpretations across cultures, we recruited two participant groups: one from the United States, and one from Japan. For purposes of this experiment, we will be referring to participants from the United States as "western," and participants from Japan as "eastern."

Participants

All western participants were recruited from Amazon Mechanical Turk, and all eastern participants from CrowdWorks. 127 western participants and 23 eastern participants were removed due to incomplete responses (n=1394 western participants, n=1174 eastern participants). The number of participants for each experimental condition ranged between 30-56. A complete breakdown of number of participants for each condition can be found in Section 7.2.1.

Stimuli

We used some of the stimuli found in Saund, Roth, Chollet, and Marsella (2019) (Section 2.1.1 of this thesis). Gestures were found by researchers by watching interviews with both trained speakers and members of the public on youtube.com in English and Japanese. Stimuli were generated by finding multi-metaphoric gestures; that is, single gestures which seem to be composed of two individual metaphors coming together. Gestures were selected by their clarity in simultaneously conveying multiple metaphors, which was determined by the surrounding context of the video. These multi-metaphoric gestures became the "Original" gestures from which manipulations were created. For the Japanese gestures, researchers identified 20 candidate gestures using translations and subtitles, then worked with native Japanese speakers to develop this context and understand what metaphors were conveyed in the original gesture. Eventually five gestures were selected from U.S. speakers, and five from Japanese speakers.

We then created two manipulations from each Original gesture, designed to push interpretation to raise or lower scores in specific Response Domains (Figure 2.8 and 2.9, Table 2.2). These manipulations were designed to influence interpretation in a particular way by manipulating elements known to be salient carriers of meaning within gesture (McNeill, 1992; Kendon, 2004). Each manipulation changes one atomic element of the gesture (i.e. the hand shape, or the direction of an arm sweep) as individuals typically change as little as they can about their gesture to convey a new concept (Calbris, 2011).

Our original gestures include five found within the western population, and five within the eastern population. This, along with the two additional manipulations for each gesture, created 30 stimuli. We call this group of an "Original" gesture and its two manipulations a "Gesture Set." An actor performed all 30 of these gestures in a neutral setting. The videos of the gestures do not have any sound, or a view of the actor's head, so no facial features may influence the interpretation of the gesture, and the actor's gender is effectively obscured. We highly recommend viewing all stimuli gestures found in the link in the Supplementary Materials (Section 7.2).

Across all stimuli, we identified six key Response Domains (Table 2.2), only a few of which applied to any particular gesture or manipulation. For brevity, we here explore only a few Gesture Sets and associated differences. Full analysis of all results can be found in the Supplementary Materials (Section 7.2). As such, this experiment is meant to demonstrate exploratory analyses and hypotheses, and does not provide generalizable evidence of the effects found.

Procedure

Participants were given instructions to view a 1.5 second video clip of an actor performing a gesture with no sound. This was accompanied by the prompt "This video is of a speaker who is mid-conversation, currently talking about a group of people." As a data-quality check, participants wrote at least 10 words about what they believe the speaker to be saying, although the video had no sound and the speaker's face was out of frame. Following this, participants filled out a 12-question survey on

Likert Item	Response Domain
This group is made up of many people	size
There are many members of this group	size
This group of people is experiencing conflict	conflict
There is tension in this group of people	conflict
This group of people is open to outsiders	openness
Non-members find this group accessible	openness
This group of people is tightly controlled	control
Someone is definitively dominant over this group of people	control
This group of people is working together	unity
There are common unifying goals within this group of people	unity
This group is very sure in their decisions	certainty
The actions of this group are taken confidently	certainty

Table 2.2: Questionnaire and Response Domain groupings administered to participants. Responses to questions are validated; correlations between responses to each question for each culture can be found in Figure 2.12.

their interpretation in six different Response Domains using a 1-7 Likert scale (anchors: 1 = "strongly disagree", 7 = "strongly agree") The full set of questions is found in Table 2.2. The questions were developed using a pilot in which participants wrote free-response answers to "She is talking about a group of people. What do you think she is saying?" These free-response answers were then coded into Response Domains, from which the survey questions were derived.

The experiment was translated into Japanese by a professional translator who worked with the experimenters to capture the nuances of the phrasing of particular questions. The survey was then reviewed by two native Japanese speakers for ease-of-comprehension and clarity.

Hypotheses

As there is little previous work done in cross-cultural universality of metaphoric gestures, this experiment is largely exploratory. We hope to compare cross-cultural and individual patterns of interpretation across multi-metaphoric and simplified gestures to reveal the extent to which different groups of people use similar physical dynamics to represent abstract concepts.

Given the theoretical grounding of this experiment in the universality of embodied cognition, we approached this experiment with two main expectations:

H1 Conceptual interpretation of metaphoric gestures is grounded in the same motion dynamics across cultures. The strong version of this claim predicts that we will find no differences in interpretations between cultures for any Response Domains, for any of these gestures.

H2 Furthermore, in each manipulation of the original gesture, the manipulations will result in differences in interpretation in the same direction for both cultures, across all Response Domains

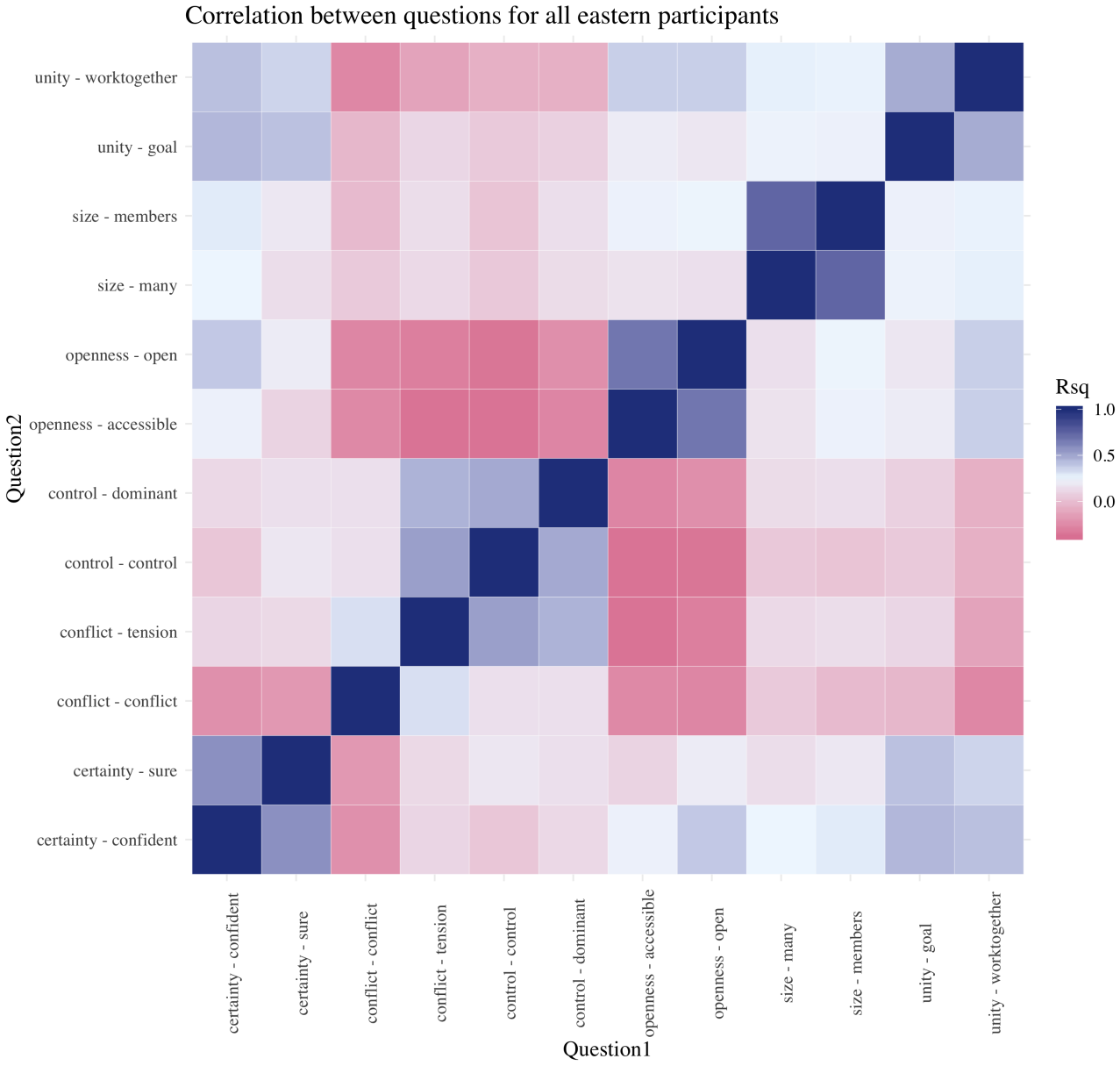


Figure 2.10: Response Domain correlation for only eastern viewers.

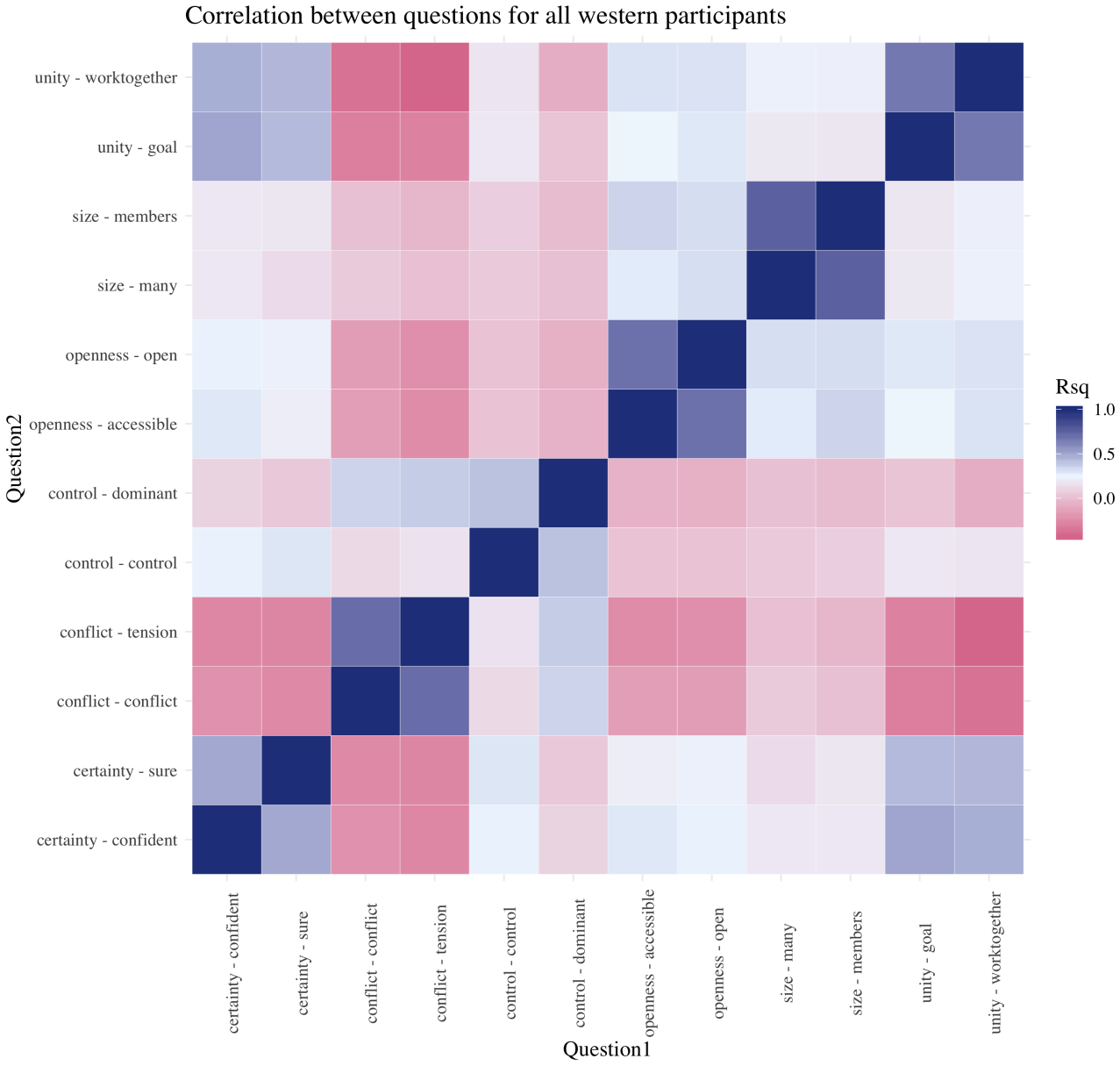


Figure 2.11: Response Domain correlation for only western viewers.

(listed in Table 2.2).

2.2.3 Results

Here we report a selection of results from the above experiment, and explain our reasoning behind and analysis of those results. A full table of results, all figures, and analysis scripts can be found using the link in the Supplementary Material (Section 7.2).

We took a variety of steps to strengthen the cross-cultural validity of the Likert scale. This was used to investigate our two hypotheses: that each gesture would be interpreted similarly between cultures (H1), and that each manipulation would have a similar effect for both cultures (H2).

Cross-Cultural Validity

Multiple steps were taken in order to ascertain how much of the response variation between cultures is due to general cross-cultural differences in Likert-style responses. In line with previous work in this area, we used parceling to combine related questions into conceptual measures we refer to as Response Domains (He, Vijver, Fetvadjiev, Carmen Dominguez Espinosa, Adams, Alonso-Arbiol, Aydinli-Karakulak, Buzea, Dimitrova, Fortin, et al., 2017) (in line with terminology from Section 2.1.1). These groupings were checked using correlation between the related item responses to ensure they are appropriate to be grouped into their respective Response Domains (Figure 2.12). For each of these, the items which compose the group are the most highly correlated with one another. The highest item-pair correlation for both cultures is the size-member pair (r^2=0.82 in western, 0.78 in eastern), the lowest item-pair correlation for western participants is the control-dominant pair (r^2=0.41), and the conflict-tension pair in eastern participants ($r^2 = 0.35$).

Previous work suggests Japanese respondents to Likert scales tend to use an Average-Response Style (ARS), and American respondents tend towards Extreme-Response Styles (ERS) (Lee, Jones, Mineyama, and Zhang, 2002). Despite trends in these directions, we do not find a significant effect of this in our results. Both the eastern and western participants showed an Average-Response Style (ARS), with eastern participants responding with the middle response 24.5% of the time, and western participants 22.5% of the time (14% expected on 7 point Likert scale). Additionally, neither participant group showed an Extreme-Response Style (ERS) (eastern participants answered 1 or 7 12% of the time, western participants 16% of the time, 28% expected). Scores were distributed similarly between cultures, and the differences between cultures in the number of responses for each score was no different from uniform (p=0.12). There also was not a significant difference in the standard deviation of individual Likert scores between groups (σ=1.67 for western, σ=1.54 for eastern). This suggests normalizing results is not necessary (Baumgartner and Steenkamp, 2001).

We bucketed Likert responses into “Low,” (1-3) “Neutral,” (4) and “High,” (5-7) agreement to the items. While this is a broad range, we see these results through the lens of the gestures being interpretable or uninterpretable by different groups, even if that interpretation is only mild. Additionally,

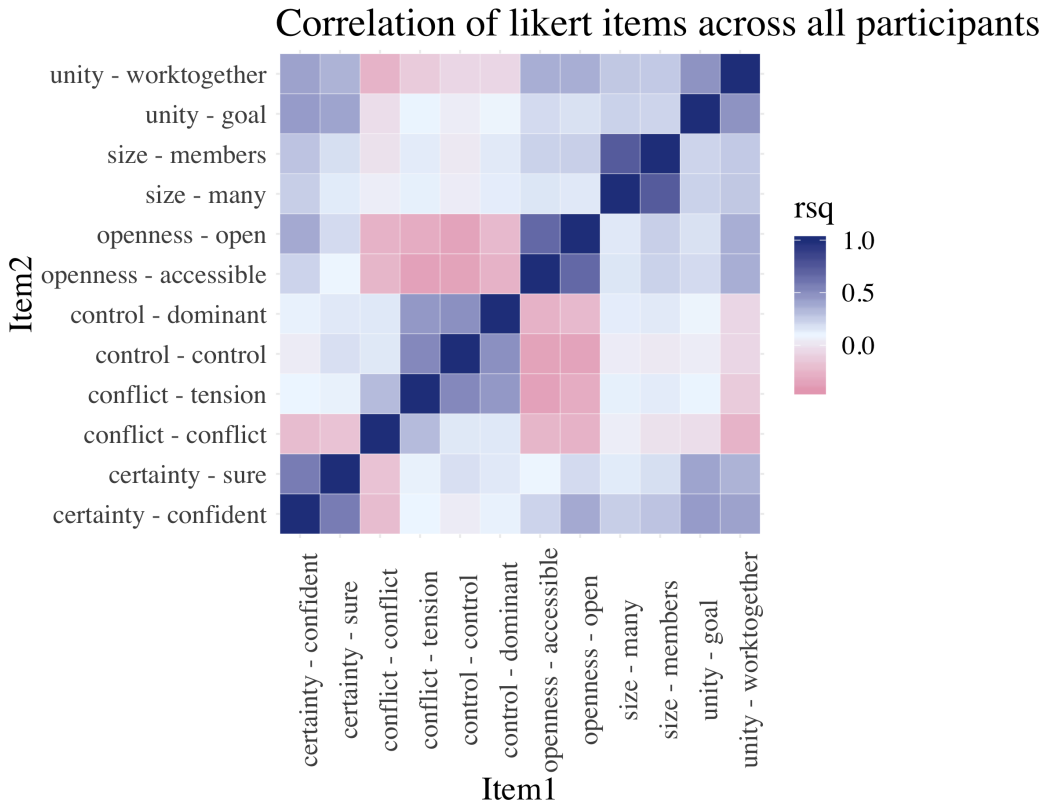


Figure 2.12: Correlations (r^2) between questions for both cultures across all gesture conditions.

the neutral response is seen as an important finding in this experiment, and so maintaining a narrow range for the middle option is as important as capturing the entire range when the viewer did have an interpretation. So, for the purposes of this analysis, a mild interpretation (High/Low) of a gesture is an interpretation nonetheless.

Between-Culture Comparisons

Contrary to H1, we found a substantial number of differences between the ways that each culture interpreted the various gestures, and our manipulations of those gestures.

We count significant differences between cultures as follows: if at least one Score Bucket is proportionally significantly different between the western and eastern cultures for a particular Response Domain, then that is seen as a difference in interpretation of the gestures between the cultures. Because there are 10 Gesture Sets (each with 3 gestures), and 6 Response Domains per condition, there are potentially a total of 180 significant differences between the cultures. Significant differences in proportion were determined using a two-proportions z-test, and applying Bonferroni adjustment within each Score Bucket for each of the Response Domains (n=6 alternative hypotheses). Full results of the ANOVA analysis to find significant effects of our gesture manipulations and of cultural interpretations is found in Table 2.3.

This method gives a total of 42 significant differences between cultures. This moderately refutes

Gesture ID	Factor	Openness	Conflict	Unity	Control	Certainty	Size
WG1	Manipulation	<0.001	<0.001	<0.001	0.185	0.246	0.360
	Culture	0.820	0.819	0.0912	<0.001	<0.001	0.233
WG2	Manipulation	0.417	0.087	0.343	0.737	0.271	0.008
	Culture	0.002	0.0837	0.009	0.154	0.056	0.011
WG3	Manipulation	<0.001	0.209	0.003	0.036	0.129	<0.001
	Culture	0.010	0.004	<0.001	<0.001	<0.001	0.012
WG4	Manipulation	<0.001	<0.001	0.0586	0.0704	0.00441	0.292
	Culture	0.107	0.265145	0.1639	<0.001	0.4652	<0.001
WG5	Manipulation	0.9566	0.0135	0.279	0.0535	0.743512	0.116
	Culture	0.0407	0.5635	<0.001	0.0887	0.000908	<0.001
EG1	Manipulation	<0.001	0.023	0.514	<0.001	0.005	0.535
	Culture	0.096	0.414	0.563	0.002	0.605	0.144
EG2	Manipulation	0.011	<0.001	<0.001	0.056	0.117	0.107
	Culture	0.758	0.769	0.429	0.034	0.338	0.003
EG3	Manipulation	0.213	0.513	0.595	0.255	0.273	0.603
	Culture	0.492	0.007	0.001	0.095	0.088	0.009
EG4	Manipulation	0.397	0.003	0.388	0.136	0.735	0.681
	Culture	0.001	0.280	<0.001	<0.001	0.696	<0.001
EG5	Manipulation	0.145	0.010	0.144	0.744	0.273	0.330
	Culture	0.023	0.066	0.041	0.15	0.090	0.452

Table 2.3: ANOVA results (F value) for all Gestures across cultures. Two-way ANOVAs were performed examining the effects of manipulation and viewer culture on each Response Domain.

RD	High	Med	Low	Total	HWM	HEM
certainty	5	1	1	7	7	0
conflict	4	3	5	12	4	8
control	7	2	7	16	14	2
openness	5	1	3	9	9	0
size	3	0	4	7	7	0
unity	5	4	3	12	11	1

Table 2.4: Number of differences in each Response Domain (RD) for High, Medium, and Low categorical responses. Of those differences, this table also indicates how many differences were due to a Higher Western Mean (HWM) or Higher Eastern Mean (HEM).

H1, which states that we will see no differences between the cultures. Interestingly, we see significantly fewer differences in the complex Original gestures than in the simplified manipulations (8 in the Original, 34 in the two manipulations, p=0.04).

Across all gestures, no particular Response Domain had significantly more or fewer differences between cultures; it does not seem to be the case that one culture was more apt to interpret any Response Domain than the other. In both cultures, we rarely see a majority neutral response in any Response Domain. Within individual responses, there was no clear pattern of responses between cultures across gestures. That is, the origin of the gesture did not seem to make that gesture "make sense," to one group over the other.

Gesture Manipulation Effects

Here we compare how each manipulation affected viewer interpretation in different Response Domains in each Gesture Set, across cultures. Within each culture, our manipulations caused a variety of different effects, only a few of which were similar across cultures (Figure 2.14), and sometimes even show opposite effects for the two cultures (Table 2.5 and Figure 2.15). This provides evidence against H2, which states that manipulations would have the same effect on viewer interpretation regardless of viewer culture.

Figures 2.13, 2.14, and 2.15 together depict some examples of cross-cultural comparison of counts of scores in each Score Bucket for all Response Domains for three specific gestures (all results can be found in the Supplementary Material in Section 7.2). Scores were bucketed to highlight the semantic meaning of the Likert scale: the gesture suggested a particular Response Domain, was neutral towards that Response Domain, or was not at all reflective of that Response Domain.

Patterns Within Individual Interpretations Across Cultures

While we observed a variety of differences in aggregated cultural interpretations across Response Domains, post hoc analyses reveal nuanced patterns of individual interpretations across cultures. Specifically, across many Response Domains and gesture manipulations, eastern viewers tended to interpret

multiple concepts (response of 4 or higher) from more gestures than western viewers. Preliminary results for one Gesture (WG1, Figure 2.8) are shown in Table 2.6. While this is not a focus for the current work, it is an interesting pattern in gesture interpretation that is suggestive of deep cognitive and/or cultural differences in interpreting meaning from gesture. This result further emphasizes the need for cultural sensitivity when designing gesture behavior in virtual agents, and illustrates a promising future avenue of exploratory research.

2.2.4 Discussion

Our results provide evidence which refutes the strong version of both of our original hypotheses: not only did individuals in each culture interpret some gestures differently (refuting H1), but our manipulations within those gestures affected their interpretations in different ways (refuting H2). However,

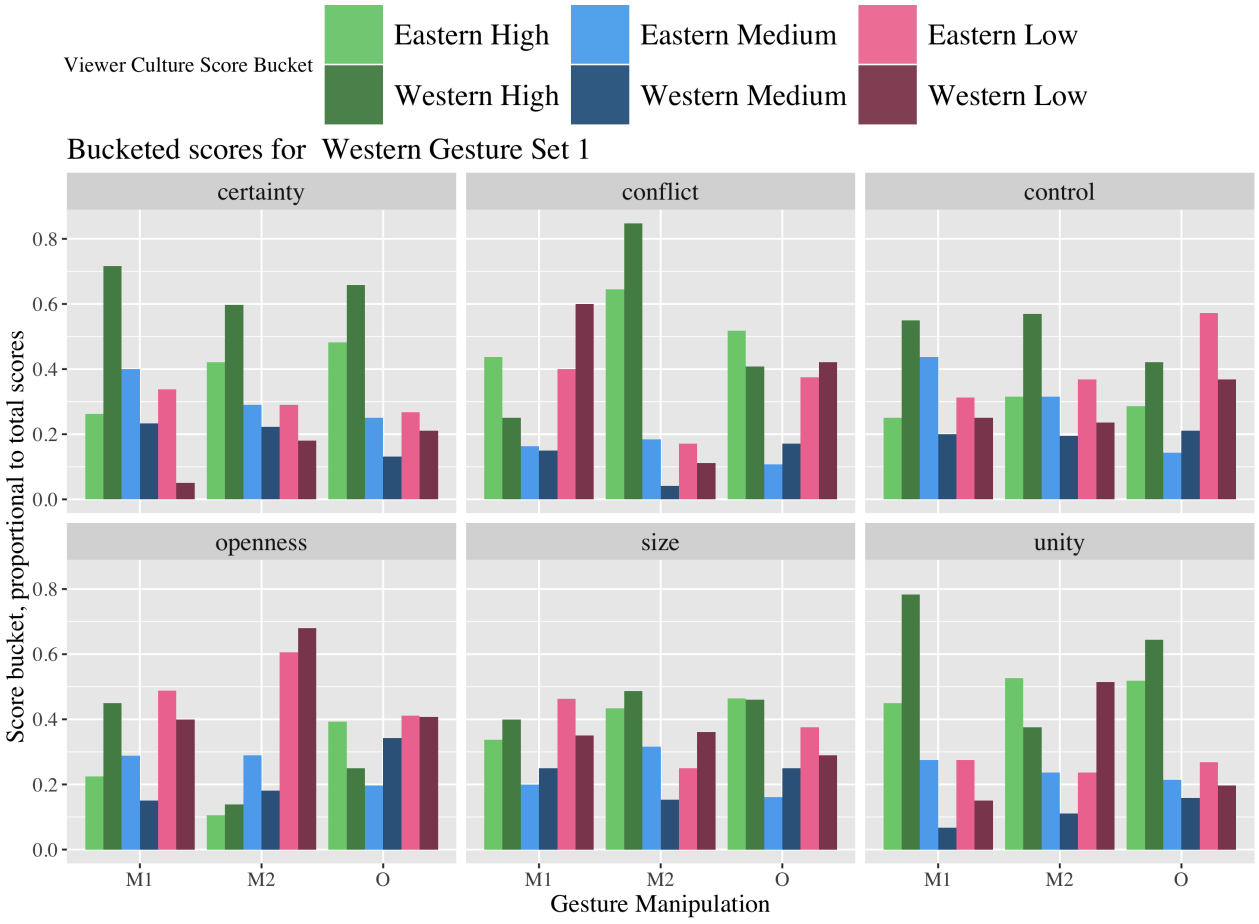


Figure 2.13: Western Gesture 1 Bucketed Results. Focusing only on the "Conflict," results, notice how Manipulation 2 significantly increases the interpretation of "Conflict" in Western participants, but not Eastern. Alternatively, if one focuses only on each culture's interpretation of "Unity," it is clear that Manipulation 1 significantly increased Western participants' perception of this concept compared to the original gesture, but not Eastern. "High"=5-7, "Medium"=4, "Low"=1-3. "M1"=Manipulation 1. "M2"=Manipulation 2. "O"=Original gesture.

other gestures and manipulations followed our predicted trends. These differences and their causes were at times subtle, and paint a complex picture of the way individual elements of a gesture interact to drive interpretation. The implication for gesture researchers, and in particular for generative gesture modeling, is that physical attributes of a gesture (such as the synchrony or velocity of movements) are not necessarily universally indicative of a particular concept.

Between-Culture Differences

Some gestures showed similar response patterns across certain Response Domains, indicating that some motion features may be more universally recognized as indicative of particular concepts. For example, when we look at the responses across manipulations for Eastern Gesture Set 1 (Figure 2.14, video of this gesture can be found in link in Section 7.2), in all three manipulations we see both cultures strongly interpret "Unity" in all of the gestures, suggesting perhaps something about the position of all three gestures – both hands palm-up in front of the speaker, regardless of open or closed fist – incites

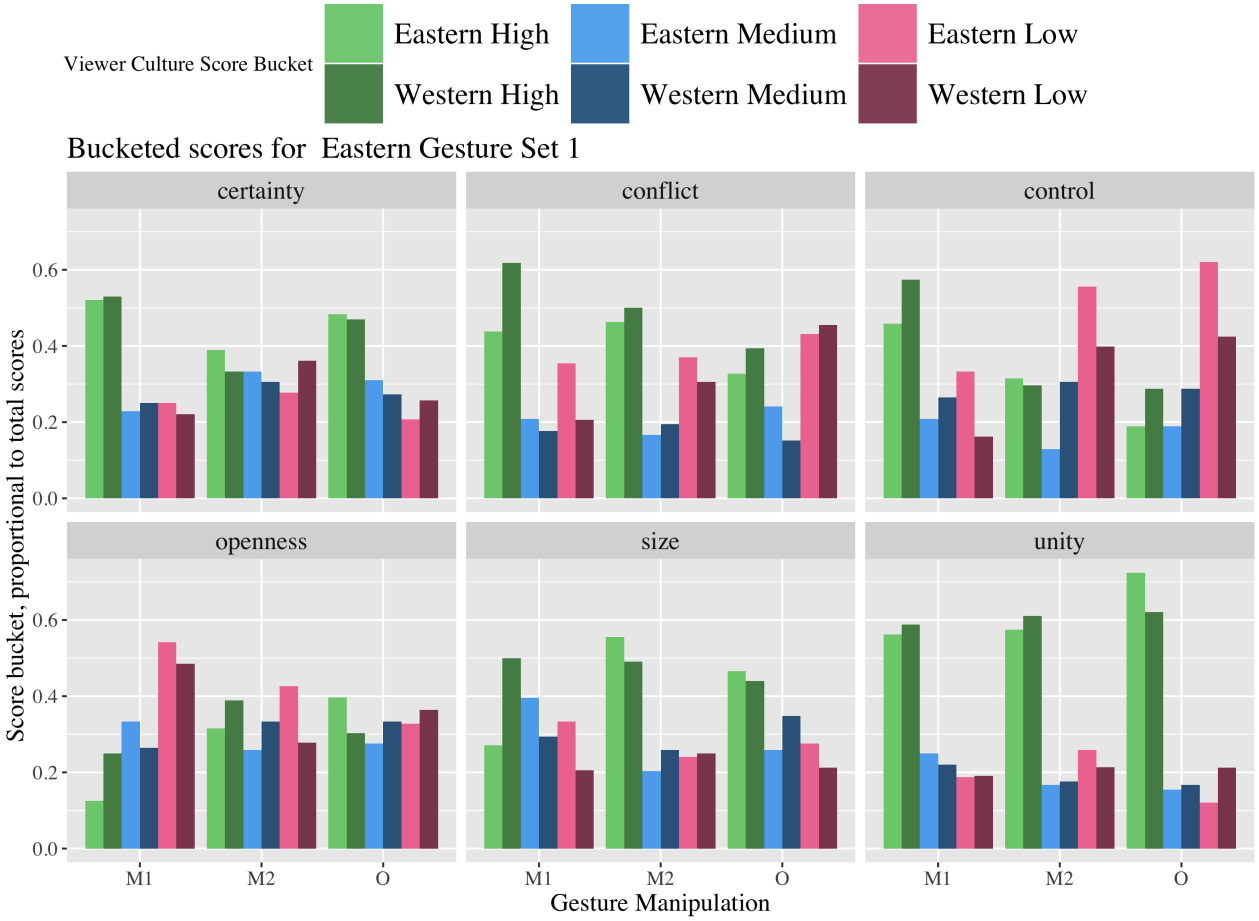


Figure 2.14: Eastern Gesture 1 Bucketed Results. Contrary to Figure 2.13, when we focus on the "Control" quadrant, notice how Manipulation 1 significantly affects interpretation of this concept for both cultures, and that this effect pushes both cultures in the same direction. "High"=5-7, "Medium"=4, "Low"=1-3. "M1"=Manipulation 1. "M2"=Manipulation 2. "O"=Original gesture.

common interpretation across cultures.

However, in that same Gesture Set the wide variance in response patterns suggests the opposite is simultaneously true: certain motion features indicate concepts for one culture, and not for another. In contrast to the previous example, we see trends that suggest the "give and take," motion of the original gesture and Manipulation 2 are indicative of low amounts of "Control" over a subject for eastern viewers, but do not seem to follow such a pattern in western viewers. And yet, Manipulation 1 in the same gesture affects viewer interpretation of "Control" similarly for both cultures.

We also observe more differences between cultures in the manipulations than in the original complex gestures. It may be the case that the original gestures were harder for individuals to interpret than the manipulations. In this case, differences between cultures could be outweighed by the variation in interpretation within cultures. Seeing more between-culture differences in the manipulations

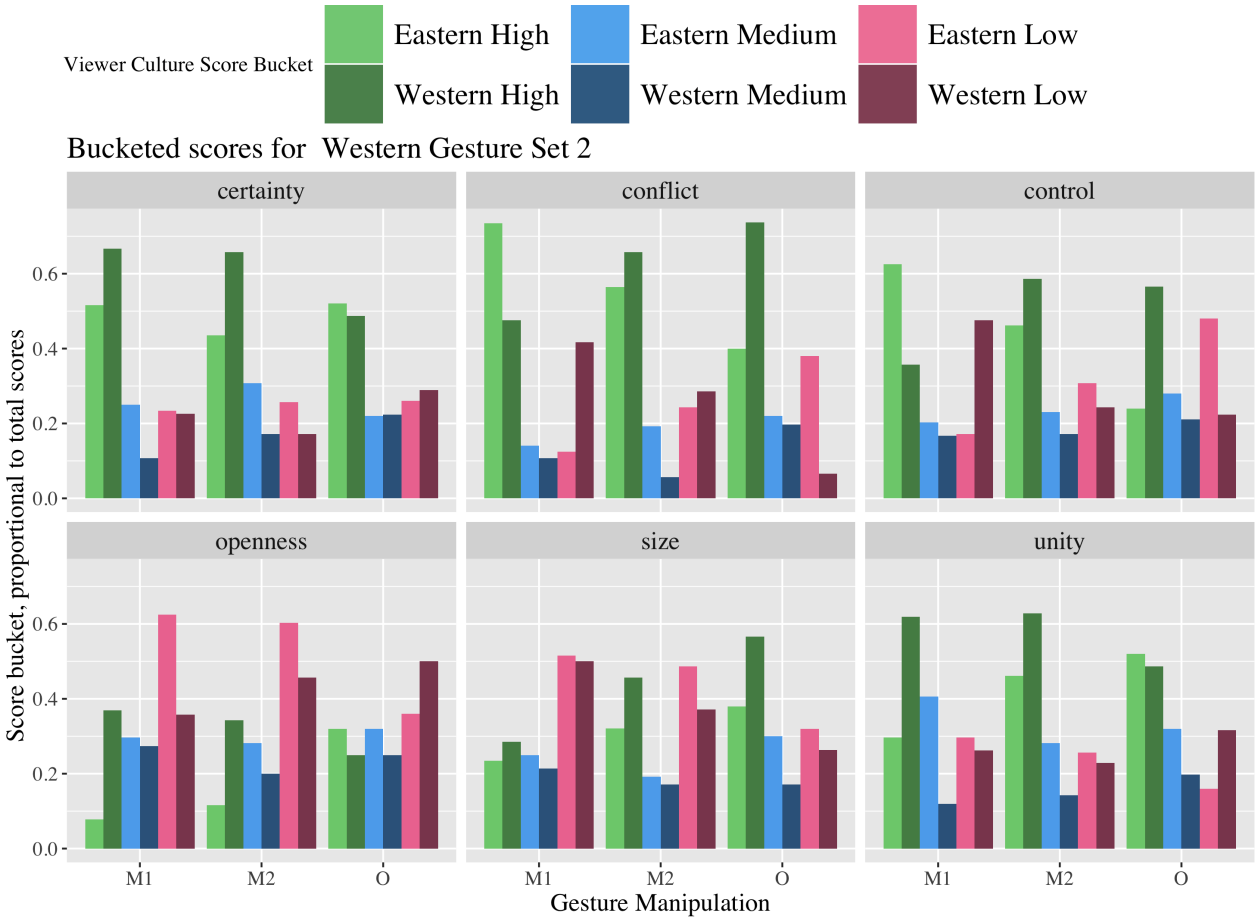


Figure 2.15: Western Gesture 2 Bucketed Results. When we focus on the quadrant for "Conflict," notice how Manipulation 1 significantly affects how each culture interprets each of this concept from the gesture, but in opposite directions. That is, Manipulation 1 caused Eastern viewers to see much more conflict than in the original gesture, but much less in Western viewers. This indicates that some minor manipulations in the form of a gesture severely impacts cultural impact in ways that may have opposite meanings in different cultures. "High"=5-7, "Medium"=4, "Low"=1-3. "M1"=Manipulation 1. "M2"=Manipulation 2. "O"=Original gesture.

	ID	VC	RD	OM	p-1	p-2
1	WG1	western	certainty	4.92	0.28 ↑	**0.05** ↓
2	WG1	eastern	certainty	4.27	0.24 ↓	0.74 ↓
3	WG1	western	conflict	3.86	**0.03** ↓	0.44 ↑
4	WG1	eastern	conflict	4.16	**0.00** ↓	**0.00** ↑
5	WG1	western	control	4.21	0.48 ↑	0.53 ↑
6	WG1	eastern	control	3.64	0.08 ↑	0.28 ↑
7	WG1	western	openness	3.71	0.50 ↑	**0.02** ↓
8	WG1	eastern	openness	4.11	**0.00** ↓	**0.00** ↓
9	WG1	western	size	4.28	0.84 ↓	0.22 ↑
10	WG1	eastern	size	4.18	0.96 ↓	0.94 ↑
11	WG1	western	unity	5.11	0.30 ↑	0.36 ↓
12	WG1	eastern	unity	4.54	**0.00** ↓	0.53 ↓
13	WG2	western	certainty	4.28	**0.04** ↑	0.75 ↑
14	WG2	eastern	certainty	4.54	**0.01** ↓	0.38 ↓
15	WG2	western	conflict	5.49	**0.00** ↓	**0.00** ↓
16	WG2	eastern	conflict	3.94	0.06 ↑	0.08 ↑
17	WG2	western	control	4.62	**0.01** ↓	**0.00** ↓
18	WG2	eastern	control	3.48	0.94 ↑	**0.05** ↑
19	WG2	western	openness	3.38	**0.01** ↑	**0.00** ↑
20	WG2	eastern	openness	3.78	0.31 ↓	**0.01** ↓
21	WG2	western	size	4.42	**0.01** ↓	**0.01** ↓
22	WG2	eastern	size	4.12	0.37 ↓	**0.04** ↓
23	WG2	western	unity	5.26	0.08 ↑	**0.03** ↑
24	WG2	eastern	unity	4.54	**0.01** ↓	0.39 ↓

Table 2.5: A selection of comparisons of within-culture differences between manipulations, grouped by eastern and western viewers. ID=the original gesture ID; VC=Viewer Culture; RD=Response Domain; OM=Original Mean (the mean of the original condition); p-1= the p value of the significant difference between the original gesture and manipulation 1 using a t-test, as well as whether the score for manipulation 1 was higher or lower than that of the original gesture, corrected for testing multiple hypotheses; p-2=the same, but for manipulation 2. Significant p values shown in **bold**. This shows the results only from two selected gestures. Full results and figures may be found in the link in Supplementary Materials (Section 7.2).

Gesture Manipulation	VC	Unity	Conflict	Both	Neither
Original	western	48	19	30	3
	eastern	22	24	50	4
Intertwine	western	63	5	15	17
	eastern	20	24	41	15
Collide	western	20	47	29	4
	eastern	18	21	58	3

Table 2.6: Percentage of participants in each culture who interpreted Unity, Conflict, Both concepts, or Neither concept from the indicated gesture manipulation. VC=Viewer Culture. Notice the consistently higher number of eastern viewers who interpret Both concepts across manipulations.

clarifies differences in the way that the particular motion dynamics expressed in those gestures are representative of particular concepts between cultures.

This does not necessarily mean that individuals from different cultures use different underlying metaphors to compose thoughts. It could be the case that two individuals are familiar with a particular physical metaphor, but their interpretation of some element of that metaphor differs. For example, two individuals could agree that "Good is Up," (Grady, 1997) but disagree on what motion dynamics constitute "up."

Gesture Manipulations

The manipulations resulted in a wide variety of differences between cultures, a selection of which are shown in Table 2.5. From the table, we see that manipulations often resulted in similar differences between cultures, but sometimes pushed interpretations in opposite directions. For example, in rows 3 and 4, we see Manipulation 2 in WG1 pushing interpretations of "Conflict," in the gesture with similar effects between cultures. However, looking at differences for WG2, our manipulations resulted in similar changes to interpretation only once (rows 21 and 22), but more often resulted in significant differences and trends that pushed viewer interpretation in opposite directions (rows 13 and 14, 17 and 18, 19 and 20; Figure 2.15). This provides further evidence that motion dynamics sometimes carry similar information across cultures, but sometimes the same movements invoke oppositional concepts within different populations.

Regarding our original inspiration from embodied cognition, these results may still support the case that concepts are represented and evoked by physical motion characteristics. It could be that the metaphoric organization of thoughts is similar, but the particular concepts take precedence when evoked by similar motions differ between cultures. Likewise, certain concepts may be reinforced or discussed more in one culture than another, bringing them to mind more frequently or easily. This could account for many of the raw differences between cultures found above.

Future Directions

The many differences in interpretation between cultures highlights the importance of going beyond this preliminary study to a more rigorous examination of individual and combined motion features, and the mapping of those motion dynamics to interpretation. Motion tracking technology gives us the power to computationally delve into these dynamics. Voice-to-text software can allow us to map these dynamics to co-speech semantic content, adding an extra dimension of intent and communication to these studies. Understanding the dynamics of motion will allow us to develop and test specific theories using constructed gestures, and create highly specific hypotheses of how individuals in a given culture may interpret these gestures.

Because these interpretations were made without any linguistic or semantic context, responses may be highly influenced by gesture elements that are meaningful in one culture, but not the other. Such differences may diminish with linguistic context, but introduce further complexities in how languages

express spatial and conceptual information. Future stimuli used to examine these questions could be improved by rigorously testing any "form-meaning associations," (Kita, 2009) in stimulus gestures.

2.2.5 Conclusion

This exploratory experiment demonstrates the complex pattern between motion dynamics and viewer interpretation of metaphoric gesture, but further work is necessary to understand how such dynamics compose to form meaning. Although the mechanism behind the mapping of gesture dynamics to concepts may be consistent across all individuals, specific motions seem to evoke different concepts in different cultures. Greater computational tools and evolving technologies may help us uncover how gestures evoke concepts, and determine to what extent metaphoric gesture interpretation is culturally dependent.

2.3 Conclusion of Chapter 2

The studies presented in this chapter demonstrate viewers' capacity to perform multi-fold interpretation of complex, multi-metaphoric gestures. Viewers readily use minute components of gesture, such as hand shape or velocity, to drastically alter their understanding of the message being communicated. Furthermore, the extent to which individuals read multiple metaphoric messages from a single gesture may be culturally dependent. Importantly, both experiments in this Chapter are case studies; they are purely illustrative of some issues around complex interpretation of metaphoric gesture.

This complexity and variation in gestural expression of metaphor is a major challenge for generative gesture algorithms. This chapter highlights the importance of understanding the mapping from motion to metaphoric meaning as a many-to-many problem. Taken together, these studies underscore the importance of creating a human-readable mapping that designers are capable of interpreting and interrogating to more holistically understand the meaning implied by the form of a gesture, rather than relying on simplistic additive or unintuitive black-box approaches. Motions cannot simply be blended or combined and be expected to carry purely additive meanings according to the viewer. Similarly, multiple motions may be illustrative of the same metaphors. The difficulty here lies in codifying and characterizing the ways in which these different motions may relate to multiple meanings. This inspires aspects of the framework presented in Chapter 4 which is capable of analyzing metaphors both individually, and in conjunction with one another.

Chapter 3

Qualitative Subjective Analysis

Chapter 2 presented two experiments which examine how specific manipulations of metaphoric and multi-metaphoric gesture may influence viewer interpretation absent of any linguistic context. However, as discussed in Section 2.1.7 and revisiting the challenges presented in Section 1.8.2, evaluating a limited set of gestures that lack linguistic context limits generalizability of findings in these experiments.

This chapter uses an expanded experimental design that includes such context to evaluate how gestures influence the interpretation of their co-speech utterances. This technique can be used to test specific hypotheses of how gestures carry subjective meaning across distinct linguistic contexts. It further exemplifies the importance of evaluating not only a viewer's subjective opinion of how "appropriate" a gesture is for a given utterance, but to probe deeply into how the gesture affects vewiers' qualitative understanding of language.

The findings from this Chapter typify the risk of false implicature in gesture. They demonstrate that gesture is not only critical for the social realism of an agent, but moreover that gestures impact viewer interpretation regardless of the intent behind them. It is therefore necessary to take into account the mapping between motion and meaning when programming an agent's gesture behavior, even – perhaps especially – if that agent's gestures are not intended to convey rich meaning.

The remainder of this chapter is largely unchanged from published work that can be cited as: **Saund, C.**, & Marsella, S. (2021, December). The Importance of Qualitative Elements in Subjective Evaluation of Semantic Gestures. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1-8). IEEE.

3.1 Current Issues in Subjective Evaluations

In this section we provide context for the current work. We first describe current algorithms used for gesture generation in virtual agents, and issues they both address and continue to face. Then, we discuss the current state of subjective evaluations of generated gestures.

3.1.1 Gesture Generation in Virtual Agents

The importance of gestures as they are used in Virtual Agents (VAs) is a growing body of research. Embodiment of avatars can increase their perceived persuasiveness (Guadagno, Blascovich, Bailenson, and McCall, 2007), likability, and trustworthiness (Kulms and Kopp, 2016), and can be highly effective in domains such as medical practice and clinical work (Fiske, Henningsen, Buyx, et al., 2019), public speaking (Chollet, Wörtwein, Morency, Shapiro, and Scherer, 2015), and teaching (Mayer and DaPra, 2012). In such sensitive situations, non-verbal behaviors, especially gestures, are crucial to the success of the agent; A gesture with adverse implied semantics at an inopportune moment may lead to a break down of rapport and trust between user and agent.

There is a long history of gesture studies in behavioral psychology (McNeill, 1985; Alibali and GoldinMeadow, 1993; Lickiss and Wellens, 1978). Within the virtual agent community, psychophysical phenomena have long been exploited to generate semantically meaningful gestures. Cerebella (Lhommet, Xu, and Marsella, 2015) uses a combination of machine learning processing of syntax and prosody content, a lexical database, and rule-based approaches with dynamic filters to generate novel gestures that are linked to the ontological structure of the utterance. Greta (Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005) has been used to implement image-schema based approaches to gesture generation (Ravenet, Clavel, and Pelachaud, 2018), resulting in metaphorically meaningful gestures.

Interest in gesture generation for virtual agents within the wider computer science community has led to an increase in available datasets, including motion extracted from publicly-available video (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019) and motion capture of natural speech (Ennis, McDonnell, and O'Sullivan, 2010; Ferstl and McDonnell, 2018). This has led to a rise in data-driven approaches to gesture generation (e.g. Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019). Notably, end-to-end machine learning approaches benefit from incorporating phrase semantics into their generative mechanisms (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Ahuja, Lee, Ishii, and Morency, 2020). Vectorization algorithms such as BERT (Devlin, Chang, Lee, and Toutanova, 2018) and USE (Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Cespedes, Yuan, Tar, et al., 2018) are popular choices to quantify the qualitative semantic aspects of speech.

Another method to generate co-speech gesture is to match gestures for novel utterances from a database of pre-parsed human-generated motions. This benefits by using natural human motion, and is capable of using co-speech linguistic context to select communicatively relevant gestures.

3.1.2 Outstanding Issues in Gesture Generation

Gesture generation techniques that rely on speech as input assume that the information conveyed in gesture is recoverable from that speech. Consider for example an exchange between two speakers which becomes combative. Speakers can convey a tremendous amount of information non-verbally:

One speaker can say "I agree with you," and in one case hold their hands forward with their palm open and facing upwards, indicating a sense of unity or collaboration in the conversation, perhaps inviting the conversation to continue. However, another plausible gesture to accompany this phrase could be to put both hands up, with the palm facing the opposite speaker in a "backing away," position, instead conveying submission or surrender, perhaps implying they feel they must protect themselves. Such off-the-record information is one example of consciously occluding information via explicit channels and instead using implicit behavior to achieve specific social goals (Kendon, 1995). Gesture and language are generated from a deep common cognitive starting point McNeill and Duncan, 2000, but what is conveyed in each may differ; one is not subset of the other (Kendon, 2000).

This example demonstrates how gestures are sometimes used to convey information explicitly not represented in speech. It is necessary to include extra-linguistic information when generating or for them to be relevant to their context. Deep learning techniques often include voice information such as pitch and prosody to identify which part of an utterance to emphasize, but this still misses extra-linguistic information that could potentially be conveyed through gesture, such as the communicative intentions of the speaker. These techniques attempt to capture the relationship between vocal expression and motion that illustrates an underlying cognitive mechanism that generates behavioral communication without explicitly modeling the speaker or viewer's cognitive state.

Speech-based end-to-end algorithms make the assumption that there is semantic information being conveyed in every gesture, and that the information being conveyed is intentional, relevant to the utterance, or both. We know however this is not the case. Speakers gesture to improve their word recall or fluency (So, Sim Chen-Hui, and Low Wei-Shan, 2012), and to strategically convey social meta-information (e.g. holding their conversational turn as discussed in Sikveland and Ogden, 2012).

Furthermore, some gestures use semantic content that is unavailable in the particular utterance in which the gesture occurs. For example, iconicity based on semantic similarity can develop throughout a conversation or a series of conversations (Pouw, Wit, Bögels, Rasenberg, Milivojevic, and Ozyurek, 2021), such that a gesture's form may have metaphoric or specific semantic meaning that is irrecoverable in the isolated context in which it occurs. The rhetorical structure of a conversation may also dictate the form or location of a gesture. For example, when discussing two options, one may use their left hand to represent one option, and their right to represent the other. The speaker can then gesturally disambiguate exophoric words (e.g. "this"), but such disambiguation is irrecoverable from the speech of a single utterance without wider conversational context.

3.1.3 Subjective Evaluations

While one may argue the merits of these different approaches to generate gestures, there remains the fundamental question of how to evaluate them. One way is the notion of judging how well the produced gesture fits its conversational context. This may ask viewers to rate a gesture's overall "appropriateness" (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Wolfert, Girard, Kucherenko, and Belpaeme, 2021) or qualify specific aspects, such as how "natural"

or "smooth" the gesture looks (Hasegawa, Kaneko, Shirakawa, Sakuta, and Sumi, 2018).

These criteria are problematic because they rely on the viewer's *explicit* impression of the gesture and its relationship to any co-speech context. But gesture influences semantic understanding in a subtle way. This is especially important for gestures which are intended to carry semantic meaning, as gesture qualitatively influences thought in the viewer (Jamalian and Tversky, 2012b). To effectively utilize gestures to convey semantic information, we must focus on the viewer's decoding of specific semantic information. This section puts forward an argument to use specific semantic dimensions that can be measured in context in order to understand not only a gesture's relevance to a particular co-speech utterance, but the impact that the gesture has on specific attributes of the viewer's interpretation.

The current work recognizes the importance of gestures fitting their conversational context by measuring specific impressions when determining the impact of semantically-related gestures. Our first experiment illustrates the importance of measuring subjective impressions of semantic and physical attributes of gesture. Specifically, we compare subjective appropriateness ratings of gestures selected through multiple techniques, including novel methods that capture the hierarchical information contained in language. We then present an exploratory experiment that suggests a method to quantify specific effects of a gesture on semantic interpretation when presented with a particular utterance.

3.2 Context for Experiments: Evaluating gestures on qualitative semantic interpretations

The semantic relationship between gesture and speech is difficult to capture. On a word-level, not only do many words have different *senses*, but the same sense of a word may have differences in meanings across different contexts. This is especially pronounced in idioms and metaphors, and further complicated by exophoric words, whose definitions are entirely dependent on context. Additionally, the mapping from gestures to semantics is many-to-many. That is, a wide variety of movements may be seen as equally appropriate to convey same meaning, and the same gesture may be seen as equally appropriate for two semantically disparate utterances. Moreover, as demonstrated by the "backing away" example in Section 3.1.2, gestures are sometimes used to explicitly convey information not represented in speech.

In this section, we put forward a novel technique to select semantically-related gestures from a gesture database, and present two experiments which highlight the importance of measuring the qualitative impact of semantically-related gestures on the viewer. In the first experiment, we demonstrate a strong correlation between subjective perception of energy level of the speaker and perception of the semantic relatedness of the co-speech transcript. In the second experiment, we attempt to measure semantic information conveyed in gesture on specific qualitative dimensions. We then discuss the implications and impacts of these findings, including the limitations of the strength of claims we can make when using the original gesture that accompanies an utterance as an evaluative baseline.

3.3 Experiment 1: Semantic "Appropriateness" vs. Perceived Energy

This experiment uses the methods described below to select gestures from a dataset for a given utterance. We show that these methods all perform roughly equally when measuring "semantic appropriateness," including the original gesture that accompanies the utterance.

3.3.1 Dataset

We use the dataset of one speaker producing continuous motion in a monologue setting released in Ferstl, Neff, and McDonnell (2021a). There are many ways to segment gesture phases into individual gesture units, including segmenting by phase (Ferstl, Neff, and McDonnell, 2021a), motion regularity (Chiu and Marsella, 2014), co-speech content, or a combination of factors. However, these methods isolate individual segments or phases of the gesture, and as such produce very short individual gestures. We are interested in preserving some conversational context, so these proved unsuitable for our use. Noting the importance of filler words in discourse (Tree and Schrock, 1999; Tottie, 2011), we instead split the gestures by filler words commonly used by the speaker ('like', 'eh', 'um', etc). This preserved a large number of gestures that were between 0.5-4 seconds, and some rhetorical structure in a phrase[1]. We only kept gestures between 0.6 - 4.1 seconds long. This resulted in a database of 5,439 gestures and utterances. We visualized these gestures using the BVH visualizer described in Kucherenko, Jonell, Yoon, Wolfert, and Henter (2021a).

3.3.2 Semantic Ontologies

Extracting semantic content from utterances presents a unique challenge in that the semantics of both the words in an utterance and the motions of the accompanying gesture may carry multiple meanings simultaneously (see Chapter 2). For this reason, utterances are given multiple semantic tags, allowing a more accurate representation of the interconnection of the semantic space of communication.

The use of semantic content in gesture also raises the problem of the levels of representation in speech that are reflected in motion. Consider for example a speaker who utters the phrase "There was an audience." In this phrase, an "audience," can be represented at the highest level as a GROUP, and thus may be accompanied by a fairly generic *container* gesture (Lhommet and Marsella, 2013; Chiu and Marsella, 2011). However, at a deeper level, an "audience," represents many more abstractions, such as SOCIAL-GROUP, VIEWER, GATHERING, etc. The mental and consequently physical representation of the concept of "audience" may then carry different attributes at different levels of

[1] We split the motion according to a number of different features, including low-velocity moments, fixed time intervals and word counts, and between procedurally identified rhetorical phrases. Splitting on filler words produced the most empirically uniform and appropriate split of gestures. For example, this method tended to produce complete rhetorical units which can be read as stand-alone phrases, suitable for evaluating even without the wider context of a full sentence. Full results of all splits are available at the link in Section 7.3.

Word	Ontology	Extended Ontology Additional Terms
There	container: + form: GEOGRAPHICAL-OBJECT mobility: FIXED origin: NON-LIVING ARTIFACT tangible: + type: LOCATION type: POS-WRT-SPEAKER	aspect: DYNAMIC aspect: UNBOUNDED locative: LOCATED mobility: NON-SELF-MOVING origin: NON-LIVING origin: NATURAL spatial-abstraction: SPATIAL-POINT scale: AREA-SCALE tangible: + type: PLACE type: GROUPING type: PERSON type: SITUATION
Was	aspect: STATIC cause: MENTAL container: - intentional: - time-span: EXTENDED type: BE type: EXISTS	aspect: INDIV-LEVEL aspect: UNBOUNDED cause: AGENTIVE cause: FORCE iobj: RECIPIENT type: ACQUIRE type: BE-INACTIVE type: CAUSE-MOVE type: PASSIVE type: STATE
Audience	container: + intentional: + object-function: OCCUPATION tangible: + type: SOCIAL-GROUP mobility: SELF-MOVING	aspect: DYNAMIC cause: AGENTIVE form: SOLID-OBJECT information: MENTAL-CONSTRCT locative: LOCATED origin: HUMAN spatial-abstraction: SPATIAL-POINT type: ABILITY-TO-HEAR type: EVENT-TYPE type: GATHERING-EVENT type: INTERVIEW type: PERSON

Table 3.1: Shallow and extended ontology for the example sentence “There was an audience.”

cognitive representation of the speaker. Within this experiment, we refer to these two levels of analysis as the **Shallow Ontology** and **Extended Ontology** (see Table 3.1).

Ontologies are calculated using SpaCy parser (Honnibal and Montani, 2017a) to extract the syntactic structure, such as part-of-speech information, noun phrases and compound noun phrases, and additional substructure. The TRIPS (Allen, Dzikovska, Manshadi, and Swift, 2007; Bose, Ritwik, An, Hannah and Valpey, Benjamin, 2021) and Wordnet (Pedersen, Patwardhan, Michelizzi, et al., 2004) ontologies are then used to identify and attach semantic meaning to phrases.

The extended ontology is a superset of the shallow ontology, and also inherits all features from the shallow ontology. That is, once an item has been identified with a particular feature (e.g. *form*) that feature cannot be described again in the extended ontology, but more features may be added to the extended ontology (see Table 3.1), for example by moving up the Wordnet and Trips ontologies. The exception to this is the *type* feature, for which the extended ontology adds as many tags as it finds. All code for deriving ontology and extended ontology is found in the link in Section 7.3.

It is unclear how much nuance in a single gesture will be captured by using a relatively small dataset (Section 3.3.1). Therefore, we compare using the Ontology and Extended Ontology to select semantically appropriate gestures.

3.3.3 Gesture Selection

To perform a gesture for an incoming utterance, we select a gesture from our database of gestures that have the desired associated co-speech elements. For each selection method described below, we compare the gesture selected with the gesture that originally accompanied the utterance; that is, we only test our selection methods with utterances which occurred in the original data, so we can compare the effects of our selected gesture with that of the gesture which originally accompanied the utterance as a baseline.

Overall, we compare four methods to the baseline gesture:

Random

We randomly select a gesture from the dataset. However, we do match the gesture based on time, within 0.2 seconds. So, if the utterance we want to match originally took 1.5s to perform, we randomly select a gesture from all gestures in the database between 1.3 and 1.7s.

BERT

We calculate the distance between the incoming utterance and all utterances in the database using BERT (Devlin, Chang, Lee, and Toutanova, 2018), and select the top 10 matches. We then select the match that most closely matches the temporal length of the original performance.

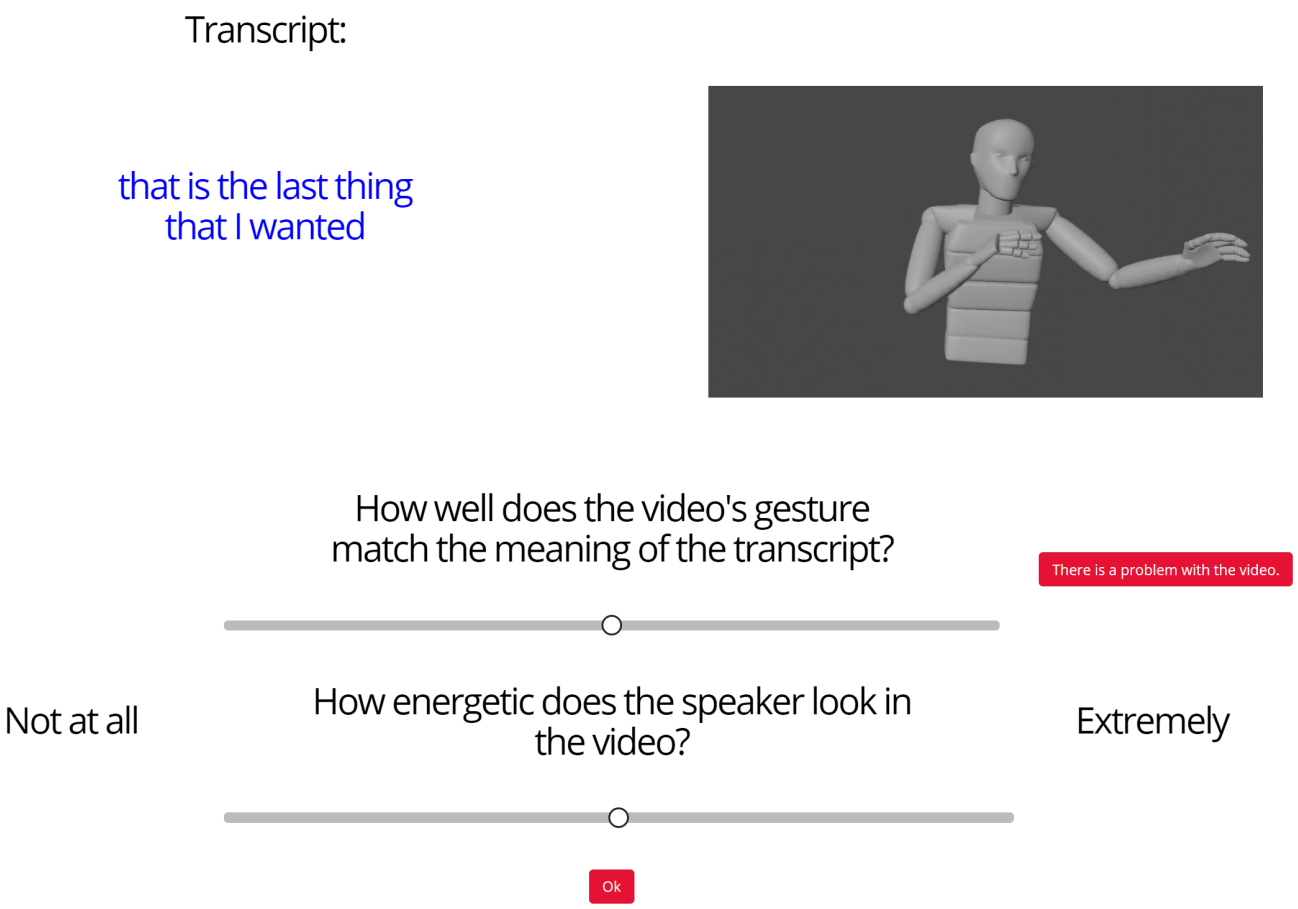


Figure 3.1: Participant view during Experiment 1.

Shallow Ontology Part-of-Speech matching

For Shallow Ontology part-of-speech (POS) matching, using the method of retrieving first-level ontological information (see Section 3.3.2), we create a set of semantic tags for each POS in an utterance. To calculate the similarity between two utterances, we calculate the intersection between each matching POS, and the percentage overlap between all matching parts of speech is used as the score between the two utterances.

Extended Ontology Part-of-Speech matching

We use the same method of calculating the intersection between sets of semantic tags for matching POS between two utterances, using all information in the Extended Ontology.

3.3.4 Procedure

Participants saw one utterance and one video of a gesture without sound, along with two sliders labeled "How well does the gesture in the video match the meaning of the transcript" and "How energetic does the speaker in the video look?" The sliders were labeled "Extremely" on the right and "Not at all" on

Selection Methodology	Semantic Appropriateness Mean (SD)	Subjective Energy Mean (SD)
Original Gesture	-7.77 (21.99)	-10.5 (21.95)
Random	-3.22 (28.90)	-2.77 (24.38)
BERT	-5.78 (28.94)	-6.0 (18.14)
Shallow ontology	-14.24 (28.36)	-0.72 (24.21)
Extended ontology	0.27 (26.04)	1.86 (20.46)

Table 3.2: Aggregated results from Experiment 1. Shows the mean and standard deviation ratings (-50 to 50) from all participants.

the left (Figure 3.1). Participants completed two training-phase trials during which they were asked to move the sliders in each direction indicating whether they agreed or disagreed with the statements to ensure they understood the meaning of the sliders. Slider values ranged from -50 to 50 (the middle being 0), and were hidden from the participant.

During the experimental phase, participants were shown one transcript along with one video selected through one of the methods described above. Each participant saw a counter-balanced mix of gesture selection techniques and transcripts. No participant saw the same transcript or video more than once, and participants saw gestures selected through each technique. Participants completed 10 trials with 1 attention check at trial 7, for which the transcript shown instructed them to click a separate button.

10 transcripts were randomly selected from the database for this experiment. Each was played with 5 different videos: the original gesture video, random, BERT-matching, shallow-ontological matching, and extended-ontological matching. This generates 50 unique combinations of transcript and video. Participants were unaware the videos were selected using different methods.

Participants

All participants were recruited from Prolific. Participants were all native English speakers who were born in and who have spent the majority of their adult life living in the UK, Australia, New Zealand, or Canada. Participants ages ranged from 19-65, N=93. Four participants' data was removed from analysis for inaccurate attention checks, which leads to 89 participants (801 total judgements). These judgements were spread out over all 50 potential utterance-gesture conditions.

3.3.5 Subjective Questions

Our motivation for the two questions asked of participants (Section 3.3.4) are to probe the relationship between subjective meaning and subjective energy. The first question – "How well does the gesture in the video match the meaning of the transcript?" – is often used to assess an algorithm's effectiveness at generating gestures that match the semantic meaning of the gesture's co-speech utterance (e.g. (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020)).

The second question – "How energetic does the speaker in the video look?" – was formulated

through ad hoc observation from participant responses during pilot experiments. In general, we observed that gestures with larger amplitudes tended to elicit stronger responses in perceived meaning. It is relatedly shown that more natural-looking gestures tend to be perceived as more meaningful (Kucherenko, Jonell, Yoon, Wolfert, and Henter, 2021a), but this effect has not thus far been investigated with relation to gesture amplitude, and in particular subjective perception of speaker energy.

Because the question of how well a gesture matches the co-utterance's meaning is subjective, we chose to also use a subjective measurement of perceived energy. However, this may have affected participant responses in perceived matching of meaning of the gesture and transcript, and future experiments are needed to determine the potential of any such effects by using objective measurements of gesture motion.

3.3.6 Results

Our results show that there are no statistically significant differences between the performance of any of the matching techniques, including randomly selecting a video. That is, a gesture that was randomly selected from the dataset was no more likely to be rated as better matched semantically as the gesture which the speaker actually produced when speaking that utterance (Figure 3.2, aggregated results in Table 3.2).

However, we see a high correlation between the subjectively rated energy level and the semantic appropriateness of a gesture ($r^2 = 0.246, p < 0.001$). When we normalize for rated energy levels (only compare gestures which were rated as higher than 0 energy) we see this trend diminish slightly, but is still present ($r^2 = 0.196, p < 0.04$).

3.3.7 Discussion of Experiment 1

In the context of the wider field of gesture generation, adding semantic information to algorithms demonstrably improves semantic appropriateness ratings (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020) and participant preference for a gesture (Wolfert, Girard, Kucherenko, and Belpaeme, 2021). However, this experiment reports no benefit to subjectively judged semantic appropriateness when we include semantic information in our matching techniques, including using word vectors. This appears counter-intuitive, though a few explanations may contribute to this outcome.

One possibility is that few of our gestures carried linguistically-related semantic information, and therefore they could all be seen to equally fit any utterance. This may explain why randomly selected gestures perform about as well as the original gesture for each of our utterances. This may therefore be an artifact of the dataset we used, as the speaker produced completely unplanned, spontaneous gesture in a monologue format, potentially leading to stylistically similar, functionally indistinguishable gestures. It is unclear if this problem is unique to this dataset, or to the technique of matching gestures from a natural database, as current matching techniques have not yet made explicit use of semantic

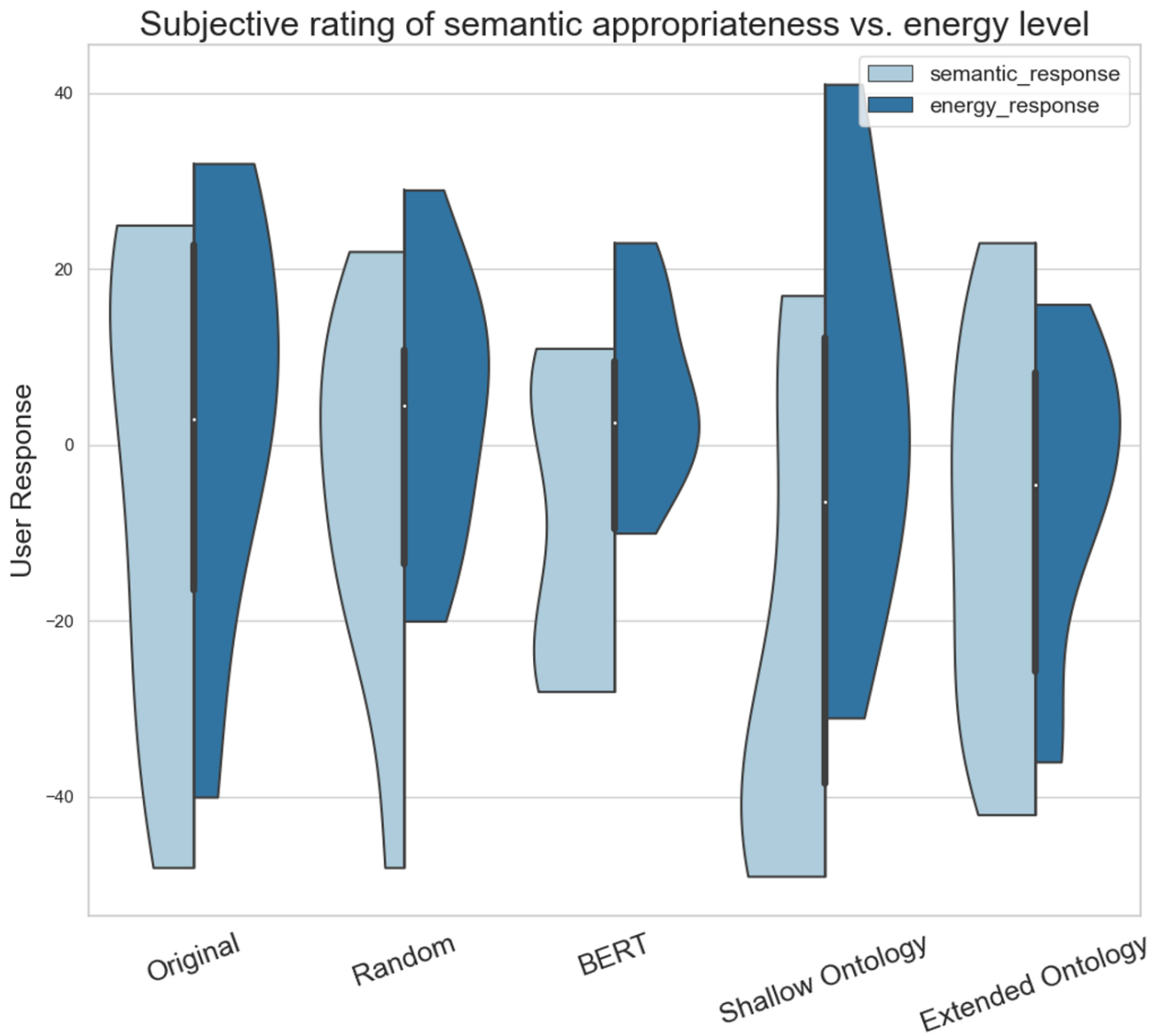


Figure 3.2: Distribution of subjective ratings of semantic match of transcript and energy for each gesture selection method.

information[2].

Another intriguing possibility emerges when we look at semantic appropriateness ratings in the context of our wider results. Namely, that the subjective ratings of semantic appropriateness are highly correlated with subjective ratings of energy levels. In other words, when a gesture seems to have more energy, viewers are more likely to read into it to report that it matches the semantics of a transcript, regardless of the specific motion of that gesture. It could be the case that asking both questions of the participants at once caused a high correlation between responses to each question. One way to disentangle this would be to ask one set of participants to what extent the transcript matched the gesture, and to ask another set how energetic the gesture was.

The implications of this for the field of gesture generation are cautionary. When evaluating the subjective impact of adding semantic information to an algorithm to generate semantically-relevant gestures, we must account for adding variation in velocity and amplitude in addition to the motion of the gesture itself[3]. Otherwise, our findings in this experiment suggest that simply adding velocity and amplitude variation is sufficient to give viewers the impression that gestures are semantically related to their co-speech utterances.

3.4 Experiment 2

Experiment 1 demonstrates the ease with which participants rated many different gestures as semantically appropriate for a particular utterance. In light of this, Experiment 2 explores how we can begin to focus on viewers' specific impressions and the importance of the relationship between gesture and semantic interpretation.

We used the same dataset (Section 3.3.1) and gesture selection methods (Section 3.3.3) as in Experiment 1.

3.4.1 Procedure

We use the same procedure of showing participants one utterance and one video at a time, including training phases and an attention check as described in Experiment 1 (Section 3.3.4). However, we use four sliders in this experiment, each labeled with questions that explore both how the speaker expresses qualitative information about the subject they are discussing, and information about the speaker's state of mind. The participants see the following setup prompt and questions on the sliders.

Judging by both the transcript and the video, the speaker is expressing...

- (1) Separation between people or ideas

[2]Although it can be argued that some semantic and rhetorical information is recoverable from acoustic speech signals.

[3]Indeed, many deep-learning gesture generation techniques do account for motion variation in some form (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Chiu and Marsella, 2011; Ferstl, Neff, and McDonnell, 2019; Ahuja, Lee, Ishii, and Morency, 2020).

- (2) An ongoing process or journey
- (3) Certainty in what they are saying
- (4) Something positive about the subject

Sliders were labeled "Very much so" on the right and "not at all" on the left. As in Experiment 1, participants underwent an explanation and two-trial training phase.

These questions were selected because gestures, especially metaphoric gestures, are known to impact these areas of interpretation (Calbris, 2011), as shown in Chapter 2. They attempt to capture a diverse range of dimensions (*semantic differentials*, Osgood, 1971) identified in the field of psycholinguistics and semantic theory (Grady, 1997).

20 transcripts were randomly selected for this experiment. Each transcript was shown with one of five potential gesture videos: the original gesture video, random, BERT-matching, shallow-ontological matching, and extended-ontological matching. As our techniques necessarily select the same video for a transcript each time, the same random video was used with each transcript to maintain consistency. Each condition was judged by 18 different participants. Participants were counter-balanced such that no participant saw the same transcript or video twice, and each participant saw videos generated by all methods.

Participants

We used the same recruitment criteria as described in Section 3.3.4. Participants ranged from age 18-71, N=210. Ten participants' data were removed from analysis for inaccurate attention checks, leaving 200 participants (1800 total judgements). These judgements were spread over 20 unique transcripts, which were each played with 5 different videos (the original gesture video, random, BERT-matching, shallow-ontological matching, and extended-ontological matching).

3.4.2 Results

When analyzing results for this experiment, we compared the interpretation of each gesture generated by our matching techniques with the interpretation of the baseline control (the original gesture). When viewing these results, it is necessary to remember that they are not an indication of whether a gesture is "good" or "bad" in a situation, or even if they add or detract from the communicative intention of the speaker. Instead, what we measure is how well each matching technique selects a gesture that influences interpretation of the transcript similarly or differently to how the original gesture influences that interpretation, across not just overall semantic appropriateness, but on specific semantic dimensions.

Overall, no technique performed significantly better than any other in this experiment. That is, no technique, including random, did especially well or poorly at selecting gestures from our database that push interpretation of an utterance in a similar way as the original gesture when shown with the same utterance (Table 3.3).

Selection Methodology	Separation (μ/σ)	Process (μ/σ)	Certainty (μ/σ)	Positive (μ/σ)
(Original Gesture)	0 (0)	0 (0)	0 (0)	0 (0)
Random	16.69 (6.53)	11.51 (9.57)	13.16 (7.88)	12.12 (9.68)
BERT	18.7841 (5.39)	10.18 (8.30)	17.20 (9.65)	20.08 (8.89)
Shallow ontology	20.05 (9.52)	13.74 (8.70)	15.99 (8.89)	10.96 (10.08)
Extended ontology	19.16 (6.97)	13.99 (6.67)	17.35 (9.89)	15.36 (7.83)

Table 3.3: Aggregated results from Experiment 2. μ (Mean) refers to the difference between the mean scores for that domain of the original gesture vs. the gesture selected in each condition. σ (Standard Deviation) is the standard deviation of these differences. Note that no selection method scores particularly closer to the original gesture than any other across all four conceptual dimensions.

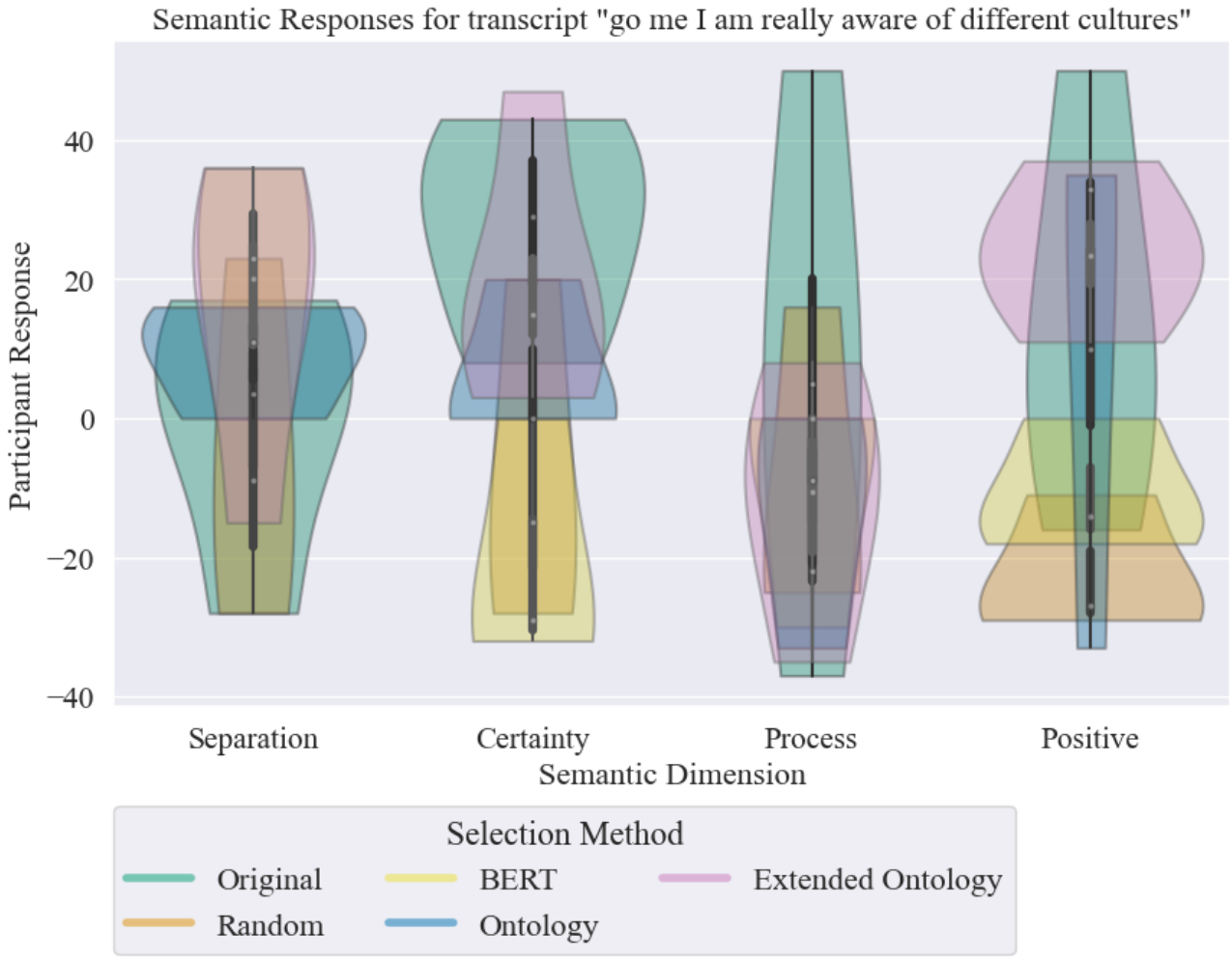


Figure 3.3: Distribution of subjective responses across semantic dimensions for the utterance “go me I am really aware of different cultures (sic)”. In this instance, the speaker is sarcastically congratulating themselves.

Because of the importance of context in gesture, however, this aggregation loses the nuance of analysing on an utterance-by-utterance basis. A more detailed analysis allows us to see specifically how each selection method influenced interpretation across each semantic dimension within a linguistic semantic context (see all results in Section 7.3). Consider an utterance pulled from the database, "go me I am really aware of different cultures." Figure 3.3 shows the distribution of responses across semantic dimensions for the transcript accompanied by gestures generated by each method. Here we see that the original gesture coupled with the utterance conveyed a degree of certainty and positivity about the subject, but highly distributed scores for Separation and Process indicate ambivalence or unclear patterns in responses for these domains. Meanwhile, the gesture selected using the BERT-matching method, for example, indicates low certainty and less positive connotations of the same utterance.

In a post hoc analysis, for each unique ontological feature identified in the database, we re-ran the analysis including only judgement instances in which both the transcript shown and the utterance of the gesture that was shown included that feature. Results indicate for some semantic features (i.e. mobility: MOVABLE, type: SYMBOLIC-REPRESENTATION, aspect: DYNAMIC) the ontological techniques pushed interpretation of the shown transcript significantly more similarly to that of the original gesture than either the random or BERT-selected gesture. When adjusting for testing multiple hypotheses this effect is no longer statistically significant, nevertheless we discuss the implications of this as an analysis technique in Section 3.4.3.

3.4.3 Discussion of Experiment 2

Overall, no gesture selection methods consistently influenced interpretation similarly to the original utterance.

One interpretation of this result is that the procedures used to identify semantic information fail to derive meaning correctly, resulting in irrelevant or misleading gestures being selected for an utterance. However, the semantic features derived for each utterance – either from our semantic ontology built on TRIPS or feature vectors from BERT – are well-validated, leading us to explore other possibilities.

An alternative to this is the understanding that semantics are necessary yet insufficient to select an appropriate gesture for a given utterance. As we reviewed in Section 3.1, this is unsurprising, as gestures are deeply intertwined with prosody, rhetorical structure, personality, and other contextual elements. However, we used silent videos to prevent timing, prosody, and other vocal cues from influencing viewer interpretation, leaving sparse other information – besides the mental state of the viewer – from influencing viewer impression of the utterance and gesture.

It is further possible that the original gesture that accompanied the utterances were ambiguous, or not specifically tied to the original utterance text on the semantic dimensions we measured. This is why we highlight the importance of a case-by-case analysis. As noted above, some semantic tags did seem to identify gestures that pushed interpretation similarly to the original gesture. This suggests that this paradigm can be used to test prescriptive hypotheses about which ontological features correlate

with viewers' impressions of gestures.

Importantly, our evaluation precludes determining whether the selected gestures are "appropriate" but instead focuses on how the gesture influences interpretation. It is possible that some of the gestures we selected may have influenced viewer interpretation in a way the speaker intended, but without knowing the cognitive state of the speaker this cannot be determined.

3.5 Discussion

None of our text-driven, semantic selection mechanisms – Ontological, Extended Ontological, or BERT – consistently selected gestures that influenced interpretation of an utterance in a similar way to the gesture which originally accompanied that utterance. This does not necessarily tell us the subjective quality or semantic appropriateness of the gestures in relation to the accompanying text, which we explored in Experiment 1.

Instead, this demonstrates that when semantic evaluations are decomposed into specific dimensions, we transform the evaluation of a gesture from simply being "appropriate" to a more specific understanding of how it influences the viewer. When evaluating gestures in context, without access to the communicative intention of the speaker, we can only evaluate a gesture's ability to influence interpretation, not its "quality." Assuming that the original gesture perfectly encapsulates the communicative intention of the speaker would be flawed: The original gestures were not methodically planned by the speaker to convey clarity, but were instead spontaneously produced. And, as in the "back away" example in Section 3.1.2, gestures do not necessarily reflect what is in the utterance.

In light of these two experiments taken together, we argue subjective analyses of gesture generation must focus on the communicative intention of a gesture in context. This is because it seems to be the case that any sufficiently energetic gesture could potentially be seen as appropriate (Experiment 1), and because different gestures, when presented in the same context, influence viewer interpretation in different ways (Experiment 2). This context includes semantic (and other linguistic) content, as well as many other aspects that dictate the form and content of social interaction, such as practical application, physical embodiment, gender, etc. We thus demonstrate support for evaluations that focus on the social human-interpretation aspect of non-verbal behavior in VAs to avoid the assumption that there is always a match between semantics and gesture. Using such evaluations would allow gesture generation researchers to focus on the viewer impact of particular gestures, as opposed to compare generated gestures to human-produced ones, which are an imperfect baseline.

In order to use such evaluation techniques, however, Virtual Agents (or their designers) need to have their own conversational, social, and behavioral goals in interactions against which to evaluate their performance. Then, gesture generation researchers can measure the effect of a generated gesture relative to the desired outcome. HCI researchers already set some such explicit goals, such as how "natural looking" a gesture is; Setting specific conversational goals that may be influenced by gesture is in line with the desire of the community to create socially compelling VAs.

But as we know, non-verbal behavior does not exist simply to affirm to one another that we are all humans. It is a natural extension of gesture generation to further ask "how well does the gesture achieve its' intended purpose?" These could include engagement metrics, or providing semantic clarity to the accompanying co-speech utterance. But they may well include extra-linguistic information, such as conveying off-the-record information. In order to evaluate such information, however, a VA would need to model it.

We are hardly the first to suggest a viewer impact-based evaluation of gesture (Pütten, Straßmann, Yaghoubzadeh, Kopp, and Krämer, 2019; Krämer, Kopp, Becker-Asano, and Sommer, 2013; Krämer, Hoffmann, and Kopp, 2010; Hoffmann, Krämer, Lam-Chi, and Kopp, 2009), or to suggest that agents deeply model their social goals to drive non-verbal and conversational behavior (Cassell, Pelachaud, Badler, Steedman, Achorn, Becket, Douville, Prevost, and Stone, 1994; Swartout, Gratch, Hill Jr, Hovy, Marsella, Rickel, and Traum, 2006). What these experiments demonstrate is an imperative to extend an impression-based paradigm to include qualitative evaluations of potentially semantically relevant non-verbal behavior.

3.5.1 Future Work

These results further lead us to argue that for virtual agents to generate rich, compelling, and linguistically relevant gestures in context, they must retain their own model of conversational context as well as social goals. If a dataset included the dimensions of communicative contexts and intentions of the speaker, the amount of data required for end-to-end models to produce the rich set of natural gestures we see spontaneously produced in human speakers is simply much larger than any datasets that are currently used by the community. Historically VAs had rich mental states including communicative intentions and emotions (Cassell, Pelachaud, Badler, Steedman, Achorn, Becket, Douville, Prevost, and Stone, 1994); the question now is how to leverage modern machine learning methods to facilitate the creation of socially advanced VAs.

Largely, this work calls for modeling an agent's goals and beliefs in order to appropriately select a gesture that is applicable in its context. Evaluations which focus on specific dimensions of interpretation and which are dynamic and agent-driven may lead to a more holistic approach to gesture generation, which focuses directly on the human impact of non-verbal behavior of virtual agents.

As we found some information in our semantic ontology correlates well to matching the interpretation of the original gesture and utterance, future work may focus on digging deeper into mapping which linguistic semantics seem correlated to particular interpretations, and extend such work into gesture forms and motion kinematics (Pouw, Wit, Bögels, Rasenberg, Milivojevic, and Ozyurek, 2021). Similarly, this work provides a foundation to reveal how families of motions, gestured in conjunction with particular co-speech semantic concepts, may influence interpretation on these semantic dimensions. One way to do this may be to select one gesture and observe how it influences semantic interpretation in the context of different co-speech utterances.

3.5.2 Conclusion

Asking participants whether a gesture is appropriate for an utterance is not specific enough for modern applications of VAs, as these findings suggest that even gestures seen as "appropriate" carry a large risk for false implicature. We must further examine what semantics viewers qualitatively read from gestures. Different approaches to generation and selection may succeed in helping agents achieve their conversational goals, depending on the co-speech context. However, text alone is insufficient to generate the rich space of natural gestures we see in human performances.

3.6 Methods Established For This Thesis

Finding related motions within a dataset and analyzing the content of their co-speech utterances inspires and is fundamental to the framework and analysis technique presented in Chapter 4. Co-speech utterance context significantly complicates viewer interpretation of gestures. Instead of relying on how *viewers* subjectively interpret gesture in conjunction with different conversational contexts, we can focus solely on how *speakers* use gesture in conjunction with co-speech utterances – specifically, the semantic and metaphoric content of their co-speech utterances.

Additionally, segmenting a monologue of motion capture into individual gestures and performing semantic analysis of each gesture's co-speech utterance provided the groundwork to actually categorize gestures by both motion and co-speech semantic content.

Chapter 4

Quantifying Links Between Gesture and Language

One major impediment to generating rich, complex, nuanced gesture in virtual agents is a lack of understanding as to how a gesture's form and motion relates to its communicative meaning. Chapter 3 shows that it is straightforward to use gesture to manipulate a viewer's interpretation of language, but what remains unclear is how to use gesture strategically to manipulate a viewer's interpretation of language *in a specific, controlled way*.

Chapters 2 and 3 concentrate on viewer interpretation of gestures, and the importance of gesture generation pipelines to utilize gesture in a meaningful, purposeful manner. Now, I shift focus towards creating a mapping between motion and meaning, with the explicit understanding that this mapping must have the following characteristics:

- It must be able to be interrogated and interpretable by human readers.
- It must account for both language and gesture's ability to convey multiple meanings simultaneously.
- It must account for the many-to-many mapping between gesture and concepts conveyed in language.
- And, similarly, it must account for the many-to-many mapping between a gesture's motion and it's *meaning*.

In this Chapter, in Section 4.1 I first present an architecture that can be used to examine the relationship between gesture and language, and briefly explore its application in the rhetorical domain. Then, in Section 4.2 I dive more deeply into an implementation of this architecture to explore a mapping between gesture and semantic content of co-speech utterance. This technical chapter presents a framework that incorporates insights from observational behavioral gesture research with data-driven analysis to produce a mapping that accounts for the vast variation in motions used to convey different concepts.

Section 4.1 is unchanged from published work that can be cited as: **Saund, C.**, Birladeanu A., and Marsella S. (2021, May). "CMCF: An architecture for realtime gesture generation by clustering gestures by motion and communicative function." In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (pp. 1136-1144) (AAMAS 2021)

Section 4.2 is unchanged from the version that is accepted for publication, and will be able to be cited as: **Saund, C.**, Matuszak H., Weinstein A. & Marsella, S. (2022, December). Motion and Meaning: Data-Driven Analyses of The Relationship Between Gesture and Communicative Semantics. *accepted for publication in Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22), December 5–8, 2022, Christchurch, New Zealand.*
https://doi.org/10.1145/3527188.3561941.

4.1 CMCF: An architecture for realtime gesture generation by clustering gestures by motion and communicative function

Gestures augment speech by performing a variety of communicative functions in humans and virtual agents, and are often related to speech by complex semantic, rhetorical, prosodic, and affective elements. In this section we briefly present an architecture for human-like gesturing in virtual agents that is designed to realize complex speech-to-gesture mappings by exploiting existing machine-learning based parsing tools and techniques to extract these functional elements from speech. We then deeply explore the rhetorical branch of this architecture, objectively assessing specifically whether existing rhetorical parsing techniques can classify gestures into classes with distinct movement properties. To do this, we take a corpus of spontaneously generated gestures and correlate their movement to co-speech utterances. We cluster gestures based on their rhetorical properties, and then by their movement. Our objective analysis suggests that some rhetorical structures are identifiable by our movement features while others require further exploration. We explore possibilities behind these findings and propose future experiments that may further reveal nuances of the richness of the mapping between speech and motion. This work builds towards a real-time gesture generator which performs gestures that effectively convey rich communicative functions.

4.1.1 Introduction

Gestures play a powerful role in human face-to-face interaction (McNeill, 1992; Kendon, 2004), and moreover reflect the relation between thought, speech, and motion (McNeill, 1992; Cienki and Koenig, 1998). Gestures are shown to mirror the fine-grained structure of dialogue, such as its underlying architecture comprised of logical and rhetorical units (Kendon, 1972). The complexity of the relationship between gesture and language is also compounded by the multiple levels at which one can observe correlations between motion and speech. Grady (1997) found that situated language is frequently used to ground abstract metaphors in concrete physical descriptors ("These fabrics aren't quite

the same, but they're *close*", p. 283), while Chiu and Marsella (2011) showed how these conceptual metaphors are readily mapped onto everyday gestures.

One example of the multiple ways in which the communicative intention can be reflected in gestures is when a person presents the option of an "important or trivial idea" using different gestural performances. In one scenario they may emphasize the rhetorical contrast of "important *or* trivial" by holding up their right hand for important and their left hand for trivial. In another situation they may focus on the semantic aspects of the contrast, making a large gestural frame to emphasize "important," and move their hands close together when they utter "trivial" in order to convey the relative significance of the ideas through the metaphorical connection between importance and size.

Because the relationship between speech and gesture is nuanced, gestures generated by virtual agents often lack the same complexity displayed in human performances. Some gesture generators are rule-based (Cassell, Vilhjálmsson, and Bickmore, 2004; Chiu and Marsella, 2011), and thus have a limited library of both gestures and understandings of when to deploy them. While many rule based approaches use acoustic data to modulate gesture (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013; Lhommet and Marsella, 2013; Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005) they are still beholden to rules which behavior designers implant in them. These rules, while effective and grounded in theory, are ultimately non-exhaustive and often prescriptive instead of reflective of gestures which occur naturally and spontaneously.

End-to-end machine learning approaches combat this, and have recently gained significant traction (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Ferstl and McDonnell, 2018). These instead take video and audio data and use it to learn a mapping of speech to gesture. One challenge here is the need for sufficient data to capture the complex multi-faceted mapping between communicative function and gestures. As a result these technique are very good at conveying prosodic elements in the speech such as emphasis through rhythmic beat gestures (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013; Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019) but they lack the sufficient data to capture more complex relationships; they assume the gesture is solely driven - or at least captured - by the acoustic properties of speech, as opposed to some deeper communicative function that may not be reflected acoustically. In addition, these techniques largely forego designer control, other than limiting the data that is the input to machine learning, to, for example, specific speakers in order to capture that speaker's style (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019).

Nevertheless, large language corpora have led to a range of evolving natural language tools, derived using machine learning, that can analyze prosody (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013), syntactic structure (Charniak, 2000), semantic and metaphoric elements (Allen, Dzikovska, Manshadi, and Swift, 2007; Miller, 1995; Ravenet, Pelachaud, Clavel, and Marsella, 2018) as well as rhetorical structure within text and dialog (Marcu, 1997; Joty, Carenini, and Ng, 2015; Oepen, Kuhlmann, Miyao, Zeman, Cinková, Flickinger, Hajic, and Uresova, 2015). In addition, there is considerable un-annotated video data that is available to analyze gestures, for example using tools such as OpenPose (Cao, Hidalgo Martinez, Simon, Wei, and Sheikh, 2019).

Our work has pursued the following ideas: (1) these different analysis techniques provide a way to extract different elements (semantic, rhetorical and prosodic) from speech while avoiding the limited data problem, (2) if we break analysis down into these elements we may be able to afford more designer control, (3) within a particular analysis element, we assume there will be a difference in gestural motion properties in order for the communicative function to be effectively conveyed, and (4) the breakdown into function and gestural motion also supports driving gestures directly from communicative functions if available. This suggests the following approach to generation: Perform these distinct analyses on speech extracted from video data. Within a particular analysis, such as rhetorical structure, cluster the associated gesture videos based on motion properties to derive clusters associated with different rhetorical elements such as contrast, elaboration, etc. These clusters then provide candidate gestural motions to convey these functions.

In this section, we present and assess a potential architectural model of gesture generation which integrates rhetorical, semantic, affective, and acoustic relationships between utterances and their accompanying gestural motions. We first present the overall architecture, and then deeply explore the implementation of the rhetorical branch of this model as a demonstration of this novel method of clustering gestures based on co-speech elements and motion. We present our method of comparing and clustering gestural motion, building off of multiple third-party ontologies and ML-based NLP tools and the motion database found in Ginosar, Bar, Kohavi, Chan, Owens, and Malik (2019), and provide techniques for evaluating the clustering of these gestures, as well as ways to overlay clusters to provide a complex picture of the relationship between motion and speech. We focus on what this clustering technique can objectively tell us about the relationships between rhetorical structure and gestural motion.

4.1.2 Architecture Overview

In this Section we present an architecture for a virtual agent which uses a pre-trained model to perform gestures, and which is agnostic about how the animations are realized. We refer to this as Clustering by Motion and Communicative Function (CMCF).

Our proposed architecture attempts to generate gestures which carry the rhetorical, semantic, and affective communicative functions of natural human gestures. While these categories are non-exhaustive, there is reason to believe that these provide an effective foundation for gesture analysis (Kendon, 1972; Ravenet, Pelachaud, Clavel, and Marsella, 2018; Chiu and Marsella, 2011). Since our demonstration and assessment in this section focuses on rhetorical structure, we provide background on its relevance to gesture generation, with the recognition that all these elements play fundamental roles in non-verbal communication (Bavelas, 1994; Cienki and Koenig, 1998; Pollick, Paterson, Bruderlin, and Sanford, 2001).

While the relationship between discourse structure and gesture has been explored in virtual agents, we explicitly explore the relationship between rhetorical structure and gesture with respect to Rhetorical Structure Theory (Mann and Thompson, 1987). Lascarides and Stone (2009) conduct similar work

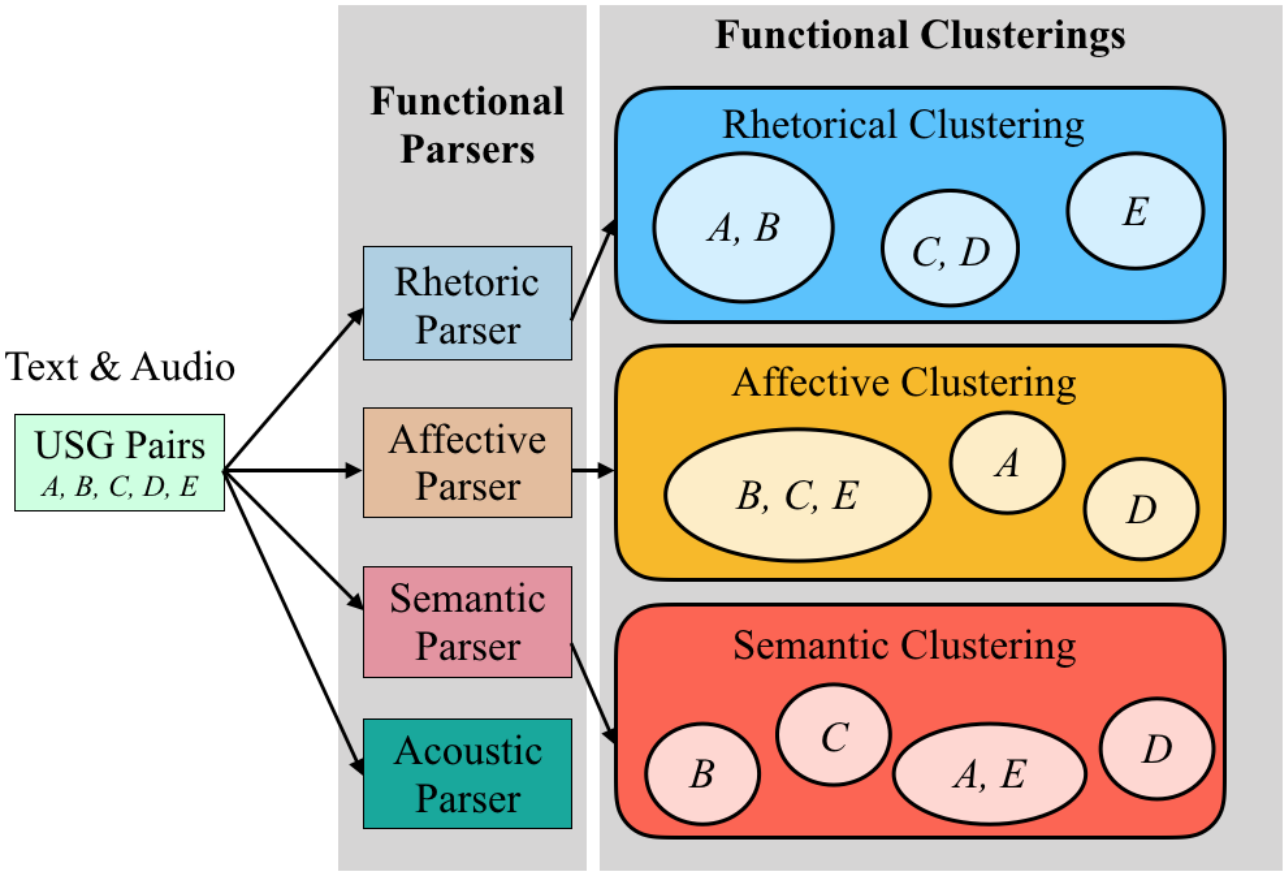


Figure 4.1: Overall architecture of generative model. During the pre-training step, example USG pairs A through E are tagged and grouped into functional clusters corresponding to the utterance. An elaboration on motion sub-clusters is found in Figure 4.2.

bridging formal analyses with pragmatic interpretation and generation mechanisms, demonstrating the importance of the shared roles of theoretical and applied work in this area.

Previous studies have used information contained in rhetorical structures to generate nonverbal behaviour in virtual agents. For example, Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro (2013) used a rule-based algorithm to extract semantic and rhetorical content from text and further applied it to generate nonverbal behavior, including gestures. By using the semantic and rhetorical content of discourse in addition to prosody to generate nonverbal behaviour, the character was shown to become more life-like, and was rated more highly on appropriateness compared to either prosodic based or random gestures. An important component of the mapping between speech and gesture thus appears to be the high-level relations between units of speech that might be projected onto specific hand movements during communication.

Clustering by Motion and Communicative Function (CMCF)

Our framework takes as input a piece of text and optionally an audio performance of that text. Its output can be used as an abstraction to an animation system. This system tags input speech with a variety of linguistic functional (discrete) labels using third-party parsers, which it then uses to derive appropriate gestures. Acoustic input is optional as each functional component of this model acts separately, and all available information is concatenated at the end.

The architectural overview for this model is shown in Figure 4.1. The pipeline contains three parallel processes for rhetorical, semantic, and affective domains[1]. It clusters gestures together categorically by tags given by the parsers, derived from the gesture's co-speech utterance. This way, when given an utterance, the agent performs a gesture that occurred with a linguistically similar utterance in the past. This works first by creating the clusters (Section 4.1.2) *offline*, then exploiting these pre-calculated clusters at run-time (Section 4.1.2).

Pre-training the model

In pre-training, the model draws from a set of gestures and their associated audio and transcription of utterances. Throughout this section, we refer to each utterance segment and the gesture that co-occurs with it temporally as an **Utterance-Segment-Gesture (USG) pair.** In this context, the Utterance Segment is the segment of the utterance which is relevant to one particular rhetorical tag. For example, the phrase "I would tell him, but it is too late" would be parsed into multiple USGs: "I would tell him," and "but it is too late," and the relevant motion (specifically the gestural stroke) associated with only the corresponding specific segment of the overall utterance.

[1] In this proposed architecture, acoustic information is to be used primarily to generate beat gestures, modulate expressive dynamics of gestures, as well as determine domain priorities over the gestural analyses spanning the utterance in the event multiple relevant domains cannot be co-articulated with a single gesture. In addition, letting acoustic information be optional allows the generator flexibility to use pre-recorded speakers or text-to-speech that may lack interesting prosodic variation.

Functional Domains

For the purposes of this architecture we define a **Functional Domain** as a level at which natural language can be analyzed. In this case, we refer to the Rhetorical, Affective, and Semantic domains.

The architecture requires an interface to third-party **Functional Parsers**, with the possible outputs from these parsers defining the set of **Functional Tags** that can be applied to USG pairs in the input dataset. This interface makes the architecture agnostic to the parser's implementations. It is thus suitably flexible to accommodate evolving rhetorical, semantic, and affective text parsers popular in NLP communities, as well as acoustic feature extractors[2]. This modularity also allows our architecture to take a communicative intent or function as input, instead of text or audio. This feature gives it flexibility and an advantage over end-to-end machine learning, and makes it compatible with SAIBA guidelines for implementing virtual agents (Kopp, Krenn, Marsella, Marshall, Pelachaud, Pirker, Thórisson, and Vilhjálmsson, 2006).

Clustering by Function and Motion

The architecture uses the functional parsers to assign functional tags to all USG pairs. Each USG pair thus has at least one functional tag within each functional domain. With these tags, it establishes a **Functional Clustering** by grouping USG pairs together with others with the same functional tag. This defines different clusterings for each functional domain, with different USG pairs grouped together in different domains.

For each cluster in each functional domain, it then creates a **Sub-clustering** based on the motion of the gesture (Figure 4.2). Each USG pair thus appears in exactly one (motion-derived) sub-cluster in at least one (functional) cluster, for each functional domain (Figure 4.1).

The motion sub-clusters are further refined through pruning and combining. It is necessary to prune out sub-clusters which are significantly larger than the rest. These can occur due to noise, because not every gesture within a USG pair with a particular functional tag is necessarily relevant to that functional domain. In the "important or trivial" example, in the *Size* cluster, this USG pair may not cluster neatly with others, instead clustering into a messier sub-cluster which can be avoided at runtime as it is unlikely to contain gestures that are meaningfully associated with the "Size" semantic aspect of speech. Accordingly, the architecture works by assigning multiple functional tags, forming a categorical clustering for each functional domain. This explicitly recognizes that the gesture may be relevant to, for example, the rhetorical structure of the utterance, but not to the semantic content.

Large motion sub-clusters also form because speakers are not constantly in motion as they speak, so many gestures have little to no motion at all and cluster together (motion does not take the speaker's static pose into account).

Following pruning of each sub-clustering, each sub-cluster is compared to each sub-cluster in the other domains to create a distance matrix for all sub-clusters that spans across functional domains.

[2] Although switching parsers would require an interface to be defined between the parser output and input to this model. For example, the output of rhetorical parsers differ according to the underlying theory on which they are based.

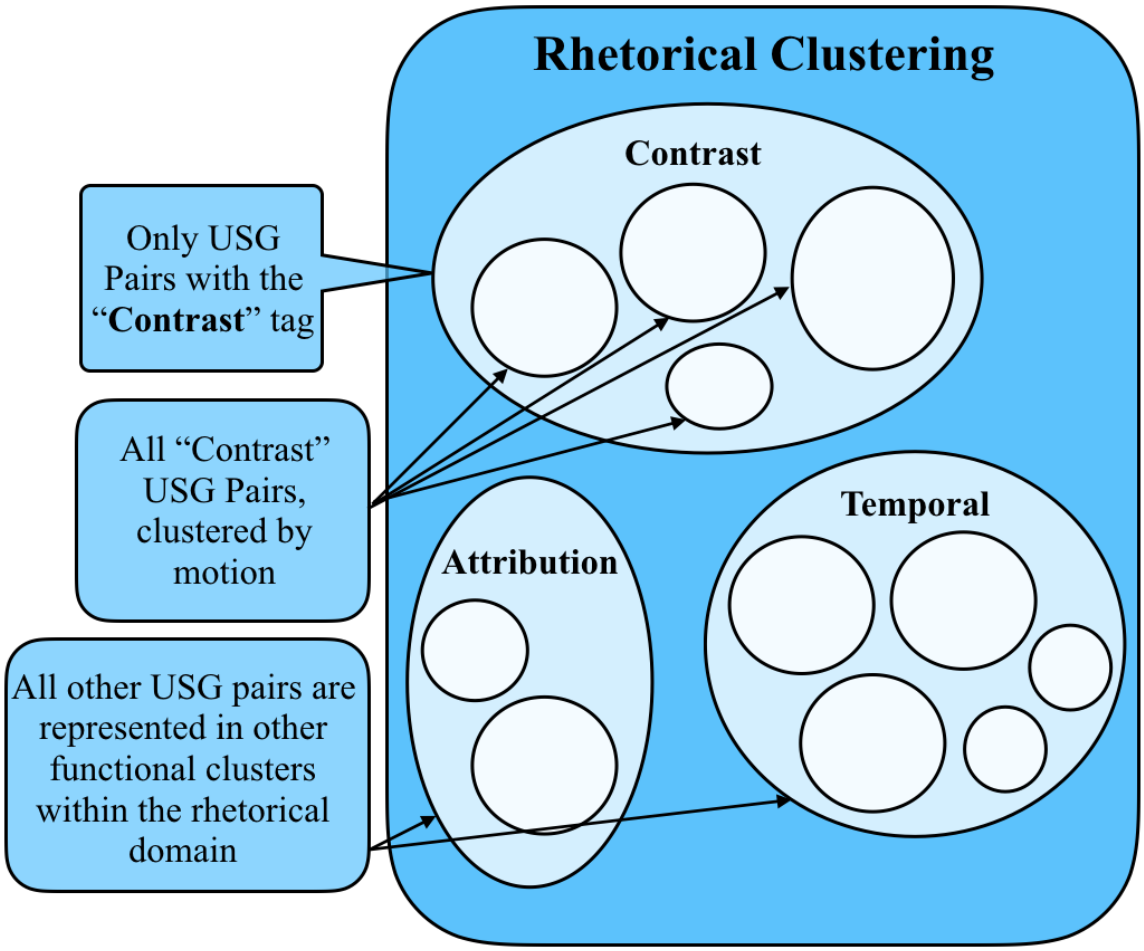


Figure 4.2: An illustration of rhetorical motion sub-clusters within tagged clusters in the functional Rhetorical domain.

This matrix describes the difference in motion between one sub-cluster and all other sub-clusters (across all functional domains), providing crucial further information with which an agent can select which sub-cluster to execute at runtime. We discuss the runtime use of this distance matrix below.

Runtime execution

At runtime, an agent uses the same parsers from pre-training to analyze an incoming utterance to perform. The agent may then select one of these functional components to emphasize according to its context, or perform a beat gesture. How an agent can choose a domain to emphasize is explored in Neff, Wang, Abbott, and Walker (2010), Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro (2013), and Chiu, Morency, and Marsella (2015). If the agent chooses to perform a gesture according to one of these functions, it selects a sub-cluster from the appropriate functional cluster to retrieve motion information and perform a gesture.

The agent must select between motion sub-clusters of its assigned functional cluster. To do this, the agent accesses the information in the distance matrix for each sub-cluster in the functional clusters. This is used to compare sub-clusters across the functional clusters that the utterance belongs to. For example, our "important or trivial," example with a rhetorical "Contrast," tag and a semantic "Size," tag (illustrated in Figure 4.3). The agent can compare the sub-clusters within these functional clusters to determine the nearest-neighbor sub-cluster, indicating that the particular motion described by these sub-clusters may be salient to multiple functional components of that utterance segment[3].

Once a sub-cluster is selected, the agent may choose to perform a gesture from it in any number of ways. Our architecture does not prescribe a specific animation but rather a family of motions which the agent's overarching architecture may interpret in a manner appropriate to that specific agent (explored in Section 4.1.3).

The labeling, clustering of the input dataset, and distance calculation of sub-clusters is done in pre-training. Because of this, the speed, and therefore feasibility of using this model in real time, is determined by the speed of the functional parsers in tagging an incoming utterance, as well as by the algorithm's contextual analysis in choosing a functional domain to emphasize.

4.1.3 Example Usage

We described the structure of this algorithm of pre-training to cluster a large corpus of gestures and select candidate gestures at runtime. We will now go step-by-step through our implementation of this architecture with an example utterance to illustrate how this model generates gestures in real time.

Let us give our example utterance "important or trivial" to an agent using this model to perform (visually illustrated in Figure 4.3). This incoming utterance is analyzed and broken up by the functional parsers and given the "Contrast" rhetorical tag, the "Neutral" affective tag, and the "Size" semantic

[3]Conversely, to reduce potential communicative ambiguity of a gesture, the agent could select a sub-cluster maximally different from other potentially relevant sub-clusters. The specifics of motion sub-cluster selection and its impact on subjective interpretation of gestures is not explored here.

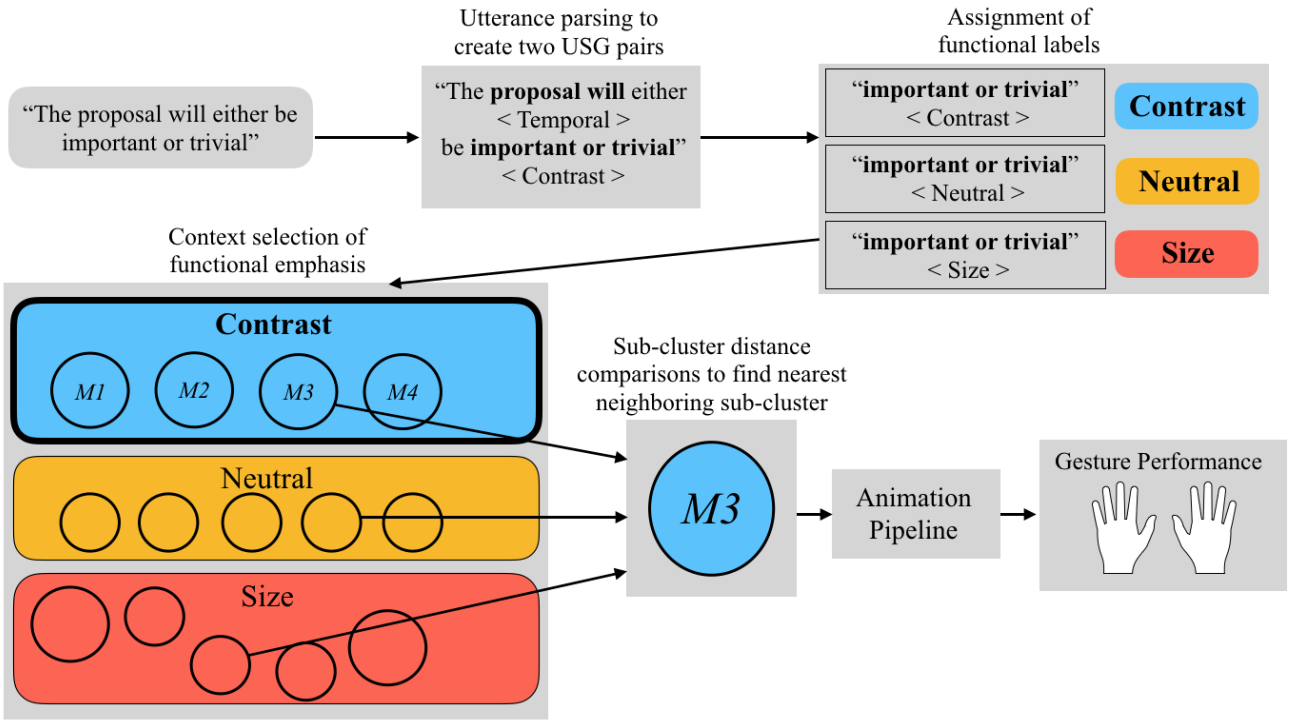


Figure 4.3: An illustration of how the architecture can select a gesture performance for the "important or trivial" example utterance.

tag. For purposes of this example, let us assume context tells the agent to emphasize the rhetorical domain of speech. We then look at all sub-clusters within the rhetorical *Contrast* cluster, and use the distance matrix calculated in pre-training to find the closest motion sub-cluster within the *Size* or *Neutral* functional clusters. The sub-cluster within the *Contrast* functional cluster which has the smallest distance to a motion sub-cluster of another domain - in this case either *Size* or *Neutral* - will be selected to perform. Potential examples of runtime animation using this pipeline are discussed in the following section.

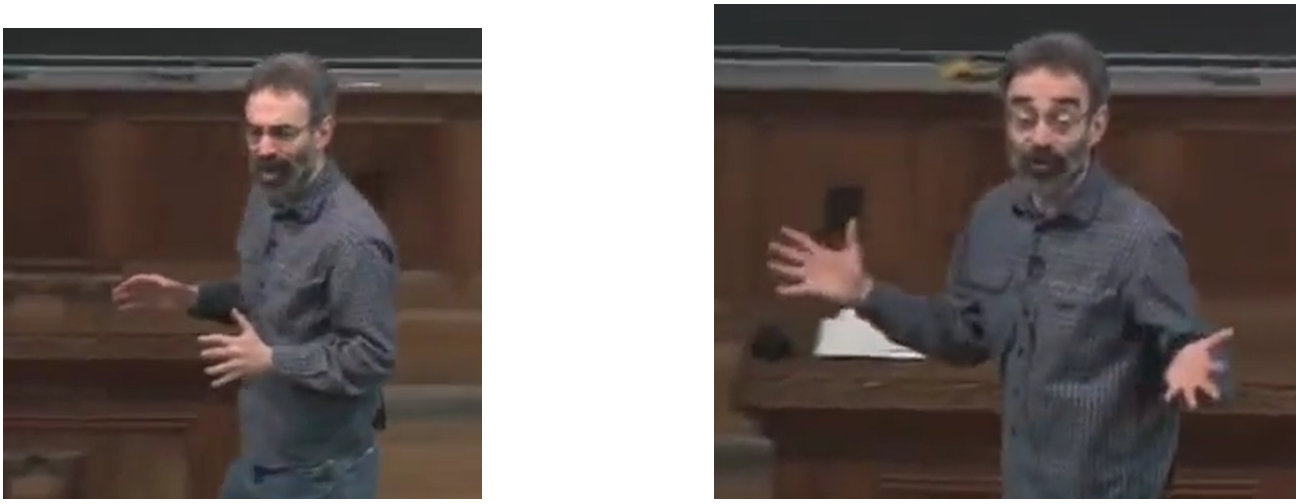
Figure 4.4 shows two specific candidate gestures from the *Contrast* motion sub-cluster obtained using this process in our specific implementation, described below in Section 4.1.4. Notice how despite different starting positions, angles, and even speakers, these two gestures follow a similar path, resulting in similar motions.

Animation Options

Although our architecture does not specify an animation pipeline, here we put forth several alternatives given the purpose of the model. As the intended use is a dynamic gesture generator to be used for on-the-fly gesture generation, these options mainly occur after pre-training and prior to runtime usage.

One option may be that an animation is created to represent each sub-cluster, with sub-clusters providing reference video for the virtual agent designer or animator. Alternatively, one can analyze a pre-defined library of animations to determine which sub-cluster they each belong to, and use this

(a) Starting and post-stroke poses for gesture by Speaker 1



(b) Starting and post-stroke poses for gesture by Speaker 2

Figure 4.4: Two gestures from the same motion sub-cluster within the "Contrast" rhetorical cluster using our implementation and input dataset.

to map sub-clusters to animations. In both cases, at runtime the agent would simply perform the animation associated with its chosen sub-cluster. This solution ensures the animation is appropriate and appears natural for the agent's form. This is feasible because the sub-clusters are clustered according to motion, and thus a sub-cluster can be represented by a single animation.

Another option could be to use the motion of USG pairs within the sub-cluster to define a "centroid" gesture. This has the added benefit of being able to be altered dynamically at runtime, although the architecture itself does not specify these alterations. However, this relies on the motion of the USG pairs to be transformed into an animation compatible format (e.g. BVH). This would therefore not be feasible with datasets that do not specify 3D motion.

4.1.4 Model Implementation for the Rhetorical Domain

In this Section we describe a method of parsing, comparing, and clustering gestures to determine their relationship to rhetorical communicative functions. First, we describe the tools used to compile our dataset. Then, we discuss our specific methods of characterising and comparing motion between gestures. We then describe how we compare the physical characteristics of gestures to feed into a clustering algorithm. Finally, we state our hypotheses underlying the use of this technique, which combines research on human gestures with computational analysis. For brevity, we only describe our implementation and objective analysis of the rhetorical domain in detail, using it to illustrate the overall approach to the different analysis pathways.

Input data

We used the pre-segmented motion and video of gestures found in Ginosar, Bar, Kohavi, Chan, Owens, and Malik (2019) as our gesture dataset. We then used Google Cloud Speech API to retrieve the transcripts that accompany each gesture. We analyzed audio from the entire video to provide context for better speech recognition, then matched the transcript section to each individual gesture based on timestamps of the original gestures and words received by the speech API. We then parsed these transcripts using the CODRA rhetorical parser (Joty, Carenini, and Ng, 2015) which is powered by the Charniak re-ranking parser (McClosky, Charniak, and Johnson, 2006) [4]. To do this, we also sent the entire transcript at once - as opposed to an individual sentence or short paragraph that would correspond to a single gesture - to achieve better rhetorical parses. Together, these tools provided a rich dataset of gesture movement and accompanying verbal communication, comprised of 11 speakers and over 500,000 minutes of frontal video.

Splitting gestures

Although in actual behavior, there is ambiguity over what constitutes an individual gesture, we can break them into the phases of the individual gesture and phrases comprised of multiple gestures (McNeill, 1992). For our analysis purposes, the key phase of a gesture is its stroke, which carries the meaning. The stroke phase can vary in length (Kendon, 2004). Gestures in this dataset were between 2 and 250 seconds (60-7500 frames), with longer gestures naturally containing more varied movement. It was therefore necessary to break these into a shorter, more standardized length.

It is common to split gestures based on motion quality (Chiu and Marsella, 2014), however we found this led to splitting gestures in the middle of spoken phrases. This created abruptly segmented and consequently confusing co-speech context for resulting gestures.

As an alternative method, we split the gestures based on the rhetorical parses of the transcripts. This preserves context in particular phrases, however it relies on gestures occurring with their relevant

[4]Current additional implementations not expanded upon in this section use the VADER sentiment analysis parser (Hutto and Gilbert, 2014) for affective parses and Spacy (Honnibal and Johnson, 2015) feeding into the TRIPS ontology (UzZaman and Allen, 2010) for semantic parses.

speech at the same time, whereas in reality gestures often precede speech (Kendon, 2000). This also introduces the problem of biasing the gestures towards being relevant for rhetorical parses, as we purposefully split gestures with respect to the rhetoric aspects of speech, as opposed to detecting a shift in semantics or changes in pitch. Furthermore, this segmentation relies on the quality of the rhetorical parser. Alternative implementations could also break gestures syntactically or affectively based on the functional parser, perhaps even splitting differently for different domains.

Splitting large gesture phrases into smaller units fails to incorporate important large-scale rhetorical structure that takes place at the paragraph (or higher) levels. For example, often in speech we reference an idea and give it "space," in our physical surroundings (Goldin-Meadow, Nusbaum, Kelly, and Wagner, 2001). We may then elaborate on that idea in various ways, referencing the physical space we created for it utterances later (McCafferty, 2004). By analyzing gestures as individual units, we knowingly fail to detect high-level rhetorical structure. We consider losing out on high-level rhetorical structure by effectively shortening our average gesture an acceptable trade-off to better cluster motion, as the purpose of this model is to generate relevant and meaningful gestures given distinct utterances, such as a turn of dialog, as opposed to, for example, a speech or lecture.

Motion Sub-Clustering

After obtaining rhetorical parses and clusters from the input data, we then create sub-clusterings based on motion for each of the rhetorical clusters. This includes determining how best to characterize the motion from keyframe values, as well as how to cluster these gestures once a suitable distance metric is determined.

Characterizing Gesture Motion

In order to cluster gestures by their motion, we developed a distance metric to determine how similar or dissimilar the motions of gestures are, which necessarily works on gestures of differing lengths. We use high-level features to create a descriptive Feature Vector of a gesture[5].

We created a 12-dimensional feature space of motion. This consists of: the maximum and minimum distance of the palms from each other, the maximum and minimum velocity and acceleration of each palm, the distance the hands move together and apart throughout the gesture, the maximum and minimum vertical and horizontal orientation of each palm, and the extent to which the hands cycle, oscillate, and change hand position over the course of the gesture. Mathematical formulas for calculations of each of these features can be found in Section 7.4. Note that these features are agnostic to the absolute position of keyframes, instead focusing on relative position between hands, and also put emphasis on two-handed gestures. We then normalized each feature across gestures, and performed K-Means clustering using the Euclidean distance of these feature vectors. The use of these features

[5]We focus on hand and arm gestures, but this architecture does not preclude analyzing full-body poses or facial gestures if such information is available.

reduces each gesture down to a single feature vector, which allows gesture comparison across different video conditions.

Although the features chosen are well-documented as meaningful in gesture literature (Calbris, 2011) they are by no means exhaustive. This technique also relies on assumptions by the implementer on the relative importance of various features of the gesture, which may be given weights (which may themselves fluctuate based on functional domain). We discuss these limitations further in Section 4.1.6.

Clustering Algorithm

We performed K-Means clustering for each rhetorical functional cluster. We used the scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011) implementation of K-Means clustering, and combined each cluster with only one gesture with its next-closest cluster. We also broke apart all clusters in the bottom 10% of silhouette scores, reassigning gestures to the next closest cluster. We determined the optimal number of clusters by running the clustering multiple times to observe the highest silhouette scores.

Hypotheses

We have illustrated a model with which to generate gestures for virtual agents in real time. The quality of these gestures relies on the assumption that the movements from USG pairs can be captured and meaningfully sub-clustered to obtain a group of gestures with similar motion profiles according to our selected features. Therefore, our analysis of the rhetorical element of the model tests the assumption that after creating categorical, functional clusters using tags obtained from the parser, we are able to effectively sub-cluster USG pairs by motion. The alternative to this would be that despite breaking gestures into functional categories, they do not cluster meaningfully using the selected motion features.

The other hypothesis to be tested is that sub-clustering by motion is in fact necessary and effective to produce communicatively meaningful gestures. It may be the case that gestures may be immediately clustered according to communicative function and naturally form families of similar motion.

4.1.5 Analysis and Results

In this section we discuss our methods of determining the efficacy of our clustering techniques, and the necessity of performing motion sub-clustering in order to generate communicatively meaningful gestures. We define objective metrics with behavioral correlates that evaluate to what extent we can expect the architecture defined above to perform gestures that are relevant to a given utterance.

Using the rhetorical splicing technique described in 4.1.4, we achieved a dataset of 66,529 gestures across 8 speakers. Of these, there were 226 unique rhetorical tags. However, as the parser only provided 20 tags, some of these are sequences of tags. For simplicity, we dropped all gestures with

multiple tags (the impact of this is discussed further in Section 4.1.6). Additionally, some of these tags do not carry gestural significance (such as the "Nucleus" tag). All such tags were grouped into one cluster with no tag. In the end this produced 43,683 gestures with 15 rhetorical clusters.

Analysis Technique

We measured sub-clustering quality with the silhouette score with respect to the Feature Vector distance metric described in 4.1.4. The silhouette score measures how well a USG pair fits within its own cluster, compared to others around it. The silhouette score s_i for one USG pair i is defined as:

$$s_i = \frac{b(i) - a(i)}{max(a(i), b(i))}, if \|C_i\| > 1 \tag{4.1}$$

Where $a(i)$ is the mean distance between i and all other points in its cluster, C_i. This is a measure of how well i fits into its cluster (the smaller the value, the better the assignment). $b(i)$ is the mean value between i and all other points in the next best fit cluster for i (with a higher value meaning a worse fit). This is a proxy for how dissimilar the next-closest cluster is (the between-cluster distance). This score necessarily falls between -1 and 1, with 1 being the best fit. We compute this value for all points in cluster C_i, and use the mean to describe the silhouette score $s(C_i)$ for the cluster. We further describe the score of a clustering as the average silhouette score for all clusters within that clustering.

High silhouette scores indicate a USG pair fits well within its own cluster, and not with the next-closest cluster. While this does lose some nuance of explicitly measuring the distance between clusters and cluster density, it is a useful proxy that is a well-established metric to measure cluster quality in the field of machine learning. This metric is also comprised of behaviorally relevant measurements: between-cluster distance, and within-cluster similarity.

A large between-cluster distance is indicative that the motions in a cluster are distinct from others near it. That is, the motion of those gestures map exclusively to their corresponding rhetorical tag, suggesting such a gesture should only be used when that tag is present or risk being confusing for the viewer. Since no other sub-clusters contain similar motions, the motion will be highly communicatively distinct within that categorical tag. Put another way, when accompanied by an utterance that falls within that functional category, the motion of a cluster that is highly distinct from other sub-clusters is likely to carry meaning.

Sub-clusters with high within-cluster similarity indicate a low variance in performance: a well-defined, specific motion. Such characteristics are relevant in the virtual-agent space because a collection of gestures with very similar movement profiles indicates the potential use of a pre-crafted library of gestures, which can be pre-loaded and run without the heavy computation of generating a completely novel gesture on-the-fly.

Functional only vs. Functional with sub-clustering

Measuring the quality of the clusters created by only using the functional co-speech elements of the gestures can indicate how well the particular motion of a gesture is relevant for the functional domain. We explored the possibility that it may be possible to skip motion sub-clustering and exclusively use functional clusters to define motion. We obtain silhouette scores for these clusters by using the Feature Vectors of each USG pair in a functional cluster to create a centroid, then measure cluster overlap using distances of these Feature Vectors to their own and other clusters' centroids.

We present two alternatives when collecting metrics: evaluating the quality of the Motion Sub-Clusterings (for example, the sub-clusters only within the *Contrast* cluster, Figure 4.5a), or evaluating the quality of the Functional Clusterings without sub-clustering (Figure 4.5b). If silhouette scores are high in the initial functional clustering, then there would be no need for motion sub-clustering as the motions defined in each category may be sufficiently distinct.Notably, these two analyses compare different sets of gestures: the former compares the motion sub-clustering only of USG pairs with a specific functional tag, while the latter compares the motion of all USG pairs by determining cluster quality using clusters defined by functional labels.

Individual vs. aggregated speaker sets

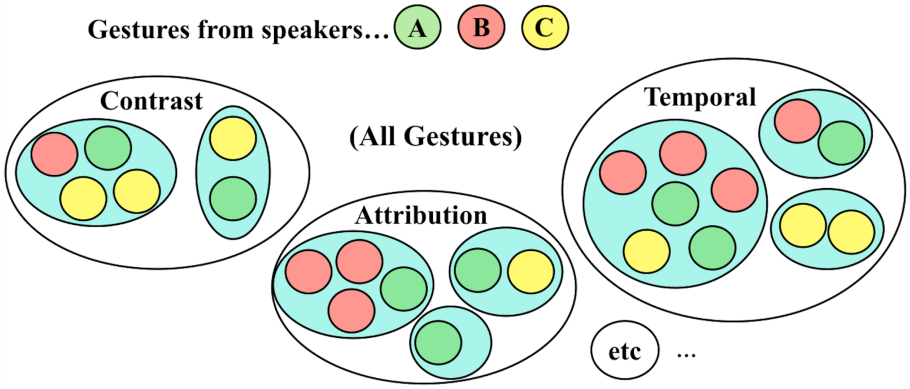
Finally, we compared the silhouette scores for both clusterings (Functional only, and Functional with Sub-Clustering) with those of individual speakers (Figure 4.5c). For this, we ran the model on all 8 speakers and found the average silhouette score for motion sub-clusterings and functional clusterings. This allows us to see trends which emerge in individuals that may be obfuscated in a dataset that aggregates all speakers. We then compare this to an aggregated dataset which contains the gestures of all speakers.

We present three evaluations of these two possible clusterings using the silhouette scores: The average silhouette score of individual speakers for functional and motion sub-clustering, the average silhouette score of the aggregated gesture set for function and sub-clustering, and the breakdown of silhouette values for sub-clusterings using the aggregated speaker dataset.

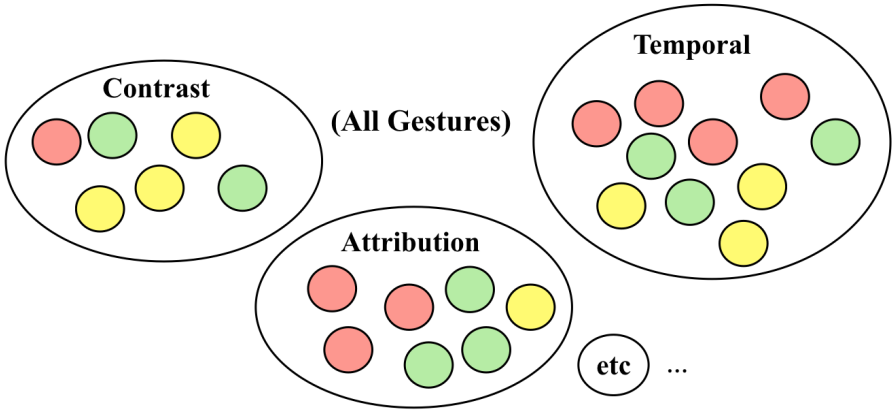
	Sub-clustering	Functional clustering
Individual speakers	0.317 (0.289)	0.018 (0.153)
Aggregated speakers	0.280 (0.304)	0.009 (0.151)

Table 4.1: Average and (standard deviation) of silhouette scores of clusterings for individuals and aggregated speakers, for sub-clustering and functional-only clustering (no motion sub-clustering). A silhouette score of 1 represents the best possible clustering for all points in that cluster, while a score of 0 is considered a very poor clustering.

(a) Visual representation of Sub-clustering analysis (aggregated).



(b) Visual representation of Functional analysis (aggregated).



(c) Visual representation of Sub-clustering analysis (individuals).

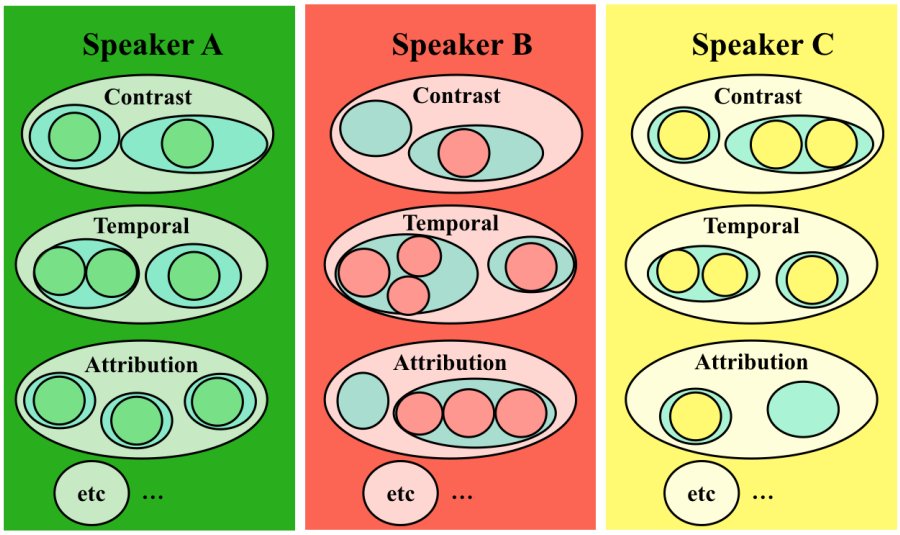


Figure 4.5: Demonstrations of Sub-clustering or Functional Clustering, and Individual and Aggregated speaker set analysis. Notice how sub-clusterings for individual speakers may result in different numbers of sub-clusters for a functional tag, and that USG pairs for the same speaker may be in the same sub-cluster when analyzed on the aggregate level but not on the individual level. Regardless, the functional tags remain constant.

Rhetorical Tag	$N_{gestures}$	$N_{clusters}$	Silhouette score
Span	20622	88	-0.25 (0.420)
Elaboration	7269	26	**0.621** (0.265)
Attribution	5795	17	**0.680** (0.275)
Joint	3404	11	0.413 (0.186)
Temporal	2081	15	0.314 (0.276)
Same-Unit	1988	88	**0.713** (0.249)
Cause	667	9	**0.685** (0.278)
Enablement	472	3	0.099 (0.391)
Background	418	27	0.280 (0.263)
Condition	360	10	0.176 (0.347)
Contrast	250	12	0.436 (0.251)
None	183	3	-0.018 (0.320)
Comparison	96	8	0.045 (0.282)
Manner-Means	52	5	0.019 (0.305)
Explanation	26	2	-0.014 (0.331)

Table 4.2: The breakdown of sub-clustering scores for each rhetorical tag when using aggregated speaker set. Number of gestures, number of motion sub-clusters, and mean and (standard deviation) of silhouette scores for sub-clusters. Selected scores over threshold of 0.6 in bold.

Interpretation of results

The improvement in scores when measuring cluster quality for motion sub-clusters instead of functional clusters (Table 4.1) indicates that clustering by motion is necessary after functional clustering. This firmly confirms our hypothesis that functional tag by itself is not enough to define a gesture, and rejects the alternative proposed in 4.1.4 that motion sub-clustering may be unnecessary. This makes intuitive sense from the "important or trivial," example as this is reflective of a phrase that may be gestured in very different ways depending on communicative function. By clustering by motion after clustering by functional tag, we separate out these very different motions and begin to establish correlational links across and within functional domains.

Furthermore, the relatively high silhouette scores for some rhetorical categories (Table 4.2) indicate that the selected features do effectively distinguish between motions within a particular rhetorical structure for some rhetorical tags, and that these motions are distinct within a rhetorical structure. These results establish moderate support for our hypothesis that clustering gestures by rhetorical structure of corresponding co-utterances, then sub-clustering by motion properties within those categories creates well-structured clusters that are relevant to the rhetorical category.

The high variation in clustering quality between rhetorical tags (Table 4.2) suggests that some structures do not have consistent canonical forms that are neatly captured by the features we used. Some of these are surprising (such as the *Comparison* cluster) and challenge our assumptions about what features may be relevant for a particular gesture.

The higher average sub-clustering silhouette scores for individual speakers (Table 4.1) suggests

this method is somewhat better suited to modeling gestures of individuals than an aggregated group, although this difference is small. This is consistent with individuals having strong tendencies to gesture in a particular style (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019). Still, the small differences in average sub-clustering scores (Table 4.1) and per tag values between 0.50-0.70 (Table 4.2) together suggest that this method is still effective when applied to aggregated speakers.

Another possibility is the higher scores for individual speakers are an artifact of different speaking conditions. While all videos are frontal views of the speaker, slightly varying angles as well as the relative size of the speaker in view makes even relational changes in position of keypoints imprecise across different video conditions. For example, if one speaker's videos lead them to be larger in relation to the overall space, their relative hand variation will look larger as well. However, this result is surprising as we would expect that if individuals gesture in a consistent manner, those motions would form a distinct sub-cluster within the larger aggregated set, and furthermore we would expect this effect to be exacerbated by variation in video settings.

4.1.6 Discussion

That some sub-clusterings achieve high silhouette scores (Table 4.2) indicates not only that there are many ways in which individuals gesture during any particular phrase, but also that rhetorical structure is indeed relevant to those motions. While there are many ways to gesture for a particular phrase, there is a limited number of families of gestures that many individuals tend to use. This is an argument for the appropriateness of gesture libraries, as one could create a specific animation for each sub-cluster (as discussed in 4.1.3).

One finding was similar silhouette scores for individuals and the aggregated group of speakers (Table 4.1). This suggests that although individuals have their own distinct and precise style of gesturing, resulting in a more effective clustering of their motions, there are some linguistic circumstances under which individuals tend to use similar motions to convey certain communicative functions.

One improvement that may further improve scores for motion sub-clustering is to explore a wider variety of motion features to map to rhetorical tags. Clusters with poor scores may perform better if we calculated motion by different features. This highlights how this architecture is not only a functional mechanism to produce gestures but may in the future also be used to test hypotheses of which features correlate to which rhetorical structures – or other high-level linguistic dimensions. Furthermore, a hybrid or weighted-feature system (particularly with automated techniques to derive weights) to determine feature vectors may lead to improvements in this domain, as certain features play a role more heavily in some communicative functions than in others. Some features may be used in the clustering of one functional domain and not in another. Further analysis must be done within each domain to determine how these features may interact to distinguish the roles a gesture plays with respect to each communicative function.

We encountered a variety of challenges which may be overcome to achieve better clustering and model performance. Constructing a dataset which is appropriate for this mapping presents the largest

obstacle. There are currently few large-scale datasets of natural social motion. Although development of recent technologies has made scraping motion from video data easier (Cao, Hidalgo Martinez, Simon, Wei, and Sheikh, 2019), these are still too noisy to effectively track certain meaningful aspects of gesture, such as precise changes in hand shape. Current datasets also do not have transcripts which accompany motion, leaving the transcription task to other third-party programs which can be error-prone and lead to difficulty for rhetorical and semantic parsers. Parsers themselves may also be improved through being trained on social conversation.

Future Work

While it is reasonable to expect that semantic and affective domains will see similar results to this domain, that assumption must first be tested. This process will also help identify relevant motion features to these different functional domains. Whereas we have implemented analyses of these domains, their mapping to motion has not been explored. This implementation also purposefully excludes rhetorical structures that occur at the paragraph or conversational level. Future analyses may explore this using different functional parsing mechanisms in combination with new motion features to provide a more holistic analysis of a wider range of input utterances.

Another avenue would be exploring the specificity of allowing clusters with multiple rhetorical tags. While these could potentially create more relevant or cleaner clusters, our initial analyses found that in practice this created hyper-specific clusters with only one gesture. Allowing multiple tags also raises a question of confidence in domain parsers; multiple tags may reveal ambiguity of the parser's analysis as opposed to specificity, and counter-intuitively lower the silhouette scores.

Although we have described one method for selecting gestures and quantitatively assessed it, a subjective analysis remains for future work. This will specifically involve crowd-sourcing opinions on traditional metrics such as naturalness of a gesture, clarity of the gesture's message, and the perceived meaningfulness of the gesture, in order to determine how well this algorithm does at selecting gestures which should then, more importantly, be tested in real-world virtual agent implementations. This inspires a variety of questions, including how well functional vs. sub-clustering selection perform on subjective metrics, such as naturalness, coherence, and appropriateness with respect to speech. While motion sub-clustering objectively produces a more consistent family of motions, whether or not human observers understand and enjoy viewing those motions remains to be seen.

The dataset used was also across a wide variety of subjects and speaker types. A subsequent experiment would be to train this model on a domain-specific set of videos in a controlled conversational setting, such as found in Ennis, McDonnell, and O'Sullivan (2010). The creation of such a dataset could further ensure high-quality motion and speech capture through use of motion capture technologies, high-quality audio equipment, and human transcription quality control.

4.1.7 Conclusion

In this section, we demonstrated a new approach through which to view the relationship between gestures and their associated utterances. We presented a novel method to map gestural motion to the gesture's co-speech properties by forming clusters based on motion properties within clusters based on communicative function. We described how an agent could make use of such a model as a generative mechanism to create socially appropriate gestures on-the-fly in conversation, and described our implementation and evaluation of the rhetorical functional domain. Our analysis finds that some rhetorical structures are often accompanied by similar gesture performances, while others are not well-defined by simple motion features. Finally, we discussed the challenges and limitations of our architecture and suggest future improvements to address them, and propose subjective evaluations building on these findings.

4.2 Motion and Meaning: Data-Driven Analyses of The Relationship Between Gesture and Communicative Semantics

Gestures convey critical information within social interactions. As such, the success of virtual agents (VA) in both building social relationships and achieving their goals is heavily dependent on the information conveyed within their gestures. Because of the precision required for effective gesture behavior, it is prudent to retain some designer control over these conversational gestures. However, in order to exercise that control practically we must first understand how gestural motion conveys meaning. One consideration in this relationship between motion and meaning is the notion of Ideational Units, meaning that only parts of a gesture's motion at a point in time may convey meaning, while other parts may be held from the previous gesture. In this section, we develop, demonstrate, and release a set of tools that help quantify the relationship between the semantics conveyed in a gesture's co-speech utterance and the fine-grained motion of that gesture. This allows us to explore insights into the complex relationship between motion and meaning. In particular, we use spectral motion clustering to discern patterns of motion that tend to be associated with semantic concepts, on both an aggregate and individual-speaker level. We then discuss the potential for these tools to serve as a framework for both automated gesture generation and interpretation in virtual agents. These tools can ideally be used within approaches to automating VA gesture performances as well as serve as an analysis framework for fundamental gesture research.

4.2.1 Introduction

Gestures convey meaning and impact attitudes of listeners towards speakers such as persuasiveness, confidence, or competence (e.g., Maricchiolo, Gnisci, Bonaiuto, and Ficca, 2009a) and can positively impact listener's learning in educational settings (Alibali, Young, Crooks, Yeo, Wolfgram, Ledesma,

Nathan, Church, and Knuth, 2013). The importance of gestures in public speaking has been known since Ancient Rome (Hall, 2004).

There are considerable differences in gesturing across individuals and conversational goals. The powerful impact of gestures within conversation makes them a focus in Virtual Agent (VA) research. In particular, automating gestures that are natural-looking and accomplish their intended communicative function.

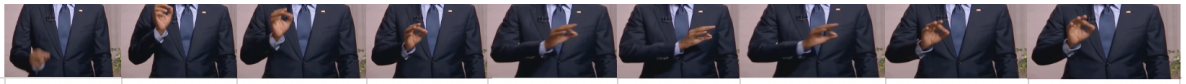
One fundamental challenge with automating this process is false implicature. Observers can ascribe intent and meaning to all gestures, even random movements (see Section 3.4.3). Ensuring an agent's gesture conveys the intended information is essential. An effective way to facilitate an agent's successful use of social gesture is to maintain some designer control over the agent's motion. However, the use of fully automated gesture generation or semi-automated approaches that retain some ability to control performance is a trade off that depends on the intended application of the agent. If the intent is simply to create the semblance of life, fully automated approaches sufficiently perform stylistic movements (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019). However, VA applications often have challenging applications, such as addressing health communications (Bickmore, Pfeifer, Byron, Forsythe, Henault, Jack, Silliman, and Paasche-Orlow, 2010), childhood bullying (Aylett, Vala, Sequeira, and Paiva, 2007), and psychological assessment (DeVault, Artstein, Benn, Dey, Fast, Gainer, Georgila, Gratch, Hartholt, Lhommet, et al., 2014). These sensitive situations require care in conveying meaning and building trust between humans and VAs. Attention must be paid to gesture to ensure meaning is conveyed memorably and appropriately, without false implicatures, and the agent builds the desired confidence in the human.

The challenge is to integrate machine-learning, data-driven processes with designer intent. We attack this challenge using (1) NLP approaches to infer how elements in the utterance convey meaning features, and (2) Clustering techniques to characterize how gestural motion conveys meaning. We seek an *explicit, inspectable mapping* of meaning elements to motion classes that supports gesture automation that a designer can influence by altering the mapping.

As a step towards such a designer tool, we present an analysis framework that derives an explicit mapping. We first discuss related work that informs the approach. Then, we describe the implementation of the framework and present example uses of its analysis. Finally, we discuss the potential use in automating performance and for basic research in human gesture, as well as discuss limitations, implications of the approach and suggest future areas of exploration and improvement.

4.2.2 Background

To contextualize the current work, we briefly discuss gesture research in psychology and how findings in this field have and have not been used in interdisciplinary research on gesture generation in virtual agents and computational gesture analysis.



Words	The	one	constitutional	office	elected	by	all	of	the people
Movement	Beat Down	Beat Down	Beat Down	Beat Down		Sweep			Beat down
Hand Shape	Precision grip								
Which Hand	Right								

Figure 4.6: Obama's hands as he speaks the phrase "the one constitutional office elected by all of the people is the presidency"

Theory

There are many useful classification schemes for co-speech gestures. McNeill (1992) talks of several dimensions under which gestures can be characterized, including deictics, beats, metaphorics and iconics. These refer to the types of co-speech utterances that occur with the gesture and the motion of the gesture. For example, beat gestures are often prosodically linked and serve to emphasize utterance content, and are widely shown to help viewers parse the rhetorical structure of speech (Leonard and Cummins, 2011; Kang, Hallman, Son, and Black, 2013).

A core concept of our model comes from Calbris (2011) and Kendon (2004), who argue that multiple gestures can be structured within larger units, what Calbris calls *Ideational Units*. An ideational unit comprises related concepts in a communication that can span multiple gestures, much like a sentence conveys interelated meaning. This influences form and co-articulation of gestures within an ideational unit and how those gestures coordinate the use of space. Consecutive gestures may share features such as handshape, handiness, and location, whereas changes in form are referential, highlighting the additional meaning being provided by the gesture in that unit. Also within the unit, there is often smooth co-articulation (blend) between gestures with an absence of rests.

Figure 4.6 shows an example of this pattern within an ideational unit in one sentence from former-president Barack Obama's 2020 speech to the Democratic National Convention. In Panel 2 he assumes a finger-pinch hand-shape (or precision grip), often associated with making a precise point. His multiple beat gestures emphasize the "constitutional office" (Panel 3-4) while maintaining this shape. Then, he draws his arm back and performs a sweep gesture across his body to create space, metaphorically illustrating "*all* of the people" that elect the president (Panel 5-7). He returns to beat gestures in the latter part of the utterance (Panel 9), but throughout the sequence of gestures, the hand-shape doesn't change. This is in line with Calbris that within the ideational unit, changes in motions or hand-shapes only occur when necessary to convey meaning.

In contrast more significant changes in form serve a demarcative function of indicating a larger shift in topic. Gesturing may come to a rest between units. The physical manifestation of grouping multiple related concepts helps punctuate conversation, underscoring the importance of how to identify and realize such informative gestures. Ideational units have significant ramifications for data driven approaches to modeling gesture and to the animation of gestures, since they imply aspects of form and motion of a gesture may be related to prior elements in the ideational unit.

Another key element of behavioral gesture research (McNeill, 1992; Kendon, 2004; Calbris, 2011) is that human conversation is often verbally and non-verbally grounded in physical metaphors. Abstract concepts and relationships between concepts can be realized via gesture (McNeill, 1992), and their physical manifestations can indicate qualities about them. For example, Grady presents the metaphors such as "importance is size" (Grady, 1997). This metaphor can be verbally realized by saying "I have a big idea,". Alternatively, an utterance such as "important idea" can be physically expressed using a metaphoric gesture that depicts a large open space between palms. These types of metaphoric gestures are ubiquitous in human-human conversation (Lhommet and Marsella, 2016), and identifying them may help impose constraints on automated VA gestures (Xu, Pelachaud, and Marsella, 2014).

Thus, the semantic concepts that underlay language can and should inform how to complement and manipulate VA non-verbal behavior in social interactions. However, the matter of how motion maps to meaning is an open question. Observational research is a helpful tool towards understanding this mapping, and often informs animations and designed behaviors for agents in high-stakes social situations. However, it lacks the precision, coverage and generalizability of data-driven approaches to gesture generation.

Gesture in Virtual Agents

The increasing popularity of Virtual Reality applies pressure to create VAs that are intuitive, complex, and capable of creating rich social experiences. It is crucial that an agent's gestures match not only the semantics of their speech, but the intended communicative function of the conversational turn. Broadly, data-driven gesture generation algorithms fall into either deep-learning, or psychologically-inspired approaches. Both have the ability to leverage prosody cues and potential semantic content from co-speech audio, though they excel at different aspects of gesture performance.

Recent advances in deep learning for motion and video generation have led to dramatic improvement in creating natural-looking, conversationally appropriate gestures (e.g. Henter, Alexanderson, and Beskow, 2020). These models produce fully automated video or animation output, and dynamically generate gesture for novel utterances. Because they are often driven wholly or largely by audio, they excel at generating rhythmic beat gestures (Bremner, Pipe, Fraser, Subramanian, and Melhuish, 2009) and effectively capture individual speaker styles (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Ahuja, Lee, and Morency, 2022; Alexanderson, Henter, Kucherenko, and Beskow, 2020). They also see improvements in qualitative performance when semantic context is added to the audio (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Liang, Feng, Zhu, Hu, Pan, and Yang, 2022; Tevet, Gordon, Hertz, Bermano, and Cohen-Or, 2022). However, semantic extraction from text for deep learning approaches relies on word or phrase vectorization (e.g. Devlin, Chang, Lee, and Toutanova, 2018) to quantify semantic input into the model. While such abstract mappings are a good first step to understand relational semantics, they lack the depth and human-readability of more grounded linguistic approaches. Additionally, while video and animation

generation has made recent vast improvements, it is still easily distinguishable and notably less natural than human motion (Kucherenko, Jonell, Yoon, Wolfert, and Henter, 2021b).

Other gesture generation models rely on a mix of automated and controlled output. Cerebella (Marsella, Xu, Lhommet, Feng, Scherer, and Shapiro, 2013) uses machine learning approaches to analyze the syntax and prosody of speech which is then fed into designer modifiable ontologies to select a gesture to perform. Greta (Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005) analyzes potential image schemas from speech to influence gesture performance. Still other models match actual human gestures to novel utterances (Ferstl, Neff, and McDonnell, 2021a; Zhou, Bian, and Chen, 2022), which avoids the problem of unnatural looking motion. However, these models must still rely on extracting information from speech or language to perform gesture matching on novel utterances, and are severely limited by the data set used to train them.

As outlined above, it is ideal to maintain designer control when creating gestures for VAs that are deployed in sensitive situations, but with some principled variation. The question of how to best integrate designer control and data-driven automation in gesture generation is certainly neither original (Wolfert, Kucherenko, Kjellström, and Belpaeme, 2019) nor straightforward.

The Current Work

This history and theoretical background informs our gesture analysis, optimizing among the following axes: (1) Gestures should carry rich metaphoric context for their co-speech utterance. (2) Each change in form of a gesture should correspond to a change in the semantics of the co-speech utterance, (3) Gestures should appear precise and natural while conveying a specific semantic concept. (4) Designers should maintain some level of control over the gesture, guided by the intention of the agent's conversational turn.

Utilizing a two-step clustering process to examine patterns of motion that correspond to co-speech semantics, as seen in Section 4.1, we analyze motion on a semantic-unit level to reveal ways in which fine-grained changes in motion correspond to derived semantics. This analysis informs clusters of natural motion that correspond to particular semantics that can be interrogated to gain insights into co-speech motion semantics, speaker and aggregate behaviors, gesture forms, and more.

4.2.3 Implementation

Full implementation and analysis code can be found in the link in Section 7.4.2

4.2.4 Utterance Meaning Analysis

To explore the relationships between utterance meaning and gestural motion, our meaning analysis breaks down syntactic, semantic, metaphoric and rhetorical structures in the utterance into **Textual Analysis Features (TAFs)**. Looking at the meaning of the phrases and words in the sentence serves

to help uncover patterns in what the speaker is trying to convey, and how those patterns relate to, and potentially predict, gestural motions. An exhaustive list of TAFs can be found in Section 7.4.2.

Syntactic Structure

The utterance is first run through the SpaCy dependency parser (Honnibal and Montani, 2017b) which identifies phrase structure, parts of speech, and dependencies between parts of the utterance. This syntactic structure is exploited by the rest of the analyses that identify semantic, metaphoric, and rhetorical content. These analyses are done at the phrase and word level as we discuss below.

Semantic Analysis

The semantic analysis processes the result of the syntactic parse using a combination of lexical and semantic databases and tools. Specifically, it currently relies on the Trips ontology (Allen and Teng, 2018) to derive semantic information. The benefit is that we can explicitly map classes of ontology features, as opposed to specific words, to gestural motion. In particular, we rely on PyTrips interface (Allen, An, Bose, Beaumont, and Teng, 2020) to the Trips ontology. Since Trips has a limited lexicon, Wordnet's larger lexicon and hypernym structure (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990; Miller, 1995) is used to map into the Trip analysis.

To illustrate how Trips is used, consider this excerpt from U.S. President Barack Obama's 2020 Democratic National Convention speech (*Barack Obama with PBS News Hour*): "Maybe you're tired of the direction we're headed, but you can't see a better path yet." The semantic analysis finds that there is a trajectory semantic feature on the word "direction" and again on the phrase "a better path." Therefore, we know that these phrases can be represented with some sort of movement through physical space, a trajectory.

We see this in Obama's gestures as well. On the word "direction," Obama sweeps his hand right to left across his body, capturing the abstract concept of path in a concrete motion. Then, on the phrase "better path," we see him travel through space again. Using his right hand, in an open palm, he motions forwards as if he is looking down a physical path before him.

Metaphor Analysis

An utterance's semantics do not always map literally to gestural motion such as the hands trajectory through space. In particular, abstract concepts can be conveyed through gestural motion using what is referred to as metaphoric gestures. To identify potential use of metaphoric gestures, we took inspiration from Grady (1997) which lays out common linguistic metaphors, such as *Abstract Concept is Concrete Object*, *Similarity is Proximity* and *Importance is Size*. These metaphors have a close relation to gesture. For example, an important idea may be conveyed by a gesture that depicts holding a large object.

ID	Semantic Key	TAF	Instances
0	The one constitutional office	Noun Lemmas	abstraction, act, duty, business, ceremony, government agency, power, work
		Adj Lemmas	essential, integral, primary, constituent
		Metaphors	Above
1	one	Ontology	Number
		Spatial	Point In Space
2	constitutional	Adj Lemmas	essential, integral, primary, constituent
3	by	Sp Ont (Speech -Prepositional Ontology)	adjacent, before, dimension, originator, topic-signal
4	all	Noun	physical object, tangible
		Metaphors	Bounded Spatial Regions, Container
5	of	Sp Ont	association-with, contain-relation, qualification, topic
6	the people	Noun Lemmas	citizenry, entity, family, group, masses, people
		Noun	plural, container, self-moving, human, phys-object, family-group, group-object, person
		Metaphors	Bounded Spatial Regions, Container

Table 4.3: Partial analysis for the phrase "The one constitutional office elected by all of the people." From this phrase we extracted several semantic keys, which each in turn contain several Textual Analysis Features (TAFs). The timestamps of the semantic keys are used to parse motion out to form individual gestures as described in Section 4.2.4. Thus this single utterance maps to seven gestures in this framework. Obama's full speech can be found at https://www.youtube.com/watch?v=oaalF5y2P0k. A complete list of TAFs can be found in the appendix.

Grady lists one hundred metaphors, but currently we focus on 56 that most likely have gestural motion correlates. As in semantic analysis, these relations are generalized so that, for example, synonyms of words like importance could be identified. In the example of *Similarity is Proximity*, we thus can map synonyms of "similar" to *Proximity*. We also incorporate antonyms to suggest *Distant*. We include the part of speech for each word to constrain the words that would return the value. For instance, object as a verb and a noun have vastly different meanings.

The metaphoric analysis operates over phrases so that metaphors of the components of the phrase are combined into overarching metaphors. We see this in a sentence such as "That is an important idea". The syntactic analysis would identify "an important idea" as a noun phrase. The metaphoric analysis takes "important" and sees that *Important is Central*, *Importance is Size*, and *Importance is Mass*. This suggests a gesture that conveys importance as big, heavy or central. The analysis of the noun phrase additionally maps "idea" using the metaphor *Abstract Concept is a Concrete Object*. This means that an idea could be shown as some sort of container or object in front of the speaker. When we combine these two metaphoric analyses, we might expect a gesture that conveys a big, heavy or central object. An alternative to this approach would be to use ML-based tools for metaphor analysis (e.g. Rei, Bulat, Kiela, and Shutova, 2017)

Rhetorical Discourse Structure

Rhetorical and discourse relations are also detected. These relations, such as elaboration, contrast, and list are related to certain kinds of gestural movements (Kendon, 1972). For instance, take a contrast indicated by a conjunction, such as "whereas" in the sentence "John works hard whereas Fred sleeps all of the time". The contrast between John and Fred can be indicated by a horizontal movement. Because we found the SpaCy analysis did not reliably find these structures, we use a database of connectives scrubbed from online repositories to re-parse the utterance. Again, an alternative would be existing ML-based rhetorical analysis tools trained on large corpora of text but few operate at the sentence level and the ones we have tried to date have been too slow for on-line use and brittle.

As this John and Fred example illustrates, discourse related gestural motions are often implicitly tied to metaphor. The metaphors *Similarity is Proximity* and *Abstract Concept is Concrete Object* together suggest such dissimilar objects would be farther apart. Since John's and Fred's behaviors are different, seen through the contrast found in the discourse and rhetorical analysis, this suggests gestures respectively denoting John and Fred would be far apart. For example, this could suggest a container or deictic gesture for John's behavior on one side of a speaker's body and then a gesture on the other side of their body to denote Fred's behavior.

Combined Analysis

As we see in the above analysis of contrast, the analysis components work in tandem to enrich the suggested gestural motion. A particularly powerful example of this can again be seen in Barack Obama's speech mentioned above. He says, "We should expect that regardless of ego, ambition or political

beliefs, the president will preserve, protect and defend the freedoms and ideals that so many Americans marched for, went to jail for, fought for, and died for." We can see a particular example of these devices working together in the phrase "regardless of ego, ambition, or political beliefs." Running this utterance through the analysis, the rhetorical analysis found that there was a "list-enumeration" tag in "ego, ambition, and political beliefs," while the metaphoric analysis found that "political beliefs" could be a *container* and *bounded spatial region*. From the semantic analysis, we can see that the ontology found that all three words were nouns that reflected "mental-constructs", and were "tangible." In his speech, Obama enumerates "ego" and "ambition" as different points in space, but on "political beliefs" he both moves to a different point while also switching to a container gesture. This is very similar to what we could expect from looking at the analysis.

Motion Analysis

Organizing gestures into both semantic categories and motion clusters is foundational to our analysis. This technique allows us to examine the physical relationships between gestures which may evoke or convey the same underlying semantic concepts. For this analysis we define a "semantic concept" as any of the features or keys that may be returned by the textual analysis described in Section 4.2.4. For example, if the feature "trajectory" is identified in the meaning analysis does a motion data set associate certain kinds of gestural motion with that feature.

Data Set

The gestures used in this analysis come from the Talking With Hands database, released in Lee, Deng, Ma, Shiratori, Srinivasa, and Sheikh (2019). This data set contains high-quality motion capture for two speakers in unscripted, spontaneous conversation, and includes data from fingers, which is often key to how gestures carry meaning and yet exceedingly rare in motion capture databases due to the difficulty of capturing hand motion. From this, we used the 124 takes that have audio associated with motion capture Bounded Volume Hierarchy (BVH) data. We used AWS Transcribe to automatically transcribe and separate the audio from each speaker because it preserves words, such as "like" and "um", as well as word repetitions. These imperfections are often important discourse markers and consequently can be highly associated with co-speech gestures (Levy and McNeill, 1992).

These 124 takes contain 14,672 conversational turns (including verbal back-channels such as "uh huh" and "yeah"). We partitioned these into individual units based on our textual analysis described in Section 4.2.4. Parsing each complete utterance led to 82,342 semantic units, which we associated with their corresponding motion by taking motion capture frames from 500ms before the start time of the first word in the phrase to 500ms after the end time of the last word in the phrase.

Our interest lies in determining how motion correlates with individual form changes within an ideational unit (Section 4.2.2), therefore the time periods that we consider for comparison are typically between 700ms-1400ms, and are intended to capture minimal and precise changes in form.

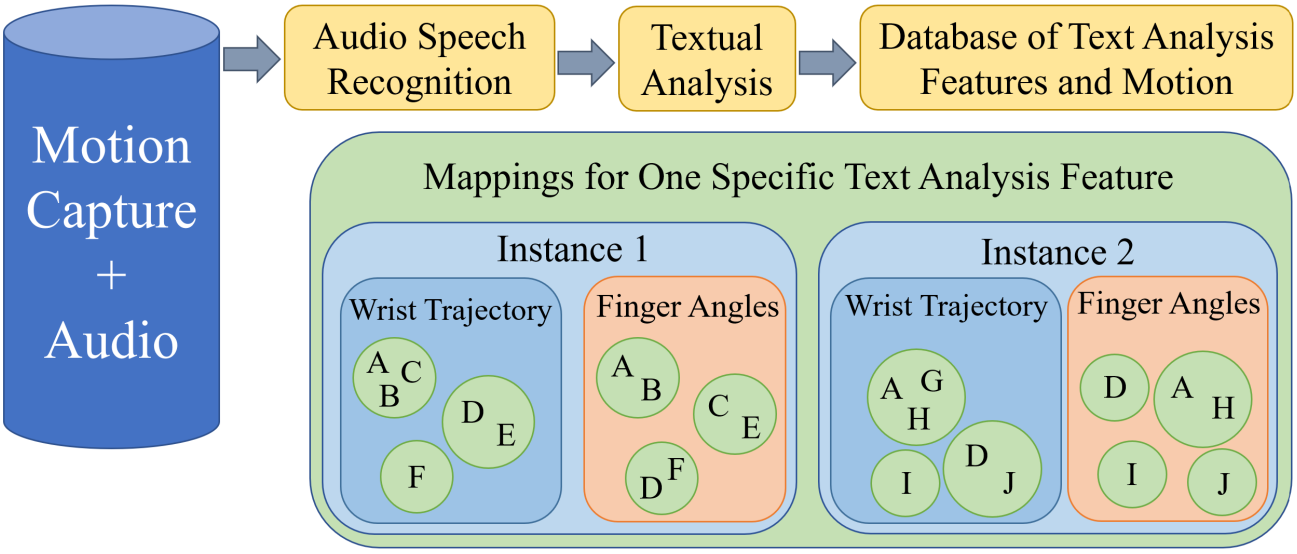


Figure 4.7: Flowchart of the analysis architecture implemented by this tool set. This shows one illustration of how hypothetical gestures A-J with co-speech utterances containing two different instances of one TAF may be clustered by Wrist Trajectory and Finger Angles. Note each gesture whose co-speech utterance contains Instance 1 appears exactly once in each of the motion clusterings. Gestures A and D appear in both instances, indicating both semantic concepts for this TAF are extracted from these gestures' utterances.

Gesture Clustering Algorithm

Our analysis relies on grouping each gesture at least twice: once categorically by an instance of a specific TAF, and once by motion. For example, one TAF extracted by our textual analysis is "Metaphor." So, we form one categorical grouping for every metaphor found across co-speech utterances, resulting in 56 different categorical clusterings. Importantly, one gesture may associated with multiple metaphors, and as such may appear in multiple metaphorical clusterings (See Figure 4.7). Then, to find patterns of motion that may be illustrative of that metaphor, we cluster the gestures of only that metaphor by motion to find sub-clusters that describe specific movements.

We use python's scikit-learn implementation of Spectral Clustering (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011). This algorithm was selected over other clustering algorithms because it does not rely on specific linear relationships between data points, but instead relative distances (affinities) of all points between one another, which is useful for extremely high-dimensional and variable motion data. Additionally, it works well to partition continuous spaces, such as the continuous, high-dimensional range of motion given by motion capture data.

Note this clustering algorithm differs from that presented in Section 4.1.4 (KMeans using a motion vector of custom extracted features). This is because of the advantages listed here for Spectral Clustering, specifically the use of relative affinities of points as opposed to absolute points in space. In this case, the distance between gesture paths as recognized by Dynamic Time Warping. This is particularly useful for high-dimensional, highly variable data, as is the case in gesture motion. This

does, however, add computational complexity to calculating motion clusters that scales exponentially with the size of the dataset – a limitation which is further discussed in Section 4.2.7.

Because of the relative sparsity of gesture in this data set, we expect many gestures for a given semantic concept to have very little motion. This may additionally leave us with only one sample of a particular motion within a given concept. This can be seen in our analysis in Figure 4.8b. Thus, we must remember that the algorithm may form one or more "catchall" clusters. In other words, not every cluster should automatically be regarded as demonstrative of a conventional pattern of motion for the semantic concept.

To determine the optimal clustering for each subset of gestures (i.e. gestures which co-occur with a specific semantic concept, or gestures performed by a specific speaker) the algorithm iteratively creates clusterings exploring the number of clusters N and chooses the clustering with the optimal silhouette score. Silhouette scores are commonly used in clustering algorithms to evaluate the quality of clustering by quantifying how well similar samples are clustered together (Rousseeuw, 1987). In order to encourage the algorithm to produce bigger clusters, it does not include clusters of size 1 in calculations for the silhouette score.

Different uses of this algorithm may want to optimize for different applications of VAs. For example, high silhouette scores may sometimes be indicative of smaller clusters, which may or may not be useful for examining certain hypotheses or for use in generation mechanisms. For this reason, for the results below, we also forced N to fall between $3 <= N <= N_g/2$ where N_g is the number of gestures used in the clustering.

4.2.5 Comparing Gesture Motion

Dynamic Time Warping

In order to compare the distance between two gestures to build a distance matrix and perform clustering, we used multi-channel dynamic time warping (DTW), which is commonly used to compare time-series data of different lengths or which map to different dimensions and has been applied to gesture motion (Dupont and Marteau, 2015; Rios-Soria, Schaeffer, and Garza-Villarreal, 2013). We use the implementation of multichannel dynamic time warping found in Wannesm, Khendrickx, Yurtman, Robberechts, Vohl, Ma, Verbruggen, Rossi, Shaikh, Yasirroni, Zieliński, Van Craenendonck, and Wu (2022).

Joint angles vs. positions

BVH data describes joint angles for individual motion channels frame-by-frame. This can be extremely useful for gesture analysis as by sub-selecting specific joints, motion analysis can be agnostic to the absolute position at which a motion takes place. Because the range of joint angles differ hugely between body positions, we first normalize each joint angle channel. In this demonstration of these tools, we only consider joint angles in the upper body.

Due to the hierarchical nature of this data, a small difference in any given angle may or may not be important to the gesture. For example, a small degree difference in shoulder rotation could dramatically affect the path of the wrist – a key component of gesture. However, depending on other joint angles up the skeleton, the wrist may actually remain stationary.

Therefore, we convert the joint angles into XYZ positions in space and use these paths as the motion data to compare using DTW. In order to remain agnostic to the gesture's absolute position we cluster based on the velocity of each channel as determined by the frame-by-frame difference of each channel. Functionally this means we cluster not on the absolute motion of the gesture but on the trajectory of each individual joint.

However, one problem with focusing on the high-level path through space of the gesture is that this fails to capture the minute but extremely informative and important component of gesture that is hand shapes. We address this by clustering the gestures according to motion in two different ways, and examining the overlap between each clustering.

Utilizing Overlapping Clusterings

A hand shape's hold during a specific motion can be illustrative of continuing a particular ideational train of thought (Figure 4.6), or conversely, its change could be an indicator of a shift in topic. To incorporate hand shape information into this analysis we, in parallel, cluster gestures by relative position of the wrists through space, and by the differences between joint angles of only the fingers. From the first clustering we see families of wrist path through space, and from the second we see specific families of hand shape changes (or holds). Then, we observe which motion clusters a specific gesture falls into within each of these motion sub-clusterings. This paints a picture of gestures which utilize high-level motion, hand-shape changes, or both to illustrate a given metaphor. This is illustrated by the example in Figure 4.7 by gestures A, B, and C being clustered together in the Wrist Path clustering, but only A and B by changes in Finger Angles, thus extracting high-level information about the gestures by virtue of their similarities and differences between other related gestures. Namely, that Gesture A and B share similar properties in terms of how their hand shape *and* wrist trajectory change over the course of the gesture, and Gesture C shares this wrist trajectory but with a different change (or hold) in hand shape.

By looking at the trajectory of the wrist and finger angles of the gesture and not absolute position, we specifically focus on changes in motion that may correspond to co-speech semantic content. That is, the absolute motions of gestures in the same cluster may be different, but the forms and motion qualities of the gestures change in similar ways. Thus we isolate the form of the gesture that changes in order to convey an idea within an ideational unit.

4.2.6 Results

In this section we demonstrate how to use the tools and techniques defined above to quantify the relationship between motion and meaning. The principles behind the results shown here could correspond with any TAFs (i.e. those discussed in Section 4.2.4 and exemplified in Table 4.3), but we use metaphors (described in Section 4.2.4, exhaustive listing in Section 7.4.2) as an example of a feature that can be analysed in conjunction with gestures. Specifically, we use this framework to address the following questions:

1. To what extent can we identify and classify the motion associated with a specific metaphor identified by the linguistic analysis?

2. How do differences in wrist trajectory and finger angles carry meaning for particular metaphors?

3. How do these patterns differ across individuals?

Throughout this exploration we focus on clusterings based on wrist trajectories and finger angles. However, one can perform a similar comparative analysis by extracting any BVH channels relevant to other specific hypotheses.

(Q1) Exploring the Relationships Between Metaphor and Motion

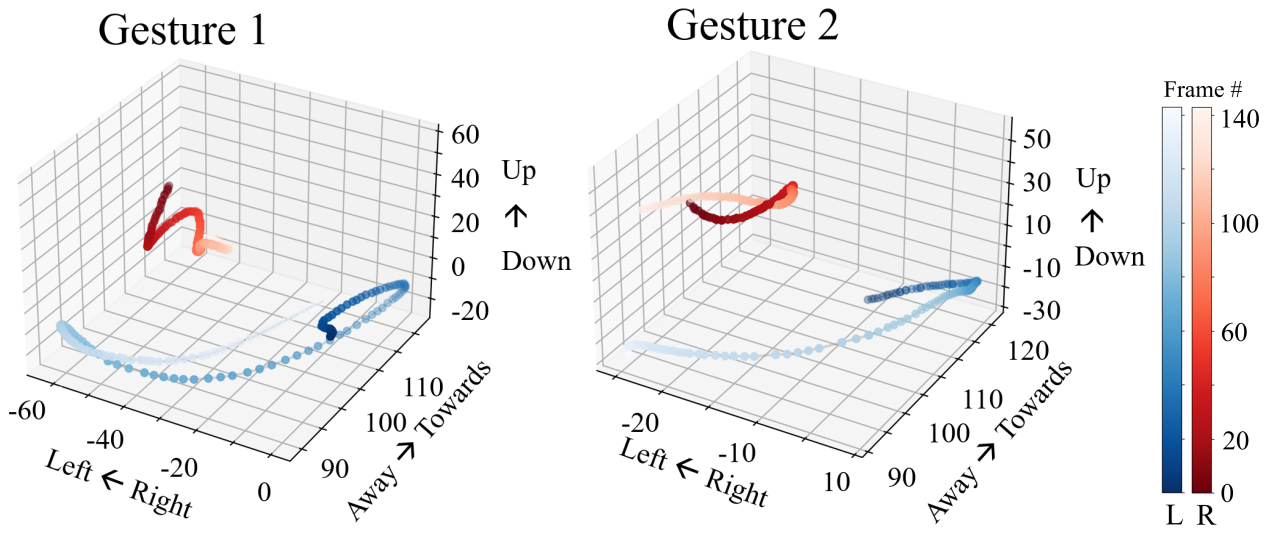
In order to understand how speakers move in conjunction with a specific metaphor, we follow the two-step clustering process described in Section 4.2.4 and first consider all gestures in the dataset for which the linguistic analysis of the co-speech utterance of the gesture suggests possible metaphors. This forms the first (categorical) clustering. Then, using only the gestures for which the co-speech utterance elicits that particular metaphor, we perform motion clustering described in Section 4.2.4.

Example: "Moments in Time are Objects in Motion Along A Path"

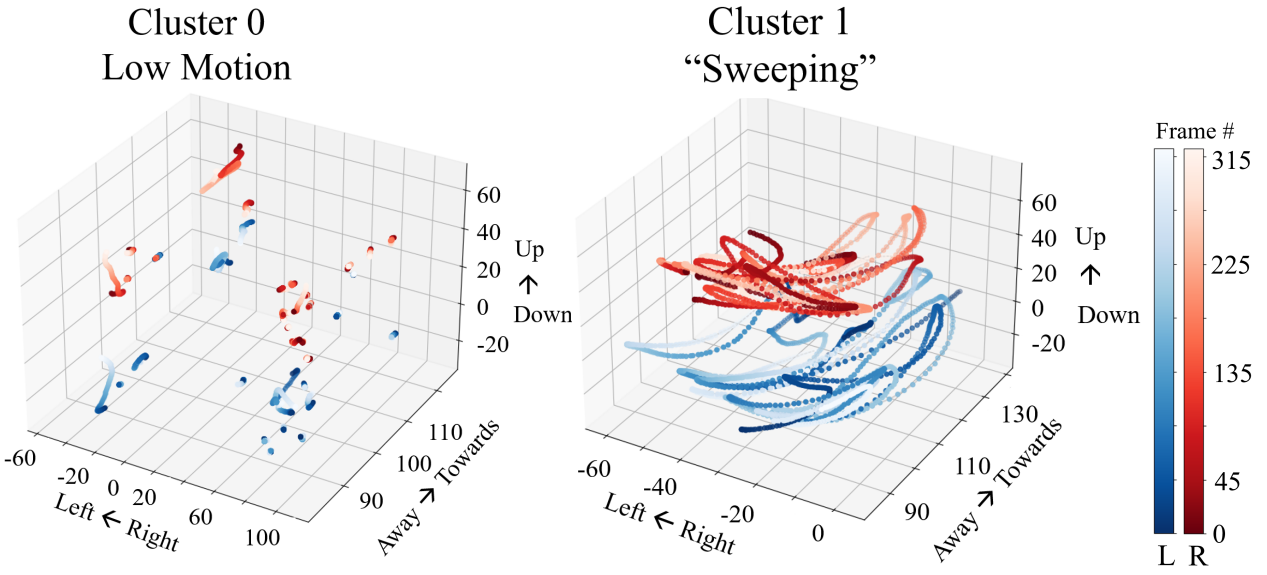
Let us dig deeply into the motion analysis of this metaphor (Grady, 1997). There are 448 unique instances of this metaphor being evoked in co-speech utterances, through phrases such as "It took us like eight hours," and "She hadn't done it by that point."[6]

We then consider the two alternative motion clusterings of these gestures. First, we cluster based on the wrist trajectory. This yields eight motion clusters, one of which is very large, acting as an "catchall" rather than a source of common trajectories. The presence of the other clusters indicates there are potentially seven common wrist trajectories used to convey this metaphor. Then, we cluster based on changes in the finger angles in the gesture. This yields 11 small clusters and one very large "catchall" cluster. The silhouette scores of each clustering indicates that a change in finger angles

[6]Note that we focus only on the motion that elicits the metaphor according to our textual analysis. In these parses, for example, we would look only at the motion that occurs during the words "eight hours," and "by that point."



(a) The wrist paths of two gestures from the same Wrist Trajectory motion cluster. Although the raw motion between the gestures differs significantly, they follow similar trajectories through space.



(b) Overlays of the wrist paths of all gestures of two different motion clusters. Notice there is significantly less motion in Figure 4.8b (0) than Figure 4.8b (1).

Figure 4.8: Selected gestures and overlays of all gestures in different wrist path motion clusters for the metaphor *Events In Time are Moving Objects Along a Path.*

(form of the hand) throughout the gesture does not seem to be consistent across gestures with this metaphor (Table 4.4).

The overlap between these two different ways to analyze motion also affords several insights into the relationship between motion and meaning. Gestures being clustered together in each clustering indicates that they share traits in both the wrist trajectory and hand shape changes. Figure 4.8a depicts the paths through space of the wrists of two such gestures. Although the raw motion of these gestures differs fairly significantly (note the differing axes indicating positional differences), they are clustered together based on how the wrist *moves* throughout the gesture. For example, they each incorporate large sweeps towards and away from the body.

When we examine in depth these seven different motion wrist trajectory clusters we see several promising patterns emerge. Firstly, one cluster seems to be fairly static (Figure 4.8b (0)). This is expected as although gestures are prevalent throughout conversation they are still relatively sparse compared to every idea conveyed. Others tend to depict sweeping motions going towards and away the torso (Figure 4.8b (1)), in line with what observational gesture researchers may expect from this metaphor.

The power of these clusters is their precision. From any cluster we can compute many informative metrics: the velocity of the wrist, the average change in hand shape or wrist rotation, the absolute and relative positions of the speaker's wrists to one another, and to the viewer, etc. Thus, we quantify the extent to which speakers naturally tend to alter their motion according to specific parameters, such that one could use these parameters as input to an external generative mechanism.

(Q2) Hand Shape and Wrist Path Changes

On the surface, the composition of the finger-angle motion clustering for the metaphor described in Section 4.2.6 may mean that changes in hand shape are not especially salient to this metaphor, as there is not much consistency between hand shape changes across all gestures that evoke this metaphor, whereas we see more defined patterns of motion in wrist trajectory.

However, when we look at other metaphors we see a different picture. Table 4.4 shows silhouette scores for Wrist Trajectory and Finger Angle clusterings for a selection of metaphors (optimal clustering determined by the method described in Section 4.2.4). For some metaphors (e.g. Forward and Location-Within) we see pronounced differences in silhouette scores between these two clusterings. This indicates that changes in finger angles throughout gestures were not consistent among gestures with this metaphor in this dataset, which may in turn be interpreted as hand shape changes not carrying salient information to this metaphor. Contrast this with metaphors Bounded Spatial Regions or Comparative (Higher), for which the clustering based on finger angles resulted in much cleaner clusters, meaning changes in hand form were consistent within these metaphors. This could mean that a change in hand shape carries meaning independently, or actually that the change is strongly correlated with a specific path or other motion in the gesture, and that two changes together are necessary to carry the semantics of the gesture. Such relationships can be explored extensively using these techniques.

Metaphor	N_g	WT	FA
Moving Objects Along A Path	448	0.65	0.38
Container	2127	0.86	0.46
Bounded Spatial Regions	902	0.56	0.61
Forward	265	0.46	0.19
Comparative-Higher	354	0.19	0.39
Location-Within	789	0.47	0.14

Table 4.4: A table of silhouette scores for Wrist Trajectory (WT) and Finger Angle (FA) clustering for a selection of metaphors. N_g represents the total number of gestures with that metaphor.

It is important to note that finger motion is often highly correlated to wrist position, and, in this dataset in particular, not every BVH file contains high-precision finger motion. It is possible that differences between wrist trajectory and finger angle clusterings are partially artifacts of the conversion from joint angles to XYZ positions, or imprecision in the motion capture data.

(Q3) Insights About Specific Speakers

Another way to examine these clusterings is to determine the variability of patterns of motion across speakers. Table 4.5a shows the speaker breakdown of each wrist trajectory motion cluster in the metaphor described in Section 4.2.6 (optimal clustering determined by the method described in Section 4.2.4). When we cluster only the gestures of one specific speaker, there are roughly similar numbers of motion clusters, but tend to see higher silhouette scores for the clusters that do form. And, when we drill down into the speaker composition of each motion cluster in the aggregate clustering (Table 4.5b), we see some clusters are dominated by one or two speakers. This supports the well-established notion that speakers have specific styles of gesturing (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019; Neff, Kipp, Albrecht, and Seidel, 2008; Gallaher, 1992), and further suggests that speakers also have specific patterns of motion they tend to use to indicate a particular metaphor.

4.2.7 Discussion

There are many potential applications for these tools in gesture motion analysis in basic research, as well as production and interpretation in Virtual Agents. Here we discuss some of the limitations and implications of this approach, and suggest future improvements.

Limitations

Utterance Analysis

These examples are not meant to suggest that the analysis is designed to predict the gestures a particular speaker will display in a particular role and context. To do that, we would need to build or train a model for the individual or role to handle the significant individual and contextual differences

Speaker ID	N Gestures	N Clusters	Silhouette Score
Aggregate	448	7	0.65
deep5	138	8	0.24
shallow26	41	12	0.86
shallow19	26	6	0.55
deep4	26	6	0.82
shallow16	24	8	0.78
shallow12	21	6	0.52
shallow18	20	8	0.80

(a) Wrist motion cluster silhouette scores for all speakers vs. a selection of individual speakers for the metaphor *Moments in Time are Moving Objects Along a Path.*

Cluster Label	N Gestures	Speaker ID	Cluster Percentage
0	304	deep5	28.4
		shallow28	11.6
		shallow26	10.4
		shallow19	6.3
		shallow16	5.8
		deep4	5.5
		& 13 More*	17.7
1	47	deep5	78.2
		& 5 More*	21.8
2	43	shallow13	58.6
		shallow12	16.3
		deep5	16.3
		& 3 More*	8.8
3	23	shallow12	23.5
		shallow20	22.5
		& 16 More*	54
4	15	deep4	66.6
		shallow14	13.3
		shallow25	13.3
		deep6	9.1
5	10	deep5	60
		deep4	40
6	6	deep5	100

(b) Speaker composition of Wrist Trajectory motion clusters. * indicates several more speakers who each make up less than 5% of gestures within the cluster.

Table 4.5: Individual and aggregate speaker silhouette scores and cluster composition for gestures that elicit *Moments in Time are Moving Objects Along a Path.*

in gesture. There is also not a one-to-one mapping in general between utterance content and gesture (Kendon, 1997). Content in an utterance may not be conveyed gesturally and a gesture may convey content not in the utterance.

The current analysis can also fail to recover important meaning. While it does well analyzing noun phrases and conjunctions, it performs poorly on verb phrases, especially in relation to participles. Consider the sentence: "My career was taking off." Here "taking off" suggests an increase in success and an upwards motion. However, the analysis does not recognize the verb and participle as a whole and only analyzes "taking." One approach here is to improve its database of verb forms (e.g., Allen, An, Bose, Beaumont, and Teng, 2020) or to use transfer learning applied to a text-to-text transformer language model.

Motion Analysis

A key limitation of this analysis is that it is based on a small motion capture data set. Although people produce thousands of words every day, many of which are accompanied by gestures, there are still not many resources to obtain data to analyze (Olugbade, Bieńkiewicz, Barbareschi, D'Amato, Oneto, Camurri, Holloway, Björkman, Keller, Clayton, et al., 2022). For example, this data set yielded around 50,000 instances of metaphors. However, because of the long-tail problem of language, this is not evenly distributed across all metaphoric concepts. Because of this we are left with many outliers that do not fit into patterns that may co-occur with the metaphor. That does not mean that these gestures are trivial, but rather that there were not enough other occurrences of this metaphor within the data set. This means that we could be throwing aside a lot of useful information that may have been caught with more data.

This analysis assumes that motion and semantic information that is conveyed in language are tightly temporally correlated, but in natural speech gesture often precedes its linguistic correlate (Kendon, 1997). We mitigate this by buffering the motion we select for a semantic parse, and by filtering gestures by the semantic information in their co-speech utterance. According to Calbris' concept of ideational units we can assume the speaker would, at worst, not perform a gesture conveying one idea while actively speaking about another. So if there are instances of speakers gesturing in conjunction with a concept unrelated to the language in their utterance at the time it is likely to occur only once in this relatively small data set, and thus be physically distinct from the patterns of motions used to convey the utterance such that it will not share any neighbors and thus not form its own cluster.

As mentioned in Section 4.2.4, Spectral Clustering has several advantages over other motion clustering methods. However, because it is based on a similarity matrix of all gestures, this technique fails to scale with larger datasets (which, as mentioned above, are challenging to acquire but necessary for large-scale gesture analysis). Several related clustering techniques based on motion similarity, such as Agglomerative or DBSCAN Clustering, should be evaluated to determine an optimal balance between computational complexity and clustering efficacy.

4.2.8 Applications

A core motivation for this technique is maintaining designer control. The utterance meaning analysis characterizes classes of meaning which are in turn mapped to classes of gestures based on the clustering. This mapping is explicit and therefore readily modifiable by a designer per application or speaker. For example, mappings can be prioritized to capture speaker style or to eliminate gestures in gesture-matching systems (e.g. Lhommet, Xu, and Marsella, 2015; Ferstl, Neff, and McDonnell, 2021a) when too many gestures are proposed for an utterance. At a finer grain, selection of a candidate gesture from a cluster may be biased by speaker style or by dynamic features to convey prosodic or affective qualities.

A significant related challenge to gesture generation in Virtual Agents is that of gesture inference. That is, how can agents use social non-verbal cues to infer and add meaning to conversation? Agents can use mappings created by this technique as an embedded model to analyse human motion, potentially allowing the agent to inform its Theory of Mind of its conversational partner. By using motion as a tool to inform semantic or thematic recognition within conversation, an agent may improve their understanding of speech, language, and speaker-specific models of communication.

Along these lines, clustering gesture based on different speakers can provide insights into nuances of how motion comes together to create personality. Quantifying the manner in which similar overall gestures differ from speaker to speaker can help dictate ways in which agents can manipulate their own motions to convey an individual sense of style.

The technique described in this section is not limited to one motion format or textual analysis approach. We invite researchers interested in further aspects of NLP, animation, and gesture to augment and modify these tools and algorithm to suit their specific hypotheses in relation to gesture and language.

This technique is also capable of a wide variety of analyses. One can leverage the ability to "average" each of these gestures clusters and determine which speaker tends to gesture most or least similarly to the average gesture, potentially quantifying to what degree each speaker uses canonical gestures to convey a particular metaphor.

Another area of exploration is conversational turn taking. In preliminary explorations we have found that gestures tend to have more similar trajectories to one another across changes in conversational turn than to other gestures that express the same concept, potentially providing a basis to quantify conversational gesture mirroring. That is, the gesture of one speaker's conversational turn following that of their partner is more similar to their partner's gesture directly prior than other gestures throughout the conversation. Similarly, a speaker's gestures tended to be more similar to their conversational partner's gestures, across conversational partners; that is, while speakers maintain an individual style of gesturing, they adjust their style to be similar to that of their partner.

A related avenue to explore is using motion to infer large shifts in conversational topic. We can include lower body data to retain information about hip rotation and orientation to help an agent infer when to pursue or shift themes in conversation (Ennis, McDonnell, and O'Sullivan, 2010), leading to

improved Human-Agent social interactions.

4.2.9 Conclusion

In this section we present an analysis technique that maps semantic concepts in language to gestural motion. We describe how clustering gestures according to changes in motion takes into account the theoretical concept of ideational units, so as to understand how nuanced semantics can be communicated across gestures. We demonstrate the application of this technique on a data set to map conversational gesture to metaphor. We invite researchers in this space to use and build upon this tool set to perform principled, data-driven analyses that grounded in behavioral theory.

Chapter 5

Data-Driven Testing of Behavioral Observations

Chapter 1 explores the history of behavioral gesture research in psychology and how theoretical gesture research both has and has not influenced computational gesture generation in virtual agents. In particular, it highlights the development of several psychologically-inspired data-driven methods of gesture generation (e.g. Poggi, Pelachaud, Rosis, Carofiglio, and De Carolis, 2005). However, deep-learning approaches rarely incorporate explicit insights from behavioral research into the design, utility, and evaluation of such algorithms. Here, I present the main goal of this thesis: to bring computational techniques to the study of gestures such that insights from behavioral psychology can be readily applied to algorithmic gesture generation.

This chapter establishes hypotheses based on theoretical gesture research, identifies gestures that may relate to these hypotheses, then, using the framework presented in Chapter 4, performs a principled, data-driven analysis of natural gesture behavior. This demonstrates explicitly how this architecture and its implementation can be used to characterize the relationship between semantics and motion. This chapter then describes how this can be exploited by both animation designers and generative algorithms for virtual agents to both understand and produce rich, complex, human-like gestures.

This Chapter shows how we can use insights and theories gleaned from observational gesture research to inform data-driven analysis methods that are sensitive to embodied metaphor and fine-grained gesture motion on the Ideational Unit level (See Section 1.1.4). The work herein demonstrates how we can exploit this relationship to generate rich, communicatively nuanced gestures on virtual agents. Although the variety of motion and co-speech utterances of the gesture space is enormous, this Chapter shows how we can start to characterize and codify that variation using modern data-driven techniques. The objective of this framework is to allow both specific hypothesis testing and exploratory analysis of how linguistic concepts conveyed in a gesture's co-utterance correspond to its motion.

5.1 Gesture Metaphors

Section 1.1.2 and Chapter 2 describe the relationship between gesture and both linguistic and abstract metaphor, particularly inspired by linguistic metaphors found in Grady (1997). Metaphors that are represented in language are often physically manifested through gesture in ways that can disambiguate or even add semantic information to the gesture's co-speech utterance (Jamalian and Tversky, 2012a; Kelly, Özyürek, and Maris, 2010).

Communication through metaphoric gesture relies on the speaker and viewer sharing a mental model of how a concept can be expressed physically via gesture. This is further complicated when speakers convey multiple ideas through multi-metaphoric gestures and when considering the cultural context of the conversation (Section 1.2 and Section 2.2). Therefore, the metaphors explored below are considered only for speakers from the dataset who are native English speakers and who grew up in the United States[1].

5.1.1 What Does It Mean For a Metaphor to be "Represented" in Language?

Metaphors – or more precisely, concepts represented through metaphors – can be communicated through multiple channels both verbally and non-verbally. The examples discussed both in Chapter 1 and in this section by Grady (1997) are all explicit linguistic metaphors: The language used to convey particular concepts is literally metaphorical. That is, in English we use the same word to convey both physical and abstract meaning. For example, the concept of an "idea" cannot be physically realized and consequently does not have a physical size, yet in English we say that an idea is "big" if it is sufficiently complex. The use of physical descriptors for abstract concepts demonstrates the explicit representation of metaphors in language.

However, the *implied* representation of metaphor in language is critical to understanding metaphoric gesture. While Grady's metaphors are explicit in language, they can also be implied using non-physical language in conjunction with non-verbal behavior. Abstract concepts can be expressed both in purely abstract terms in language while evoking a physical metaphor in gesture. Human communication is thus deeply enriched by Embodied Cognition; Our capacity to blend literal and abstract mental representations of concepts enhances our ability to precisely convey our intended message in conversation. Consequently, a metaphor can be inferred from language without being literally in the language. I present an example of metaphoric representation being evoked in this way in Section 5.2.1. Specifics of how the parser extracts concepts from speech are provided in Section 5.1.3.

[1]This information was not provided explicitly as meta-data, but was gathered by myself by listening to the dataset and using culturally relevant information provided by the speakers, e.g. that they celebrate Thanksgiving. If the native culture of the speaker is ambiguous from the provided recordings I excluded them from the data used to test hypotheses in Section 5.3.1.

5.1.2 Metaphors Implied in Gesture

However, not all metaphoric concepts are represented in speech. Gestures often rely on metaphors that are not represented in language at all, even implicitly. As discussed in Section 1.1.1, not all gestures require accompaniment by a co-speech utterance. Although Table 1.1 does state that metaphoric gestures necessitate a co-speech utterance, it is not the case that the utterance itself must carry metaphoric content.

Consider an example from the data set described in Section 4.2.4. One exchange contains a speaker saying "are you going to choose this one, or this one? Obviously this one" while gesturing using two hands. They hold their left hand loosely open with palm upright at their waist and beat while uttering the first "this one," and hold their right hand in the same orientation but at the height of their head and beat during the second "this one." When they say "obviously this one," they lower their left hand to a rest position and keep their right hand open and upright, indicating the hand that was previously elevated was the "obvious" (Good) choice. This is an example of evoking the one of Grady's metaphors *Good is Up* without a linguistic counterpart. This example demonstrates a significant challenge to automated gesture analysis, as the metaphoric concept evoked by the gesture does not have any linguistic complement that can be automatically recognized through semantic parsing. Therefore, this gesture, despite being illustrative of semantically meaningful metaphoric gesture, is challenging to capture using automated methods. I discuss this problem more in Section 5.6.1.

5.1.3 Extracting Metaphors Using Semantic Parsing

Because conceptual metaphors may not manifest explicitly in language, automated parsing must consider a wide variety of ways a metaphor may be represented. This includes the examples such as the one shown in Section 5.2.1 in which the language implies that a conceptual metaphor could potentially be evoked with gesture. Grouping gestures by concepts is only possible if those concepts are recoverable in language (see Section 5.1.2), but it is necessary to highlight that the process of extracting this information is transparent and human-readable.

The details of the semantic parser are described in Section 4.2.4, and are inspired from Cerebella (Lhommet, Xu, and Marsella, 2015). In order to filter gestures with potentially metaphoric content this framework uses *Functional Metaphors* implemented in the semantic dictionary deployed in Cerebella, which were originally chosen and designed specifically because of their observed relevance to gesture behaviors. Fundamentally, they search for extensions of key metaphoric concepts described in Grady (1997) leveraging the hierarchies in WordNet (Pedersen, Patwardhan, Michelizzi, et al., 2004). For example, detection of the concept of *Happy* uses semantic dictionaries of synonyms constructed from WordNet that suggest positive sentiment. Similar dictionary entries are used for the other metaphors assessed here. Additional inferencing is done at the phrase and clause level to get compound constructions or rhetorical structures, leveraging a syntactic parse by the Spacy parser (Honnibal and Johnson, 2015). This means that conceptual semantic analysis is not restricted to the word-level and can extract

multiple concepts from phrases like "an important idea."

The implementation used in these analyses was chosen due to its transparency, as WordNet makes clear the reasoning behind relationships of words and (some) phrases to overarching concepts. This allows consumers of the resulting mapping to inspect precise relationships between raw utterances and concepts, and edit those relationships if desired.

5.1.4 Multi-Metaphoric Analysis

Of course, both language and gesture may convey multiple abstract concepts simultaneously, with many different potential manifestations (see Chapter 2). A sentence may convey multiple potentially metaphoric concepts which may each in turn inform how a speaker may emphasize or clarify elements of their speech. For example, the sentence "my mind was wandering off" conveys the past ("Time is a Line"), information about mental states ("Mental States are Places"), and even movement ("Process is a Path"). Each of these concepts may be relevant to the speaker's gesture while speaking this sentence, and so it is important to include this gesture in the analysis for each of these metaphors.

The beauty of this analysis technique is indeed its ability to consider metaphors both individually and in combination. The analysis itself operates on sets of gestures, but is agnostic to how those sets are determined. Thus, we can compare not only individual metaphors represented in a co-speech utterance, but multiple metaphoric concepts and sets of combinations of concepts. This further affords the possibility to disentangle different influences and potential hierarchies of metaphoric gesture content. I discuss the potential of this technique for multi-metaphoric analysis more in Section 5.6.2.

5.2 Metaphor Examples

Several embodied metaphors are commonly implemented in rule-based gesture generation systems. They are well-established in observational research, and often accompanied by "canonical" gestures that imply hypotheses around how humans gesture while evoking them.

5.2.1 Happy Is Up

The concept of "Happiness" has several abstract relationships in US vernacular of the English language. Grady identifies some such as "Happy is Bright" ("I was feeling bad, but she really *brightened* my day") and "Happy is Healthy" ("Only a *sick* man would do a thing like that."). However, these abstract comparisons are difficult to map to physical gesture spaces. However, one metaphor which is easily manifests physically is "Happy is Up." This maps well both to a linguistic space ("I was feeling *low* yesterday but the weather has really *picked me up*.") as well as gesturally.

One example can be found in Megyn Kelly's 2017 interview with Business Insider (Kelly and Business Insider, 2017) shown in Figure 5.1. In this, she says "It was good for me, my career was going well," while closing her hand in a pinch grip (Figure 5.1b) and moving her wrist upwards to

the right in a "pulling" motion (Figure 5.1c). She "pulls" an abstract *object* (Lhommet and Marsella, 2013) (her career) upwards as part of the indication that it is "going well." Thus she evokes this physical metaphor while literally describing a situation that was "good for her." And, this is an example of a metaphor represented in speech because it is *potentially implied* by the language, as "good" is potentially synonymous with "happy." While the concept of "Good" in conveyed literally in the language, the *metaphor* is only present because of the accompanying gesture.

This example also elegantly showcases human's ability to seamlessly blend multiple literal and metaphoric concepts across language and gesture. Not only does this gesture evoke the metaphor "Happy is Up," but both the language ("my career was *going* well") and gesture (upwards horizontal pulling) demonstrate the metaphor of "Process is Movement Along a Path," as discussed in Section 5.1.4.

(a) Beginning of the phrase "it was good for me." She positions her hand and prepares for a grasping motion.

(b) Motion while uttering "my career." She pinches her thumb and fingers together in a grasping motion and elevates her closed hand upwards.

(c) Motion while uttering "was going well." She releases her hand from a pinch position at the top of the path she has made with her wrist.

Figure 5.1: Megyn Kelly evoking the metaphor "Happy is Up" while uttering "it was good for me, my career was going well."

5.2.2 Time Is A Line

Grady proposes several metaphors in relation to the linearity of time. These include "Now is Here" ("Spring is almost *here*, Autumn is a *long way off*."), but more specifically "Moments In Time Are Objects In Motion Along A Path" ("Summer always *passes* too quickly."). The latter can physically manifest with virtually any reference to the past, present, or future.

The abstract metaphor of time existing on a line is pervasive throughout many languages and many cultures, though not all (Le Guen and Balam, 2012). However, the orientation and directionality of the past, present, and future varies from language to language (Kita, 2009) and even within languages from culture to culture. For instance, Calbris (2011) observes that US English speakers project time on a left-right axis relative to the speaker, with the past on the left and the future on the right. This effect has also been observed quantitatively in Spanish speakers (Santiago, Lupáñez, Pérez, and Funes, 2007). French speakers tend to project time on a forward-backwards axis, with the future in front of the speaker and the past behind themselves. This is directly opposite to native Polynesian cultures (Bender and Beller, 2014). Furthermore, native Mandarin speakers have been shown to position time either on vertical orientation with the future higher and the past lower, or horizontally perpendicular to the speaker, with the future in front and the past behind (Chui, 2022; Chui, 2018; Yu, 2012).

Hence, behavioral gesture researchers have long observed the ubiquity of this metaphor throughout speech and gesture, and have studied its affect on viewer interpretation. Jamalian and Tversky (2012b) and Kendon (1972) demonstrate that gesturing along these axes of time affects how viewers interpret ambiguous semantic co-speech content. Bender and Beller (2014) provide an extensive review of how many cultures and languages gesturally conceptualize time, and the many ways gesture can be used to manipulate this abstract conceptualization. This further exemplifies the nuance, complexity, and deep relationship between gesture and the mental metaphoric conceptualization of time.

5.2.3 Analyzing Multiple Metaphors: Quantity, Importance, and Abstract Objects

The topic of gesture in relation to quantity is of particular interest to many behavioral and computational gesture researchers. Grady's proposed metaphors with respect to quantity include both qualitative ("Quantity is Size" ["She assigned us a *huge* amount of work."]) and locational ("Quantity is Position" ["These two numbers are very *close*."]) elements. Moreover, size is linguistically and often abstractly related to importance ("Importance is Size" ["Tomorrow is a *big* day for this organization."]). Thus, the abstract concepts of both *Quantity* and *Importance* may be realized through similar physical gestures.

These linguistic observations have led gesture researchers to explore implementations in which the importance of abstract objects or events is emphasized by depicting an object that is large, heavy, or both. Lhommet and Marsella (2014a) conceptualize abstract "objects" as having "physical properties such as a size [and] weight" which can be depicted through physical properties of the motion of a

gesture. They elaborate on this idea (Lhommet and Marsella, 2016) by hand annotating a dataset to find that up to 25% of specialized referential gestures in their data relate specifically to the metaphors "Quantity is Size" and "Importance is Size." And, they also incorporate an implied third metaphor: "Abstract Concept is Concrete Object."

Ravenet, Pelachaud, Clavel, and Marsella (2018) explores this idea in conjunction with image schemas (Cienki, 2005), and suggests that gestural characteristics are tied to the Image Schema that underlies the production of metaphorical reasoning. Indeed, it has long been shown that, in addition to a gesture's location in space, metaphoric gestures and their characteristics serve to illustrate and demonstrate particular physical properties of the concept being communicated by the speaker (Cienki and Koenig, 1998; McNeill and Levy, 1980; Reddy, 1979).

5.2.4 Sets, Categories, Containers

A common motion used to convey a set of elements, ideas, or concepts is a *container* gesture, computationally explored by Lhommet and Marsella (2014b), Chiu, Morency, and Marsella (2015), and Ravenet, Pelachaud, Clavel, and Marsella (2018). Containers are linked to Grady's metaphors "Categories/Sets are Bounded Spatial Regions" ("Are tomatoes *in* the fruit or vegetable category? In any case they're *among* my favorites.") and "Constituents are Contents" ("I'm going to *take* several scenes *out* of the play."). These can be folded into the metaphor "Ideas are Objects," the implication being actions can be taken on these individual objects, such that constituent (abstract) objects can be added to or removed from (metaphorical) sets (Lhommet and Marsella, 2014b; Ravenet, Pelachaud, Clavel, and Marsella, 2018). Functionally, this is often associated with descriptors of *Inclusivity*, as seen in the example sentences.

People often physically manifest this metaphorical set membership through gesture. In Chapter 1, Figure 1.1 (copied here as Figure 5.2 for reference) a therapist outlines the boundaries of a space while uttering the phrase "anything at all." In this example the word "anything" is a representation of the concept *Inclusivity* – "anything" can be included in the set she defines. Once this abstract set is physically realized through metaphoric gesture, many motions may depict adding or removing abstract objects, increasing or decreasing the size of the set, or reinforcing one set's boundaries in relation to another. *Inclusivity* is the Functional Metaphor implemented by the semantic parser used in this implementation of this framework. In this case, *Inclusivity* is synonymous with the evocation of categories and sets. Importantly, this can be triggered from either the implication of a container (e.g. "we can talk about *anything*..."'), or an abstract deictic (see Section 1.1.1, e.g. by continuing the phrase "... *besides* what he wants.")

5.3 Testing Specific Metaphors

Through this framework we can begin to systematically explore the extent to which specific motions are used in conjunction with semantic metaphors present in a gesture's co-speech utterance. Specif-

(a) The beginning of the movement as she says "Anything at all."

(b) The second part of this gesture, creating the space where "anything" may metaphorically be.

Figure 5.2: The motion of the metaphoric gesture accompanying the phrase "Anything at all." This example originally appears in Lhommet and Marsella (2014a).

ically, given discussion of the metaphors described in Section 5.1, I formulate testable hypotheses around how speakers will move while evoking these metaphors.

5.3.1 Hypotheses

As discussed in Section 5.1.1, metaphoric gestures do not always coincide with literal metaphors in language. Instead, conceptual metaphors may be represented implicitly and realized physically. Furthermore, conceptual metaphors may by conveyed exclusively via gesture and have no explicit representation in speech. Since this analysis technique relies on filtering gestures by co-speech semantic concept, the hypotheses described below is only capable of considering gestures in which metaphors are represented (literally or implicitly) in language.

Happy Is Up

The specific potentially metaphoric concepts extracted from language by the semantic parser used in these hypotheses are "Positive Emotion" and "Negative Emotion." These correspond with those provided by Lhommet, Xu, and Marsella (2015) in Cerebella.

Based on the behavioral observations above, I examine two hypotheses about this metaphor:

H1) When the concept of *Happy* is represented in a gesture's co-speech utterance, the motion of the gesture will trend significantly upwards with at least one hand.

H2) Conversely, when the concept of *Sad* is represented in a gesture's co-speech utterance, the motion of the gesture will trend significantly downwards with at least one hand.

Time Is A Line

As described above, the orientation of time with relation to the speaker varies enormously from culture to culture. Therefore, any hypothesis which attempts to capture patterns of how this metaphor may be

performed via gesture must be particularly sensitive to the culture and conversational context of the speaker and the situation.

This metaphor can be used to determine both whether there are patterns that suggest speakers in this dataset use gesture to physically evoke the linearity of time, and also explore on which axis each individual speaker evokes it. Thus, I use this method both to quantify the extent to which this metaphor is physically evoked with the presence of co-speech metaphor in a gesture's utterance, and also to explicitly test the hypothesis proposed by Calbris, Montredon, and Zaü (1986) that English speakers gesture on the left-right axis more so than on the front-back axis in relation to time.

The potentially metaphoric concepts extracted from language by the semantic parser used in these hypotheses are "Time (Before)," "Time (Now)," and "Time (After)."

This leads to the following hypotheses:

H3) When the concept of *Before* is represented in a gesture's co-speech utterance, at least one hand will trend towards the left and/or towards from the speaker's torso.

H4) When the concept of *After* is represented in a gesture's co-speech utterance, at least one hand will trend towards the right and/or away from the speaker's torso.

H5) When the concept of *Now* is represented in a gesture's co-speech utterance there will be less motion than in cases in which *Before* or *After* is present.

Analyzing Multiple Metaphors: Quantity, Importance, and Abstract Object

The potentially metaphoric concepts extracted from language by the semantic parser used in these hypotheses are "Quantity (All)," "Importance is Size (Big)," and "Abstract Concept is Concrete Object." Importantly, the concept of "all" of something is not necessarily metaphoric, but can be physically realized. This exploration thus potentially combines both literal and implied concepts, which may have iconic representations in language. For example, the phrase "all of my ideas," literally describes a complete group (though of abstract elements), whereas the phrase "her mind is full of good ideas," implies it. In the first example, "all" is explicit in the language. In the second, it is implied by the language by the word "full," which does not literally describe a physical object. Thus this exploration mixes linguistic, physical, and metaphorical concepts which may be represented in language either explicitly or implicitly.

Previous work (e.g. Lhommet and Marsella, 2016) indicates that gestures depicting quantity or importance will scale accordingly with size. The implication in these instances is that the metaphor for *Abstract Object* is also present in these situations. This framework allows us to dig deeper into the question of how the presence of these metaphors influences gesture. Namely, we can examine how these three metaphors may separately inform motion, but furthermore how the *combination* of these metaphors correlate with motion. This informs our specific hypotheses around motion for these metaphors:

H6) When the concept of *Importance* or *Quantity* are represented in a gesture's co-speech utterance the distance between the wrists will increase.

H7) When the concepts of *Importance* **and** *Quantity* are represented in a gesture's co-speech utterance, this increased distance between the wrists will be more pronounced than when they represented individually.

H8) In each of these cases, if the concept of *Abstract Object* is also represented in the gesture's co-speech utterance, the increased distance between wrists should be more pronounced than when they are represented individually, or in combination without this concept.

It is important to note that when considering combinations of metaphors, due to the long-tail nature of language, the number of examples of various combinations quickly diminishes, even with relatively large datasets. The impact this has on this method's ability to address these hypotheses is acknowledged in Section 5.5.4, and this limitation is discussed further in Section 5.6.1.

5.3.2 Characterizing a Specific Metaphor: Sets

In addition to organizing and analyzing motion whose co-speech utterances may contain representations of multiple specific concepts, we can also simply characterize the motion gestures whose co-speech utterance contains a representation of one concept. This can be approached either by considering gestures whose utterances may contain representations of *only* that one concept, or which may contain *at least* that one concept. For example, we can compare gestures from whose utterances only the concept *Inclusivity* is represented, or include others which include a variety of any other concepts, such as *Quantity (All)*, *Contrast*, *Positive Emotion*, *Rejection*, etc.

Characterizing *Sets* is an exercise in demonstrating the illustrative power of this framework to establish associations and relationships between aggregate motion for this co-speech linguistic concept. For this reason I omit specific hypotheses and instead focus on an in depth exploration of motion that co-occurs with this concept.

Inclusivity is chosen as the filtering concept for this metaphor as it is most widely evoked by the parser in conjunction with a wide variety of evocations of the concepts *Sets* and *Categories*.

5.4 Quantitative Metrics

In this section I describe the implementation of three metrics I use to test the hypotheses presented in Section 5.3.1. All code for processing and analyzing gestures utterances and motion can be found at https://github.com/UofGSocialRobotics/CCFM-generation/.

5.4.1 Directional Wrist Movement

All of the metaphoric concepts presented in Section 5.3.1 include hypotheses concerning the position and/or direction of the wrists relative to the orientation of the speaker, and to each other. We expect to see differences in motion on the vertical axis for H1 and H2, and the horizontal axis in H3-8. This is relatively straightforward to visually interpret, but requires careful consideration in how to measure movement over the course of the gesture.

Measuring Along a Single Axis

Projection of the wrists onto a single axis (i.e. paying attention to only of the X, Y, or Z positions of the wrist) determines the extent to which wrists move to the left or right, forward or backwards, upwards or downwards, or remain stationary during the course of the gesture. This can be measured either by the start and end position of the wrist, or the overall distance traveled in each direction over the course of the gesture. Semantically, gestures that start and end in the same place may mean something different to viewers than gestures that begin in one place and end in another, even if the overall distance traveled in a given direction is the same.

Using raw measurements of the motion of a gesture fails to accurately compare gestures of different amplitudes that may both evoke the same metaphor. Instead, a gesture's trajectory along the axis of interest must be measured in relation to the length and amplitude of that specific gesture.

In order to preserve the true extent to which a gesture's motion travels along the axis of interest and to accurately compare gestures of different lengths and amplitudes, I take the average difference between each frame of the gesture along each axis individually. Functionally, this is the average distance the wrist travels in each frame. This accounts for variation in length between gestures, as well as gestures that include a post-stroke hold or that travel along multiple axes. It also penalizes gestures that travel back and forth along the axis of interest, which would be problematic for hypotheses which include such types of repetitive motion. I expand on this effect in Section 5.6.1.

As many metaphoric gestures can be performed by either hand or two-handedly, I measure the distance the wrist has traveled in the axis of interest using this implementation for both wrists independently.

5.4.2 Wrist Distance & Trajectory

A gesture that increases or decreases the distance between the hands is of particular interest for hypotheses presented in Section 5.3.1, as this may illustrate a large or small object, or depict an object that is growing or shrinking. This is specifically relevant to H6-8 which explore multiple metaphors which may either together or separately embody a "large" object.

Unlike directional wrist movement, this metric explicitly measures the distance of the wrists between one another. This is measured in two ways: (1) the overall average raw distance between wrists, and (2) the average change in distance between wrists across each frame of the gesture. This accounts

for gestures of different magnitudes, and also penalizes gestures in which the wrists move apart and then come together.

Another important metric to report in this paradigm is the Maximum minus Minimum wrist distance throughout the gesture. This indicates a "sweeping" motion, and adds depth to the characterization of the motion of wrists relative to one another.

5.4.3 Wrist Path Symmetry

The symmetry of the path the wrists make through space is one way to explore the degree to which one or two hands are involved in gestures. Examining this feature can help disambiguate whether people tend to use one or two hands when conveying certain concepts. Many concepts can be conveyed with either one or two hands. For example, a speaker could indicate a set (*Inclusivity*) via either a sweep of their hand, or by outlining a space (as seen in Figure 5.2). However, some concepts may show preference for single-handed gestures, such as in comparative clauses. Furthermore, measuring the extent to which the wrists move in synchrony may reveal performative preferences by speakers in this dataset, such as performing one-handed gestures with their dominant hand. Ideational Unit theory (Section 1.1.4) suggests that once a speaker gestures using one hand, they will not switch hands or move to a two-handed gesture in order to avoid confusion, since doing so may imply unintended conceptual meaning. There may also be differences in symmetry in the expression of concepts that canonically occur in a particular gesture space, such as to the left or right on the horizontal axis in H3-5.

The symmetry of the hands can be measured by performing DTW on the path of each wrist of the gesture, provided that one of the wrists is flipped on the left-right axis. This is a straightforward way to measure similarity in paths between the wrists, as even if the hands move out-of-sync but in the same pattern (e.g. in a "cyclical" motion) their paths through space are similar and will result in a low DTW distance.

5.5 Results

In this section I present the results of testing each hypothesis presented in Section 5.3. Metaphor-by-metaphor, I use the analysis technique provided in Chapter 4 (described in detail in Section 4.2.3) to explore several long-standing theories developed by observational gesture research. Importantly, this technique cannot disprove behavioral theories, as even if the stated hypotheses are not upheld in these results the absence of specific observations in one data set is not sufficient to disprove general behavioral phenomena.

5.5.1 Calculating Significance

The mapping from gesture to meaning is a many-to-many problem. As demonstrated in Chapter 3, many different utterances can reasonably accompany the same gesture, and the same utterance can be accompanied by different gestures. However, the conceptual meaning of the communicative performance changes. Similarly, it is unclear how concepts blend when communicating multiple metaphors in a single gesture (Chapter 2.1). Because of the nature of this many-to-many mapping and the difficulty in disentangling the role different metaphoric concepts may play in a single gesture performance, it is unreasonable to expect a precise mapping between gesture and semantic concept. Additionally, this linguistic context may or may not even be relevant to the gesture performance (see Section 5.6.1).

Instead, the goal in this analysis is to determine the extent to which different motions are representative of particular concepts in a relative context (Chapter 3). In other words, an "upwards" motion may represent many things besides the potential presence of the metaphor "Happy is Up," but it should at least disambiguate "Happy" and "Sad." Therefore, significance values are determined by comparing the value of interest (XYZ distance traveled, or path symmetry) in gestures whose co-speech potentially contains conceptual representation (literal or implied) of one metaphor to those with the opposite metaphor. The idea behind this is that while metaphoric gestures do not carry *precise* meaning on their own, they complement language in such a way that the whole gestural performance – including conceptual metaphors represented in speech – form patterns that can be established using this analysis. For comparison, a group of gestures for which none of the metaphors of interest are represented in the co-speech (but which do potentially contain representations of other metaphors) is included. This helps inform whether a motion that coincides with a metaphor is sufficient for disambiguation, or if it is further a potentially weak but reliable signal even when compared to a wider variety of gestures. These are the two baselines used when comparing the motion of gestures whose co-speech contains representations of a concept.

For Tables 5.1, 5.2, and 5.3 all significance values of $p < 0.05$ are shown in **bold**, while cells that could explicitly support or refute the stated hypotheses are highlighted in yellow. In this analysis, the X-axis is sagittal (positive is away from the torso), the Y-axis is horizontal (smaller is left), and the Z-axis is vertical (smaller is down).

Significance values are calculated by first using a one-way ANOVA to confirm significant differences in motions between gestures whose co-speech contains representations of the metaphors being compared, treating each output metric as a separate signal. All main effects between groups as calculated by ANOVAs for each metric are significant to at least 3 decimal places. This confirms main effects, which validates the use of the post hoc tests. I therefore use a Tukey HSD test to determine which groups of gestures differ significantly from one another. This test is used instead of Welch's t-test in order to account for the many pairwise comparisons between gesture groups. One test is performed for each dependent variables (i.e. one for Left X, one for Left Y, one for Left Z, one for Wrist Distance, etc.) comparing the groups of co-speech concepts. Key assumptions of the Tukey test

include normally distributed samples and equal variances across samples. Normalcy within groups and homoscedasticity between groups is established by Bartlett's tests and Pearson's tests.

Because of the long tail of language we expect group sizes vary drastically, and we cannot expect one metaphor to have potential linguistic representation in a high percentage of all utterances. To mitigate this effect, instead of comparing the metaphors of interest with all other gestures, I randomly sample a number of gestures from the overall dataset, excluding gestures from the groups of interest, equal to the largest group of interest. This step is performed multiples times, and results shown represent the average result over 10 iterations of random sampling. Group sizes are shown in Tables 5.1, 5.2, and 5.3.

The results for each hypothesis are two-fold. While I propose several hypotheses that are rooted in behavioral observation, this framework also affords an exploratory lens to examine aggregate motions of gestures in conjunction with co-speech linguistic concepts. Each section begins by strictly evaluating the hypotheses posed in 5.3, then considers possible effects and relationships observed in post hoc metrics. It is debatable whether to adjust for multiple comparisons across these post hoc metrics (Gelman, Hill, and Yajima, 2012; Rothman, 1990). The results shown below use Tukey tests to adjust for multiple comparisons between means for each individual metric, and further correction is **not** applied between metrics. This undeniably demands "subsequent study with preplanned hypotheses should be conducted to confirm the observed association" (Althouse, 2016) in post hoc explorations.

5.5.2 Happy is Up (H1 and H2)

Results for this hypothesis are shown in Table 5.1, column 4 (Z axis). On aggregate, the motions of the gestures that co-occur with the linguistic metaphor "Happy is Up" do not move significantly upwards on the vertical axis, although, we do see non-significant trends downwards for "Sad is Down." Interestingly, one significant trend is that for the concept *Happy* the wrists are further apart from one another, compared to gestures which do not co-occur with either metaphor, with the opposite being true for *Sad*. Although we see trends which support our initial hypotheses, H1 and H2 are ultimately rejected from these results as the wrists moved significantly along the horizontal but not the vertical axis.

Now turning towards exploratory analyses, this significance along the Y-axis (left-right) may be because the proxy concept used to collect this data was the detection of positive and negative emotions. One interpretation of this significant motion along the horizontal axis may be that different metaphors are expressed – namely, "Happiness is Inflation" ("It *filled* me with joy") and "Sadness is Deflation" ("I was *crushed* when I heard the news"). These "inflating" and "deflating" are reflected in the relative wrist distances in gestures that co-occur with each concept: The wrists are furthest apart in "Happy" gestures, significantly closer in gestures without either affect, and significantly closer again in "Sad" gestures. This is in line with similar observational findings in which positive affective signs and co-speech utterances correlate with outward motion of the wrists (Börstell and Lepic, 2020).

Similarly, negative affective states have been shown to correlate with body positions orienting

Metaphoric Concept	Hand	X	Y	Z	Wrist Distance	N
Happy is Up	left	0.055	0.104	-0.187	53.95	4498
	right	0.069	0.158	-0.018		
Sad is Down	left	0.227	-0.145	-0.290	51.45	752
	right	0.064	-0.503	-0.304		
Neither Metaphor	left	0.019	0.510	0.148	52.61	4498
	right	-0.075	0.644	0.061		

(a) Raw Values for X, Y, Z, and Symmetry metrics. The X-axis is sagittal (positive is away from the torso), the Y-axis is horizontal (smaller is left), and the Z-axis is vertical (smaller is down).

Compared Concepts	Hand	P_x	P_y	P_z	P Wrist Distance
Happy vs. Sad	left	0.594	0.106	0.827	**<0.001**
	right	0.998	**0.024**	0.556	
Happy vs. Neither	left	0.823	**0.043**	0.061	**0.031**
	right	0.358	0.180	0.201	
Sad vs. Neither	left	0.496	**0.030**	0.300	**0.002**
	right	0.708	**0.010**	0.090	

(b) Significant differences between values for X, Y, Z, and Symmetry metrics, as calculated by a Tukey HSD test.

Table 5.1: Raw values and significant differences for gestures whose co-speech contains representations of the concepts *Happy* and *Sad*, and for gestures which contain neither concept. The highlighted column indicates significance values relevant to H1 and H2.

away from the conversational partner (Noroozi, Corneanu, Kamińska, Sapiński, Escalera, and Anbarjafari, 2018). This could be responsible for the significant differences observed in wrist position between affective and non-affective gestures. Notably, many other aspects of motion are related to affective expression (Glowinski, Dael, Camurri, Volpe, Mortillaro, and Scherer, 2011; Nam, Lee, Park, and Suk, 2014; Lhommet and Marsella, 2014b; Camurri, Lagerlöf, and Volpe, 2003) which could be analyzed in this framework and potentially related to the expression of linguistic concepts.

It may further be the case that certain motions co-occur with different combinations of concepts. For example, there may be a stronger effect on the vertical axis for gestures whose co-speech contains representations of **both** *Happy* and *Process is a Path*. This dataset contains no utterances in which both of these concepts are represented. I elaborate on this phenomenon in Section 5.6.1 and on multi-metaphoric analysis in Section 5.5.4.

The existence of these different linguistic metaphors for affect expression and the results of this computational analysis reflect the extremely complex and multi-faceted mapping between language, motion, and communicative intent. It demonstrates the power of this technique to identify when and how people use certain metaphors. It furthermore powerfully underscores the essential need for further development and sophistication of semantic parsers, which I elaborate on in Section 5.6.2.

5.5.3 Time is a Line (H3, H4, H5)

From Table 5.2 we see strong support for H3 and H4. Specifically, with both *Before* and *After* metaphors there is strong movement to the left and right, respectively. Moreover, the path of the wrists are significantly less symmetrical when compared to non-time related gestures. Supporting the hypothesis posed by Calbris, we see this effect strongly on the horizontal axis, as she suggests is typical of culturally US English speakers. This analysis technique could however be applied to datasets of users from differing cultural or linguistic backgrounds. This would provide comparisons of the projection of the hands across axes between these groups, and could be used to demonstrate robust objective support for theories of gesturing in accordance to temporal relationships across cultures (Kita, 2009; Santiago, Lupáñez, Pérez, and Funes, 2007; Chui, 2018).

We see support for H5 in rows 3, 4, and 6 of Table 5.2b. Specifically, while both *Before* and *After* differ significantly in their wrist position relative to non-time gestures, *Now* gestures show very little difference relative to the rest of the data set. This is not to say that these sets of gestures do not differ from one another, simply that the metrics analyzed thus far are not sensitive to their differences. This highlights the importance of choosing and implementing appropriate metrics to categorize motion for the specific hypotheses in question.

Looking at the exploratory results, we further see interesting aggregate effects when we consider the handedness of each gesture. Gestures for both *Before* and *After* are significantly less symmetrical than gestures which do not contain representations of time in the language of their co-speech utterance. When we focus on this asymmetry, we see further noteworthy objective observations: the difference between the motion of the right wrist is more significant than that of the left wrist. This indicates that when this metaphor is physically evoked it tends to emphasize the use of one hand. Although the dominant hand of each speaker is not provided in the dataset, this may explain why the right hand shows stronger trends[2].

The difference in wrist position is also more pronounced for *Before* and *After* on the left and right wrist, respectively. This supports Calbris' idea presented in Section 1.1.4 and discussed further in Section 4.2.2 (utilized by Lhommet and Marsella, 2013) that ideational units of gesture optimize for the least amount of change in movement possible to convey the metaphor. Plainly, a speaker switching hands to convey an element of time could be confusing. If they were already using their right hand to convey a concept, reaching their right hand across their body to the left to indicate the past is more visually confusing than simply using their left hand. Since moving their left or right hand to the left or right respectively to indicate the past or future are functionally, communicatively the same, it is simpler to use the hand closest to the abstract space of the metaphorical concept – in this case, "Before is **Left**" or "After is **Right**."

[2]However, this hypothesis would need to be explicitly tested in situations in which speaker handedness is known. Furthermore, we would not expect this phenomenon to be specific to gestures that co-occur with temporal concepts.

Metaphoric Concept	Hand	X	Y	Z	Symmetry	N
Before is Left	left	0.368	-0.614	0.771	711.4	854
	right	0.142	-0.044	0.080		
After is Right	left	-0.138	1.524	0.231	698.0	3742
	right	0.028	2.038	-0.439		
Now is Here	left	0.058	1.691	0.011	686.4	400
	right	-0.531	1.880	-0.996		
None	left	0.025	0.376	0.097	685.6	3742
	right	-0.068	0.488	-0.052		

(a) Raw Values for X, Y, Z, and Symmetry metrics. The X-axis is sagittal (positive is away from the torso), the Y-axis is horizontal (smaller is left), and the Z-axis is vertical (smaller is down).

Compared Concepts	Hand	P_x	P_y	P_z	P Symmetry
Before vs. After	left	0.229	**0.006**	0.262	0.211
	right	0.78	**<0.001**	0.695	
Before vs. Now	left	0.970	**0.047**	0.757	0.082
	right	0.335	**0.005**	0.5484	
Before vs. None	left	0.742	0.776	0.353	**0.050**
	right	0.931	**0.021**	0.993	
After vs. None	left	0.867	**0.011**	0.876	**0.049**
	right	0.914	**0.049**	0.169	
After vs. Now	left	0.758	0.842	0.988	0.916
	right	0.383	0.861	0.473	
Now vs. None	left	0.999	0.295	0.999	0.999
	right	0.863	0.279	0.505	

(b) Significant differences between values for X, Y, Z, and Symmetry metrics, as calculated by a Tukey HSD test.

Table 5.2: Raw values and significant differences for gestures whose co-speech contains representations of the concepts *Before*, *After*, and *Now*, and for gestures which contain none of these concepts. The highlighted column indicates significance values relevant to H3, H4, and H5.

5.5.4 Mutli-Metaphoric Analysis of Quantity, Importance, and Abstract Object (H6, H7, H8)

Table 5.3 demonstrates one of many possible ways to analyze interactions between metaphors when considering multi-metaphoric gestures (I elaborate on other potential analysis techniques in Section 5.6.2). Here I perform comparisons between gestures whose co-speech contains representations of each metaphor (absent the others), each combination of metaphors, and none of these metaphors. I use a Tukey HSD test to explore the compounded effects of each metaphor, however it must be highlighted that the skewed sample sizes invalidate the statistical power of this test. Although I discuss the trends seen in the results in Table 5.3a, the p values shown in Table 5.3b rows 5-7 are unreliable. Yet, the ability to analyze such a small number of gestures, even if not statistically significant, illustrates this framework's utility by revealing patterns and identifying gestures that occur with specific co-occurring linguistic concepts.

Table 5.3a shows mixed support for H6. Although the wrist distance for *Quantity* gestures is not larger than gestures that whose co-speech language does not contain representations of this concept, the wrists do move out significantly throughout the gesture (row 2, column 5 "Wrist Distance Change"). However, wrist distance in gestures whose co-speech contains a representation of *Importance* is both smaller than non-Importance gestures, and wrist distance decreases significantly more over the course of the gesture. This is not necessarily surprising, as a concept that often coincides with *Importance* is "Precision," which may call for the hands to move together. This concept is not represented in the current semantic parser, so this post hoc hypothesis cannot be tested at this time.

These results show interesting effects related to H7, evaluating the effect of the presence of *Importance* and *Quantity* as co-occurring semantic concepts. From Table 5.3a we see that gestures whose co-speech utterance contains representation of both *Importance* and *Quantity* (hereby referred to as *Importance+Quantity*) have significantly larger wrist distance than gestures which contain neither concept, and the wrists tend to move outwards. Further investigation would be needed to determine whether this relationship is additive (*Quantity* and *Importance* have equal effect), combinatory (*Quantity* + *Importance* has different effects than simply additive), or even hierarchical (e.g. the presence of *Quantity* overrides the presence of *Importance*). The fact that *Importance* + *Quantity* leads to significantly more symmetrical gestures than either concept individually suggests combinatory effects on wrist path symmetry, whereas the significantly higher wrist distance in this combined case indicates a more hierarchical relationship in which *Quantity* overrides *Importance* in terms of wrist distance and direction.

As stated in Section 5.5.1, H8 is extremely difficult to evaluate due to the small sample size of co-occurring metaphors, particularly *Abstract Object*. Although these concepts do co-occur with relative frequency compared to other concept-triads (e.g. in phrases such as "a big idea," or "a huge goal," which interweave abstract, concrete, explicit, and implicit linguistic representations), the long-tail effect of language and communication means that despite this relative frequency we see very few examples of this combination in this dataset of over 120k utterances. This effect is discussed further in

Section 5.6.1. In practice, this skewed sample size is likely the reason we see fewer significant effects of combinations that involve *Abstract Object*. However, despite this small sample size, gestures whose co-speech utterance contains representations of all three concepts do have amplified effects in terms of larger wrist distance, less wrist path symmetry, and greater change in wrist distance throughout the gesture. This relationship again requires further investigation to disentangle patterns of motion which may correspond to these concepts in combination with one another.

5.5.5 Characterizing *Inclusivity*

Table 5.4 shows all metrics for gestures whose co-speech contains representations of the concept *Inclusivity*. There are several striking differences between all gestures which contain this concept and those which do not. The wrists in *Inclusivity*+ gestures are significantly further apart, and have a significantly larger Maximum-Minimum wrist distance, indicating a large amplitude. However, the average change in distance frame-by-frame is not significantly different from gestures without this concept. In other words, although these gestures have a large amplitude, they tend to return to their starting position, potentially indicating a "sweeping" motion. Positionally, this effect appears to occur with the right hand, as the right wrist travels significantly further outwards over the course of *Inclusivity*+ gestures than other gestures. This effect is diminished though still present in the left hand, again perhaps alluding to speaker preference to use the dominant hand to indicate certain concepts.

Contrast this to *Inclusivity_Only* gestures, in which there are no significant differences observed in these metrics. This pattern suggests, as discussed in Section 5.2.4, that *Inclusivity* is tightly coupled with many other semantic concepts that correlate both with physical actions (such as adding or removing items from a container), operations (such as growing or shrinking), and properties (such as size or weight). Without the co-occurrence of potentially physical concepts, this linguistic concept does not appear to form a characteristically distinct family of motions – echoing results in Section 5.5.4. *Inclusivity*+ is the typical occurrence of this concept.

This pattern of results is especially compelling when looked at in the context of the other concepts reviewed above. The motion for *Inclusivity*+ shares characteristics with *Importance+Quantity+Abs_Obj* and with *Happy*: larger overall wrist distance, larger outward movement with the wrists (indicated by the Y axis) and amplitude (indicated by Max-Min wrist distance), though not a larger cumulative wrist distance change compared to other gestures. This demonstrates that many concepts can be represented by similar movements, and that this framework can be used to compare aggregate motion characteristics across linguistic concepts. Further exploration regarding how the co-occurrence of these concepts influences their physical performances may reveal more informative patterns.

This exploration emphasizes that concepts combine both linguistically and physically to form distinctive patterns of motion. This framework's ability to characterize and distinguish such patterns within and across concepts is discussed further in Section 5.6.1.

Metaphoric Concept	WD	Symmetry	Max-Min	WDC	N
Importance is Size	*47.679*	*708.859*	*16.102*	*-3.377*	*86*
Quantity	54.611	709.488	14.041	0.536	6382
Abstract Object (Abs_Obj)	52.828	744.914	15.148	0.418	756
Importance+Quantity	56.454	679.865	19.284	1.341	246
Importance+Abs_Obj	*55.143*	*822.932*	*21.158*	*-6.589*	*4*
Quantity+Abs_Obj	*54.323*	*827.323*	*17.224*	*0.988*	*87*
Importance+Quantity+Abs_Obj	*57.495*	*888.384*	*20.284*	*2.899*	*11*
None	53.700	723.571	14.120	0.039	6305

(a) Raw Values for Wrist Distance (WD), Symmetry, Max-Min, and Wrist Distance Change (WDC) metrics for the Concepts *Importance, Quantity* and *Abstract Object*. Note that N indicates the number of gestures whose co-speech contains representations of **only** the metaphor(s) listed, e.g. there are 86 gestures whose co-speech contains representations for *Importance* and does not also contain representations for *Quantity* or *Abstract Object*, there are 246 gestures whose co-speech contains representations for *Importance* and *Quantity* but not *Abstract Object*, etc.

Compared Concepts	*p* WD	*p* Symmetry	*p* Max-Min	*p* WDC
Importance vs. None	***0.008***	*0.065*	*0.441*	***0.012***
Quantity vs. None	0.938	**0.049**	0.787	**0.024**
Object vs. None	0.699	0.094	0.380	0.425
Importance+Quantity vs. None	**0.049**	**0.032**	**0.032**	0.092
Importance+Abs_Obj vs. None	*0.176*	*0.153*	*0.318*	*0.279*
Quantity+Abs_Obj vs. None	*0.698*	***0.048***	*0.143*	*0.487*
Importance+Quantity+Abs_Obj vs. None	***0.044***	***0.012***	***0.038***	*0.479*

(b) Significant differences between values for Wrist Distance (WD), Symmetry, Max-Min, and Wrist Distance Change (WDC) as calculated by a Tukey HSD test.

Table 5.3: Raw values and significant differences for gestures whose co-speech contains representations of the concepts *Importance*, *Quantity*, and *Abstract Object*, and for gestures which contain combinations of or none of these concepts. Please note the skewed sample sizes when considering multiple metaphors in conjunction with one another. The highlighted column indicates significance values relevant to H6, H7, and H8. Please note the rows for which the sample size is too small to draw statistical inferences are indicated by *italics*.

Metaphoric Concept	Wrist Distance	Symmetry	Max-Min	Wrist Distance Change	N
Inclusivity_Only	53.366	693.465	13.837	0.766	113
Inclusivity +	54.390	715.061	15.006	0.334	2673
None	53.604	722.358	14.080	0.306	2673

(a) Raw Values for Wrist Distance and Symmetry metrics for the Concept *Inclusivity*, either with or without other concepts.

Compared Concepts	p Wrist Distance	p Symmetry	p Max-Min	p Distance Change
Inclusivity_Only vs. None	0.884	0.339	0.861	0.310
Inclusivity+ vs. None	**0.021**	0.242	**0.002**	0.936

(b) Significant differences between values for Wrist Distance and Symmetry, as calculated by a Tukey HSD test.

Metaphoric Concept	Hand	X	Y	Z
Inclusivity_Only	left	-0.325	1.101	-0.499
	right	-0.685	-0.116	-1.052
Inclusivity+	left	0.0735	0.217	0.237
	right	-0.058	1.043	0.135
None	left	0.262	0.615	0.0594
	right	-0.118	0.077	-0.001

(c) Raw Values for X, Y, Z, and metrics. The X-axis is sagittal (positive is away from the torso), the Y-axis is horizontal (smaller is left), and the Z-axis is vertical (smaller is down).

Compared Concepts	Hand	P_x	P_y	P_z
Inclusivity_Only vs. None	left	0.633	0.728	0.618
	right	0.586	0.303	0.349
Inclusivity+ vs. None	left	0.516	0.262	0.570
	right	0.828	**0.008**	0.680

(d) Significant differences between values for X, Y, Z, and Symmetry metrics, as calculated by a Tukey HSD test.

Table 5.4: Raw values and significant differences for gestures whose co-speech contains representation of only the concepts *Inclusivity*, those which contain this concept and other concepts, and for gestures which do not contain this concept.

5.6 Discussion

5.6.1 Limitations

Language vs. Communicative Intent

Fundamentally, this technique only analyses the relationship between motion and co-speech utterance. It does not consider either the viewer interpretation of the performance (combined gesture and utterance) or the actual communicative intent of the speaker. This limits the extent to which we can explore hypotheses around communication, as it compares only what people actually do to perform the gesture and does not attempt to measure how the conversational partner interprets it. Such a study would be difficult in this paradigm, as any effects would need to be demonstrated across different conversational and semantic contexts. For example, would a viewer, consciously or unconsciously, extract the metaphor "Good is Up" from a gesture if the linguistic context of a gesture is something very bad, or in that case would they perhaps interpret the physically vertical element of the gesture as a quantity or the magnitude of badness?

This issue is exacerbated by metaphoric gestures whose co-speech utterances do not contain any representation of an abstract concept. Recall the example provided in Section 5.1.2. This gesture, despite being illustrative of semantically meaningful metaphoric gesture, fails to be considered in this analysis framework. It is worth noting, however, that an extension of this framework that combines rhetorical and semantic parsing (for example, that presented in Section 4.1.1) may be able to capture this instance by detecting *Contrast*.

Conversely, a motion that co-occurs temporally with a concept represented by the language may or may not be relevant to that spoken language. Gestures vary hugely in their temporal correlation to co-speech relevance (Leonard and Cummins, 2011). Although we attempt to account for this by segmenting individual components of a gesture we expect to have high temporal correlation to language, this process is not validated. This problem presents a major challenge in large-scale automated gesture analysis.

Language Parser

The depth of analysis this framework is capable of is dependent on the quality of the semantic parser. Although the semantic dictionaries and parser are editable in order to fine-tune the extent to which ontological hierarchies are traversed when mapping utterances to concepts, the concepts that are triggered are ultimately the result of a third-party system. Furthermore, it lays out the space of concepts as combinatorial discrete units. One could instead use a Deep-ML based approach by first performing unsupervised clustering of co-speech utterances using a phrase vectorizer (e.g. Word2Vec (Church, 2017), BERT (Devlin, Chang, Lee, and Toutanova, 2018), or Universal Sentence Encoder (Cer, Yang, Kong, Hua, Limtiaco, John, Constant, Guajardo-Cespedes, Yuan, Tar, et al., 2018)) and use the resulting clusters as proxies for linguistic concepts. However, this disregards a principle intention of

this mapping: to allow interaction designers to understand *explicit* relationships between co-speech linguistic concepts and gestural motion.

Aggregate Analyses

It is crucial to keep in mind that all of these evaluations are severely limited by analyzing aggregate metrics. Chapter 4, Section 4.1.6 reiterates the utility of relying not just on aggregated metrics, but on sub-clustering gestures by their motion properties. Including a sub-clustering step in this analysis to further conceptualize how co-speech semantic concepts combine, and how those combinations manifest physically in the gesture is the next step in utilizing this analysis framework. By doing this we can then begin to understand how combinations of concepts form families of motion, and consequently can help us better understand the complex mapping between meaning and motion.

For example, consider the results shown in Section 5.5.2 concerning "Happy is Up." It may be the case that in some gestures whose co-speech contains representations of *Happy*, this metaphor is explicitly reflected in upwards motion, but that these gestures are outweighed by other physical and semantic influences when aggregated across all instances of *Happy*. This phenomenon could be found and further decomposed by first sub-clustering all *Happy* gestures by motion, then analyzing each sub-cluster using the metrics presented here.

A key element of this analysis is that the TAFs which correspond to motion represent metaphors which may be potentially present in speech. For example, although we see a signal in the analysis for the metaphor "Happy is Up," it is impossible to tell exactly how many of the gestures clustered in this analysis contain explicit representations of this metaphor. In this case, performing motion sub-clustering may help elucidate which of these gestures contain representations of this metaphor.

By organizing gestures into families of motion that can then be quantitatively and qualitatively analyzed, researchers can potentially then classify gestures into categories of explicit metaphoric representation, implied representation, or no representation. For example, co-utterances that are semantically parsed to potentially contain the metaphor "Happy is Up," but with corresponding gestures that become outliers when sub-clustered by motion may be disregarded as not representative of this metaphor. This may be especially revealing for metaphors which are less common, or less commonly elicit similar gestures, as in such a case a signal in motion may be obscured by gestures for which the co-speech utterance is not intended to convey the extracted metaphor. Future work using this analysis technique would benefit from first performing motion sub-clustering on all gestures which potentially contain a given metaphor, then employing motion analysis to determine trends not across all instances of the metaphor, but of specific families of motion that tend to coincide with that metaphor.

Long Tail Problem

Section 5.5.4 illustrates the difficulty of multi-metaphoric analysis due to the extreme variety and sparsity of language and conceptual space. Despite over 120 hours of video and motion capture,

our parser found only eleven instances of utterances that contain representations of all three concepts examined.

This is a general issue with data-driven gesture generation and analysis. Specifically, deep learning relies on huge amounts of data, and is fundamentally limited by the variety of examples it is given to train. With only eleven instances of a combination of concepts only very specialized ML models (such as One-Shot models, e.g. Li, Qin, Lu, Xu, and Hu (2020)) are capable of capturing meaningful patterns, and these are only feasible in niche applications, and will always by severely limited when drawing inferences between motion and meaning. Although this framework lacks the ability to make statistical claims about such small samples, it is able to identify trends and combinatory affects even with small sample sizes.

This problem thus presents a key strength of this analysis framework. While all ML approaches will have difficulty inferring meaning from so few examples, it highlights the utility of harnessing psychological behavioral background and theory in conjunction with data-driven analyses.

Data Set

Furthermore, this technique is limited by the data set on which the analysis is based. As noted in Section 5.1, the cultural background of speakers in the data set of interest will directly inform any results, and consequently limit any conclusions that can be drawn. Additionally, as described in Section 5.1.1, this analysis is incapable of considering metaphoric gestures in which a metaphor is not represented in language of the co-speech utterance. Consequently, this technique may exclude entire many types of gesture – even frequent or trivial gestures – that are simply not handled by the semantic analysis.

Metrics

The similarity in motion observed across concepts, particularly as discussed in Section 5.5.5, broadcasts this framework's reliance on which metrics are implemented to explore and characterize motion. The metrics described and implemented in Section 5.4 are tailored to the specific hypotheses posed in Section 5.3.1. They omit many properties that are relevant to speakers' ability to convey information, including motion oscillations, hand shape, finger angle, and finger spread, which are all key components to Calbris' notion of Ideational Units (Calbris, 2011), as well as properties relevant to conveying affect, such as motion kinematics (Pollick, Paterson, Bruderlin, and Sanford, 2001), and presenting individual speaker style (Pollick, 2003). Some metrics which are not implemented to characterize motion in this analysis could perhaps be distinguishing for the concepts explored.

This reliance is the reason for feature-agnostic motion sub-clustering implemented in Chapter 4, but fundamentally, in order to begin to describe motions that co-occur with linguistic concepts, those motions must be characterized in human-understandable terms. When forming testable hypotheses, adding more comparative metrics further increases the need to control for multiple comparisons. However a key strength of this tool is its ability to perform exploratory analyses as a prelude to analyzing testable hypotheses.

5.6.2 Implications and Future Directions

Framework Improvements

The current analyses rely only on patterns that are identifiable through the motion and path of the wrists. However, Calbris (1990) makes clear (and I emphasize in Chapter 4, Section 4.2.5) the importance of hand shape and changes therein to how gestures carry meaning. Not only finger angle and movement, but palm orientation, gesture amplitude, alternative wrist paths, and relational position to the speaker's torso and head may lead to useful insights into how motion relates to co-speech meaning.

Additionally, while the current analysis presents intriguing results, other statistical methods are available for this type of data. For example, Linear Mixed Effects Models may be able to reveal deeper relationships between co-speech concept combinations and our desired metrics. This would also provide both Main and Interaction effects for the given concepts. However, one issue with this approach is it quickly becomes convoluted with many independent variables and output metrics. Additionally, without a holistic mapping of independent and dependent variables, and with the extensive range of concepts which may potentially correlate to however many motion properties one chooses to implement, it is difficult to define what regressions such a model should perform. This type of analysis also suffers from effects of the long-tail problem described in Section 5.6.1.

This framework also assumes that changes in gesture form are closely temporally aligned with any associated introduction of concepts, whereas from behavioral research we know this is not always the case (Kendon, 1995). Motion and concept alignment could be improved in a variety of ways, including using a specialized gesture parser (Ferstl, Neff, and McDonnell, 2021a).

Implications

A significant motivation for this combined semantic and motion analysis is to lessen the need to hand-annotate data. Lhommet and Marsella (2016) hand-annotate gesture to determine their metaphorical meaning, but this technique can analyse the text in relation to motion to identify the presence of particular gestures. Additionally, because of the resources required to perform manual annotations, they focus their efforts only on particular metaphors and particular types of gesture. Automating this type of annotation allows gesture labeling at a much finer grain and much more exhaustively. For example, the technique of filtering for gestures whose co-speech contain representations of only certain concepts can be combined with the sub-clustering technique described in Chapter 4, Section 4.2.5 to automatically organize all gestures into overlapping groups based on shared semantic and physical properties. Combined with motion sub-clustering, this technique can be used to determine and characterize alternative ways of conveying metaphoric concepts both individually, and in combination with one another.

Similarly, instead of testing specific hypotheses, this technique can be used to exhaustively find patterns among metaphors. For example, calculations can be performed to determine which metaphor shows the highest levels of hand synchrony across all gestures, or which metaphor coincides with the

highest wrist distance. These features can be combined to determine, for example, which concepts result in the least wrist synchrony and the largest distance traveled along a specific axis. Thus, this technique is a tool not only for hypothesis testing, but also exhaustive exploration over an aggregated search space of gesture motion in relation to co-utterance semantic content. This can be extended with multi-metaphoric analysis by, for example, finding which metaphoric concepts co-occur most frequently and discerning patterns amongst only gestures which contain both concepts. This framework provides opportunities for many compelling exploratory analyses. Conversely, this framework offers the alternative to carve the gesture-concept space first by motion[3], and then map co-speech concepts onto motion patterns.

This technique can also act as a test against already existing gesture generators. It can compare the ways in which conversational partners actually gesture in conjunction with co-speech concepts in language with gestures that a generator would perform for those same concepts. While this may not be a test of clarity or motion smoothness, it provides objective metrics that directly compare the output of a generation algorithm with that of gestures "in the wild."

In addition to using this technique to test or generate specific hypotheses around gesture, this mapping can be used by agents as an aid to interpret gestures from human conversational partners. Using the metaphorical linearity of time as an example, agents could potentially use the axis or style a speaker uses to gesture about time to determine the cultural or cognitive context of the speaker. Thus, this method can help agents use gesture interpretation to perform rudimentary Theory of Mind.

Applications to Gesture Generation

While the focus of this work is on understanding and navigating the mapping between motion and linguistically communicated meaning, it is relevant to, and may be harnessed to produce, gestures on virtual agents.

Firstly, it builds a mapping which can be interrogated ad hoc by animation and application designers. Animators can thus fully appreciate the potential of their hand-crafted gestures to convey particular concepts, and furthermore automatically annotate potential concepts to which an animation should be mapped. This mapping can also be reversed by first clustering by motion, then performing a conceptual analysis to get a picture of how similar motions correspond to co-occurring concepts. By explicitly and holistically examining the potential motions that correspond with linguistic concepts, this framework can be used as a tool by designers to avoid false implicatures.

Secondly, the raw components of this system are not in and of themselves particularly novel. Parsing and analyzing gestures, then using extracted components as a database for generation is a well-explored technique (Ferstl, Neff, and McDonnell, 2021a; Chiu and Marsella, 2014; Yoon, Cha, Lee, Jang, Lee, Kim, and Lee, 2020; Habibie, Elgharib, Sarkar, Abdullah, Nyatsanga, Neff, and Theobalt, 2022). The unique aspect of this framework is in the precision of the motion parsing into

[3]although the computational complexity of comparing DTW between all gestures would require further exploration into appropriate clustering algorithms.

largely individual changes in forms of gesture, thus inferring the role of these changes in conveying meaning across phrase and clause structures. This means that the gestures analyzed in this framework cannot simply be played in conjunction with a concept.

While this framework could be extended to select a sequence of gestures to perform based on the semantic parse of an utterance to be conveyed, the process would involve building up and weaving together the many concepts represented within the utterance, and disambiguating which gesture forms are suitable to convey those concepts. A generative system based on this framework would furthermore have to decide which concept(s) in an utterance to prioritize, how to effectively convey those concepts given the changes in form indicated by the mapping generated in this framework, and then, finally, how to co-articulate between those changes in form. Each of these steps represent considerable challenges in and of themselves, including at the level of the animation framework, and were not a focus in this work.

5.7 Conclusion

In this chapter I used the frameworks described in Chapter 4 to perform data-driven analysis to test theory-driven hypotheses. The creation of a mapping between gesture and semantic content in language affords new avenues through which to examine the long history of behavioral observation of motion and semantic communication.

Chapter 6

General Discussion

6.1 Summary of main findings and contributions

This thesis presented an interdisciplinary exploration that combines human subject studies and data-driven techniques to gesture classification, analysis, and generation. Throughout this work a central focus is on gestures that convey meaning through metaphor. This thesis explored how metaphoric gestures are interpreted by an observer, how one can generate such rich gestures using a mapping between utterance meaning and gesture, as well as how one can use data driven techniques to explore the mapping between utterance and metaphoric gestures.

Chapter 1 of this thesis outlined the importance of gesture in social interaction and outlined a brief history of the field of gesture in behavioral psychology. It discusses how this both has and has not informed current computational models of gesture generation on virtual agents. Chapters 2-5 then present original research in this interdisciplinary field.

Chapters 2 and 3 focus on the impact of gesture on viewer interpretation. Chapter 2 uses case studies in two experiments to paint a picture of the complexity of viewer interpretation in multi-metaphoric gestures. The results of the first experiment show that small changes individual components of the gesture – e.g. the hand shape or the velocity – can lead to large changes in interpretation in the semantic meaning of the gesture. The second experiment furthermore shows that such changes in the form of the gesture are not culturally universal. Interestingly, it also suggested that viewers in some cultures may more readily interpret multiple semantic meanings from some gestures than others. I explore how these insights could be more rigorously tested using the data-driven analysis technique in Chapter 5.3.1.

Chapter 3 elaborates on this finding by comparing methods of measuring the subjective communicative message of gestures. It shows that viewers tend to interpret meaning from any sufficiently energetic gesture, underscoring the importance of agents using gestures with intention and not as simple random motion. This reinforces the importance of measuring specific concepts communicated through gesture.

Chapters 4 and 5 shift focus from viewer interpretation of gestures. They instead use understand-

ings of how gesture impacts communication to move towards data-driven analysis techniques that can be used to test and influence generative mechanisms. Chapter 4 presents an architecture and framework to evaluate data sets of motion capture and audio to objectively explore the relationship between gesture and co-speech utterance. This framework captures the rhetorical and semantic structure of speech, and can be extended to include affective and audio cues. It uses separate unsupervised clusterings of text and motion to create mappings between gesture and language which can then be used to inform models of gesture in virtual agents. Finally, Chapter 5 exploits this property to tie together theoretical gesture research and data-driven analysis to test specific hypotheses generated by observational research around how gestures relate to the semantic content of language.

This thesis contributes several objective analysis techniques to the interdisciplinary field of gesture generation, and discusses the deep complexity on both ends of the gesture generation process in both speakers and viewers. It demonstrates how these techniques are used to glean insights into the relationship between gestural motion, communicative intent, and viewer interpretation. Fundamentally, in order to achieve a pipeline to generate complex, convincing, communicatively rich gestures on virtual agents, researchers must systematically explore how humans use gesture as a conversational tool.

6.2 Limitations and Future Directions

6.2.1 Deep Cognitive Gesture Modeling

The techniques presented in this thesis form a mapping between the motion and co-speech utterance of gestures. However, for virtual agents to truly gesture and communicate non-verbally with the same richness, nuance, and complexity as humans, it is likely that they need to use the same mental representations. That is, human communication is not a thought-to-language model, in which gesture is purely additive. Per Growth Point Theory (Section 1.1.3) language and gesture are generated from the same deep cognitive starting point, and therefore are complementary rather than redundant in their communicative capacity. While the analyses in this thesis form objective behavioral models and can be used to *test* cognitively-informed theories of behavior, they cannot be used to comment on or model deeper underlying cognitive processes used in social communication.

6.2.2 Generation

The architecture presented in Section 4.1 can be used map novel utterances to a database of gestures, and therefore can be extended towards a mechanism to generate gesture in virtual agents. However, I did not implement and test this mechanism due to the technical animation challenges. This thesis plays a role in determining what behaviors to exhibit. However, realizing the animation of behaviors was not its focus.

The framework presented in Chapter 4 and used in Chapter 5 is in fact not intended to be immediately applicable to a generative algorithm. Unlike other gesture database matching architectures

(e.g. Ferstl, Neff, and McDonnell, 2021a), the units this framework analyzes are far from complete gestures, and blending between gestures as they are parsed in Section 4.2.4 is non-trivial. The purpose of clustering gestures together – by either co-speech concepts or motion – is to understand how individual changes in form correspond to linguistic concepts. Ultimately, the goal of this framework is not to generate gestures, but to analyze them.

Subjective Interpretation & Viewer Impact

Another challenge in using this pipeline as a generative mechanism, as explored in Chapter 3, is how to then evaluate the viewer impact of the gesture performance. Ultimately, the goal of these analysis techniques is to measure the impact and communicative effects of gestures. In contrast, the techniques that are presented and evaluated in Chapters 4 and 5 present novel ways to map the relationship between gesture and co-speech utterance, which operate on a surface level of communicative intent rather than impact. In order to truly test whether a gesture has its intended impact on the viewer, as discussed in Section 6.2.1, one would need to have access to the communicative intention of the speaker. In the case of generating gestures on virtual agents, that means that an algorithm would need to explicitly model the designer's intent for the agent. This could potentially include its cognitive state, communicative goals, and contextual understanding of the conversation.

The framework presented in Chapter 4 could be extended with minimal changes if the communicative impact were readily accessible as annotated data. Either the semantic conceptual parser could be swapped out to group gestures not by co-speech concept, but by viewer impact, or viewer impact could be added to the mapping, further enriching the interplay between motion, language, and communicative impact.

6.2.3 Context

The mappings between co-speech metaphor and motion generated in Chapters 4 and 5, as well as the subjective evaluations in Chapters 2 and 3, are all deeply context-dependent. As Chapter 3 explicitly demonstrates, a viewer's qualitative interpretation of gestures change depending on the content of the gesture's co-speech utterance. This is only one element of any number of relevant contextual cues that may influence how a viewer interprets a gesture. Some are imminently testable, such as the race and gender presentation of the speaker's voice and appearance, the conversational context of an utterance, or the physical (or virtual) location of the conversation. However, other contextual cues are more difficult to manipulate, control for, and measure. These include the viewer's own mental model of how an agent understands a given topic, how that agent uses gesture to communicate, and simply the cultural background of the viewer. Manipulating a viewer's theory of mind of an agent is non-trivial and indeed an entire field of research in and of itself (See Perez-Osorio, Wiese, and Wykowska, 2021).

6.2.4 Semantic Parser

It is important to emphasize that the semantic analysis described in Section 4.2.4 is heavily informed by that implemented in Cerebella (Lhommet and Marsella, 2014b), although this implementation significantly extends Cerebella's ability to analyze phrase and clause structure, as well as the semantic analysis using the work of Grady (1997) and the TRIPS ontology (Allen, Dzikovska, Manshadi, and Swift, 2007). The analysis used in Chapter 5 benefits from being human-readable and informed by psychological theory, however it is potentially constrained. One could, in theory, first perform an open, unsupervised clustering on all co-speech utterances in the dataset (e.g. using a phrase vectorizer such as Devlin, Chang, Lee, and Toutanova, 2018), then manually label each cluster with the concept expressed therein. This benefits from not prescribing concepts that can be expressed. However, it disregards the benefit of foundational research in behavioral gesture research that establishes concepts that are often relevant to gestural motion.

6.3 The Future of Gesture Research

6.3.1 Big data and gesture

It is impossible to talk about the future of gesture research without addressing the research field of Big Data. Pervasive data collection and novel data acquisition methods (e.g. VR rooms and motion sensing and extraction technologies) grant researchers previously unthinkable amounts and quality of data off of which to base theories and generate agent behavior. This skyrocketing rate of data acquisition has led to an inflection point in gesture research, and brought this interdisciplinary field from psychologically-driven to a spotlight on more data-driven machine learning generation techniques.

Using neural networks to create generative models of gesture for individual speakers is a present reality. Consider for example entries into the 2022 GENEA gesture challenge by Yoon, Wolfert, Kucherenko, Viegas, Nikolov, Tsakov, and Henter (2022) in which ten out of eleven entries leveraged deep learning techniques. Several models herein produce gestures that are nearly-indistinguishable from the original speaker in many cases, but which are also driven exclusively by audio inputs. Notably, the only entry which does not use deep neural networks to generate motion scores the highest on human-likeness of motion, but utilizes text only for timing purposes, not for semantic meaning (Zhou, Bian, and Chen, 2022). The gestures which result from these entries were furthermore judged by "appropriateness," with a particular piece of audio, not communicated message. Similar to the findings presented in Chapter 3, the authors note that "the [appropriateness] evaluation was not altogether successful, since the mismatched condition M – which paired natural motion segments with unrelated speech segments, intended as a bottom line – attained the second-highest appropriateness rating, above all synthetic systems. This suggests a significant dependence between the human-likeness of a motion segment and its perceived appropriateness for speech, confounding the evaluation."

Fundamentally, a black-box, deep-learning approach to gesture generation simplifies its inherently

cognitively-driven and complex nature. Deep-learning models generate gesture from audio, not communicative intent. This attempts to drive gesture behavior from smaller spaces (e.g. prosody) because the entire space of gesture meaning does not have a neat mapping. The models in this competition, for example, do not handle the complexity of semantics, rhetoric, or affect (aside from how those elements are expressed in voice qualities). As the authors from another prominant deep-learning based gesture generation algorithm state, "audio does not directly encode high-level language semantics that may allow us to predict certain types of gesture (e.g. metaphorics)" (Ginosar, Bar, Kohavi, Chan, Owens, and Malik, 2019). It could be argued that middle layers of these networks implicitly derive other salient features (Kucherenko, Nagy, Jonell, Neff, Kjellström, and Henter, 2021). Additionally, some hybrid approaches that incorporate acoustic variables do take advantage of psychological insights in terms of explicitly deriving salient gesture features (Ferstl, Neff, and McDonnell, 2021a), and some recent deep learning generation systems do attempt to retain some level of user control over generated motion (Habibie, Elgharib, Sarkar, Abdullah, Nyatsanga, Neff, and Theobalt, 2022).

The rise in gesture generation techniques that are exclusively audio-based is problematic, as gestures have the ability to change the interpretation of the same audio (Jamalian and Tversky, 2012a; Lhommet and Marsella, 2013). While some deep learning generation algorithms have attempted to incorporate semantic information to improve gesture "appropriateness" (Kucherenko, Jonell, Waveren, Henter, Alexandersson, Leite, and Kjellström, 2020; Liang, Feng, Zhu, Hu, Pan, and Yang, 2022), the influence of semantics on generation remains opaque and evaluation of semantically-informed gestures remains shallow (see Chapter 3). Without a principled way to deal with semantics, machine learning techniques currently remove meaning and communicative intention out of the equation when it comes to gesture generation. So the challenge remains to enrich deep learning approaches that have the potential to generate not only extremely natural beat gestures, with more complex, nuanced, and subtle gestures as well.

6.3.2 Using gesture to make inferences about cognition

Using deep learning to generate gestures misses the deeper complexity of gesture research: the cognitive relationship between thought and behavior. While neural networks given sufficient data may produce extremely high quality behavior, it sheds less light on the way humans actually store, process, generate, and then transmit thoughts. For artificial social agents to be truly human in their expression, an alternative view is to assume they must abide by the same cognitive processes and limitations as we do – although it is left to context whether the goal of an agent is to be human-like, or communicatively efficient, or agreeable to talk to, etc.

This possibility is eloquently expressed by the theory of Embodied Cognition, outlined and explored extensively by Wilson (2002). The theory of Embodied Cognition states that many features of cognition are shaped by the human experience of a physical body. This includes both high level mental constructs (such as concepts and categories, (Lakoff and Johnson, 1980)) as well as performance on various cognitive tasks (such as reasoning or judgment). According to this hypothesis, the organiza-

tion of human thought is limited by the constraints of our body not only neurologically, but by our mental incapacity to imagine what it would be like to exist without our body. This drives our physical metaphors, both gestural and in language, and indeed may be reflected in a hierarchy of metaphors in our own thoughts. With this in mind, it may be impossible to create a perfectly human-like gestural model for social artificial agents unless their thoughts are organized like ours.

In this view, part of the goal modeling gestures is to make inferences about our own cognition that may be applied to social artificial agents. By demonstrating correlations between expressed thoughts and physical motions, we may uncover elements of this mental hierarchy to learn about the structure and organization of our own thoughts. These insights can propel both the field of cognitive science as well as human-computer social interaction. Understanding gesture advances our understanding of linguistic, social, and embodied cognition.

6.3.3 The Critical Role of Interdisciplinary Collaboration

As is evident throughout this thesis, generating convincing, natural, social non-verbal behavior on virtual agents is an extremely complex task. It requires weaving together insights and understandings from a broad range of disciplines, from psychology to software development to philosophy to interaction design. Historically, traditional research establishments maintain knowledge silos across departments. As a result, researchers in gesture generation, and non-verbal behavior more broadly, tend to explore methods and hypotheses with relatively narrow perspectives which hinders potential growth of the field.

For example, black-box machine learning has contributed hugely to the field of artificial motion generation (Henter, Alexanderson, and Beskow, 2020). However, these systems do not incorporate or inform any real insights from behavioral psychology. Psychologists and machine learning scientists can work in tandem to use heuristics gained from behavioral observation (including computational behavioral observation as implemented in Chapter 5) to boost both training speed and performance outcomes, and potentially use far less data than is traditionally required in fully automated approaches (Ferstl, Neff, and McDonnell, 2021a) while maintaining some level of designer control (Habibie, Elgharib, Sarkar, Abdullah, Nyatsanga, Neff, and Theobalt, 2022).

Context-specific agents require even more careful thought and collaboration. Therapists, doctors, teachers, etc must be consulted when deploying agents to perform certain social interactions in highly skilled domains with the necessary sensitivity. These specialized occupations already incorporate social skills training, and performance in such specialized situations benefit from conscious effort put into non-verbal social behavior (Riggio and Throckmorton, 1988; Riemer and Jansen, 2003; Pawlikowska, Zhang, Griffiths, Van Dalen, and Vleuten, 2012). Even beyond the realm of research and study, practical behavior experts can and should be consulted when considering how to program agents to act authentically in certain social situations.

These challenges become even more important and intricate when expanding beyond virtual agents into social robotics. Robots come with their own technical challenges. Beyond animation, robots must

worry about mapping affective and communicative behaviors onto limited degrees-of-freedom (Perre, Van Damme, Lefeber, and Vanderborght, 2015), physical constraints in speed and acceleration of motion, their own physical surroundings including their path through space (Sariff and Buniyamin, 2006), and the safety and preferences of their conversational partner (Mumm and Mutlu, 2011). As the digital world becomes increasingly pervasive – and physically realized – we see more and more social robots expected to interact with humans "in the wild," but often with suboptimal outcomes (Kwon, Jung, and Knepper, 2016; Sabanovic, Michalowski, and Simmons, 2006; Jung and Hinds, 2018). Researchers across mechanical and electronic robotics, human-computer interaction, and social affective behavior must come together to facilitate seamless, socially fulfilling experiences in physically embodied agents.

Given these vast challenges and applications, the complexity of gesture generation and the processes of non-verbal behavior interpretation raise real concerns with severe ethical consequences. This thesis focuses on the technical and analysis challenges of gesture generation, but forefront in the minds of technologists must always be the question of whether or not a virtual agent is appropriate for any given application. One must consider not only the behavior of an agent, but meta-issues around Human-Agent interactions such as replacing the potential for human-human connection, removing a job from an actual human, data privacy from virtual interactions, and mediating interactions between humans and VAs.

Behaviorally, it is deeply problematic if an agent's communication is poor and its actions affect the behavior of a real person. This is one reason why we cannot simply deploy a black-box machine learning approach that lacks explainability and transparency, even with something as seemingly inconsequential as conversational gestures. More broadly, it is hardly settled whether we, as a society, should seek social comfort and familiarity with artificial agents at all (Slater, Gonzalez-Liencres, Haggard, Vinkers, Gregory-Clarke, Jelley, Watson, Breen, Schwarz, Steptoe, et al., 2020; Pan and Hamilton, 2018; Luxton and Hudlicka, 2021). Frankly, due to ethical considerations (e.g. privacy and the potential for data harvesting by private corporations, which could then be used for coercive social manipulations) it is ambiguous whether socially reactive agents as digital interfaces should even be a goal for society (Krämer and Manzeschke, 2021). It is profoundly important that VA researchers are literate in the ethical consequences of the technology to which we dedicate ourselves.

Instead of working across or between disciplines, gesture generation – and social non-verbal behavior more broadly – is an inherently *multi*-disciplinary field. Virtual and augmented reality, digital tele-presence, and embodied avatars are already becoming increasingly ubiquitous. Immersive, convincing social agents are only possible with a multidisciplinary understanding of social interactions.

6.4 Concluding Remarks

In conclusion, this thesis contributes to the foundational understanding of the complex mapping between gesture, language, and communication. It highlights the complexity of measuring viewer inter-

pretation and impact of gesture, and uses modern data sets and acquisition tools to marry psychological behavioral gesture research and computational analysis techniques. This creates a novel objective mapping between motion and language that can be used to investigate insights into how spontaneous gesture is used in conjunction with speech. This is a key step in the process of engineering fulfilling, social, natural, communicative, nuanced experiences between humans and virtual agents.

Chapter 7

Supplementary Material

7.1 Additional Experimental Material for Section 2.1

7.1.1 Videos and Analysis For Section 2.1

Full videos of example gestures, complete stimuli set, raw data, and analysis code can be found at the following address:
`https://osf.io/txv7g/?view_only=bccaefcd70f44e5e8d06f27accb1893a`
All results are also shown below.

7.1.2 Experimental Procedure Statements for Study 1, Experiment 1

- People in this group are working together to solve a problem.
- She is referring to everybody in the group.
- There is tension between people in the group.
- People in this group disagree with one another.
- This group consists of many people.
- People in this group generally get along.
- The speaker likes the people in this group.
- The speaker is annoyed with the group.
- The speaker is in control of the group.
- The speaker is open to feedback from the group

7.1.3 Experimental Procedure Statements for Study 1, Experiment 2

On a scale from 1 to 7 how much do you believe...

- This group is made up of many people.
- There are many members of this group.
- This group of people is experiencing conflict.
- There is tension in this group of people.
- This group of people is open to outsiders.
- Non-members find this group accessible.
- This group of people is tightly controlled.
- Someone is definitively dominant over this group of people.
- This group of people is working together.
- There are common unifying goals within this group of people.
- This group is very sure in their decisions.
- The actions of this group are taken confidently.

7.2 Additional Experimental Material for Section 2.2.1

Full videos of all gesture stimuli, along with all results, raw data, analysis code, and graphs can be found at the following address:
`https://osf.io/y76na/?view_only=2bba82a76cb74e29a6df1dd705172dc3`

Full data and participant information is available at the above link. In addition, the number of participants who viewed each video are shown below.

Gesture Condition	Western N	Eastern N
WG1-0	48	48
WG1-1	40	44
WG1-2	56	48
WG2-0	48	45
WG2-1	42	44
WG2-2	55	49
WG3-0	47	47
WG3-1	48	43
WG3-2	56	40
WG4-0	45	40
WG4-1	43	41
WG4-2	47	47
WG5-0	52	42
WG5-1	49	40
WG5-2	44	42
EG1-0	43	39
EG1-1	44	34
EG1-2	54	37
EG2-0	46	34
EG2-1	47	36
EG2-2	51	31
EG3-0	41	38
EG3-1	49	31
EG3-2	42	30
EG4-0	46	39
EG4-1	40	34
EG4-2	48	30
EG5-0	42	31
EG5-1	40	34
EG5-2	38	36

All bucketed results are also shown below.

7.3 Additional analyses for Chapter 3

All raw data, data processing code, analysis, and implementation code can be found at `https://osf.io/c9mqw/?view_only=01df07d39aee455fa1b26745572051e7`

All violin graphs for response distributions across semantic domains are also shown below.

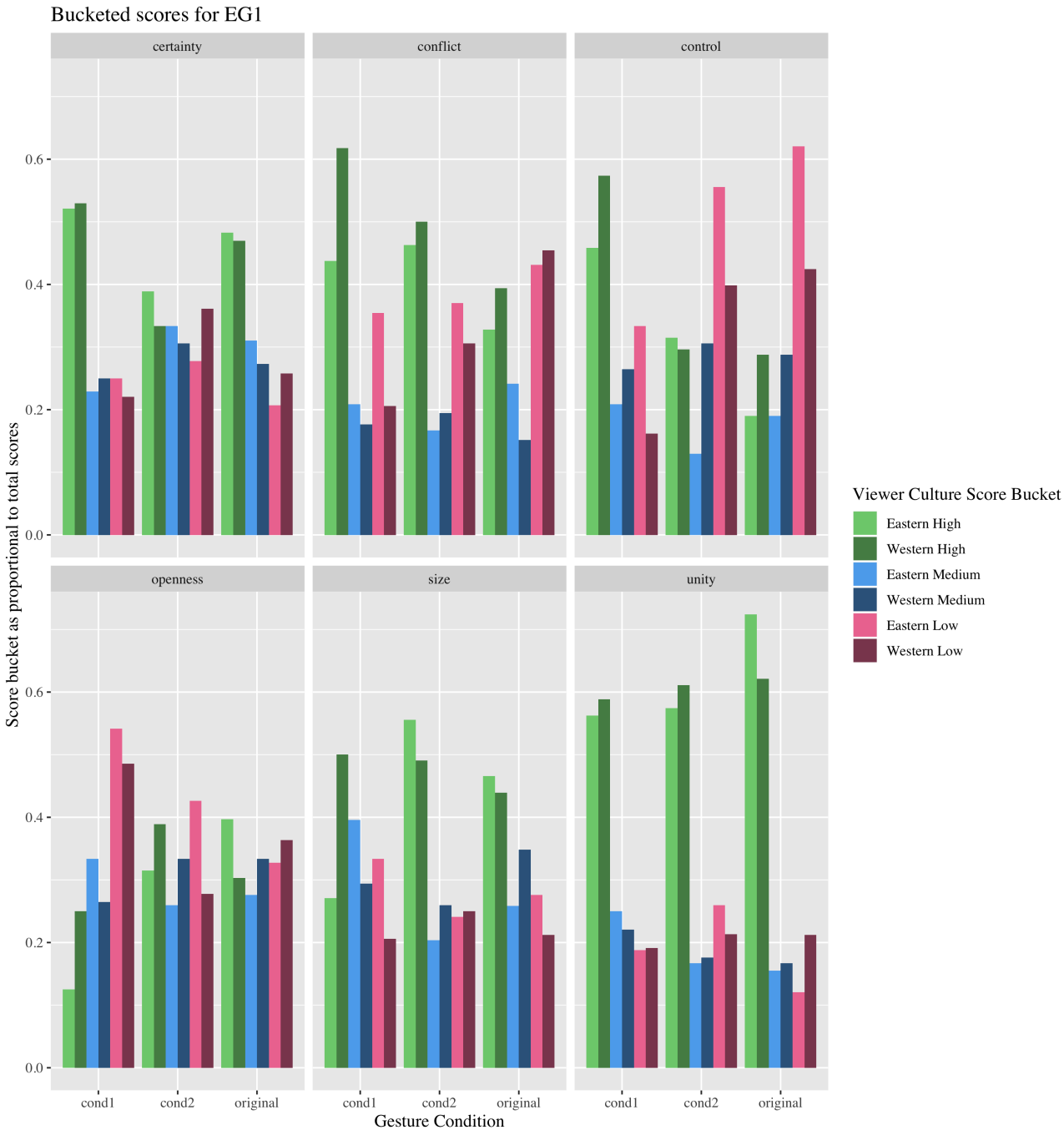


Figure 7.1: Bucketed results for Eastern Gesture 1.

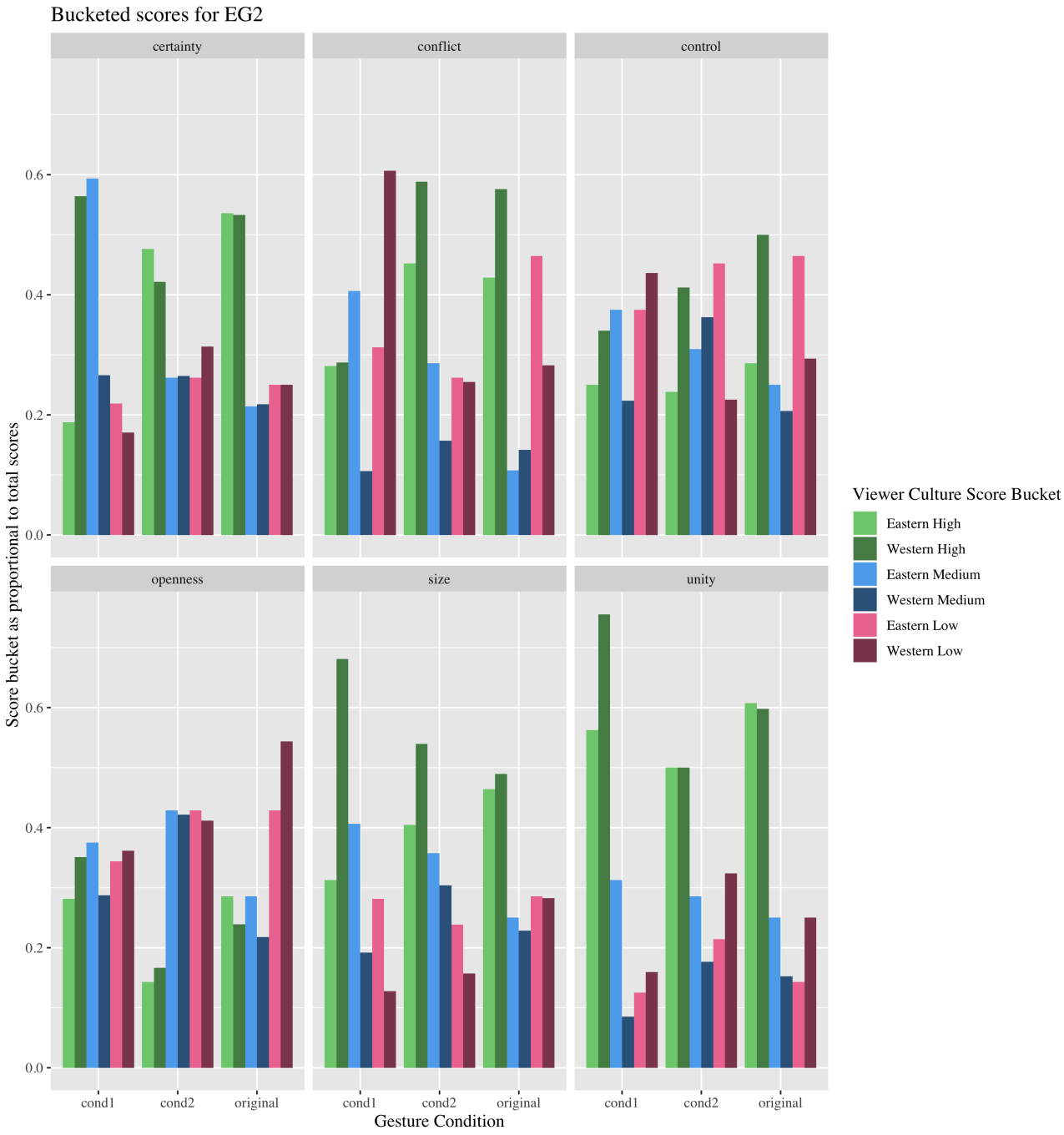


Figure 7.2: Bucketed results for Eastern Gesture 2.

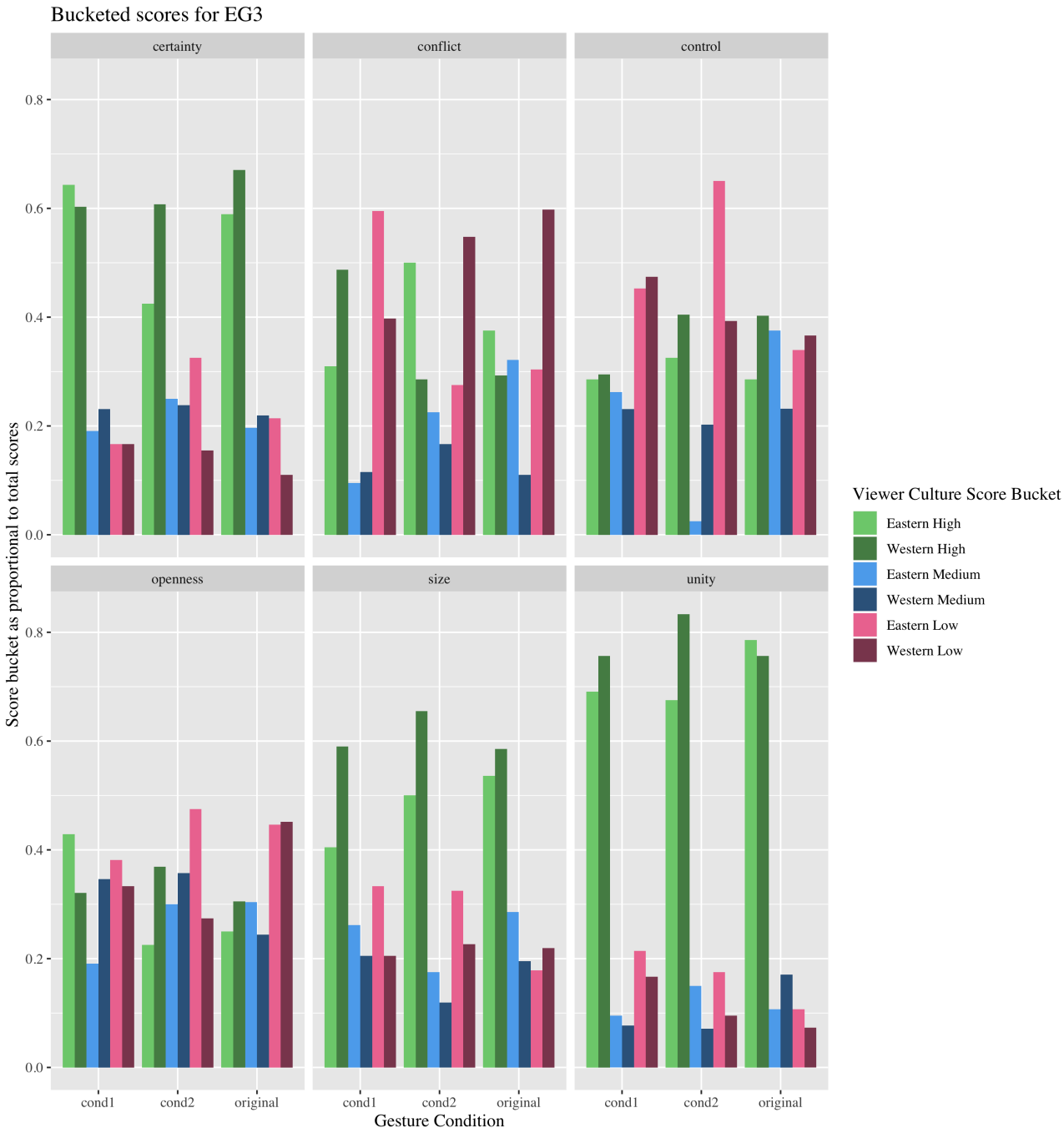


Figure 7.3: Bucketed results for Eastern Gesture 3.

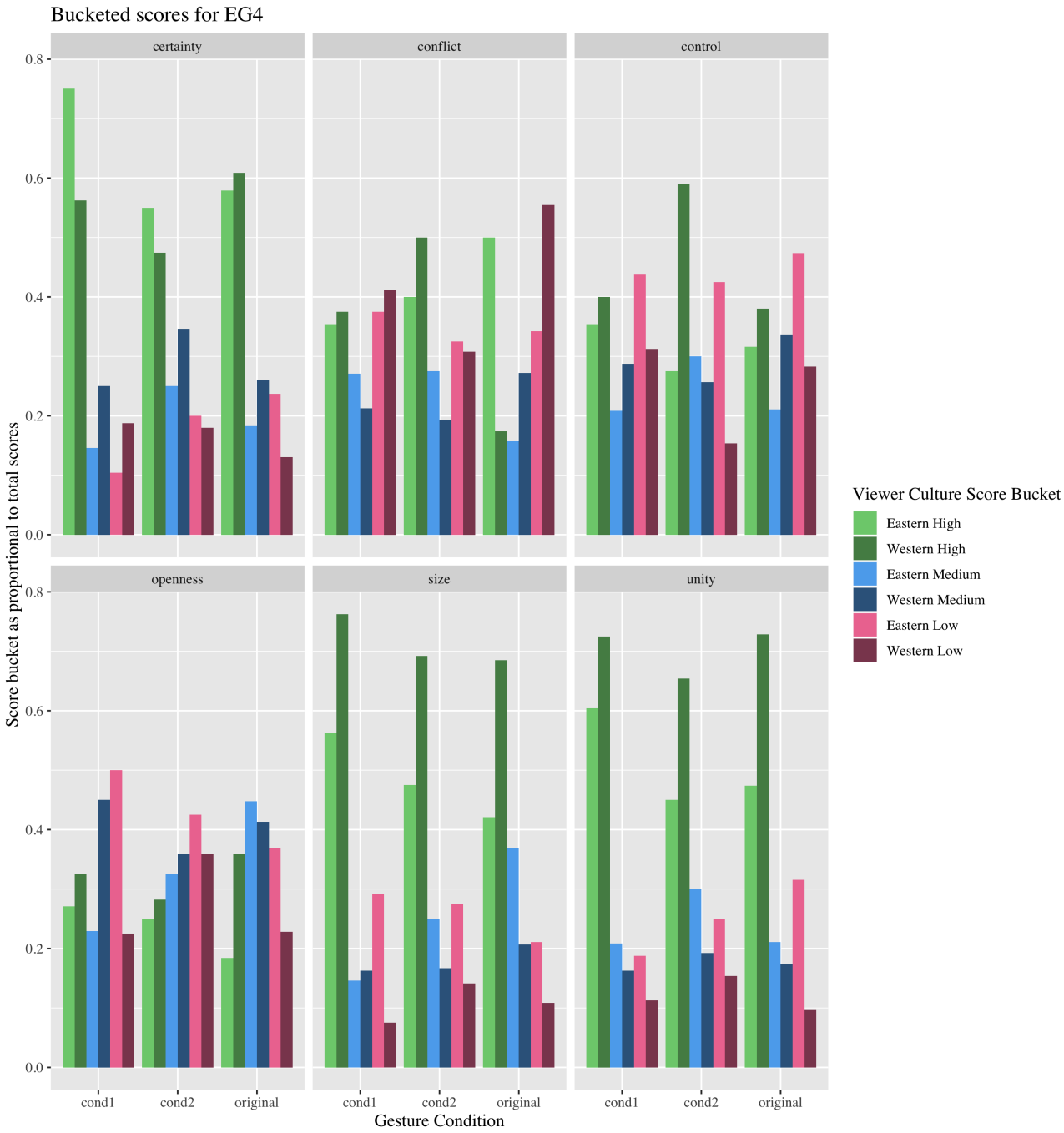


Figure 7.4: Bucketed results for Eastern Gesture 4.

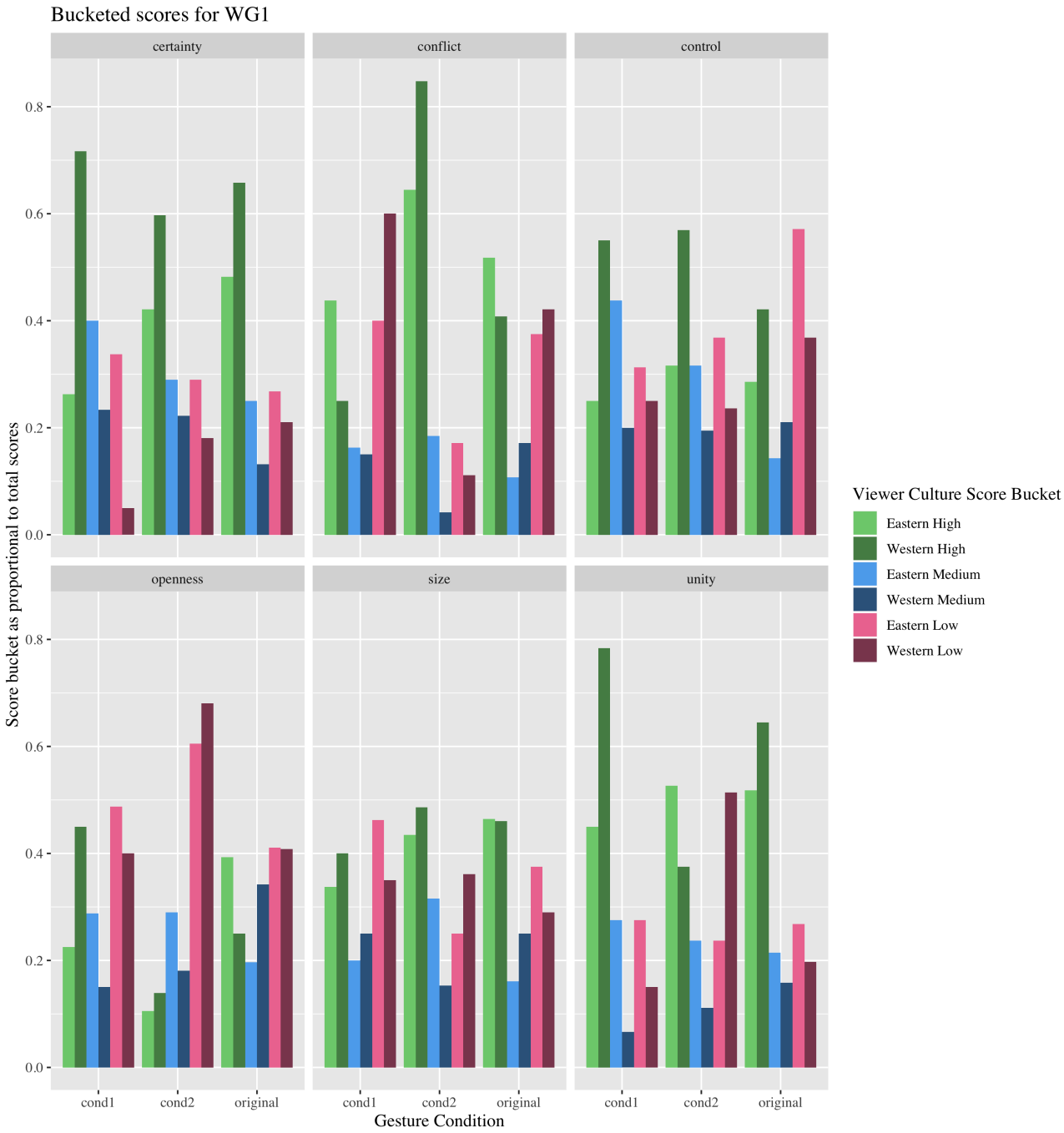


Figure 7.5: Bucketed results for Western Gesture 1.

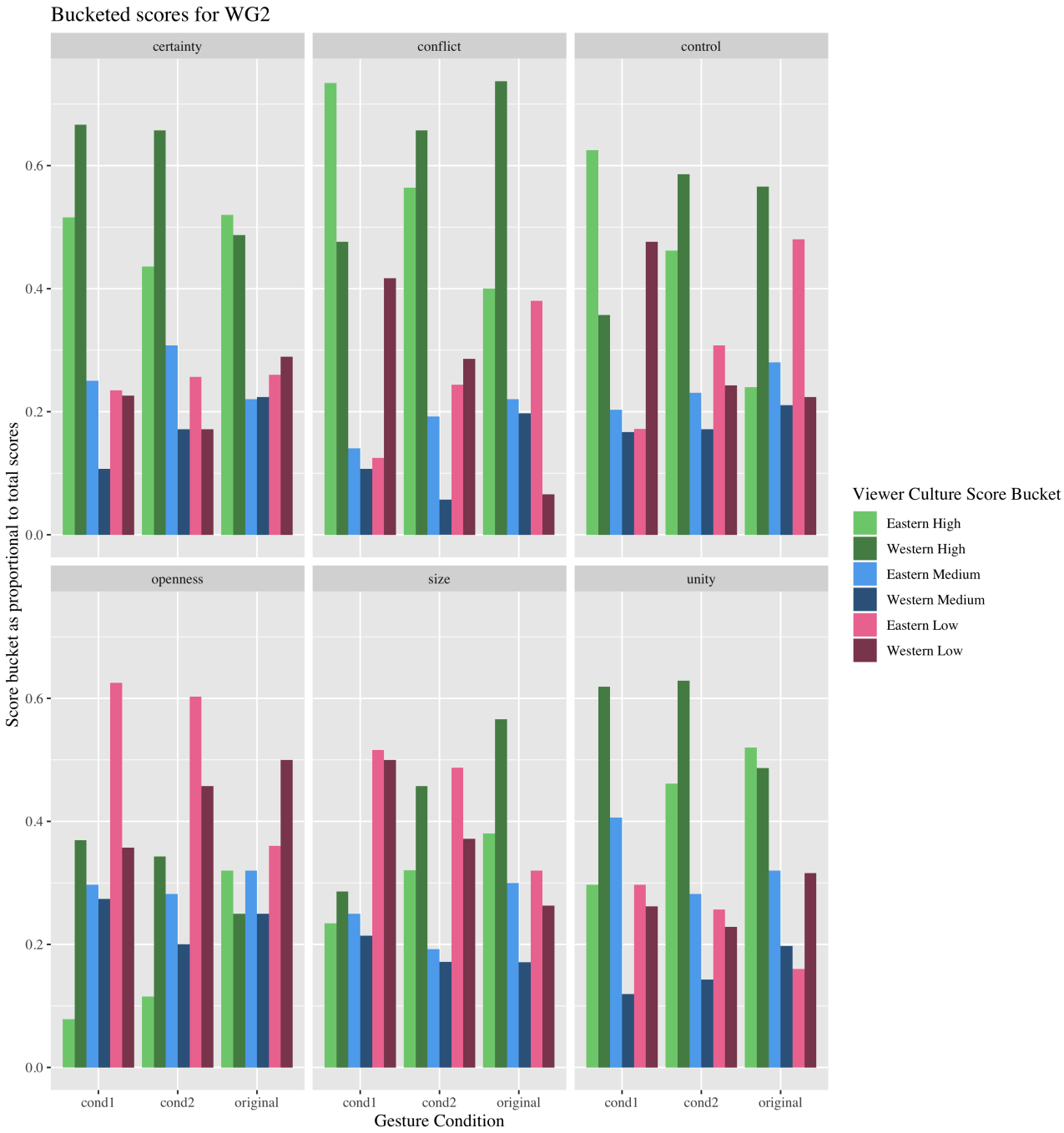


Figure 7.6: Bucketed results for Western Gesture 2.

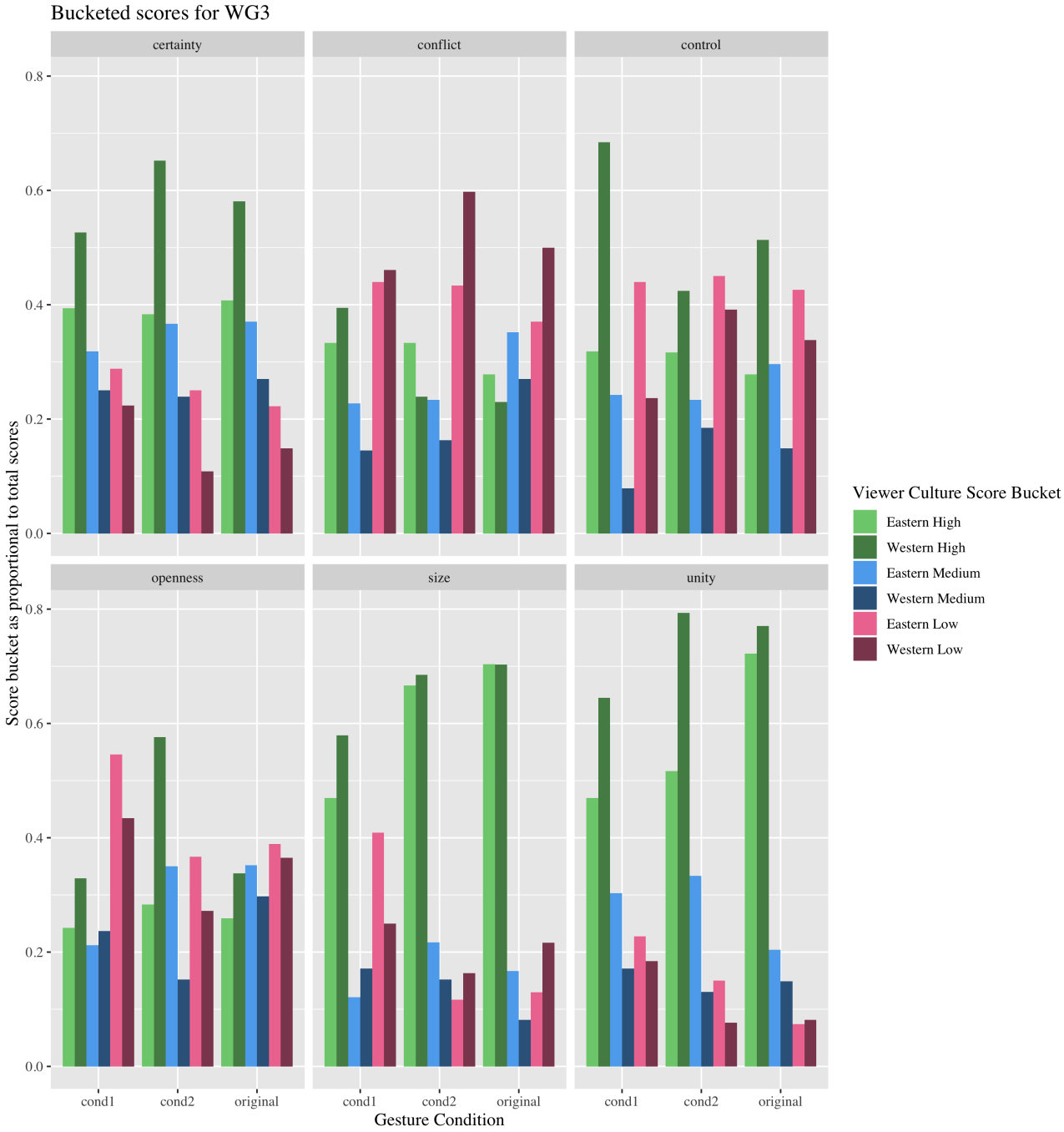


Figure 7.7: Bucketed results for Western Gesture 3.

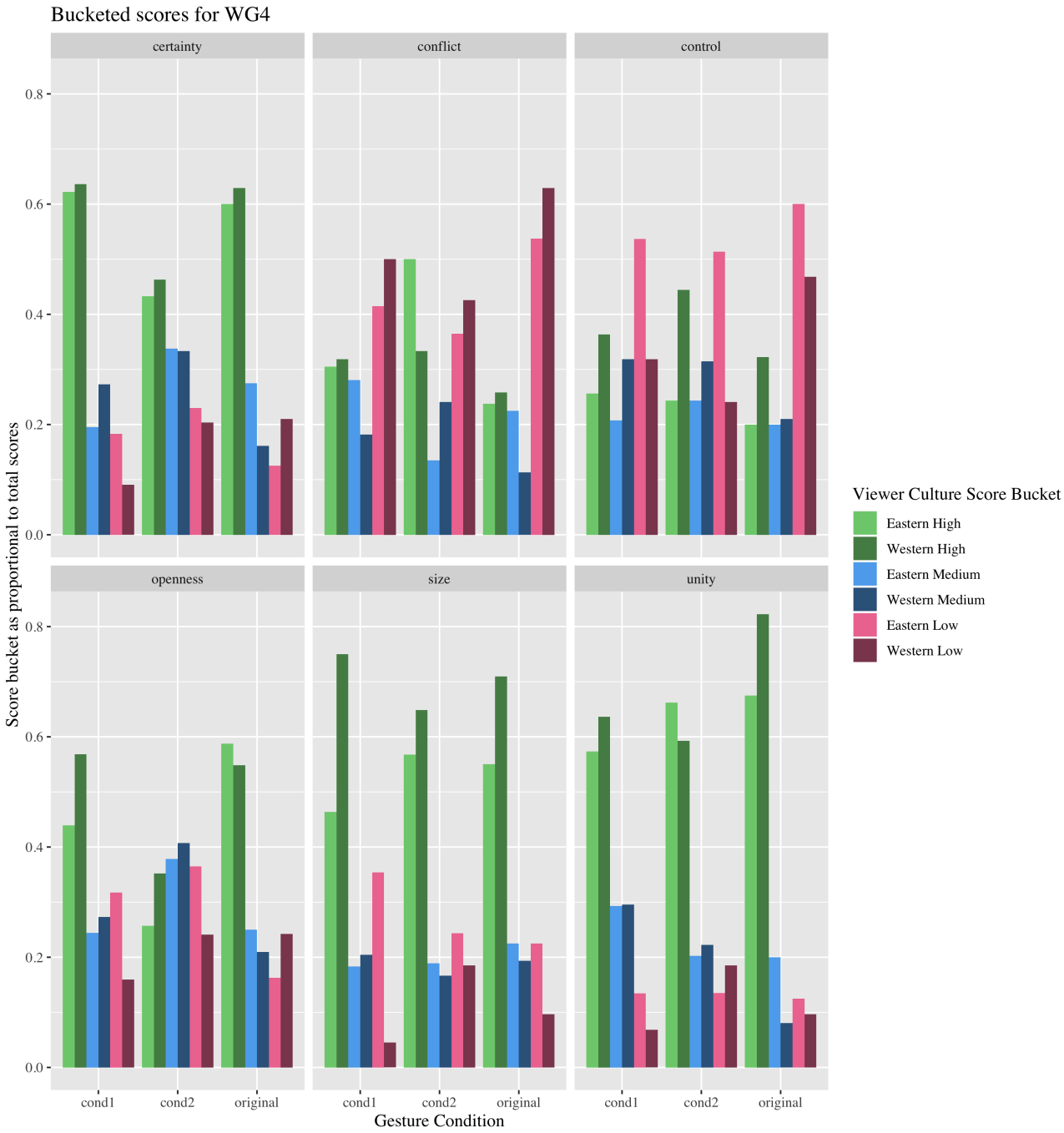


Figure 7.8: Bucketed results for Western Gesture 4.

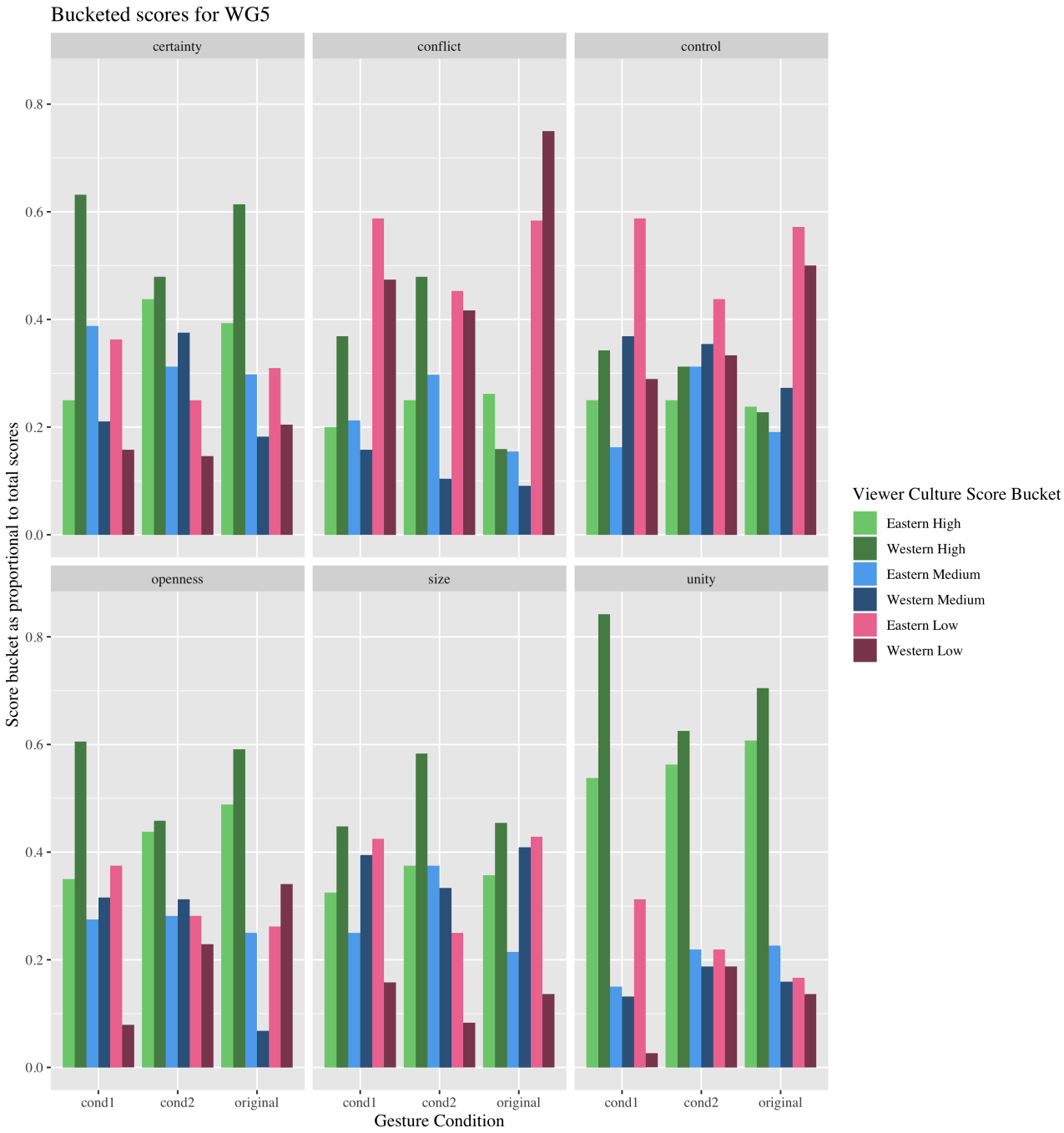


Figure 7.9: Bucketed results for Western Gesture 5.

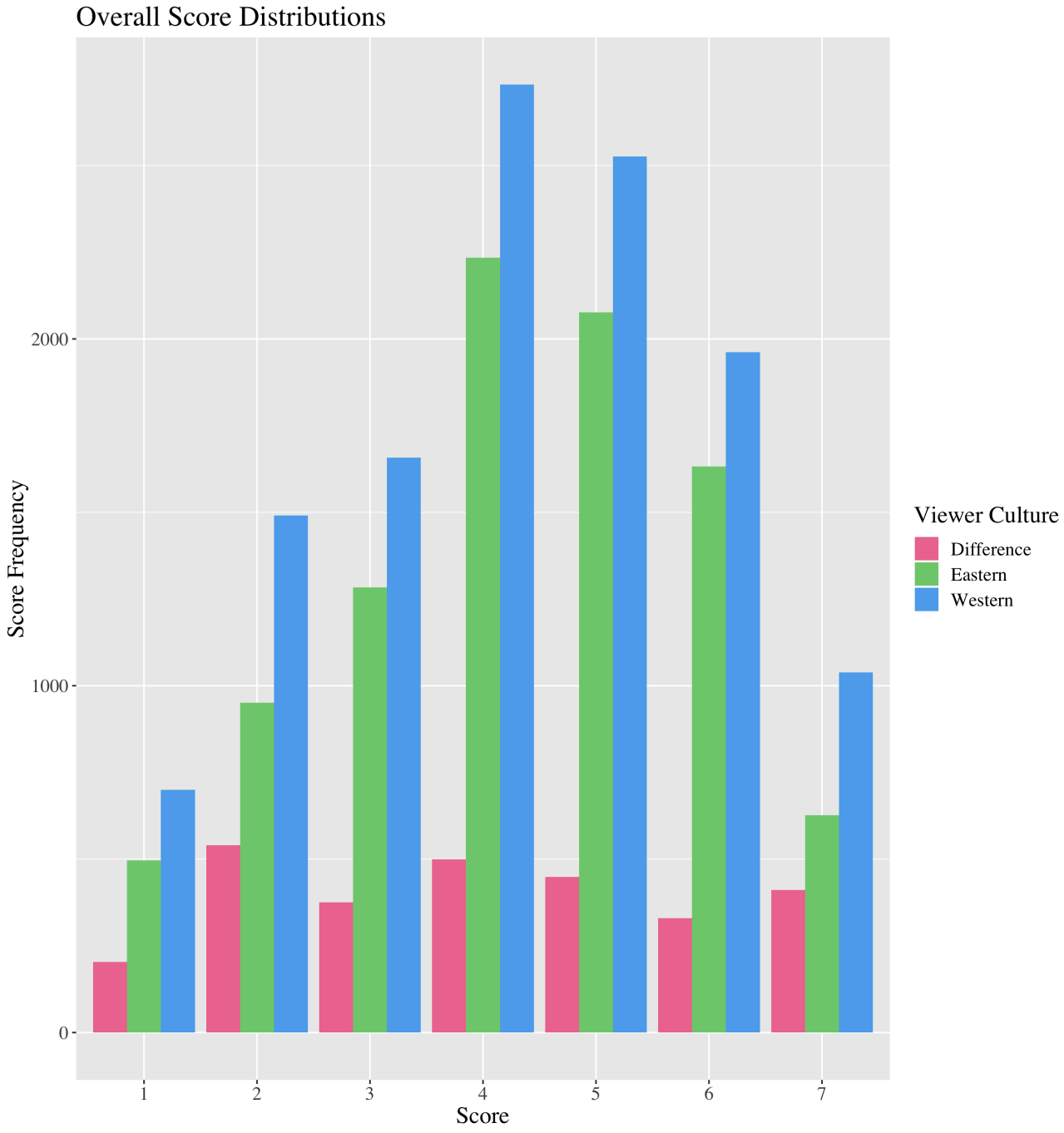


Figure 7.10: Overall distribution of responses across cultures.

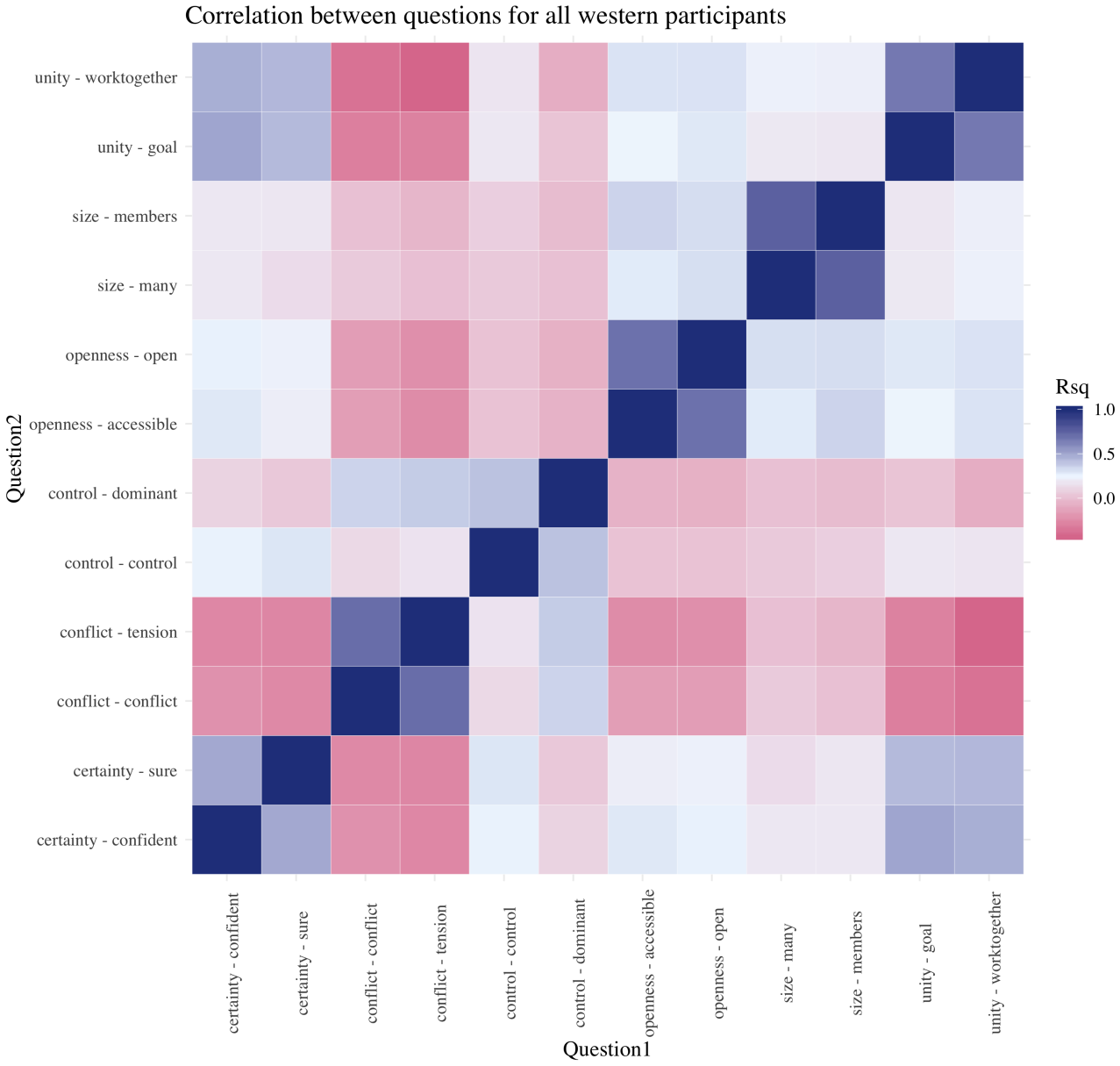


Figure 7.11: Overall correlation between response domains for western viewers.

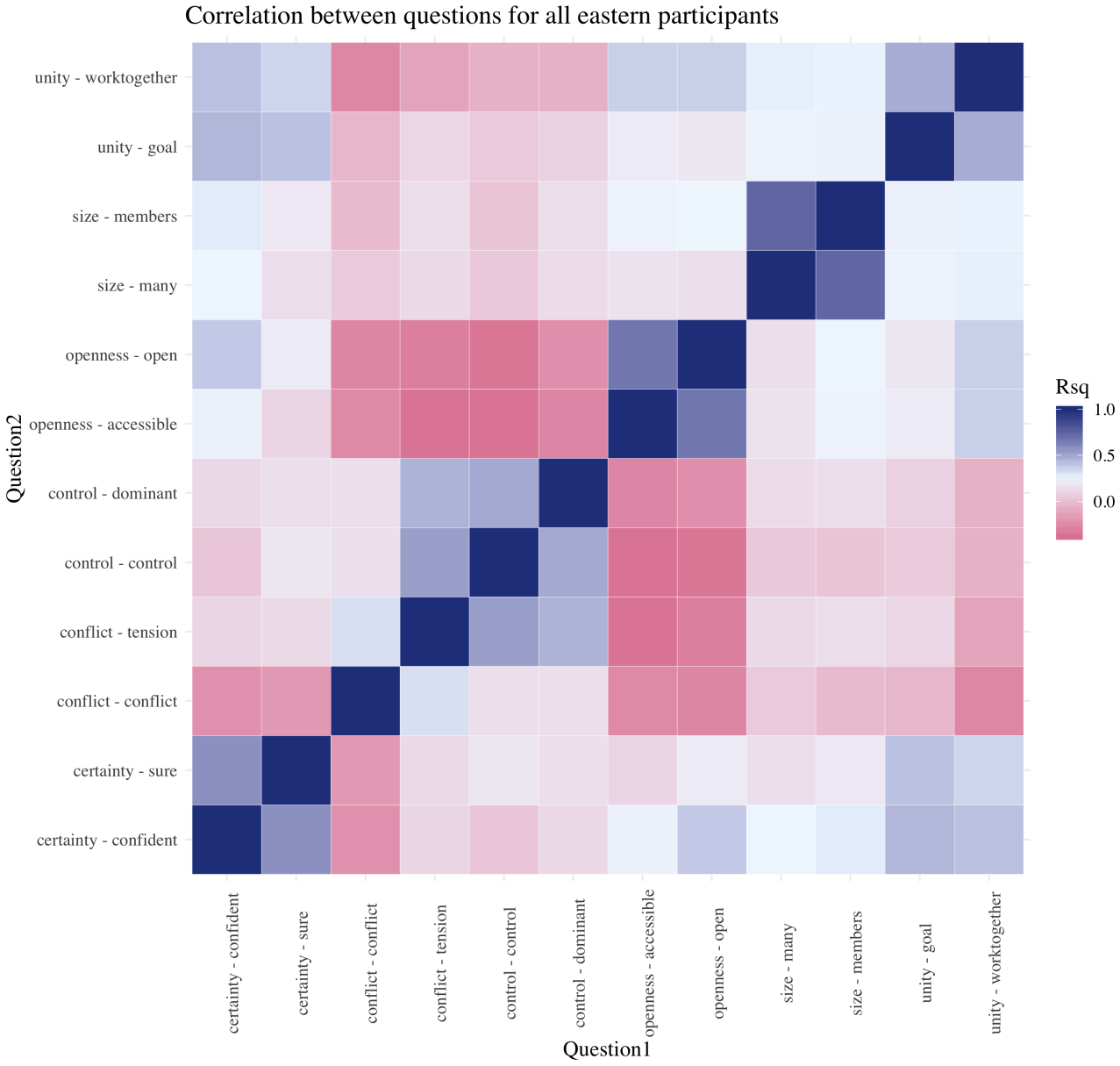


Figure 7.12: Overall correlation between response domains for eastern viewers.

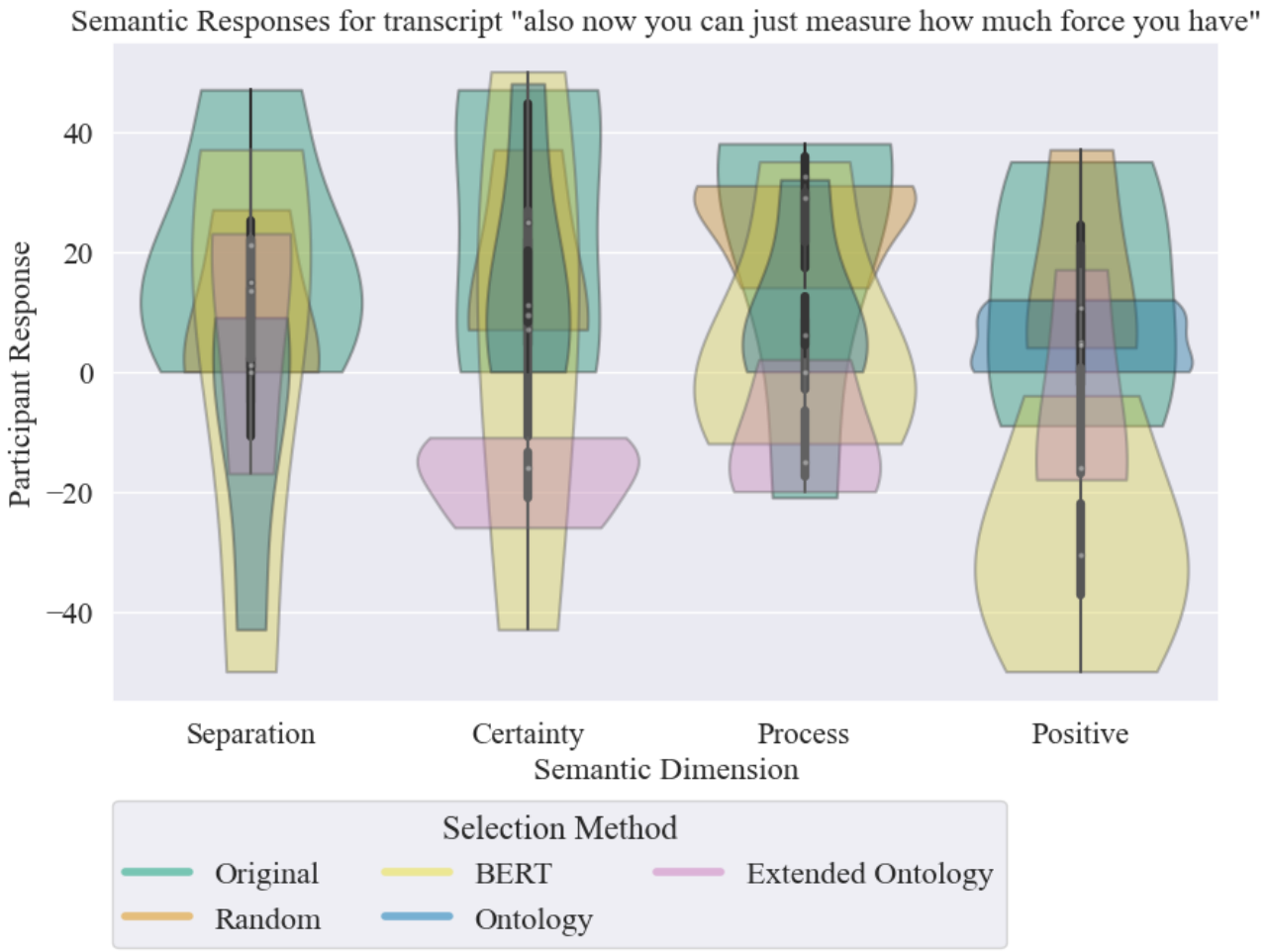


Figure 7.13: Semantic response distributions for the transcript shown in the title.

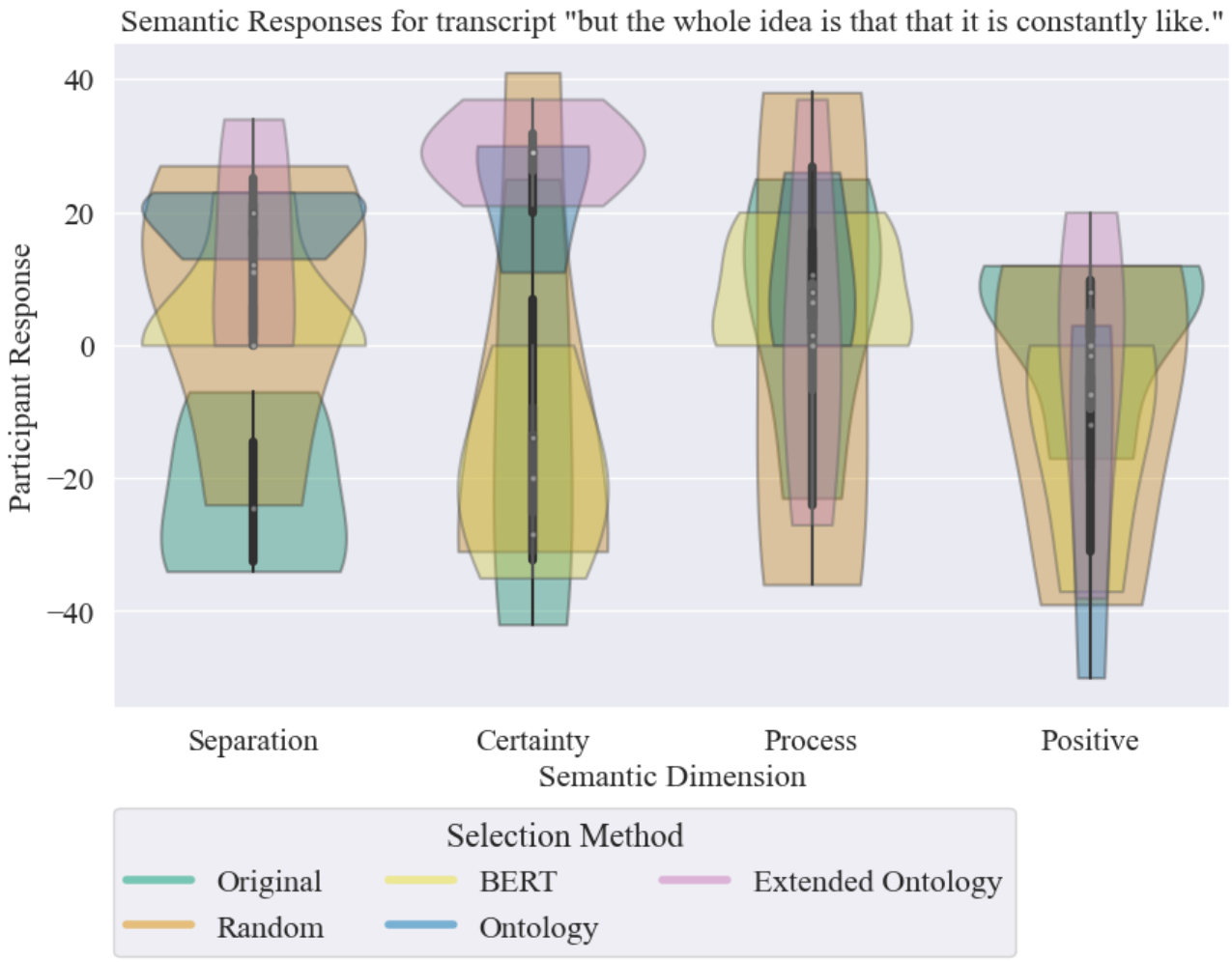


Figure 7.14: Semantic response distributions for the transcript shown in the title.

7.4 Architecture Implementation for Chapter 4

All tools for this implementation and analysis code can be found at
`https://github.com/UofGSocialRobotics/CCFM-generation/`

This link also includes a README to go from raw data to analysis results.

7.4.1 Feature Derivation for Section 4.1.4

Maximum and Minimum wrist distance

Distance between wrists is measured using Euclidean distance of the 3D keypoint found at the wrist joint of each hand. We calculate the wrist distance at each frame and use the maximum and minimum distance found throughout the gesture as these features.

Maximum and Minimum velocity for each wrist

Velocity for a given frame is calculated as the change in wrist position (using Euclidean distance) over the course of three frames, and is thus calculated beginning at the third frame of a gesture. We use three frames (as opposed to a frame-by-frame difference) to smooth very minor changes in hand position that tend to be present in our data. We perform this calculation for each frame (and for each wrist) and use the minimum and maximum of this calculation as these features.

Maximum and Minimum acceleration for each wrist

After calculating the velocity of each frame (for each wrist), we calculate the difference between velocities on a frame-by-frame basis. This gives us the acceleration of each wrist at each frame. We then use the maximum and minimum of these calculations as these features.

Maximum and Minimum vertical orientation of each palm

For each frame, we create a plane of the palm orientation based on keyframes from five positions, where the fingers meet the palm. We add the differences in X (horizontal) positions between each point, and use this sum as a measurement of the difference between the orientation of these points and a vertical plane.

Maximum and Minimum horizontal orientation of each palm

For each frame, we create a plane of the palm orientation based on keyframes from five positions, where the fingers meet the palm. We add the differences in Y (vertical) positions between each point, and use this sum as a measurement of the difference between the orientation of these points and a horizontal plane.

The distance the wrists move together and apart throughout the gesture

This is calculated by taking the sum of the absolute value of the difference between wrist distances over each frame of the gesture. This sum is used as this feature.

Cyclical motion

We first project the path of each wrist onto the Y (vertical) axis. Then, we calculate the angular displacement (θ) for the path of the wrist using a sliding window of three frames $0, 1, 2$. We do this by estimating the radius of a potential circle made by the path of the wrist over this time r by finding the midpoint m between the wrists at frame 0 and frame 2, then adding the difference between the distance between m and frame 1. The sum of θ across each window is used as the feature for this gesture.

7.4.2 Textual Analysis Features

All TAFs:

- NOUN_Ont (Noun Ontology)
- metaphor (extracted feature)
- SPhr (Spatial Phrase)
- VERB_Ont (Verb Ontology)
- PROPN_Ont (Proper Noun Ontology e.g. “Statue of Liberty” –> Sculpture
- adj (adjective)
- NUM_Ont (Numerical Ontology e.g. “Fourth” –> Four
- advLemmas (Adverbial Lemma)
- noun (Noun Desriptors e.g. plural)
- funcMetaphors (Functional Metaphor, extracted feature)
- adjLemmas (Adjectival Lemma)
- nounLemmas (Noun Lemma)
- AUX_Ont (Auxiliary Ontological Features)
- possession (possession)
- adv (Adverb)
- SP_ONT (Speech-Preposition Ontology)

All instances of Functional Metaphor:

- affirmation
- affirmation_weak
- approximation

- around
- aside
- await
- between
- causal
- certainty
- comparative_bigger
- comparative_longer
- comparative_notspecified
- comparative_shorter
- comparative_smaller
- comparative_taller
- contrast_notspecified
- contrast_weak
- cycle
- dampen
- deixis_here
- deixis_me
- deixis_notspecified
- deixis_us
- deixis_you
- dir_low
- dir_notspecified
- dir_side (Directionality, to the side)
- emonegative
- emopositive
- empty

- focus_center
- greeting
- important
- inclusivity
- indifference
- intensification_notspecified
- intensification_weak
- location_around
- location_aside
- location_distant
- location_here
- location_within
- movement
- negation_notspecified
- negation_strong
- object_abstract
- obligation
- process_abstract
- process_concrete
- process_frequent
- process_stop
- quantity_all
- quantity_approximation
- quantity_empty
- quantity_large
- quantity_nothing
- question_while

- reject
- scary
- stop
- surprise
- surround
- time_after
- time_before
- time_now
- uncertainty

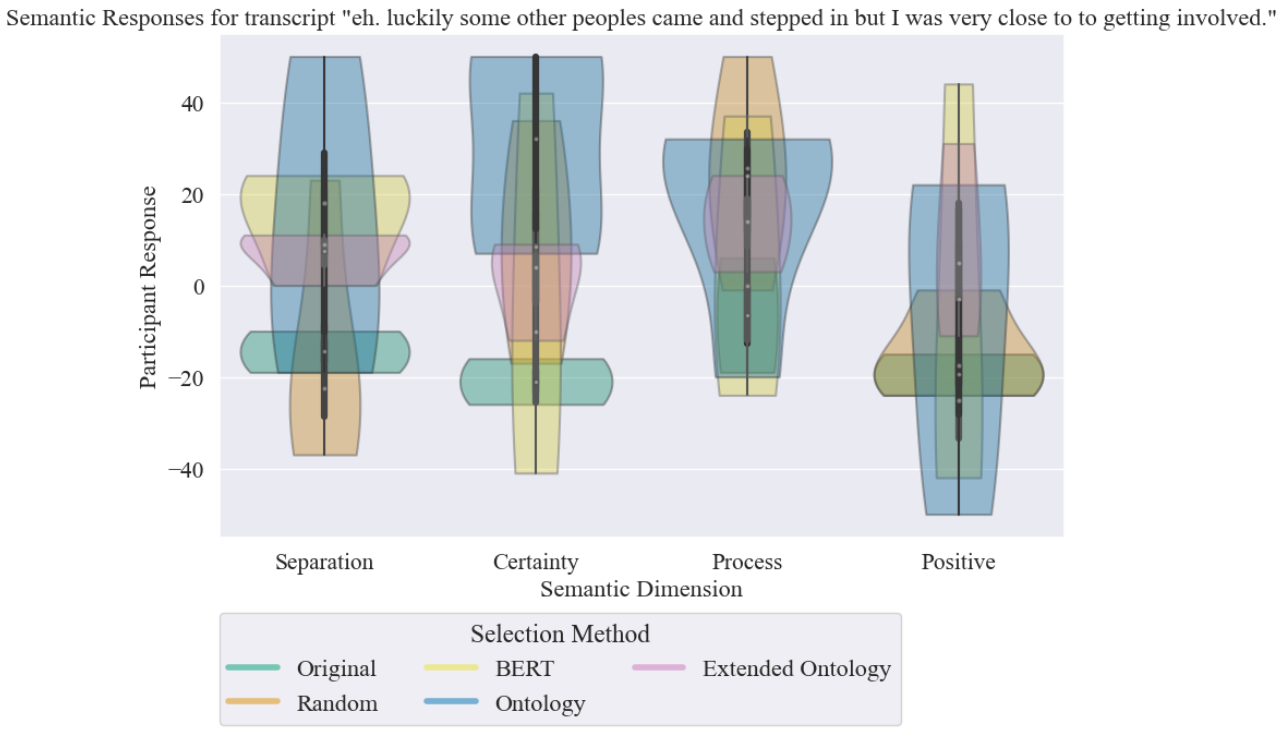


Figure 7.15: Semantic response distributions for the transcript shown in the title.

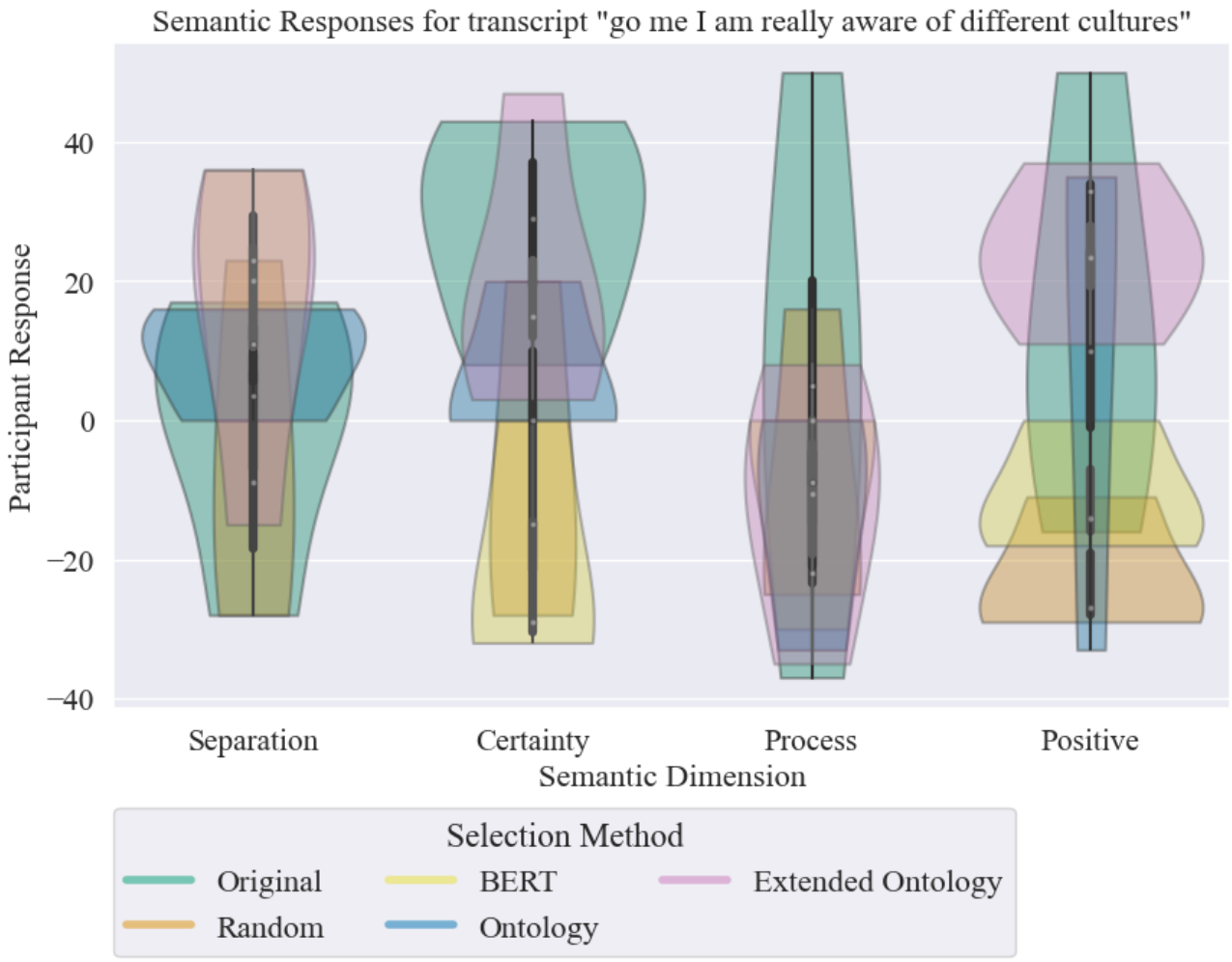


Figure 7.16: Semantic response distributions for the transcript shown in the title.

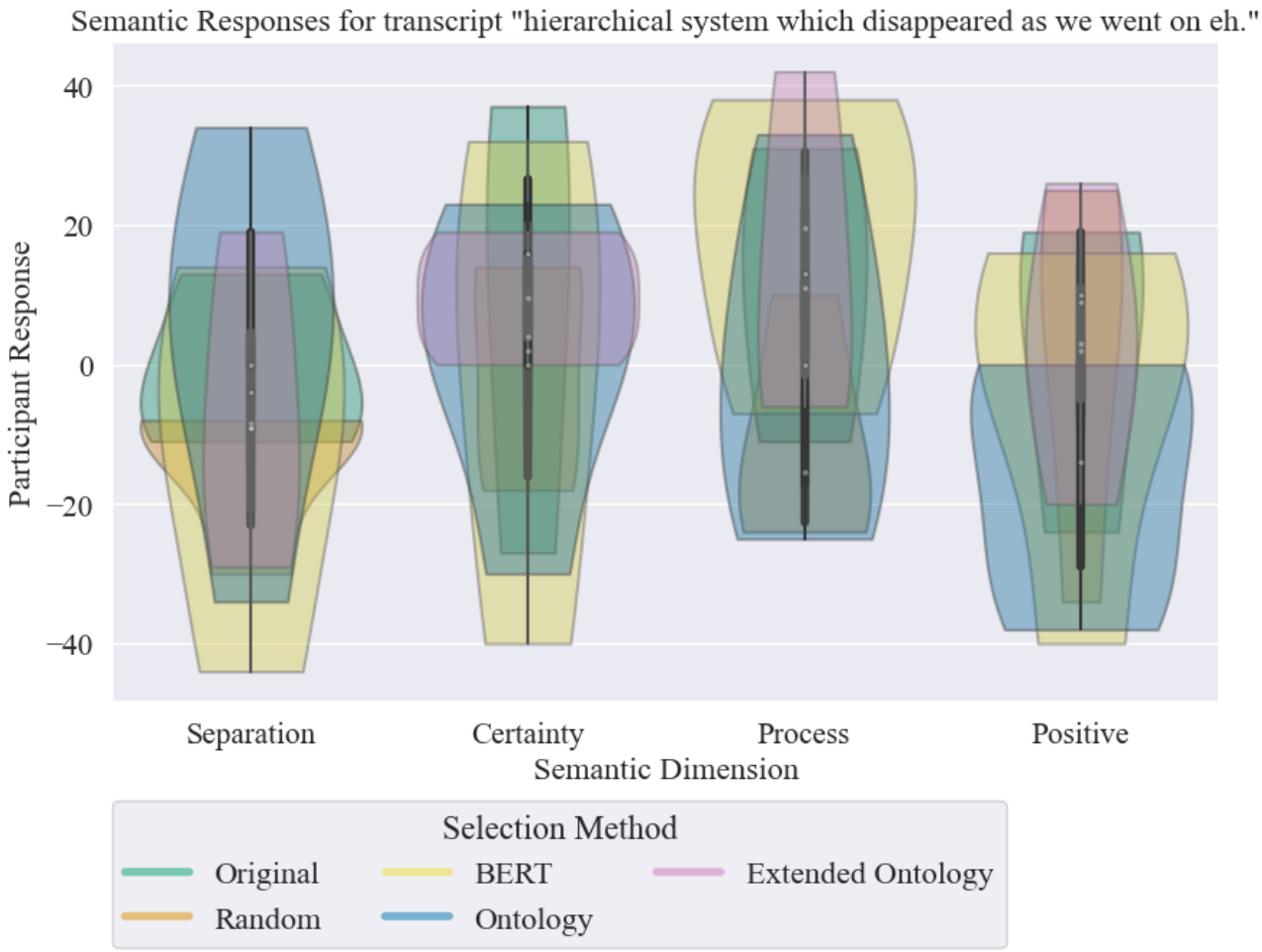


Figure 7.17: Semantic response distributions for the transcript shown in the title.

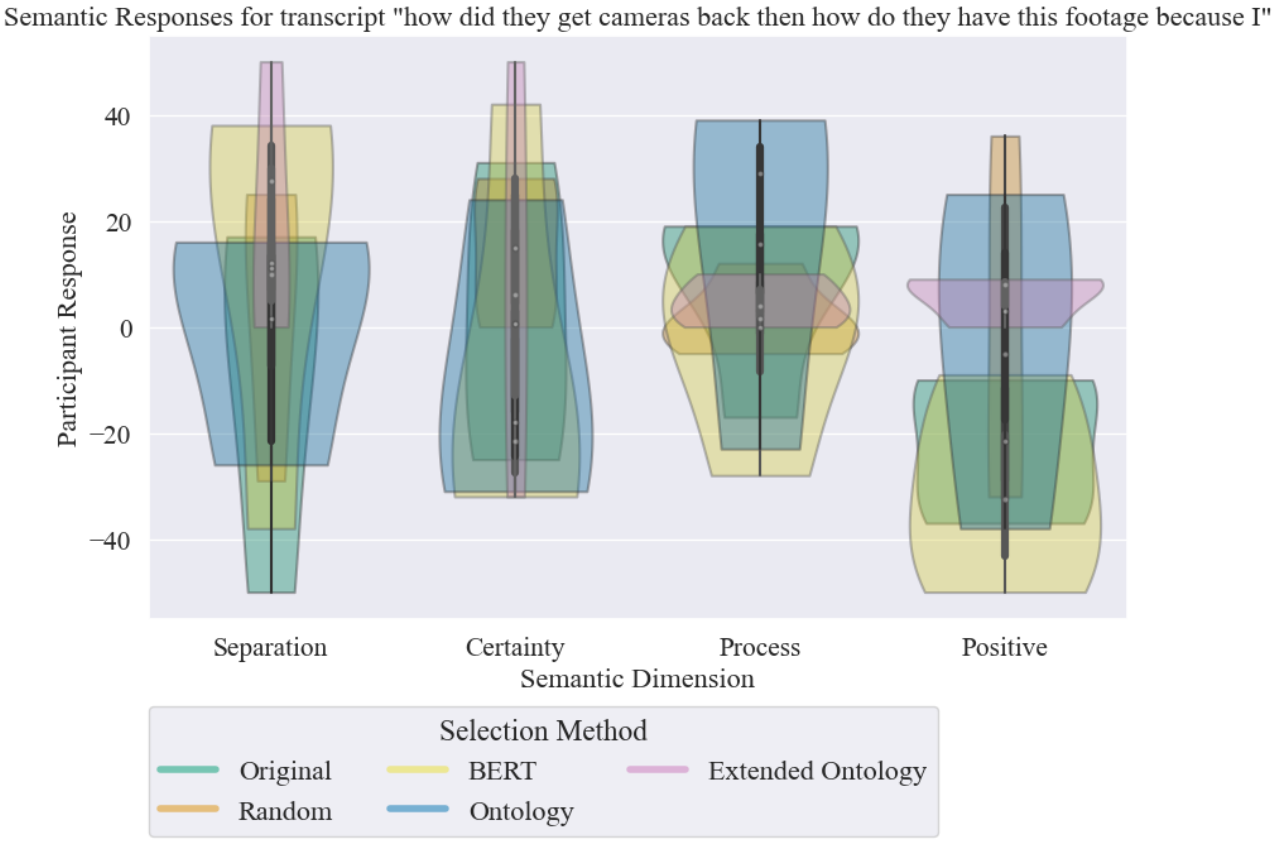


Figure 7.18: Semantic response distributions for the transcript shown in the title.

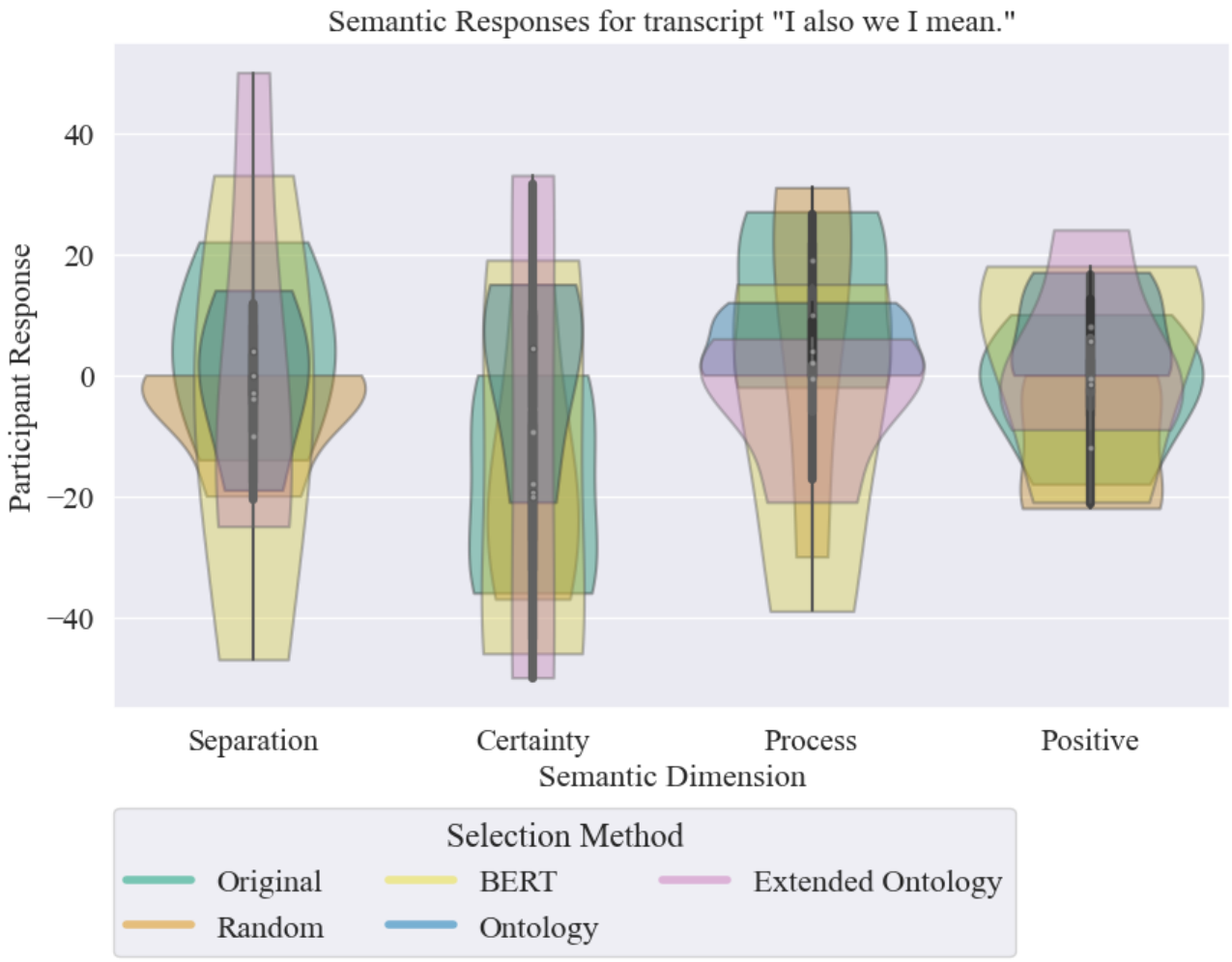


Figure 7.19: Semantic response distributions for the transcript shown in the title.

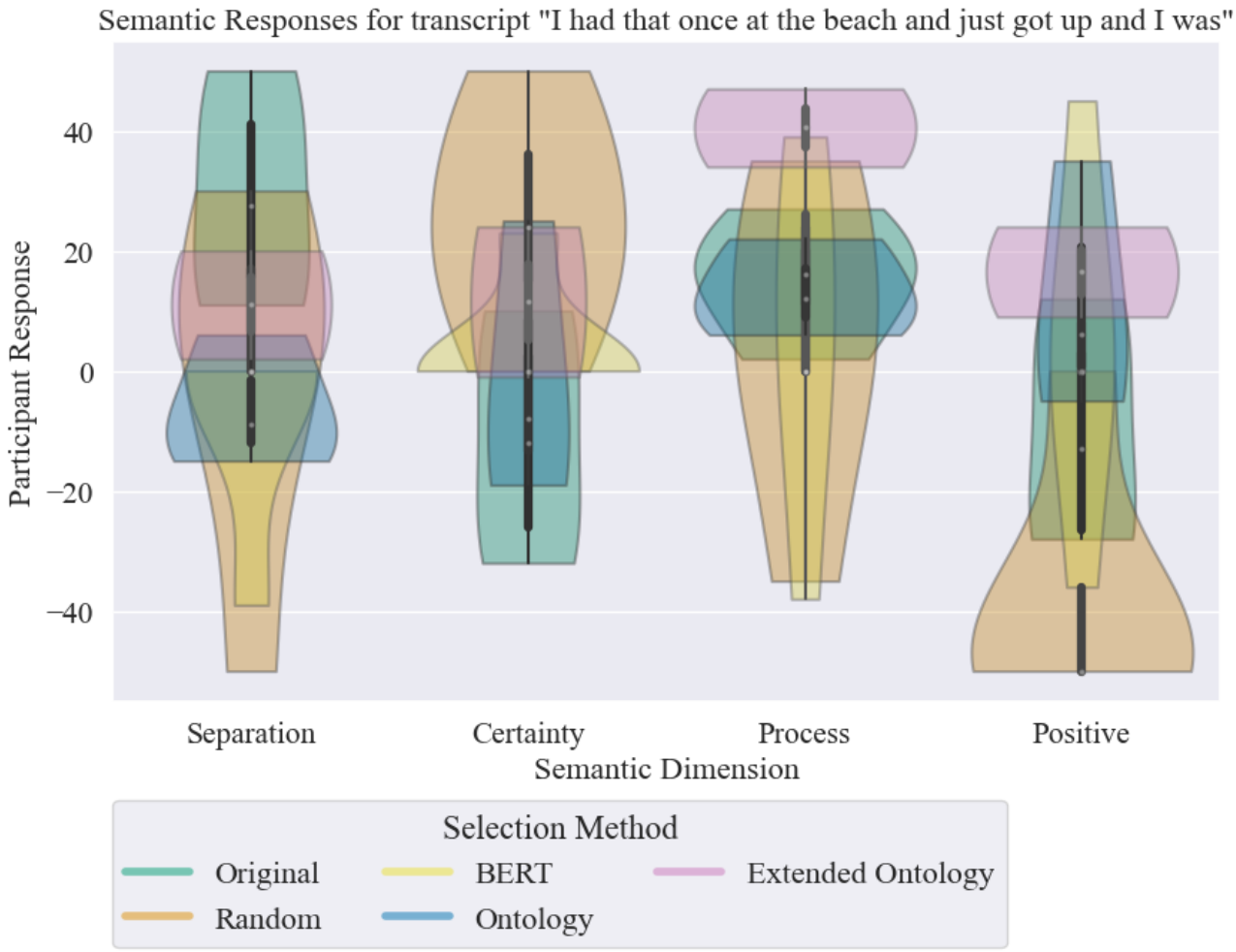


Figure 7.20: Semantic response distributions for the transcript shown in the title.

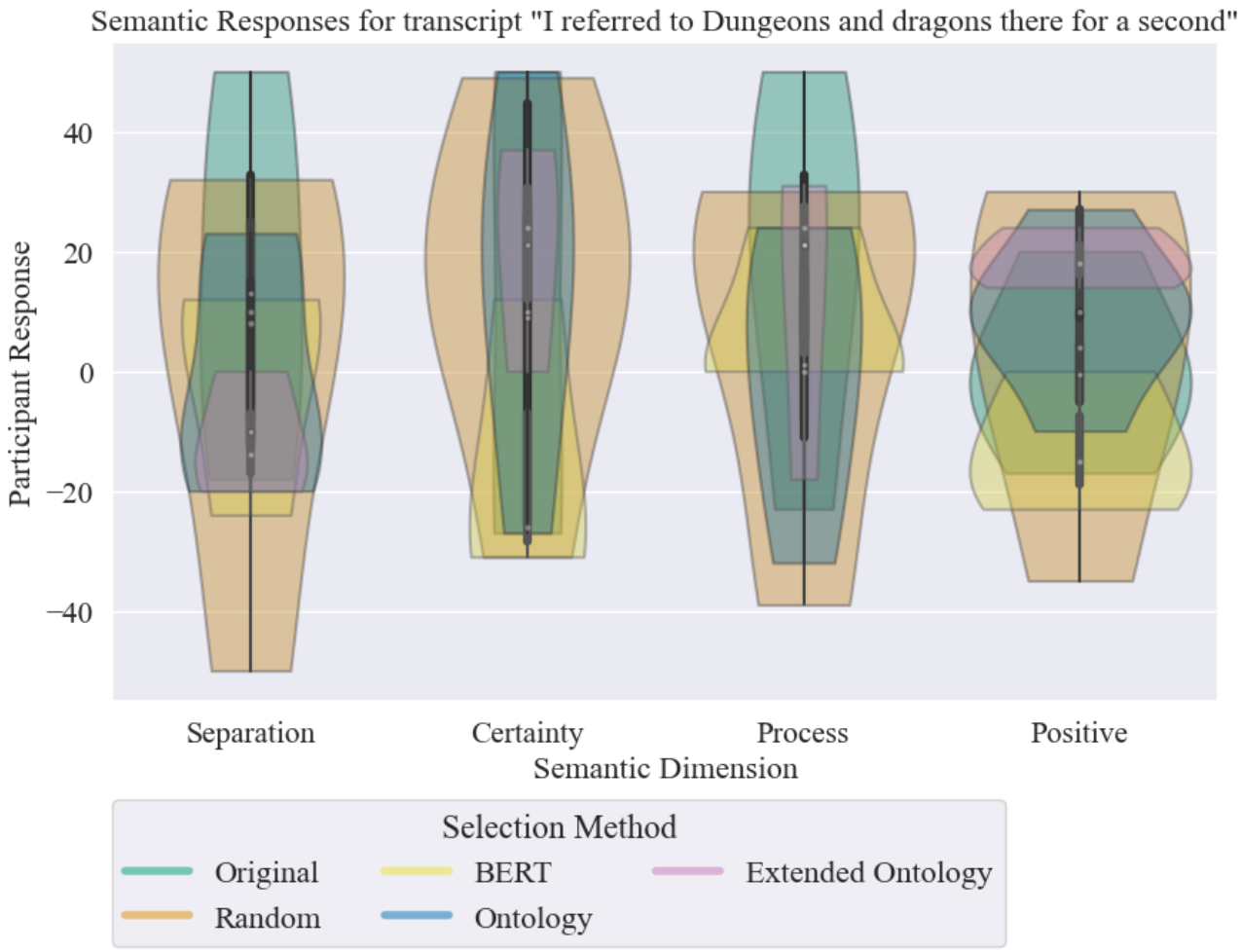


Figure 7.21: Semantic response distributions for the transcript shown in the title.

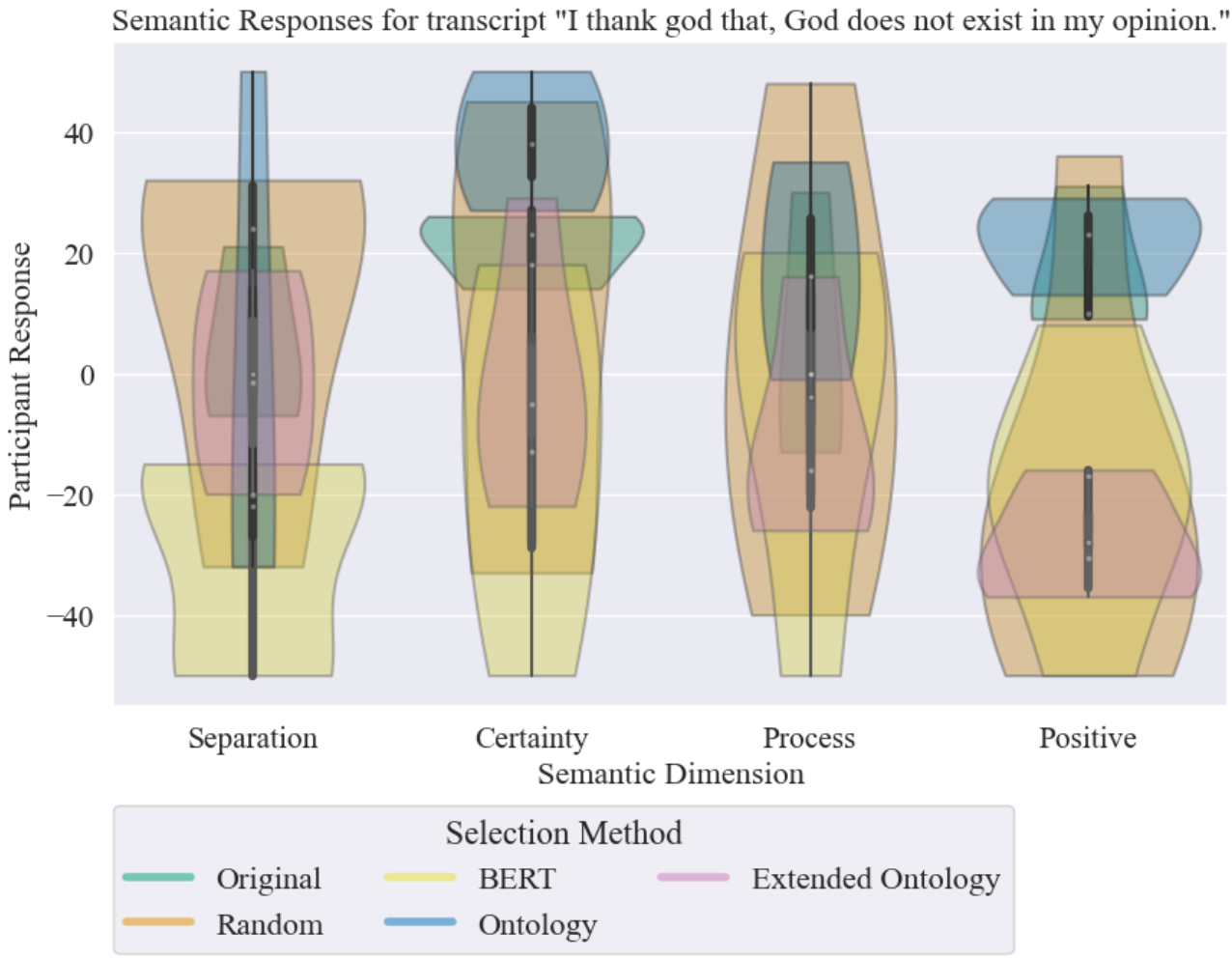


Figure 7.22: Semantic response distributions for the transcript shown in the title.

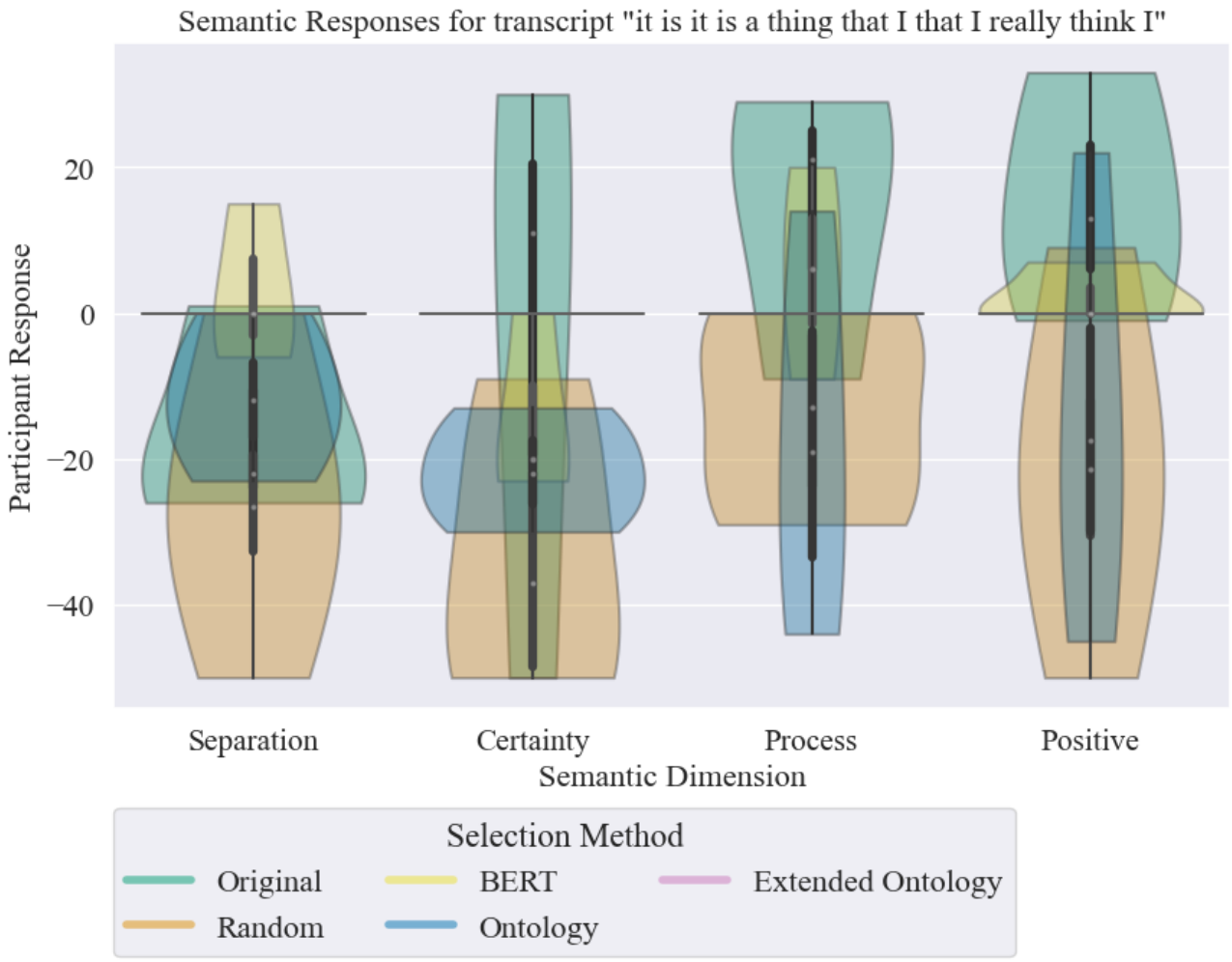


Figure 7.23: Semantic response distributions for the transcript shown in the title.

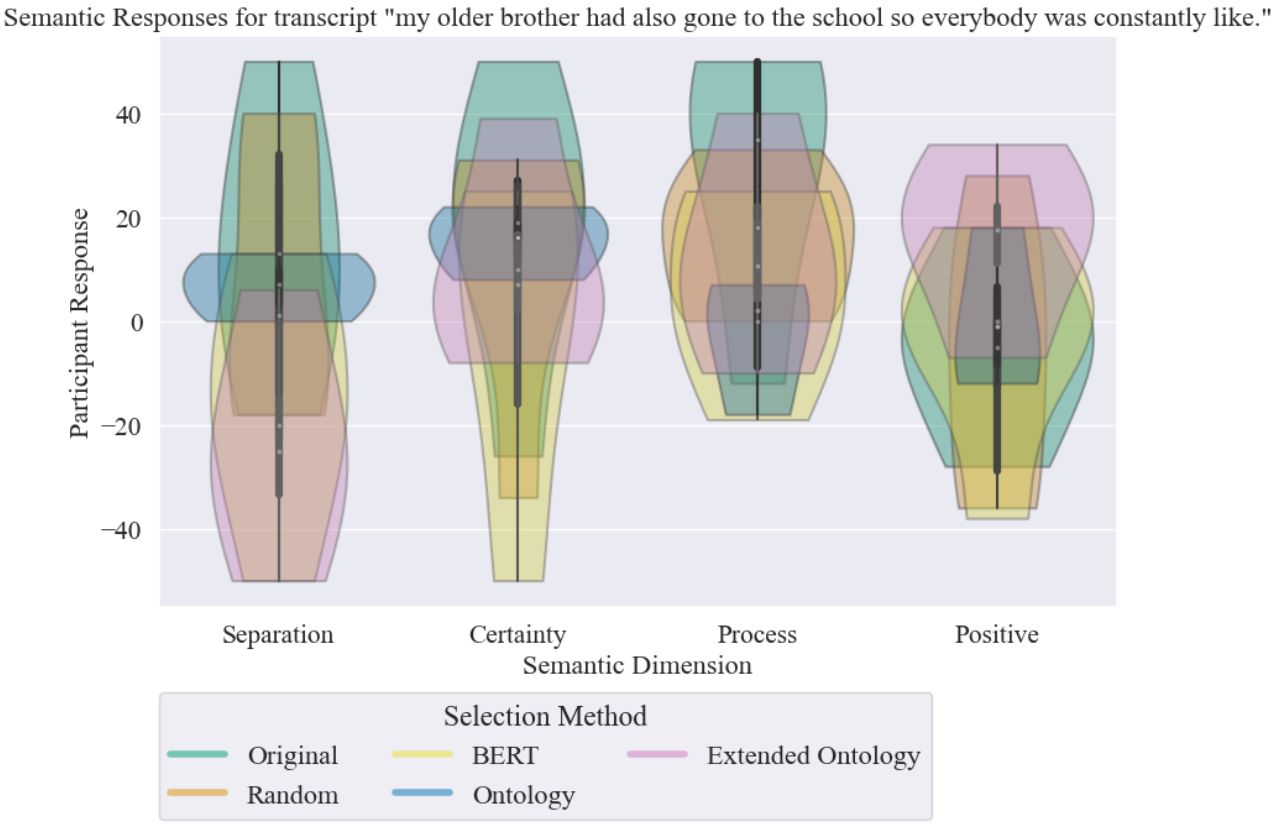


Figure 7.24: Semantic response distributions for the transcript shown in the title.

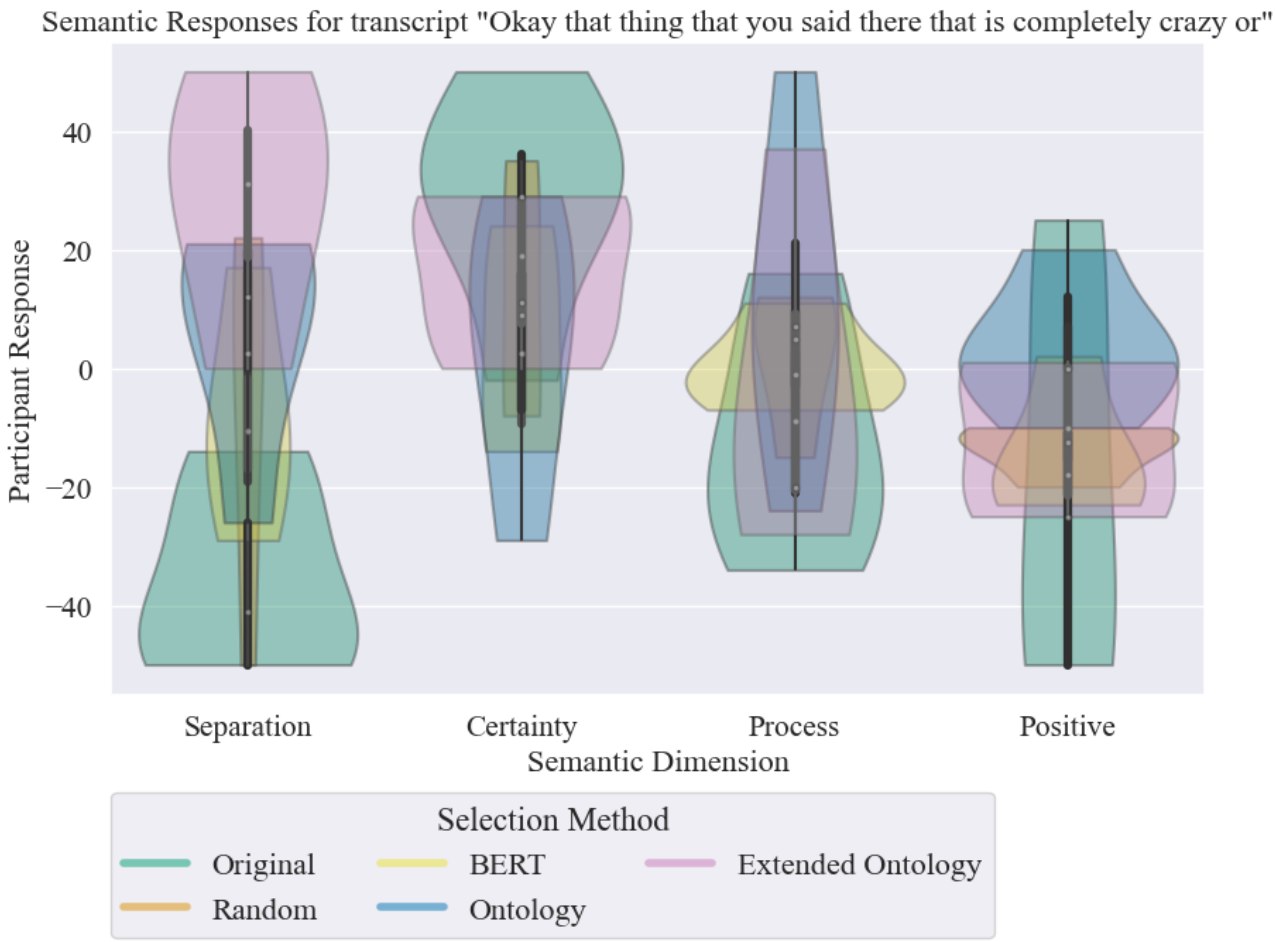


Figure 7.25: Semantic response distributions for the transcript shown in the title.

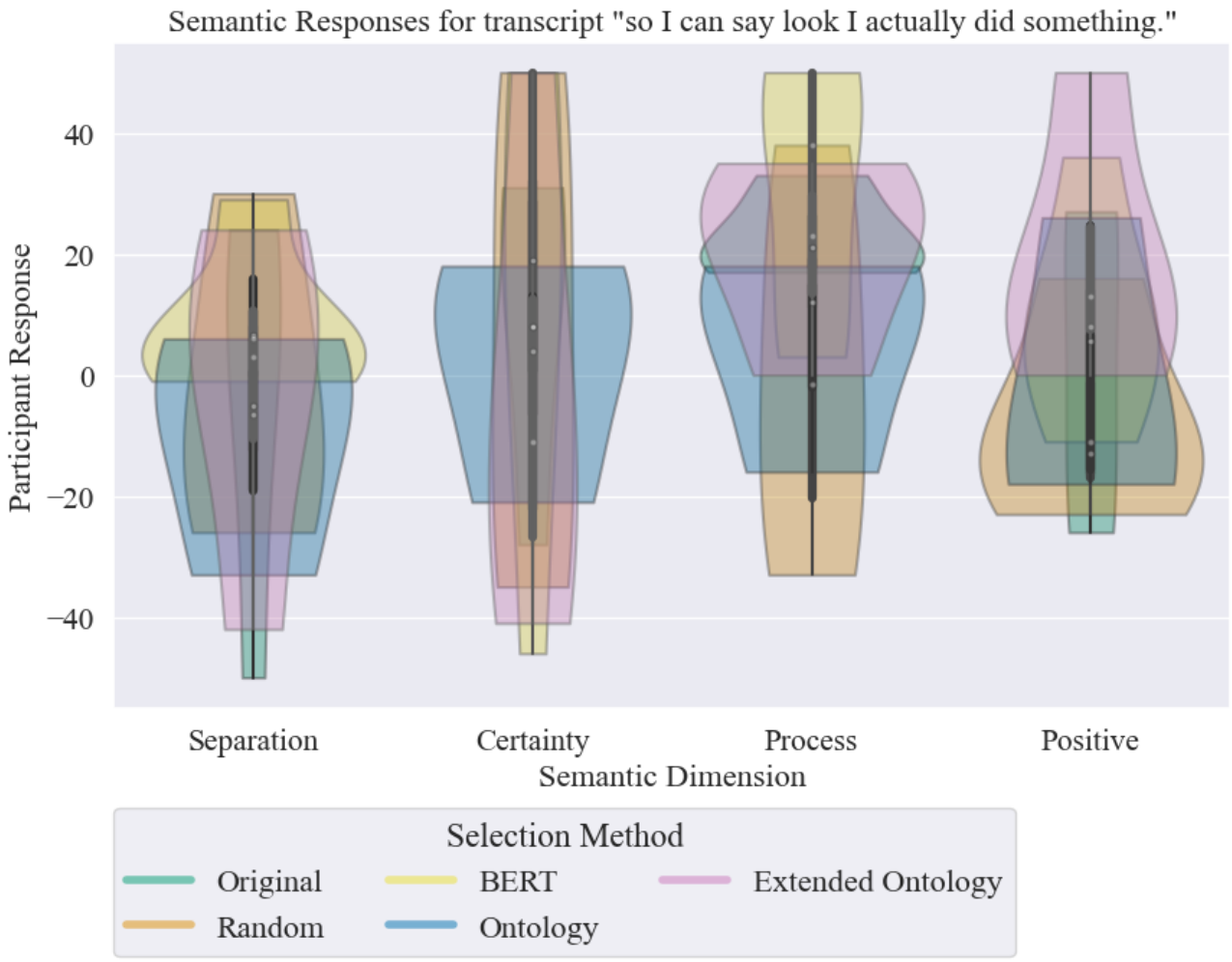


Figure 7.26: Semantic response distributions for the transcript shown in the title.

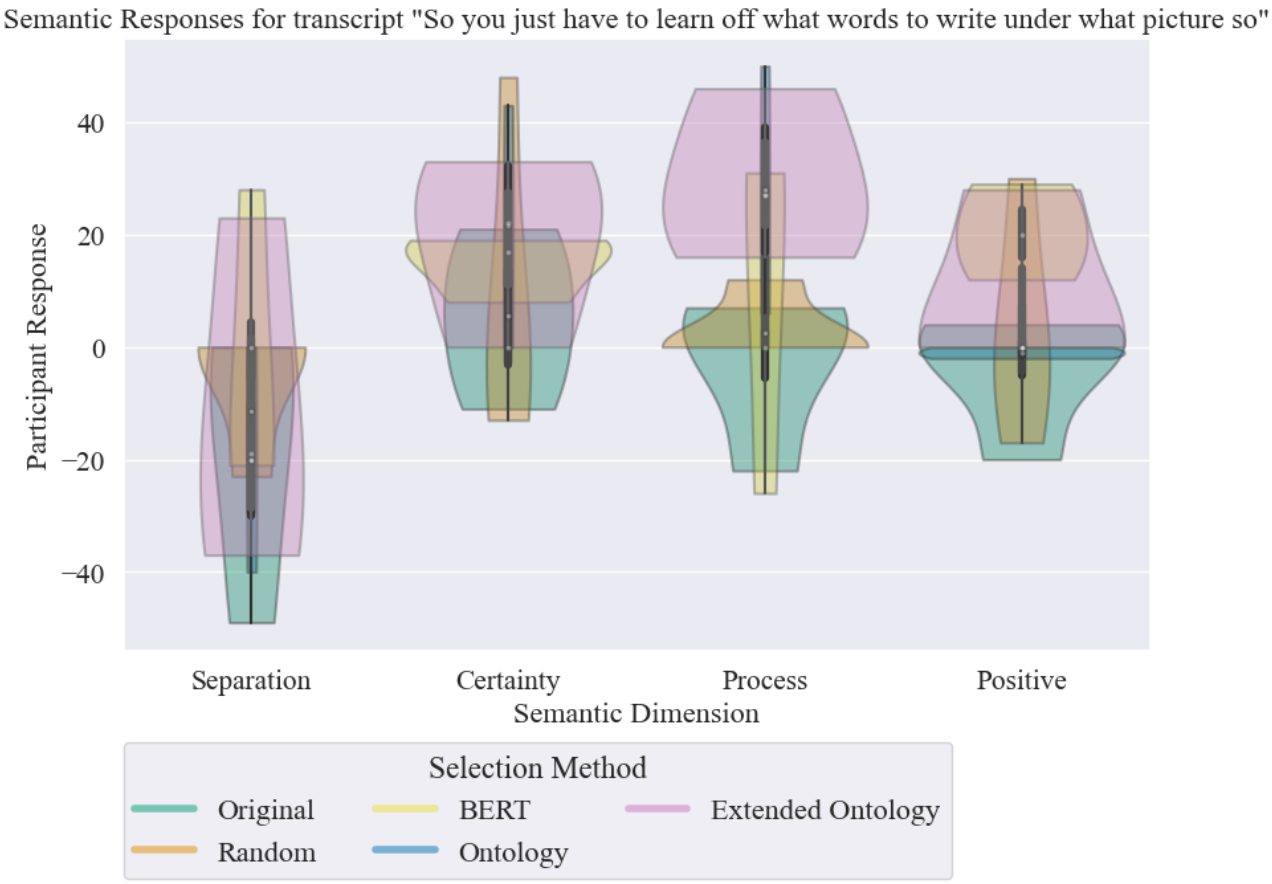


Figure 7.27: Semantic response distributions for the transcript shown in the title.

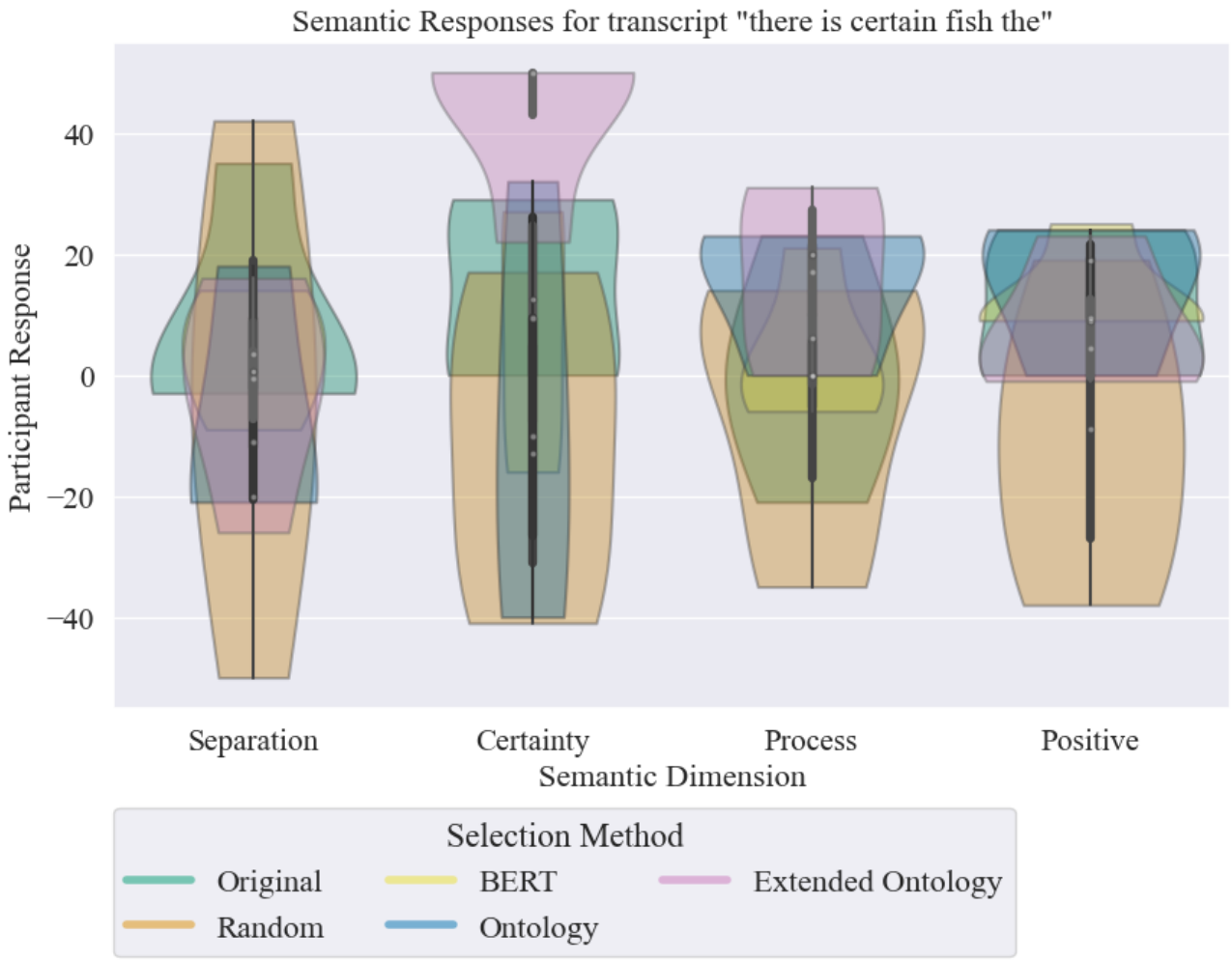


Figure 7.28: Semantic response distributions for the transcript shown in the title.

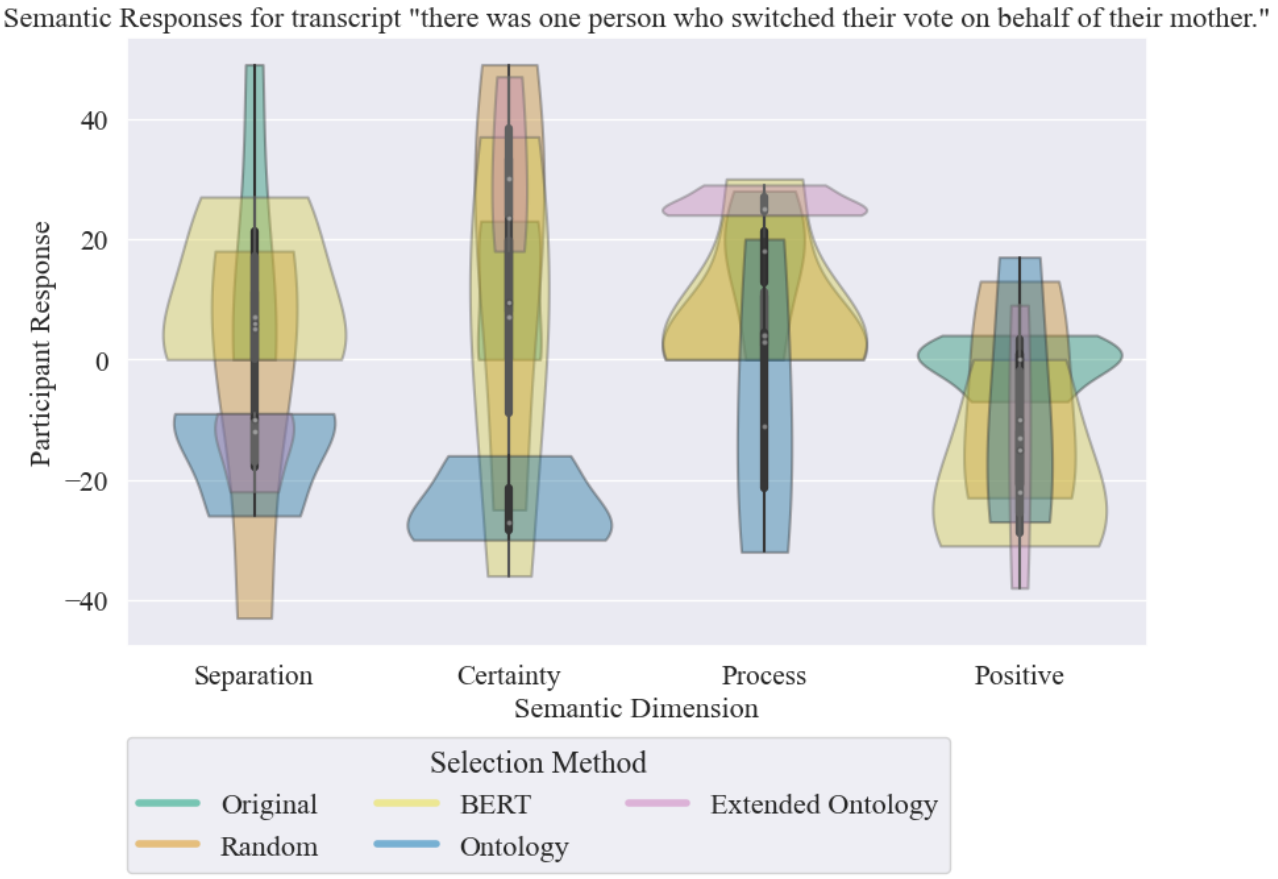


Figure 7.29: Semantic response distributions for the transcript shown in the title.

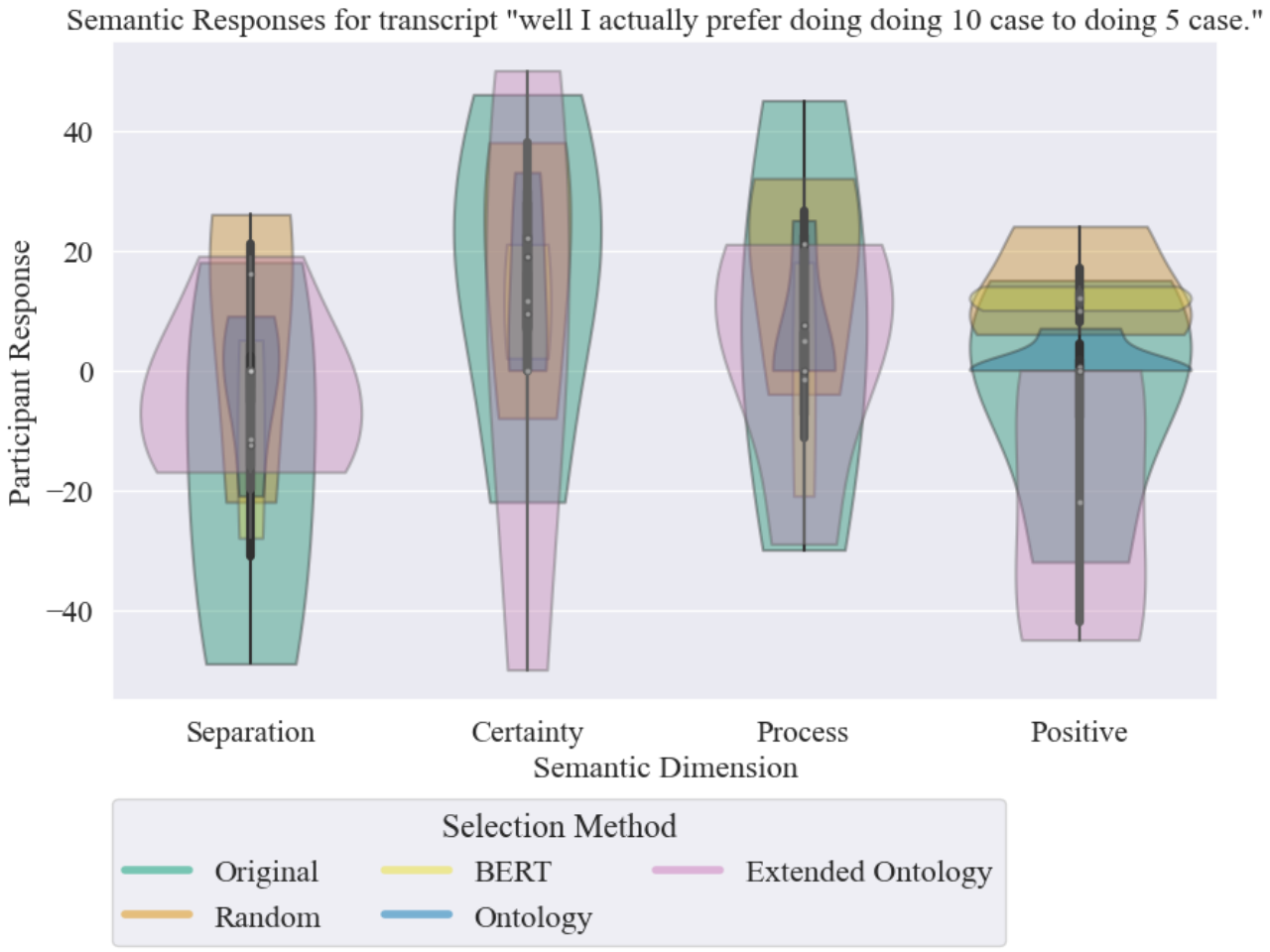


Figure 7.30: Semantic response distributions for the transcript shown in the title.

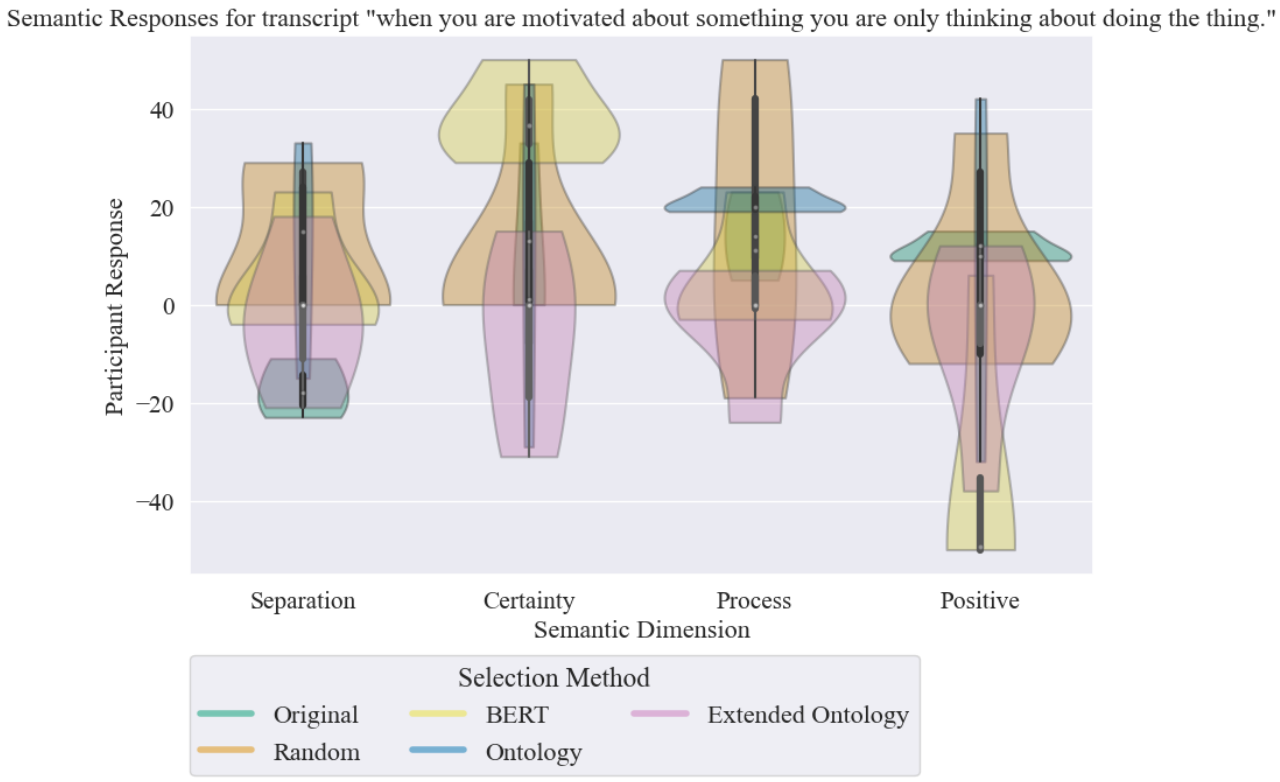


Figure 7.31: Semantic response distributions for the transcript shown in the title.

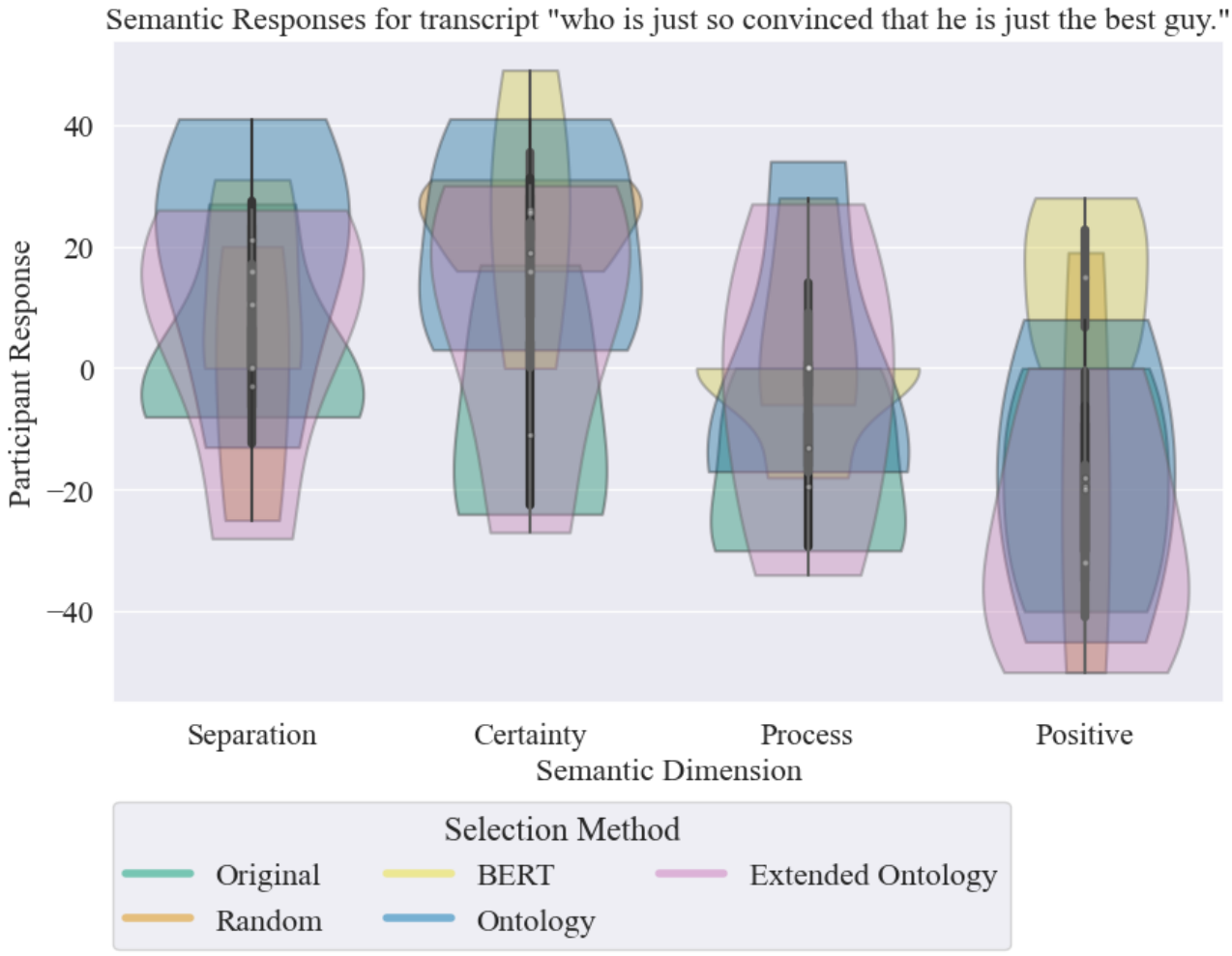


Figure 7.32: Semantic response distributions for the transcript shown in the title.

Bibliography

Ahuja, Chaitanya, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency (2020). "No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1884–1895.

Ahuja, Chaitanya, Dong Won Lee, and Louis-Philippe Morency (2022). "Low-Resource Adaptation for Personalized Co-Speech Gesture Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20566–20576.

Alexanderson, Simon, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow (2020). "Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows". In: *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library, pp. 487–496.

Alibali, Martha W and Susan GoldinMeadow (1993). "Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child s state of mind". In: *Cognitive psychology* 25.4, pp. 468–523.

Alibali, Martha W, Andrew G Young, Noelle M Crooks, Amelia Yeo, Matthew S Wolfgram, Iasmine M Ledesma, Mitchell J Nathan, Ruth Breckinridge Church, and Eric J Knuth (2013). "Students learn more when their teacher has learned to gesture effectively". In: *Gesture* 13.2, pp. 210–233.

Allen, James, Hannah An, Ritwik Bose, Will de Beaumont, and Choh Man Teng (2020). "A broad-coverage deep semantic lexicon for verbs". In: *arXiv preprint arXiv:2007.02670*. URL: `https://pypi.org/project/pytrips/`.

Allen, James, Myroslava O Dzikovska, Mehdi Manshadi, and Mary Swift (2007). "Deep linguistic processing for spoken dialogue systems". In: *ACL 2007 Workshop on Deep Linguistic Processing*, pp. 49–56.

Allen, James and Choh Man Teng (2018). "Putting semantics into semantic roles". In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 235–244.

Althouse, Andrew D (2016). "Adjust for multiple comparisons? It's not that simple". In: *The Annals of thoracic surgery* 101.5, pp. 1644–1645.

Autodesk, INC. (Jan. 15, 2019). *Maya*. Version 2019. URL: `https:/autodesk.com/maya`.

Aylett, Ruth, Marco Vala, Pedro Sequeira, and Ana Paiva (2007). "Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education". In: *International Conference on Virtual Storytelling*. Springer, pp. 202–205.

Barrett, Deirdre (2010). *Supernormal stimuli: How primal urges overran their evolutionary purpose*. WW Norton & Company.

Baumgartner, Hans and Jan-Benedict EM Steenkamp (2001). "Response styles in marketing research: A cross-national investigation". In: *Journal of marketing research* 38.2, pp. 143–156.

Bavelas, Janet Beavin (1994). "Gestures as part of speech: Methodological implications". In: *Research on language and social interaction* 27.3, pp. 201–221.

Beaudoin-Ryan, Leanne and Susan Goldin-Meadow (2014). "Teaching moral reasoning through gesture". In: *Developmental science* 17.6, pp. 984–990.

Bender, Andrea and Sieghard Beller (2014). "Mapping spatial frames of reference onto time: A review of theoretical accounts and empirical findings". In: *Cognition* 132.3, pp. 342–382.

Bergmann, Kirsten and Stefan Kopp (2009). "Increasing the expressiveness of virtual agents: autonomous generation of speech and gesture for spatial description tasks". In: *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 361–368.

— (2012). "Gestural alignment in natural dialogue". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34.

Bergmann, Kirsten, Hannes Rieser, and Stefan Kopp (2011). "Regulating Dialogue with Gestures-Towards an Empirically Grounded Simulation with Conversational Agents". In: *Proceedings of the SIGDIAL 2011 Conference*, pp. 88–97.

Bickmore, Timothy W, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow (2010). "Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials". In: *Journal of health communication* 15.S2, pp. 197–210.

Börstell, Carl and Ryan Lepic (2020). "Spatial metaphors in antonym pairs across sign languages". In: *Sign Language & Linguistics* 23.1-2, pp. 112–141.

Bose, Ritwik, An, Hannah and Valpey, Benjamin (Feb. 19, 2021). *PyTrips*. Version 0.5.22. URL: `https://github.com/mrmechko/pytrips`.

Breazeal, Cynthia, Nick DePalma, Jeff Orkin, Sonia Chernova, and Malte Jung (2013). "Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment". In: *Journal of Human-Robot Interaction* 2.1, pp. 82–111.

Bremner, Paul, Anthony G Pipe, Mike Fraser, Sriram Subramanian, and Chris Melhuish (2009). "Beat gesture generation rules for human-robot interaction". In: *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, pp. 1029–1034.

Bryman, Alan (2017). "Quantitative and qualitative research: further reflections on their integration". In: *Mixing methods: Qualitative and quantitative research*. Routledge, pp. 57–78.

Calbris, G (1995). "Anticipation du geste sur la parole". In: *Dins Verbal/Non Verbal, Frères juneaux de la parole. Actes de la journée d'études ANEFLE, Besançon: Université de Franche-Comte*, pp. 12–18.

Calbris, Geneviève (1990). *The semiotics of French gestures*. Vol. 1900. Indiana University Press.

— (2003). "From cutting an object to a clear cut analysis: Gesture as the representation of a preconceptual schema linking concrete actions to abstract notions". In: *Gesture* 3.1, pp. 19–46.

— (2011). *Elements of meaning in gesture*. Vol. 5. John Benjamins Publishing.

Calbris, Geneviève Zao in, Jacques Montredon, and Paul Woolfehden Zaü (1986). *Des gestes et des mots pour le dire, p. 145*. Clé international Paris.

Camurri, Antonio, Ingrid Lagerlöf, and Gualtiero Volpe (2003). "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques". In: *International journal of human-computer studies* 59.1-2, pp. 213–225.

Cao, Z., G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh (2019). "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cassell, J, H. Vilhjalmsso, and T. Bickmore (2001). "BEAT: The behavior expression animation toolkit". In: *Proceedings of ACM SIGGRAPH*, 477–486.

Cassell, Justine, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone (1994). "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents". In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 413–420.

Cassell, Justine, Hannes Högni Vilhjálmsson, and Timothy Bickmore (2004). "Beat: the behavior expression animation toolkit". In: *Life-Like Characters*. Springer, pp. 163–185.

Castellano, Ginevra, Santiago D Villalba, and Antonio Camurri (2007). "Recognising human emotions from body movement and gesture dynamics". In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 71–82.

Cavicchio, Federica and Sotaro Kita (2013). "Bilinguals switch gesture production parameters when they switch languages". In: *Proceedings Tilburg Gesture Research Meeting (TIGeR) 2013. Retrieved from http://www. researchgate. net/publication/236899431_Bilinguals_Switch_ Gesture_Production eters_when_they_Switch_Languages*.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. (2018). "Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175*.

Chafai, Nicolas Ech, Catherine Pelachaud, and Danielle Pelé (2007). "A case study of gesture expressivity breaks". In: *Language resources and evaluation* 41.3-4, pp. 341–365.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci (2014). "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers". In: *Behavior research methods* 46.1, pp. 112–130.

Charniak, Eugene (2000). "A maximum-entropy-inspired parser". In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Chi, Diane, Monica Costa, Liwei Zhao, and Norman Badler (2000). "The EMOTE model for effort and shape". In: *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 173–182.

Chiu, Chung-Cheng and Stacy Marsella (2011). "How to train your avatar: A data driven approach to gesture generation". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 127–140.

— (2014). "Gesture generation with low-dimensional embeddings". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 781–788.

Chiu, Chung-Cheng, Louis-Philippe Morency, and Stacy Marsella (2015). "Predicting co-verbal gestures: a deep and temporal modeling approach". In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 152–166.

Chollet, Mathieu, Torsten Wörtwein, Louis-Philippe Morency, Ari Shapiro, and Stefan Scherer (2015). "Exploring feedback strategies to improve public speaking: an interactive virtual audience framework". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1143–1154.

Chu, Mingyuan, Antje Meyer, Lucy Foulkes, and Sotaro Kita (2014). "Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy." In: *Journal of Experimental Psychology: General* 143.2, p. 694.

Chui, Kawai (2018). "Spatial conceptualization of sequence time in language and gesture". In: *Gesture* 17.1, pp. 176–195.

— (2022). *Language and Gesture in Chinese Conversation: Bǐshǒu-shuōhuà*. Routledge.

Church, Kenneth Ward (2017). "Word2Vec". In: *Natural Language Engineering* 23.1, pp. 155–162.

Cienki, Alan (2005). "Image schemas and gesture". In: *From perception to meaning: Image schemas in cognitive linguistics* 29, pp. 421–442.

Cienki, Alan J and Jean-Pierre Koenig (1998). "Metaphoric gestures and some of their relations to verbal metaphoric expressions". In: *Discourse and cognition: Bridging the gap*, pp. 189–204.

Cook, Susan Wagner and Susan Goldin-Meadow (2006). "The role of gesture in learning: Do children use their hands to change their minds?" In: *Journal of cognition and development* 7.2, pp. 211–232.

Cooperrider, Kensy (2014). "Body-directed gestures: Pointing to the self and beyond". In: *Journal of Pragmatics* 71, pp. 1–16.

Corera, Sheran and Naomi Krishnarajah (2011). "Capturing hand gesture movement: a survey on tools, techniques and logical considerations". In: *Proceedings of chi sparks*.

Crowder, Elaine M (1996). "Gestures at work in sense-making science talk". In: *The journal of the learning sciences* 5.3, pp. 173–208.

De Melo, Celso M, Liang Zheng, and Jonathan Gratch (2009). "Expression of moral emotions in cooperating agents". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 301–307.

DeVault, David, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. (2014). "SimSensei Kiosk: A virtual human interviewer for healthcare decision support". In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1061–1068.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

DiMaggio, Paul (1997). "Culture and cognition". In: *Annual review of sociology* 23.1, pp. 263–287.

Downs, Julie S, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor (2010). "Are your participants gaming the system? Screening Mechanical Turk workers". In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2399–2402.

Dupont, Marc and Pierre-François Marteau (2015). "Coarse-DTW: Exploiting Sparsity in Gesture Time Series." In: *AALTD@ PKDD/ECML*.

Efron, David (1941). "1972". In: *Gestures, race and culture*.

Ekman, Paul and Wallace V Friesen (1969a). "Nonverbal leakage and clues to deception". In: *Psychiatry* 32.1, pp. 88–106.

— (1969b). "The repertoire of nonverbal behavior: Categories, origins, usage, and coding". In: *Nonverbal communication, interaction, and gesture*, pp. 57–106.

Ennis, Cathy, Rachel McDonnell, and Carol O'Sullivan (2010). "Seeing is believing: body motion dominates in multisensory conversations". In: *ACM Transactions on Graphics (TOG)* 29.4, pp. 1–9.

Eyben, Florian, Martin Wöllmer, and Björn Schuller (2009). "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit". In: *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, pp. 1–6.

Feng, Andrew, Yazhou Huang, Marcelo Kallmann, and Ari Shapiro (2012). "An analysis of motion blending techniques". In: *International Conference on Motion in Games*. Springer, pp. 232–243.

Feng, Dan, Pedro Sequeira, Elin Carstensdottir, Magy Seif El-Nasr, and Stacy Marsella (2018). "Learning Generative Models of Social Interactions with Humans-in-the-Loop". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 509–516.

Ferstl, Ylva and Rachel McDonnell (2018). "Investigating the use of recurrent motion modelling for speech gesture generation". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 93–98.

Ferstl, Ylva, Michael Neff, and Rachel McDonnell (2019). "Multi-objective adversarial gesture generation". In: *Motion, Interaction and Games*, pp. 1–10.

— (2020). "Adversarial gesture generation with realistic gesture phasing". In: *Computers & Graphics*.

Ferstl, Ylva, Michael Neff, and Rachel McDonnell (2021a). "ExpressGesture: Expressive gesture generation from speech through database matching". In: *Computer Animation and Virtual Worlds* 32.3-4, e2016.

— (2021b). "It's A Match! Gesture Generation Using Expressive Parameter Matching". In: *arXiv preprint arXiv:2103.03130*.

Fiske, Amelia, Peter Henningsen, Alena Buyx, et al. (2019). "Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy". In: *Journal of medical Internet research* 21.5, e13216.

Gallaher, Peggy E (1992). "Individual differences in nonverbal behavior: Dimensions of style." In: *Journal of personality and social psychology* 63.1, p. 133.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima (2012). "Why we (usually) don't have to worry about multiple comparisons". In: *Journal of research on educational effectiveness* 5.2, pp. 189–211.

Gibbs Jr, Raymond W (2008). *The Cambridge handbook of metaphor and thought*. Cambridge University Press.

Ginosar, Shiry, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik (2019). "Learning individual styles of conversational gesture". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506.

Glowinski, Donald, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer (2011). "Toward a minimal representation of affective gestures". In: *IEEE Transactions on Affective Computing* 2.2, pp. 106–118.

Goldin-Meadow, Susan and Martha Wagner Alibali (2013). "Gesture's role in speaking, learning, and creating language". In: *Annual review of psychology* 64, pp. 257–283.

Goldin-Meadow, Susan, Howard Nusbaum, Spencer D Kelly, and Susan Wagner (2001). "Explaining math: Gesturing lightens the load". In: *Psychological science* 12.6, pp. 516–522.

Grady, Joseph (1997). "Foundations of meaning: Primary metaphors and primary scenes". In:

Guadagno, Rosanna E, Jim Blascovich, Jeremy N Bailenson, and Cade McCall (2007). "Virtual humans and persuasion: The effects of agency and behavioral realism". In: *Media Psychology* 10.1, pp. 1–22.

Gunes, Hatice and Massimo Piccardi (2006). "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 1. IEEE, pp. 1148–1153.

Gurney, Daniel J, Karen J Pine, and Richard Wiseman (2013). "The gestural misinformation effect: skewing eyewitness testimony through gesture". In: *The American journal of psychology* 126.3, pp. 301–314.

Habibie, Ikhsanul, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt (2022). "A Motion Matching-based Framework for Con-

trollable Gesture Synthesis from Speech". In: *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–9.

Hadar, Uri (1989). "Two types of gesture and their role in speech production". In: *Journal of Language and Social Psychology* 8.3-4, pp. 221–228.

Hall, Jon (2004). "Cicero and Quintilian on the oratorical use of hand gestures". In: *The Classical Quarterly* 54.1, pp. 143–160.

Hartmann, Björn, Maurizio Mancini, and Catherine Pelachaud (2005). "Implementing expressive gesture synthesis for embodied conversational agents". In: *International Gesture Workshop*. Springer, pp. 188–199.

Hasegawa, Dai, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi (2018). "Evaluation of speech-to-gesture generation using bi-directional LSTM network". In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 79–86.

He, Jia, Fons JR Van de Vijver, Velichko H Fetvadjiev, Alejandra de Carmen Dominguez Espinosa, Byron Adams, Itziar Alonso-Arbiol, Arzu Aydinli-Karakulak, Carmen Buzea, Radosveta Dimitrova, Alvaro Fortin, et al. (2017). "On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures". In: *European Journal of Personality* 31.6, pp. 642–657.

Heloir, Alexis, Michael Neff, and Michael Kipp (2010). "Exploiting motion capture for virtual human animation". In: *Proceedings of the workshop on multimodal corpora: Advances in capturing, coding and analyzing multimodality*. Citeseer, pp. 59–62.

Henter, Gustav Eje, Simon Alexanderson, and Jonas Beskow (2020). "Moglow: Probabilistic and controllable motion synthesis using normalising flows". In: *ACM Transactions on Graphics (TOG)* 39.6, pp. 1–14.

Hirth, J, K Berns, and K Mianowski (2012). "Designing arms and hands for the humanoid robot Roman". In: *Advanced Materials Research*. Vol. 463. Trans Tech Publ, pp. 1233–1237.

Hoffmann, Laura, Nicole C Krämer, Anh Lam-Chi, and Stefan Kopp (2009). "Media equation revisited: do users show polite reactions towards an embodied agent?" In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 159–165.

Holle, Henning, Christian Obermeier, Maren Schmidt-Kassow, Angela D Friederici, Jamie Ward, and Thomas C Gunter (2012). "Gesture facilitates the syntactic analysis of speech". In: *Frontiers in psychology* 3, p. 74.

Holler, Judith and Geoffrey Beattie (2003). "Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener?" In: *Gesture* 3.2, pp. 127–154.

Honnibal, Matthew and Mark Johnson (2015). "An improved non-monotonic transition system for dependency parsing". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1373–1378.

Honnibal, Matthew and Ines Montani (2017a). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear.

Honnibal, Matthew and Ines Montani (2017b). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7.1, pp. 411–420.

Hutto, Clayton J and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Eighth international AAAI conference on weblogs and social media*.

Hwang, Bon-Woo, Sungmin Kim, and Seong-Whan Lee (2006). "A full-body gesture database for automatic gesture recognition". In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, pp. 243–248.

Iverson, Jana M and Susan Goldin-Meadow (1997). "What's communication got to do with it? Gesture in children blind from birth." In: *Developmental psychology* 33.3, p. 453.

— (1998). "Why people gesture when they speak". In: *Nature* 396.6708, pp. 228–228.

— (2001). "The resilience of gesture in talk: Gesture in blind speakers and listeners". In: *Developmental Science* 4.4, pp. 416–422.

Jacobs, Naomi and Alan Garnham (2007). "The role of conversational hand gestures in a narrative task". In: *Journal of Memory and Language* 56.2, pp. 291–303.

Jamalian, Azadeh and Barbara Tversky (2012a). "Gestures alter thinking about time". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34.

— (2012b). "Gestures alter thinking about time". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34, pp. 503–508.

Jokinen, Kristiina, Costanza Navarretta, and Patrizia Paggio (2008). "Distinguishing the communicative functions of gestures". In: *International Workshop on Machine Learning for Multimodal Interaction*. Springer, pp. 38–49.

Joo, Hanbyul, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. (2017). "Panoptic studio: A massively multiview system for social interaction capture". In: *IEEE transactions on pattern analysis and machine intelligence* 41.1, pp. 190–204.

Joty, Shafiq, Giuseppe Carenini, and Raymond T Ng (2015). "Codra: A novel discriminative framework for rhetorical analysis". In: *Computational Linguistics* 41.3, pp. 385–435.

Jung, Malte and Pamela Hinds (2018). *Robots in the wild: A time for more robust theories of human-robot interaction*.

Kang, Seokmin, Gregory L Hallman, Lisa K Son, and John B Black (2013). "The different benefits from different gestures in understanding a concept". In: *Journal of Science Education and Technology* 22.6, pp. 825–837.

Kay, Paul and Willett Kempton (1984). "What is the Sapir-Whorf hypothesis?" In: *American anthropologist* 86.1, pp. 65–79.

Kelly, Megyn and Alyson Shontell at Business Insider (2017). *Megyn Kelly Talks Matt Lauer, Fox News, Donald Trump, Roger Ailes*. Business Insider IGNITION conference. URL: https://www.youtube.com/watch?v=TpsEQckXwYI.

Kelly, Spencer D, Dale J Barr, R Breckinridge Church, and Katheryn Lynch (1999). "Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory". In: *Journal of memory and Language* 40.4, pp. 577–592.

Kelly, Spencer D, Aslı Özyürek, and Eric Maris (2010). "Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension". In: *Psychological science* 21.2, pp. 260–267.

Kendon, Adam (1972). "Some relationships between body motion and speech". In: *Studies in dyadic communication* 7.177, p. 90.

— (1995). "Gestures as illocutionary and discourse structure markers in Southern Italian conversation". In: *Journal of pragmatics* 23.3, pp. 247–279.

— (1997). "Gesture". In: *Annual review of anthropology* 26.1, pp. 109–128.

— (2000). "Language and gesture: Unity or duality". In: *Language and gesture* 2, pp. 47–63.

— (2004). *Gesture: Visible action as utterance*. Cambridge University Press, pp. 108–126.

Khooshabeh, Peter, Cade McCall, Sudeep Gandhe, Jonathan Gratch, and James Blascovich (2011). "Does it matter if a computer jokes". In: *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 77–86.

Kipp, Michael (2014). "Anvil: A universal video research tool". In: *Handbook of corpus phonology*, pp. 420–436.

Kipp, Michael and Jean-Claude Martin (2009). "Gesture and emotion: Can basic gestural form features discriminate emotions?" In: *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, pp. 1–8.

Kipp, Michael, Michael Neff, and Irene Albrecht (2007). "An annotation scheme for conversational gestures: how to economically capture timing and form". In: *Language Resources and Evaluation* 41.3-4, pp. 325–339.

Kita, Sotaro (2009). "Cross-cultural variation of speech-accompanying gesture: A review". In: *Language and cognitive processes* 24.2, pp. 145–167.

Kita, Sotaro and Asli Özyürek (2003). "What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking". In: *Journal of Memory and language* 48.1, pp. 16–32.

Kock, Ned (2005). "Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools". In: *IEEE transactions on professional communication* 48.2, pp. 117–130.

Kopp, Stefan, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson (2006). "Towards a common framework

for multimodal generation: The behavior markup language". In: *International workshop on intelligent virtual agents*. Springer, pp. 205–217.

Kopp, Stefan, Herwin van Welbergen, Ramin Yaghoubzadeh, and Hendrik Buschmeier (2014). "An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing". In: *Journal on Multimodal User Interfaces* 8.1, pp. 97–108.

Krahmer, Emiel and Marc Swerts (2007). "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception". In: *Journal of Memory and Language* 57.3, pp. 396–414.

Krämer, Nicole, Stefan Kopp, Christian Becker-Asano, and Nicole Sommer (2013). "Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior". In: *International Journal of Human-Computer Studies* 71.3, pp. 335–349.

Krämer, Nicole and Arne Manzeschke (2021). "Social reactions to socially interactive agents and their ethical implications". In: *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pp. 77–104.

Krämer, Nicole C, Laura Hoffmann, and Stefan Kopp (2010). "Know Your Users! Empirical Results for Tailoring an Agent s Nonverbal Behavior to Different User Groups". In: *International conference on intelligent virtual agents*. Springer, pp. 468–474.

Kron, Frederick W, Michael D Fetters, Mark W Scerbo, Casey B White, Monica L Lypson, Miguel A Padilla, Gayle A Gliva-McConvey, Lee A Belfore II, Temple West, Amelia M Wallace, et al. (2017). "Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial". In: *Patient education and counseling* 100.4, pp. 748–759.

Kucherenko, Taras, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström (2020). "Gesticulator: A framework for semantically-aware speech-driven gesture generation". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 242–250.

Kucherenko, Taras, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter (2021a). "A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020". In: *26th International Conference on Intelligent User Interfaces*. IUI '21. College Station, TX, USA: Association for Computing Machinery, pp. 11–21. ISBN: 9781450380171. DOI: 10.1145/3397481.3450692. URL: https://doi.org/10.1145/3397481.3450692.

— (2021b). "A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020". In: *26th international conference on intelligent user interfaces*. Accepted for publication., pp. 11–21.

Kucherenko, Taras, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter (2021). "Speech2Properties2Gestures: Gesture-Property Prediction as a Tool for Gener-

ating Representational Gestures from Speech". In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pp. 145–147.

Kulms, Philipp and Stefan Kopp (2016). "The effect of embodiment and competence on trust and cooperation in human–agent interaction". In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 75–84.

Kwon, Minae, Malte F Jung, and Ross A Knepper (2016). "Human expectations of social robots". In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 463–464.

Lakoff, George and Mark Johnson (1980). "The metaphorical structure of the human conceptual system". In: *Cognitive science* 4.2, pp. 195–208.

— (2008). *Metaphors we live by*. University of Chicago press.

Lascarides, Alex and Matthew Stone (2009). "A formal semantic analysis of gesture". In: *Journal of Semantics* 26.4, pp. 393–449.

Le Guen, Olivier and Lorena Ildefonsa Pool Balam (2012). "No metaphorical timeline in gesture and cognition among Yucatec Mayas". In: *Frontiers in Psychology* 3, p. 271.

Lee, Dong Yul, Max R Uhlemann, and Richard F Haase (1985). "Counselor verbal and nonverbal responses and perceived expertness, trustworthiness, and attractiveness." In: *Journal of Counseling Psychology* 32.2, p. 181.

Lee, Gilwoo, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh (2019). "Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 763–772.

Lee, Jerry, Patricia Jones, Yoshimitsu Mineyama, and Esther Zhang (Aug. 2002). "Cultural Differences in Responses to a Likert Scale". In: *Research in nursing health* 25, pp. 295–306. DOI: 10.1002/nur.10041.

Lee, Jina and Stacy Marsella (2006). "Nonverbal behavior generator for embodied conversational agents". In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 243–255.

Leonard, Thomas and Fred Cummins (2011). "The temporal relation between beat gestures and speech". In: *Language and Cognitive Processes* 26.10, pp. 1457–1471.

Leong, Chee Wee, Beata Beigman Klebanov, and Ekaterina Shutova (2018). "A Report on the 2018 VUA Metaphor Detection Shared Task". In: *Workshop on Figurative Language Processing*, pp. 56–66.

Levinson, Stephen C (1996). "Language and space". In: *Annual review of Anthropology* 25.1, pp. 353–382.

Levy, Elena T and David McNeill (1992). "Speech, gesture, and discourse". In: *Discourse processes* 15.3, pp. 277–301.

Lhommet, Margaux and Stacy Marsella (Sept. 2014a). "Metaphoric gestures: towards grounded mental spaces". In: *International Conference on Intelligent Virtual Agents*. URL: http://www.ccs.neu.edu/~marsella/publications/pdf/Lhommet_IVA2014.pdf.

— (Sept. 2016). "From embodied metaphors to metaphoric gestures". In: *Proceedings of the Cognitive Science Society*. URL: http://www.ccs.neu.edu/~marsella/publications/pdf/16-Cogsci-Lhommet.pdf.

Lhommet, Margaux and Stacy C Marsella (2013). "Gesture with meaning". In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 303–312.

— (2014b). "19 Expressing Emotion Through Posture and Gesture". In: *The Oxford handbook of affective computing*, p. 273.

Lhommet, Margot, Yuyu Xu, and Stacy Marsella (2015). "Cerebella: automatic generation of nonverbal behavior for virtual humans". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 4303–4304.

Li, Lianwei, Shiyin Qin, Zhi Lu, Kuanhong Xu, and Zhongying Hu (2020). "One-shot learning gesture recognition based on joint training of 3D ResNet and memory module". In: *Multimedia Tools and Applications* 79.9, pp. 6727–6757.

Liang, Yuanzhi, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang (2022). "SEEG: Semantic Energized Co-Speech Gesture Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10473–10482.

Lickiss, Karen P and A Rodney Wellens (1978). "Effects of visual accessibility and hand restraint on fluency of gesticulator and effectiveness of message". In: *Perceptual and Motor Skills* 46.3, pp. 925–926.

Lloyd, Kenneth E (1980). "Do as I say, not as I do." In: *New Zealand Psychologist*.

Luo, Pengcheng, Michael Kipp, and Michael Neff (2009). "Augmenting gesture animation with motion capture data to provide full-body engagement". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 405–417.

Luo, Pengcheng, Victor Ng-Thow-Hing, and Michael Neff (2013). "An examination of whether people prefer agents whose gestures mimic their own". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 229–238.

Luxton, David D and Eva Hudlicka (2021). "Intelligent Virtual Agents in Behavioral and Mental Healthcare: Ethics and Application Considerations". In: *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*. Springer, pp. 41–55.

Maatman, RM, Jonathan Gratch, and Stacy Marsella (2005). "Natural behavior of a listening agent". In: *International workshop on intelligent virtual agents*. Springer, pp. 25–36.

Maestripieri, Dario (2005). "Gestural communication in three species of macaques (Macaca mulatta, M. nemestrina, M. arctoides): use of signals in relation to dominance and social context". In: *Gesture* 5.1-2, pp. 57–73.

Mann, William C and Sandra A Thompson (1987). *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marchena, Ashley B de and Inge-Marie Eigsti (2014). "Context counts: The impact of social context on gesture rate in verbally fluent adolescents with autism spectrum disorder". In: *Gesture* 14.3, pp. 375–393.

Marcu, Daniel (1997). "The rhetorical parsing of unrestricted natural language texts". In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 96–103.

Maricchiolo, Fridanna, Augusto Gnisci, Marino Bonaiuto, and Gianluca Ficca (Feb. 2009a). "Effects of different types of hand gestures in persuasive speech on receivers evaluations". In: *LANGUAGE AND COGNITIVE PROCESSES* 24, pp. 239–266. DOI: 10.1080/01690960802159929.

— (2009b). "Effects of different types of hand gestures in persuasive speech on receivers' evaluations". In: *Language and cognitive processes* 24.2, pp. 239–266.

Marsella, Stacy, Jonathan Gratch, Paolo Petta, et al. (2010). "Computational models of emotion". In: *A Blueprint for Affective Computing-A sourcebook and manual* 11.1, pp. 21–46.

Marsella, Stacy, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro (2013). "Virtual Character Performance from Speech". In: *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. SCA '13. Anaheim, California: ACM, pp. 25–35. ISBN: 978-1-4503-2132-7. DOI: 10.1145/2485895.2485900. URL: http://doi.acm.org/10.1145/2485895.2485900.

Marsella, Stacy C, Sharon Marie Carnicke, Jonathan Gratch, Anna Okhmatovskaia, and Albert Rizzo (2006). "An exploration of delsarte's structural acting system". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 80–92.

Mayer, Richard E and C Scott DaPra (2012). "An embodiment effect in computer-based learning with animated pedagogical agents." In: *Journal of Experimental Psychology: Applied* 18.3, p. 239.

McCafferty, Steven G (2004). "Space for cognition: Gesture and second language learning". In: *International Journal of Applied Linguistics* 14.1, pp. 148–165.

McCall, Cade, Debra P Bunyan, Jeremy N Bailenson, Jim Blascovich, and Andrew C Beall (2009). "Leveraging collaborative virtual environment technology for inter-population research on persuasion in a classroom setting". In: *PRESENCE: Teleoperators and Virtual Environments* 18.5, pp. 361–369.

McClosky, David, Eugene Charniak, and Mark Johnson (2006). "Reranking and self-training for parser adaptation". In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 337–344.

McNeill, David (1985). "So you think gestures are nonverbal?" In: *Psychological review* 92.3, p. 350.

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

McNeill, David (2006). “Gesture: a psycholinguistic approach”. In: *The encyclopedia of language and linguistics*, pp. 58–66.

McNeill, David, Justine Cassell, and Elena T Levy (1993). “Abstract deixis”. In: *Semiotica* 95.1-2, pp. 5–20.

McNeill, David and Susan Duncan (2000). “Growth points in thinking-for-speaking”. In: *Language and gesture* 1987, pp. 141–161.

McNeill, David and Elena Levy (1980). “Conceptual Representations in Language Activity and Gesture.” In:

Metallinou, Angeliki, Zhaojun Yang, Chi-chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan (2016). “The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations”. In: *Language resources and evaluation* 50.3, pp. 497–521.

Miller, George A (1995). “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11, pp. 39–41.

Miller, George A, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller (1990). “Introduction to WordNet: An on-line lexical database”. In: *International journal of lexicography* 3.4, pp. 235–244.

Morris, Desmond (2015). *Bodytalk: A world guide to gestures*. Random House.

Morris, R, Daniel McDuff, and R Calvo (2014). “Crowdsourcing techniques for affective computing”. In: *The Oxford handbook of affective computing*. Oxford Univ. Press, pp. 384–394.

Mubin, Omar and Christoph Bartneck (2015). “Do as I say: Exploring human response to a predictable and unpredictable robot”. In: *Proceedings of the 2015 British HCI Conference*, pp. 110–116.

Mumm, Jonathan and Bilge Mutlu (2011). “Human-robot proxemics: physical and psychological distancing in human-robot interaction”. In: *Proceedings of the 6th international conference on Human-robot interaction*, pp. 331–338.

Murphy, Keith M (2003). “Building meaning in interaction: rethinking gesture classifications”. In: *Crossroads of Language, Interaction, and Culture* 5, pp. 29–47.

Nam, Tek-Jin, Jong-Hoon Lee, Sunyoung Park, and Hyeon-Jeong Suk (2014). “Understanding the relation between emotion and physical movements”. In: *International Journal of Affective Engineering* 13.3, pp. 217–226.

Neff, Michael and Eugene Fiume (2008). “From performance theory to character animation tools”. In: *Human motion*. Springer, pp. 597–629.

Neff, Michael, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel (2008). “Gesture modeling and animation based on a probabilistic re-creation of speaker style”. In: *ACM Transactions on Graphics (TOG)* 27.1, pp. 1–24.

Neff, Michael, Yingying Wang, Rob Abbott, and Marilyn Walker (2010). “Evaluating the effect of gesture and language on personality perception in conversational agents”. In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 222–235.

Niewiadomski, Radoslaw, Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud (Jan. 2009). "Greta: An interactive expressive ECA system". In: vol. 2, pp. 1399–1400. DOI: 10.1145/1558109.1558314.

Nishio, Shuichi, Kohei Ogawa, Yasuhiro Kanakogi, Shoji Itakura, and Hiroshi Ishiguro (2018). "Do robot appearance and speech affect people's attitude? Evaluation through the ultimatum game". In: *Geminoid Studies*. Springer, pp. 263–277.

Nobe, Shuichi (2000). "Where do most spontaneous representational gestures actually occur with respect to speech". In: *Language and gesture* 2, p. 186.

Noroozi, Fatemeh, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari (2018). "Survey on emotional body gesture recognition". In: *IEEE transactions on affective computing* 12.2, pp. 505–523.

Novack, Miriam A and Susan Goldin-Meadow (2017). "Gesture as representational action: A paper about function". In: *Psychonomic Bulletin & Review* 24.3, pp. 652–665.

Núñez, Rafael E and Eve Sweetser (2006). "With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time". In: *Cognitive science* 30.3, pp. 401–450.

Ochs, Magalie, Grégoire de Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, C Pelachaud, D Mestre, and Philippe Blache (2017). "An architecture of virtual patient simulation platform to train doctors to break bad news". In:

Oepen, Stephan, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, and Zdenka Uresova (2015). "Semeval 2015 task 18: Broad-coverage semantic dependency parsing". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 915–926.

Olugbade, Temitayo, Marta Bieńkiewicz, Giulia Barbareschi, Vincenzo D'Amato, Luca Oneto, Antonio Camurri, Catherine Holloway, Mårten Björkman, Peter Keller, Martin Clayton, et al. (2022). "Human Movement Datasets: An Interdisciplinary Scoping Review". In: *ACM Computing Surveys (CSUR)*.

Osgood, Charles Egerton (1971). *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*. CUP Archive.

Özçalışkan, Şeyda and Susan Goldin-Meadow (2005). "Gesture is at the cutting edge of early language development". In: *Cognition* 96.3, B101–B113.

Özyürek, Asli (2002). "Do speakers design their cospeech gestures for their addressees? The effects of addressee location on representational gestures". In: *Journal of Memory and Language* 46.4, pp. 688–704.

Pan, Xueni and Antonia F de C Hamilton (2018). "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape". In: *British Journal of Psychology* 109.3, pp. 395–417.

Pawlikowska, Teresa, Wenjuan Zhang, Frances Griffiths, Jan Van Dalen, and Cees van der Vleuten (2012). "Verbal and non-verbal behavior of doctors and patients in primary care consultations–How this relates to patient enablement". In: *Patient education and counseling* 86.1, pp. 70–76.

PBS News Hour, Barack Obama with. *Barack Obama with PBS News Hour*. PBS News Hour, Barack Obama's full speech at the 2020 Democratic National Convention. URL: https://www.youtube.com/watch?v=oaalF5y2P0k.

Pedersen, Ted, Siddharth Patwardhan, Jason Michelizzi, et al. (2004). "WordNet:: Similarity-Measuring the Relatedness of Concepts." In: *AAAI*. Vol. 4, pp. 25–29.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001). "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001, p. 2001.

Perez-Osorio, Jairo, Eva Wiese, and Agnieszka Wykowska (2021). "Theory of Mind and Joint Attention-The Handbook on Socially Interactive Agents". In:

Perre, Greet Van de, Michaël Van Damme, Dirk Lefeber, and Bram Vanderborght (2015). "Development of a generic method to generate upper-body emotional expressions for different social robots". In: *Advanced Robotics* 29.9, pp. 597–609.

Poggi, Isabella and Catherine Pelachaud (2008). "Persuasion and the expressivity of gestures in humans and machines". In: *Embodied communication in humans and machines*, pp. 391–424.

Poggi, Isabella, Catherine Pelachaud, Fiorella de Rosis, Valeria Carofiglio, and Berardina De Carolis (2005). "Greta. a believable embodied conversational agent". In: *Multimodal intelligent information presentation*. Springer, pp. 3–25.

Poggi, Isabella and Laura Vincze (2008). "Gesture, gaze and persuasive strategies in political discourse". In: *International LREC Workshop on Multimodal Corpora*. Springer, pp. 73–92.

Pollick, Frank E (2003). "The features people use to recognize human movement style". In: *International gesture workshop*. Springer, pp. 10–19.

Pollick, Frank E, Helena M Paterson, Armin Bruderlin, and Anthony J Sanford (2001). "Perceiving affect from arm movement". In: *Cognition* 82.2, B51–B61.

Pouw, Wim, Jan de Wit, Sara Bögels, Marlou Rasenberg, Branka Milivojevic, and Asli Ozyurek (2021). "Semantically related gestures move alike: Towards a distributional semantics of gesture kinematics". In: *Proceedings of the 23rd International Conference on Human-Computer Interaction*.

Pütten, Astrid M Rosenthal-von der, Carolin Straßmann, Ramin Yaghoubzadeh, Stefan Kopp, and Nicole C Krämer (2019). "Dominant and submissive nonverbal behavior of virtual agents and its effects on evaluation and negotiation outcome in different age groups". In: *Computers in Human Behavior* 90, pp. 397–409.

Radden, Günter (2003). "The metaphor TIME AS SPACE across languages". In: *Zeitschrift für interkulturellen Fremdsprachenunterricht* 8.2.

Rauscher, Frances H, Robert M Krauss, and Yihsiu Chen (1996). "Gesture, speech, and lexical access: The role of lexical movements in speech production". In: *Psychological science* 7.4, pp. 226–231.

Ravenet, Brian, Chloé Clavel, and Catherine Pelachaud (2018). "Automatic nonverbal behavior generation from image schemas". In: *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, pp. 1667–1674.

Ravenet, Brian, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella (2018). "Automating the production of communicative gestures in embodied characters". In: *Frontiers in psychology* 9, p. 1144.

Reddy, Michael (1979). "The conduit metaphor". In: *Metaphor and thought* 2, pp. 285–324.

Rei, Marek, Luana Bulat, Douwe Kiela, and Ekaterina Shutova (2017). "Grasping the finer point: A supervised similarity network for metaphor detection". In: *arXiv preprint arXiv:1709.00575*.

Reidsma, Dennis, Iwan de Kok, Daniel Neiberg, Sathish Chandra Pammi, Bart van Straalen, Khiet Truong, and Herwin van Welbergen (2011). "Continuous interaction with a virtual human". In: *Journal on Multimodal User Interfaces* 4.2, pp. 97–118.

Ren, Liu, Alton Patrick, Alexei A Efros, Jessica K Hodgins, and James M Rehg (2005). "A data-driven approach to quantifying natural human motion". In: *ACM Transactions on Graphics (TOG)* 24.3, pp. 1090–1097.

Riemer, Marc J and Detlev E Jansen (2003). "Non-verbal intercultural communication awareness for the modern engineer". In: *World Transactions on Engineering and Technology Education* 2.3, pp. 373–378.

Riggio, Ronald E and Barbara Throckmorton (1988). "The relative effects of verbal and nonverbal behavior, appearance, and social skills on evaluations made in hiring interviews 1". In: *Journal of Applied Social Psychology* 18.4, pp. 331–348.

Rios-Soria, David J, Satu E Schaeffer, and Sara E Garza-Villarreal (2013). "Hand-gesture recognition using computer-vision techniques". In:

Rothman, Kenneth J (1990). "No adjustments are needed for multiple comparisons". In: *Epidemiology*, pp. 43–46.

Rouse, Steven V (2015). "A reliability analysis of Mechanical Turk data". In: *Computers in Human Behavior* 43, pp. 304–307.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Sabanovic, Selma, Marek P Michalowski, and Reid Simmons (2006). "Robots in the wild: Observing human-robot social interaction outside the lab". In: *9th IEEE International Workshop on Advanced Motion Control, 2006.* IEEE, pp. 596–601.

Santiago, Julio, Juan Lupáñez, Elvira Pérez, and María Jesús Funes (2007). "Time (also) flies from left to right". In: *Psychonomic Bulletin & Review* 14.3, pp. 512–516.

Sariff, N and Norlida Buniyamin (2006). "An overview of autonomous mobile robot path planning algorithms". In: *2006 4th student conference on research and development*. IEEE, pp. 183–188.

Saund, Carolyn, Marion Roth, Mathieu Chollet, and Stacy Marsella (2019). "Multiple metaphors in metaphoric gesturing". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 524–530.

Scassellati, Brian (2002). "Theory of mind for a humanoid robot". In: *Autonomous Robots* 12.1, pp. 13–24.

Schuller, Björn, Anton Batliner, Stefan Steidl, and Dino Seppi (2011). "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". In: *Speech Communication* 53.9-10, pp. 1062–1087.

Schwartz, Barry, Abraham Tesser, and Evan Powell (1982). "Dominance cues in nonverbal behavior". In: *Social Psychology Quarterly*, pp. 114–120.

Sikveland, Rein Ove and Richard Ogden (2012). "Holding gestures across turns: Moments to generate shared understanding". In: *Gesture* 12.2, pp. 166–199.

Slater, Mel, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, et al. (2020). "The ethics of realism in virtual and augmented reality". In: *Frontiers in Virtual Reality* 1, p. 1.

Smith, Rebecca A and Emily S Cross (2022). "The McNorm library: creating and validating a new library of emotionally expressive whole body dance movements". In: *Psychological research*, pp. 1–25.

So, Wing Chee, Colin Sim Chen-Hui, and Julie Low Wei-Shan (2012). "Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall?" In: *Language and Cognitive Processes* 27.5, pp. 665–681.

Stone, Matthew, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler (2004). "Speaking with hands: Creating animated conversational characters from recordings of human performance". In: *ACM Transactions on Graphics (TOG)* 23.3, pp. 506–513.

Swartout, William R, Jonathan Gratch, Randall W Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum (2006). "Toward virtual humans". In: *AI Magazine* 27.2, pp. 96–96.

Takeuchi, Akikazu and Taketo Naito (1995). "Situated facial displays: towards social interaction". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 450–455.

Takeuchi, Kenta, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi (2017). "Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM". In: *Proceedings of the 5th International Conference on Human Agent Interaction*, pp. 365–369.

Talmy, L (1985). "Grammatical categories and the lexicon". In: *Language typology and syntactic description* 3, pp. 57–149.

Tevet, Guy, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or (2022). "MotionCLIP: Exposing Human Motion Generation to CLIP Space". In: *arXiv preprint arXiv:2203.08063*.

Thiebaux, Marcus, Stacy Marsella, Andrew N. Marshall, and Marcelo Kallmann (2008). "SmartBody: Behavior Realization for Embodied Conversational Agents". In: *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*. AAMAS '08. Estoril, Portugal: International Foundation for Autonomous Agents and Multiagent Systems, pp. 151–158. ISBN: 978-0-9817381-0-9. URL: http://dl.acm.org/citation.cfm?id=1402383.1402409.

Tottie, Gunnel (2011). "Uh and um as sociolinguistic markers in British English". In: *International Journal of Corpus Linguistics* 16.2, pp. 173–197.

Tree, Jean E Fox and Josef C Schrock (1999). "Discourse markers in spontaneous speech: Oh what a difference an oh makes". In: *Journal of Memory and Language* 40.2, pp. 280–295.

Turchyn, Sergiy, Inés Olza Moreno, Cristóbal Pagán Cánovas, Francis F Steen, Mark Turner, Javier Valenzuela, and Soumya Ray (2018). "Gesture Annotation with a Visual Search Engine for Multimodal Communication Research". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tversky, Barbara and Bridgette Martin Hard (2009). "Embodied and disembodied cognition: Spatial perspective-taking". In: *Cognition* 110.1, pp. 124–129.

USC Institute for Creative Technologies (May 25, 2020). *SmartBody*. Version 2020. URL: https://smartbody.ict.usc.edu/download2.

UzZaman, Naushad and James Allen (2010). "TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text". In: *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 276–283.

Vasic, Milos and Aude Billard (2013). "Safety issues in human-robot interactions". In: *2013 ieee international conference on robotics and automation*. IEEE, pp. 197–204.

Wachsmuth, Ipke and Stefan Kopp (2001). "Lifelike gesture synthesis and timing for conversational agents". In: *International Gesture Workshop*. Springer, pp. 120–133.

Wannesm, Khendrickx, Aras Yurtman, Pieter Robberechts, Dany Vohl, Eric Ma, Gust Verbruggen, Marco Rossi, Mazhar Shaikh, Muhammad Yasirroni, Todd, Wojciech Zieliński, Toon Van Craenendonck, and Sai Wu (2022). *wannesm/dtaidistance: v2.3.5*. Zenodo. DOI: 10.5281/ZENODO.5901139. URL: https://zenodo.org/record/5901139.

Whiten, Andrew and Richard W Byrne (1988). "The Machiavellian intelligence hypotheses". In:

Wilson, Andrew D, Aaron F Bobick, and Justine Cassell (1996). "Recovering the temporal structure of natural gesture". In: *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*. IEEE, pp. 66–71.

Wilson, Jason R, Nah Young Lee, Annie Saechao, Sharon Hershenson, Matthias Scheutz, and Linda Tickle-Degnen (2017). "Hand gestures and verbal acknowledgments improve human-robot rapport". In: *International Conference on Social Robotics*. Springer, pp. 334–344.

Wilson, Margaret (2002). "Six views of embodied cognition". In: *Psychonomic bulletin & review* 9.4, pp. 625–636.

Wolfert, Pieter, Jeffrey M Girard, Taras Kucherenko, and Tony Belpaeme (2021). "To rate or not to rate: Investigating evaluation methods for generated co-speech gestures". In: *arXiv preprint arXiv:2108.05709*, pp. 494–502.

Wolfert, Pieter, Taras Kucherenko, Hedvig Kjellström, and Tony Belpaeme (2019). "Should beat gestures be learned or designed?: A benchmarking user study". In: *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*. IEEE conference proceedings.

Wolfert, Pieter, Nicole Robinson, and Tony Belpaeme (2022). "A review of evaluation practices of gesture generation in embodied conversational agents". In: *IEEE Transactions on Human-Machine Systems*.

Wolff, Charlotte (2015). *A psychology of gesture*. Routledge.

Xu, Yuyu, Catherine Pelachaud, and Stacy Marsella (2014). "Compound gesture generation: a model based on ideational units". In: *International Conference on Intelligent Virtual Agents*. Springer, pp. 477–491.

Yoon, Youngwoo, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee (2020). "Speech gesture generation from the trimodal context of text, audio, and speaker identity". In: *ACM Transactions on Graphics (TOG)* 39.6, pp. 1–16.

Yoon, Youngwoo, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee (2019). "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 4303–4309.

Yoon, Youngwoo, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter (2022). "The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation". In: *Proceedings of the ACM International Conference on Multimodal Interaction*. ICMI '22. ACM.

Yu, Ning (2012). "The metaphorical orientation of time in Chinese". In: *Journal of Pragmatics* 44.10, pp. 1335–1354.

Zhou, Chi, Tengyue Bian, and Kang Chen (2022). "GestureMaster: Graph-based speech-driven gesture generation". In: *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Challenge*. GENEA, pp. 1–4.