7 Gesture Generation

Carolyn Saund and Stacy Marsella

Gestures accompany our speech in ways that punctuate, augment, substitute for, and even contradict verbal information. Such co-speech gestures draw listeners' attention to specific phrases, indicate the speaker's feelings toward a subject, or even convey "off-the-record" information that is excluded from our spoken words. The study of co-speech gesture stretches at least as far back as the work of Quintilian in 50 AD, 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. In this chapter, we discuss the importance, selection, production, challenges, and future of co-speech gestures for artificial social intelligent agents.

7.1 The Importance of Gesture in Social Interaction

7.1.1 What are Gestures?

Gestures as we discuss them here are the spontaneous movements that accompany speech. Generally, these are limited to hand and arm movements [McNeill 1992] but can occasionally extend to the head, feet, or other body parts [Kendon 2000]. Our focus here, however, is on hand and arm movements.

Specifically, this chapter 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. As we describe below, 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.

Table 7.1 **Table of gesture classical types and co-speech properties**

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

Importantly, the classifications provided here are by no means exhaustive. In this section, in addition to introducing one prevailing taxonomy (Section 7.1.1.1), we discuss weaknesses and alternative proposals to classifying gestures using these dimensions (Sections 7.1.1.2 and 7.1.1.3), as well as many other factors that determine how researchers tend to group gestures, both physically and functionally.

7.1.1.1 Classification Dimensions

A common method of classifying co-speech gestures is by the five types or dimensions described in Table 7.1. 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 the communicative intentions of the speaker.

Emblems are gestures that may essentially be thought of as replacements for spoken language. A prominent example is the "thumbs up" gesture that is common in several cultures, but often with strikingly different meanings. In North American and European cultures, for example, if somebody asks a question, a "thumbs-up" response unambiguously means "yes," with or without verbal affirmation. They carry equivalent meaning to their linguistic counterpart. Importantly, the interpretation of these gestures are culturally and linguistically dependent; the "OK" symbol in Western cultures 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.

Deictic gestures are pointing gestures that direct attention toward 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.

7.1.1.2 Multiple Classifications

As McNeill [2006] has argued, these classifications are not strict types but rather dimensions that 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-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 that 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

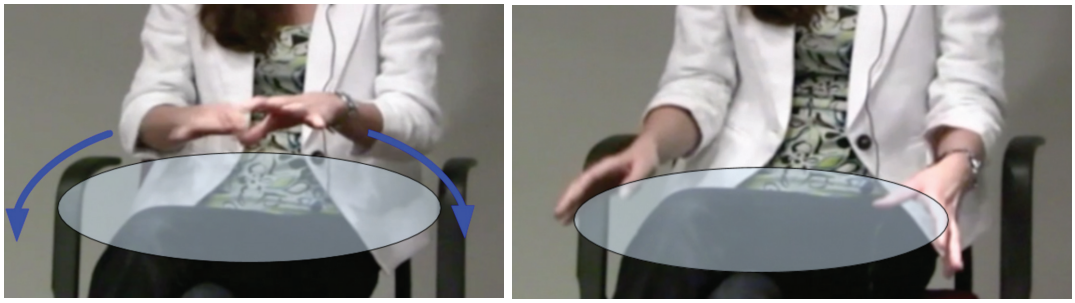

Figure 7.1 The motion of the metaphoric gesture accompanying the phrase "Anything at all." (a) The beginning of the movement as she says "Anything at all." (b) The second part of this gesture, creating the space where "anything" may metaphorically be.

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 7.1(a) 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 the metaphor use in spoken language and accompanying gestures, the catalog of metaphors used in speech provides 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.

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

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.

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 et al. 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 [1997] *categories/sets are bounded spatial regions*), thus incorporating yet another element of the dimensions described above into a single gestural motion.

7.1.1.3 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 analyzing gestures not by abstract representation but instead by the production of those representations themselves. That is, gestures can be analyzed exclusively by their body movements as opposed to attempting to interpret what those movements represent. He argues that movement-based analysis is less prone to researcher bias and less likely to leave out body movements that 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 are 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.

Still more schemes that suggest classifying gestures using both principles of form and function also attempt to address this problem. Saund et al. [2019] discusses the possibility of delineating and classifying gestures according to both conversational context (the function of the gesture) in tandem with the novel physical spaces they occupy (physical form of the gesture). Additionally, because of these overlapping dimensions, the process of describing and classifying the motion of gestures themselves is often decoupled from the meaning the gesture carries [Kipp et al. 2007]. This allows other schemes to break down gesture classification into linguistic and motion sub-problems [Cassell 1998]. It is only by considering the full picture of gesture production, from intention, to function, to physical action, that we can begin to create socially compelling gesture in artificial agents.

7.1.2 Timing

These axes of gesture vary as well 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 [Calbris 1995, Nobe 2000, Gibbs 2008]. However, perception of appropriateness for different gestures with respect to co-speech timing is not fixed [Leonard and Cummins 2011].

The window of time for gestures to be relevant to corresponding speech is similarly fluid, depending 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 that initiate gesture production [Kendon 2000].

While the development of social artificial agents have a ways to go before these artifacts can form rich coherent conversational speech from a communication intention alone, it is important to keep in mind that such a pipeline that truly possesses the spontaneity, creativity, and expressive substance of human gestures must similarly be responsible for producing meaningful co-speech gestures. We discuss current implementations of various gesture generators in relation to speech in Section 7.2.

7.1.2.1 Gestural Phases and Units

At the level of individual gestures, there is a complex feature structure. There are the 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 (a.k.a. gesture units [Kendon 2004]) 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].

One such name to refer to a sequence of related ideas that can span multiple gestures is an ideational unit [Calbris 1990]. Calbris argues that ideational units structure the discourse and the kinesic segmentation of gestures, and serve to impose requirements on gestural features both within and across ideational units in an overall performance.

Within a gesture performance, some features such as hand shape, movement trajectory, or location in space may be coupled across gestures while other features serve at times a key role 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.

7.1.3 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 [Talmy 1985, Kita 2009] are a few examples of components of gesture that 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 that are positive signals in one culture may be rude insults in another [Calbris 2011].

But beyond this, different cultures' concepts of physical space and indeed 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 speaker's 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 also that the underlying conceptual representation of time may differ between cultures as well. A further review may be found in Kendon [1997]. For an overview of the implementation of culture in SIAs, please refer to Chapter 13 of this handbook.

7.1.4 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 et al. 2009, Krämer et al. 2013].

7.1.4.1 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 et al. [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 includes clarification requests, establishing a confidence level in the content of conversation, assessments of relevance, and indications and connections of topical information within the conversation. Content-agnostic behavior, however, has to do with the social rules of the conversation. Content-agnostic gestures may include next-speaker selection or handling of anti-social or non-canonical discourse behavior, such as interrupting.

7.1.4.2 Observer's Internal Beliefs

The gestures that accompany face-to-face spoken interaction convey a wide variety of information and stand in different relations to the verbal content. For the observer, gestures serve a wide variety of communication functions, such as commenting, requesting, protesting, directing attention, showing, and rejecting [Jokinen et al. 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] and 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]. Gestures 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]. A downward stroke of a gesture is often used to emphasize the significance of a word or phrase in the speech or enumerate points.

Gestures are also used to reinforce and clarify their co-speech utterances. Jamalian and Tversky [2012] show that different gestures in coordination 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 et al. 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 [2012] showed that using different types of metaphoric gestures changes the way that individuals qualitatively describe certain systems and processes. Gestures can also present information about the speaker's state and views toward the subject of conversation. Pollick et al. [2001] show 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.

Similarly, gestures have also been shown to influence memory recall in cases of eye-witness testimony [Gurney et al. 2013], opening up discussion of gestures providing leading answers in a similar off-the-record manner.

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 et al. 2009].

7.1.4.3 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 the underlying beliefs, intentions, and processes of the speaker [Cienki and Koenig 1998].

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

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].

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] (we discuss the ways in which the form of gesture carries

meaning in Section 7.2.1). 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.

7.1.4.4 Speaker Impact

While gesture is an invaluable tool for 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, 1998]. 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 et al. 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 et al. 1996]. This phenomenon points to deep relationships between physical body movements and cognition, discussed in the next section.

7.2 Models and Approaches

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. Broadly, this difficulty can be 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.

However, before we go further into how gestures may be generated and acted by socially intelligent agents, we must elaborate on how gestures carry meaning in the

first place in order to discuss how the components of gesture may be manipulated based on communicative intent.

In this section, we focus on broad approaches and their similarities and differences. While we provide contemporary examples of these various architectures, we do not deal with implementations of computational models or gesture generation mechanisms. For a more extensive look at the implementation of such architectures, please refer to Chapter 16.

7.2.1 How Gestures Carry Meaning

As we saw earlier, gestures play a variety of functions in face-to-face interaction and further there may be multiple such functions that are relevant during a specific utterance. However, there is a limit to the complexity of information they can reliably convey [Saund et al. 2019]. In this section, we discuss the traits of gesture that have been shown to carry meaning to viewers.

There are many individual components of a gesture that 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 that guide those components. Non-exhaustively, this includes velocity and amplitude of arm motions, orientation of the speaker toward the subject, the direction and symmetry of the hands, 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 the gestural components specified in Zao in Calbris et al. [1986]: movement, localization, body part, orientation, and configuration. 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 are integral in conveying these functions [Castellano et al. 2007]. Calbris et al. [1986] also explores how varying parameters of a gesture may result in multiple gestural representations of a single idea, and how, because of the parameter space of gestures, one idea may be presented by many different conceivable gestures.

7.2.2 Challenges of Gesture Generation

The two challenges of selection and execution come with two important constraints that plague all aspects of intelligent social agent research: processing time

and realization (animation or hardware) constraints. An acceptable pause between utterances is anywhere from 100–300ms [Reidsma et al. 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 5s, 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 et al. 2006] with stages that represent intent planning, behavior planning, and behavior realization. SAIBA interfaces with two markup languages, functional markup language and behavior markup language, 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. 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, we dive deeper into the considerations of the process going from communicative intent to gesture performance.

7.2.2.1 Selection

Selecting a gesture comes with a range of considerations. Some driving factors may be the communicative intent of the speaker, from the motivation and subgoal of a particular utterance to any driving goals of 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 et al. 2017], this function may be considered unnecessary or even detrimental to an agents 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 2014] but can also lead to critical misunderstandings [Gurney et al. 2013].

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 agents' concern with respect to gesture is not only "what does my gesture mean?" but "what does my gesture mean *to them*?" Scassellati [2002] provides an overview of how these challenges might be addressed in artificial agents, including implementations to find ways that can be used to predict internal state and, consequently, potential user responses. For an overview of theory of mind for SIAs, please refer to Chapter 9 of this handbook.

Moreover, what may still be more relevant to an agent's gestures is its own internal emotional state. Gesture can also be used to portray emotion in a way that is detectable by viewers [Pollick et al. 2001, Kipp and Martin 2009]. There is considerable literature dedicated to computational models of emotion, with a summary found in Marsella 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 that 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 is a gesture performance 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 communication (or hinder it) [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.

7.2.2.2 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, Section 7.2.1), they can be extremely challenging

or impossible to replicate exactly, especially in physical robots with limited degrees of freedom compared to people or non-humanoid forms.

One area of concern in terms of the 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]. 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 [Autodesk, INC.] to character animation engines [Niewiadomski et al. 2009, USC Institute for Creative Technologies] exist to both hand animate, use motion capture, or procedurally generate gestures on virtual agents.

As discussed in Section 7.1.2.1, 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 7.1.2.1). 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 et al. 2014]. Xu et al. [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 co-articulation between gestures within gesture units, including determining whether individual gestures go into phases of relax, rests, or holds. The work of Xu et al. drew on Calbris' [2011] concept of an ideational unit.

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. 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. A variety of motion blending techniques used specifically in the context of gesture generation are discussed in Feng et al. [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 very different and severely limited compared to graphics-based humanoid models. Specifically, robots suffer from the 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. These challenges are discussed further in Section 7.3.

7.2.2.3 Gesture Catalogs Versus Dynamic Generation

Broadly, we can characterize approaches to gesture generation as either using a set catalog of gestures or a set of parameters that drives dynamic generation of gestures on the fly. Here we provide an overview of these approaches, while below we will instantiate them with existing implementations.

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.

Alternatively, an option of greater complexity is to allow agents to generate gestures entirely from a more complete parameterization of the motion such as 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. The first approach 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 lookup table approach, where the context selects a set of pre-specified parameter values. For example, Poggi et al. [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 "mild" or "extreme" versions of a gesture, all at run time.

7.2.3 Broad Approaches in Current Implementations

We have discussed the ways in which gestures carry meaning and the challenges facing researchers who implement generative models of gesture. Now, we present implementations that attempt to overcome these challenges to create compelling gestures in socially intelligent agents.

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, however, is that of going from mental states to gestural performance. As we noted, human gesturing is influenced by a wide variety of 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 [Swartout et al. 2006, Kopp et al. 2014]). Consequently, the proxy input in gesture models is often reduced to the text and/or audio of the utterance that 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 to such information when selecting or recording the voice, respectively.

7.2.3.1 Rule-based Models

One of the earliest, if not earliest, generators is the behavior expression animation toolkit (BEAT) [Cassell et al. 2004], which works by analyzing the relation between surface text and gestures. Text is parsed to attain information such as clauses, themes/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) [Chiu and Marsella 2011] extends the BEAT framework by making a clearer distinction between the communicative intent embedded in the surface text (e.g., affirmation, intensification, negation) 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 that utilizes real-world utterance analysis is by Stone et al. [2004]. They proposed a framework to extract utterances and gesture motions from recorded human data, and then generate animations by synthesizing these utterances and motion segments. This framework includes an authoring mechanism to segment utterances and gesture motions and a selection mechanism to compose utterances and gestures. Similar to this, Neff et al. [2008] created a comprehensive list of mappings between gesture 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. Notably, 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 that should be present in a gesture (as seen in Poggi et al. [2005]). Additionally, various contextual information, such as speech prosody or detected listener attention, can determine other elements of gestural performance such as speed (or 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 [Lhommet and Marsella 2013, Marsella et al. 2013] is a system currently in use in both virtual agent and social robotics applications. which dynamically generates gestures that appropriately correspond to speech both auditorily and semantically.

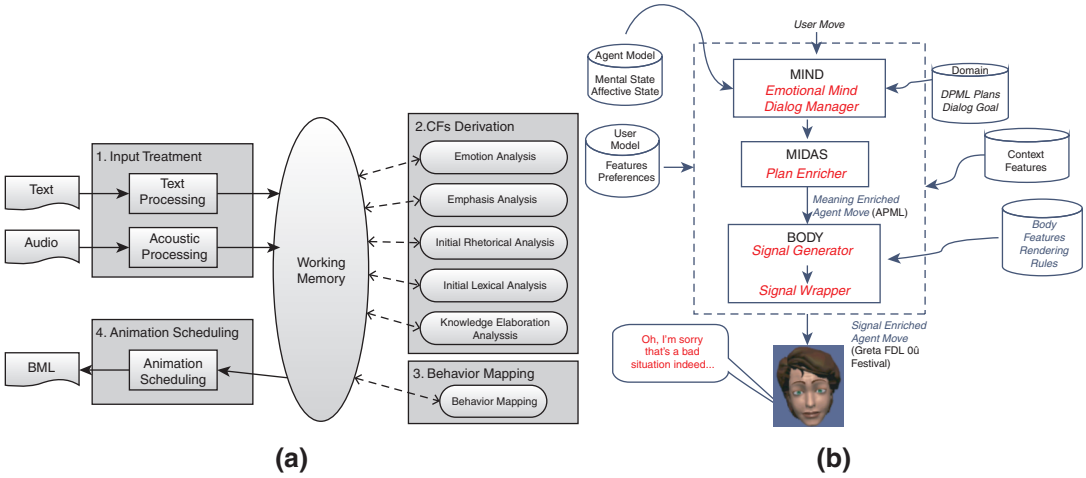

Figure 7.2 The architectures of two generative gesture models. (a) Cerebella architecture and (b) GRETA architecture.

Greta [Poggi et al. 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 7.2(a) and (b).

7.2.3.2 Data-driven Techniques

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 et al. 2019] to extract motion data to correlate audio to speaker movements. Using varying combinations of adversarial networks and regression, models are able to produce extremely natural gestures over a wide variety of speech-audio inputs [Ferstl et al. 2020]. This approach undeniably leads to impressively natural results, particularly in the context of generating gestures based on an individual speaker [Ginosar et al. 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 more natural, are less nuanced and complex than those we see in human performance. While it has been argued that the middle layers of these networks can derive some of these aspects [Takeuchi et al. 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 the co-utterance text of gestures [Yoon et al. 2019]. These have resulted in gestures that are judged as related to co-utterance, as well as life-like and likeable. 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.

The possibility of hybrid systems can offer the best of both worlds in terms of flexibility, novelty, and performance. From the examples above, it is easy to see how these two approaches exist on a continuum. 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 et al. 2001, Hutto and Gilbert 2014] and speech [Eyben et al. 2009, Schuller et al. 2011]. Similarly, we can detect transcripts from audio input and parse these using rhetorical and semantic cues through text parsers (e.g., Charniak [2000], Pedersen et al. [2004], and Joty et al. [2015]), 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 non-verbal behavior when that information is modeled by the agent [Marsella et al. 2013, Lhommet et al. 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 non-verbal behavior.

7.2.4 Gesture Collection and Analysis

To study and understand naturally occurring gestures, researchers use a variety of techniques, tools, and analyses.

Like many fields of behavioral psychology, researchers have used natural observation since the 1970s and 1980s. In the lab, however, 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]. Recently, researchers have begun using more subjective techniques such as conversational scenarios [Ennis et al. 2010] and questions, explicitly designed to elicit a variety of metaphoric gestures

[Chu et al. 2014]. 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 have provided troves of real-life examples of gestures by a huge variety of speakers in different contexts [Ginosar et al. 2019, Yoon et al. 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 [Chafai et al. 2007, Neff et al. 2010, Kipp 2014]. 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 non-verbal behavior [Luo et al. 2009]. However, this equipment is also 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 datasets 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 et al. 2017] to one central camera [Cooperrider 2014], and from set gestures in tightly controlled staging conditions [Gunes and Piccardi 2006, Hwang et al. 2006] to spontaneous recordings collected completely outside laboratory settings [Ginosar et al. 2019, Yoon et al. 2019]. Along with a growing interest in open science and dataset production, new annotation tools such as the Visual Search Engine for Multimodal Communication Research [Turchyn et al. 2018], which allows researchers to rapidly search datasets for specific types of motion, are becoming more sophisticated and widely used.

7.2.5 Evaluation

Evaluations of these 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 et al. 2010] as well as common factors of interest such as

trustworthiness, persuasiveness [Poggi and Pelachaud 2008], and naturalness [Maatman et al. 2005], often 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 [Wilson et al. 1996, Wachsmuth and Kopp 2001, Chiu and Marsella 2014]. 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 et al. 2005], for example.

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]. 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 that 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.

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 non-verbal behavior on a human participant's behavior. prisoner's dilemma [De Melo et al. 2009], the ultimatum game [Nishio et al. 2018], and the desert survival task [Khooshabeh et al. 2011] are a few examples.

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

Tools to deploy evaluations are also advancing rapidly. Whereas researchers previously 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 on this medium challenging [Breazeal et al. 2013]. Additionally, crowdsourced participants may be non-naive "expert survey-takers," which can skew study results [Downs et al. 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 et al. 2014, Rouse 2015].

Ultimately, the evaluation of a model must be specific to both its implementation and application.

7.3 Similarities and Differences in Intelligent Virtual Agents and Social Robots

Both social robots and virtual agents are discussed when considering the future of human–computer interactions. The application domains that researchers in each field aim to apply these artificial social agents largely overlap, and include personal assistance, companionship, education, leisure, and clerical work [Riek 2014].

The importance of co-speech gesture in both domains has been strongly established, albeit with discrepancies as to the impact of physical embodiment [Li 2015]. Gesture is widely acknowledged as vital in initiating social conversation [Satake et al. 2009], building rapport [Riek et al. 2010], and increasing human-likeness [Salem et al. 2013] for both virtual agents and social robots. Non-verbal behavior in social robots also increases users' abilities to maintain mental models of the robot's internal state [Breazeal et al. 2005], which is vital in co-operative tasks [Hiatt et al. 2011].

So far the algorithms we have described have been agnostic to the agent that may employ them. In this section we explore the similarities between gesture generation in virtual agents and social robots, but more pressingly the acute challenges that come with realizing gestures on physical devices.

7.3.1 Physical Presence

A significant body of literature suggests that robots gain some benefit to social interaction over virtual agents [Thellman et al. 2016]. Techniques that require physical presence, such as user mimicry and attention-grabbing motions [Fridin and Belokopytov 2014], may give robots an edge on virtual agents in terms of boosting learning outcomes in tutoring settings [Leyzberg et al. 2012, Belpaeme et al. 2018],

particularly for children [Jost et al. 2012]. Social robots have also been shown to be more helpful and enjoyable in interactions than their virtual counterparts for adults who are familiar with robots [Wainer et al. 2007]. However, robots also suffer from very high user expectations with respect to physical interaction and ability to sense the environment [Lee et al. 2006].

Many of these evaluations are task-based or based solely on physical embodiment and not about specific movements of gestures on robots versus virtual agents. It is unclear how these physical properties transfer to gestures' communicative properties.

7.3.2 Challenges of Physicality

There are many reasons why it is difficult to compare human-like gestures on virtual agents and robots due to robot form, function, movement capabilities, environmental limitations, and the high stakes of making movement mistakes in a robot. These limitations require creativity, artistry, and thorough exploration to realize communicative expression in new ways on physically limited robots. Ultimately, individual use cases must be taken into account when determining the tradeoff between utilizing a virtual agent or a social robot for specific purposes.

Humans have many more degrees of freedom in motion than most commercially available robots, and especially the social robots seen today [Leite et al. 2013]. High degrees of freedom robots are costly and more difficult to program than simpler counterparts. While a few humanoid robots with potentially full expression do exist [Robotics 2019, Shigemi et al. 2019], many more exist with humanoid shapes but severely limited expression [Gouaillier et al. 2009, Robotics 2018], and still more bypass any attempt at humanoid presentation in favor of more abstract forms [Anki, Breazeal 2014, Embodied]. For this reason, most generative algorithms designed for virtual agents must be re-mapped onto a robot's more limited expressive abilities, which can make gestures appear awkward or mis-timed [Bremner et al. 2009, Ng-Thow-Hing et al. 2010].

In most cases industrial robots are equipped with a set of pre-recorded gestures that are not generated online but simply replayed during human–robot interaction, as seen in Gorostiza et al. [2006], Sidner et al. [2003], or Salem et al. [2012]. Aligning speech to motion is particularly difficult in robots due to the path-planning required for novel gestures [Kopp et al. 2008].

Existing in the physical environment, while potentially more compelling and certainly with a wider range of physical tasks that may be accomplished, comes with distinct challenges when it comes to gesture. Problems unique to robots extend from motion planning to design, control, sensing, biomimetics, and

Figure 7.3 Some examples of contemporary social robots, ranging from humanoid with arms and legs, to less humanoid but still distinctly human with torso and arms, to more abstract but retaining a head and torso shape, to completely un-humanoid (and object-like). (a) ASIMO by Honda. (b) Pepper by SoftBank Robotics. (c) Jibo, photo courtesy of NTT DISRUPTION US Inc. (d) Cozmo, photo courtesy of Digital Dream Labs. Photos retrieved from global.honda/innovation/robotics; https://www.softbankrobotics.com/emea/en/pepper; https://www.jibo.com; https://www.anki.com

complex software [De Santis et al. 2008]. Additionally, robots must be consistently aware of their environment, including the people with whom they interact. Peri-personal space is a long-studied phenomenon in human–human interactions [Burgoon and Aho 1982, Sussman and Rosenfeld 1982, Burgoon 1991], and well-documented in virtual agents in AR interactions [Slater et al. 2000, Ennis et al. 2010], but establishing a "social safety zone" seems to be an especially salient issue when involving heavy or unfamiliar robots [Truong and Ngo 2016]. The problem of keeping robots at a socially acceptable distance from humans during interactions in itself requires knowledge of computer vision, psychology, and robotic path-planning [Gupta et al. 2018]. Despite the importance of proprioception and path-planning, most robots on the market today do not have robust full-body sensors capable of pro-actively avoiding collision, which means that some gestures could put the robot at risk of hurting itself or others.

Another ongoing challenge in gesture research for social robotics is the mapping of communicative intent to expression onto the many abstract forms of existing devices (e.g., those found in Figure 7.3) [Hoffman and Ju 2014]. Attribution of internal states from abstract motions has long been chronicled and analyzed

[Dittrich et al. 1996, Pollick et al. 2001], but the field is currently in the earliest stages of developing a framework that is capable of mapping the many elements of expression onto abstract frames [Van de Perre et al. 2018]. The art of mapping communication onto abstract bodily forms that are human-understandable is yet to be mastered.

7.3.3 Reach and Market Penetration

One of the fundamental distinctions between VAs and SRs is the ease of reaching users. VAs have been deployed on computers, web pages, tablets, and phones. Any device with a screen can be used to realize a VA application. The fact that they can be deployed so widely has special relevance for less-wealthy countries where the market penetration of cell phones is very high due to limitations in traditional landlines for telecommunications. For the user, there may be no significant additional hardware cost in using a VA application. SRs in comparison require the purchase of the robot and therefore are more of a luxury as opposed to a necessity given limited budgets. This is especially true of the current crop of SRs that can socialize but are incapable of performing useful physical actions that could justify the cost.

7.3.4 Interdisciplinary Collaboration

The fields of social robotics and virtual agents largely overlap. Both attempt to facilitate natural, socially fulfilling, and productive interactions in a wide range of fields, including medicine, teaching, and leisure. Both are concerned with the artificial agent's theory of mind [Breazeal and Scassellati 1999] and see agents as tools to study wider psychological phenomenon under tight controls, such as gender effects of gestures in human–computer interactions [Siegel et al. 2009, Feng et al. 2017]. Additionally, some properties known to be important in human interpretation of gesture, such as smoothness, shape, and timing, are shown to transfer to gestures in robots [Bremner et al. 2009].

The need and call for collaboration is not new [Holz et al. 2009]. Some researchers have begun using generative models originally developed on virtual agents with social robots, notably Salem et al. [2010] and Le and Pelachaud [2011]. This is made possible through common frameworks such as the dominant SAIBA framework [Kopp et al. 2006], described in detail in Chapter 16 [ICMI 2012], which may be combined to create an agent-agnostic generative pipeline [Le et al. 2012].

However, work in this area needs much more exploration. Collaborations need more than experts in robotics and virtual agents, and must include professionals in interaction and aesthetic design, animation, market research, and other artists. Without a holistic team, robots continue to be designed according to physical

constraints, with behaviors, animations, and designs then being forced to work within the physical constraints of the robot. Rather than separate disciplines, for commercial success all aspects of a social robot or agent must be included when considering specific use cases and audiences.

This is especially true in gestures, for which studies of interpretation of non-humanoid motions are academically limited but anecdotally extremely expressive. Consider Disney's many non-humanoid and non-verbal characters. In addition to actual robot characters Wall-E and Eve, animators use many cues to portray both character traits about animal characters as well as express a wide variety of communicative functions in non-humanoid ways. The transference of gesture *properties* onto non-humanoid characters without humanoid gesture *components* (described in Section 7.2.1), both virtual and robotic, is something that seems to be mastered by artists storytellers but not yet rigorously harnessed by academic researchers in either robotics or virtual agents.

7.4 Current Challenges

The technology and tools for modeling and generating gestures continues to advance. Further, larger datasets 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, we discuss what are some of those challenges.

7.4.1 Gestures and the Context that Informs Their Use

One of the key challenges we face in realizing gestures for social agents is the complex relation of gestures to the context of the interaction and overall structure of the discourse. As has been pointed out repeatedly by gesture researchers (e.g., Kendon [2000]), 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 the 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 whether the approach is end-to-end machine learning, rule-based, or some hybrid. Fundamentally capturing the above 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 internal mental states and roles of the participants will not be apparent in the individual utterance text or prosody, making 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, there must be some way of modeling this information over the course of interaction.

7.4.2 Complex Gesturing

A related challenge concerns complex gesturing. As illustrated above, 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.

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 will need to 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.

7.4.3 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

[de Marchena and Eigsti 2014], which can have significant effects on how the speaker is perceived [Lee et al. 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, this aspect of the context must affect both the selection as well as production of gestures.

This raises the question of how an agent infers a conversational partner's 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. This brings up issues of detecting user engagement and inferring mental state, as well as a growing issue of concern in gesture research: cross-cultural interpretation. As the world becomes more interconnected and developers of social agents become increasingly interested in international marketplaces, the importance of gesturing in a culturally sensitive way is gaining 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 as a line" metaphor discussed in Section 7.1.3). This means that metaphoric gestures that 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.

7.4.4 Ambiguity

On the other hand, one might well argue that human-like or "natural" behaviors may bring ambiguity. 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, some work suggests that when gestures are too complex [Saund et al. 2019] 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 tradeoffs 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, divergence from the naturalness of face-to-face interaction, broadly speaking but specifically here in terms of non-verbal behavior, can lead to an increase in cognitive effort, an increase in communication ambiguity, and a decrease in physiological arousal [Kock 2005].

7.4.5 The Application

Unquestioningly, these tradeoffs will be context-dependent, specifically application dependent. In a social skills training application to train doctors to break bad news to patients [Kron et al. 2017, Ochs et al. 2017], naturalness is a paramount consideration in part because people are being 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. Notably, social agent researchers [Marsella et al. 2006, Neff et al. 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 et al. 2000].

7.4.6 Impact

This discussion underscores the critical challenge of understanding and measuring the impact of gestures on human participants.

One way to evaluate this impact across large demographic populations is through increasingly popular crowdsourcing platforms [Breazeal et al. 2013, Morris et al. 2014]. In addition to evaluating a social agent's gesture performance, crowdsourcing opinions makes a combined approach to gesture generation possible: generative models that use crowd or expert input to create and refine generative models of dialog for a social agent [Feng et al. 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.

7.5 Future Directions

While this chapter has discussed state-of-the-art implementations of gestures in social agents, there are many promising horizons for future research that will allow still better gesture performances as well as insights into the cognitive processes behind gesture production.

7.5.1 Big Data and Gesture

It is impossible to talk about the future of gesture research without addressing the research field of big data. Using neural networks to create generative models of gesture for individual speakers is a present reality. Ginosar et al. [2019] present a model that produces gestures built off of L1 regression and adversarial neural networks. This model produces gestures that are nearly indistinguishable from the original speaker in many cases, but which are also driven exclusively by audio inputs.

This approach simplifies the inherently cognitively driven and complex nature of gesture. This model generates 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. This model, for example, does not handle the complexity of semantics, rhetoric, or affect (aside from how those elements are expressed in voice qualities). It could be argued that the middle layers of these networks implicitly derive other salient features. However, the gestures that result from these methods have been judged by naturalness with a particular piece of audio, not communicated message.

This is problematic as gestures have the ability to change the interpretation of the same audio [Jamalian and Tversky 2012, Lhommet and Marsella 2013]. 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 move to deep learning approaches that have the potential to generate not only extremely natural beat gestures but also more complex, nuanced, and subtle gestures as well.

7.5.2 Using Gesture to Make Inferences About Cognition

Using deep learning to generate gestures, however, 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 that they must abide by the same cognitive processes and limitations as we do.[2]

This possibility is eloquently expressed by the theory of *embodied cognition* [Hostetter and Alibali 2008]. 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) as well as performance on various cognitive tasks (such as reasoning or judgment). According to this hypothesis, the organization 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.

7.5.3 "Better than Human"

One of the common assumptions in the design of virtual agents is that human appearance and behavior is a gold standard for effective face-to-face interaction. This assumption is based on several factors. The non-verbal behaviors of human–human interaction are both our evolutionary heritage and socially learned. Therefore, an agent using these behaviors will be able to leverage the various deliberate inferences and automatic processes that are in play when we perceive these behaviors.

Human–human interaction is also often a guiding principle informing the design of social robots. Of course, the behaviors invariably get distilled down when realized in a robot, often due to mechanical constraints. For example, subtlety in dynamics may be removed, degrees of freedom may be removed such as not having fully functional hands. Some channels may be removed altogether such as eliminating eyebrows.

2. 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.

Figure 7.4 Examples of interactive wearables from Behnaz Farahi. (a) Iridescence. (b) Caress of the Gaze.

The use of human–human interaction as a design goal or even a guiding principle risks ignoring several factors. We are very adaptive, and in a persistent relation we could adapt to an artificial agent's behavior. That adaptation in turn may even help to build a stronger bond with an agent, for example, as a child requires a shared secret mode of interaction with an agent. Additionally, human non-verbal behavior is often ambiguous, and we may want to avoid that ambiguity in a particular application. Rather the focus may be on the most effective way to communicate the information, most effective in terms of an application's goals. Finally, by limiting ourselves to human non-verbal modalities, we ignore that we could employ novel non-human modalities.

For example, the work of Behnaz Farahi [2016, 2018, 2019], investigates novel modalities in interaction. Her "bio-inspired" work on the interactive installation Iridescence (Figure 7.4(a), [Farahi 2019]) draws inspiration from the gorget of the male Anna's Hummingbird that changes color during courtship. Iridescence changes colors and make patterns in response to observer's movements and facial expressions. Similarly, Caress of the Gaze is a wearable that explores how "clothing could interact with other people as a primary interface [Farahi 2016]." It uses eye-gaze tracking technologies to respond to the observer's gaze. Such work explores the potential of opening up new modalities in face-to-face interaction.

7.6 Summary

In this chapter we discussed the many ways that gesture enhances communication. Gesture acts as a guide for dialog, an influence on the observer, and a

reflection of the speaker's internal beliefs. We briefly summarized a long history of gesture studies, including myriad ways to classify gesture by both motion and communicative function. We discussed 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.

We then discuss 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 that 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.

These models may be deployed on either virtual agents or social robots, with the latter presenting great physical challenges but offering potentially greater impact on the viewer. Abstractions over gesture architectures are necessary to foster interdisciplinary collaboration between these two closely related mediums.

Despite recent advancements, gesture generation still faces many challenges, such as generating conversationally (semantically) relevant movements, incorporating 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.

New technology constantly advances techniques for studying gesture for both data collection and computational modeling of the physical gesture performance. In particular, superhuman stimuli offer unique avenues through which to study the impact of gesture, going beyond the possibilities of human–human studies. Additionally, collaborations in machine learning and the advancement of computational hardware and infrastructure allow more resources to use big data and end-to-end modeling of gesture behavior. These new technologies present opportunities to understand gesture's relationship to the semantic context in which it is produced, which will lead to new insights in human behavior, communication, and cognition.

References

M. W. Alibali and S. GoldinMeadow. 1993. Gesture–speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cogn. Psychol.* 25, 4, 468–523. DOI: https://doi.org/10.1006/cogp.1993.1012.

Anki. Cozmo. https://anki.com/en-gb/cozmo.html.

Autodesk, INC. Maya. https://autodesk.com/maya.

D. Barrett. 2010. *Supernormal Stimuli: How Primal Urges Overran their Evolutionary Purpose*. WW Norton & Company.

J. B. Bavelas. 1994. Gestures as part of speech: Methodological implications. *Res. Lang. Soc. Interact.* 27, 3, 201–221. DOI: https://doi.org/10.1207/s15327973rlsi2703_3.

T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. 2018. Social robots for education: A review. *Sci. Robot.* 3, 21, eaat5954. DOI: https://doi.org/10.1126/scirobotics.aat5954.

K. Bergmann and S. Kopp. 2012. Gestural alignment in natural dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

K. Bergmann, H. Rieser, and S. Kopp. 2011. Regulating dialogue with gestures—Towards an empirically grounded simulation with conversational agents. In *Proceedings of the SIGDIAL 2011 Conference*. 88–97.

C. L. Breazeal. 2014. Jibo, the world's first social robot for the home. *Indiegogo*. https://www.indiegogo.com/projects/jibo-the-world-s-first-socialrobot-for-the-home, checked on, 1, 22, 2019.

C. Breazeal and B. Scassellati. 1999. How to build robots that make friends and influence people. In *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)*, Vol. 2. IEEE, 858–863. DOI: http://doi.org/10.1109/IROS.1999.812787.

C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human–robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713.

C. Breazeal, N. DePalma, J. Orkin, S. Chernova, and M. Jung. 2013. Crowdsourcing human–robot interaction: New methods and system evaluation in a public environment. *J. Hum.-Robot Interact.* 2, 1, 82–111. DOI: https://doi.org/10.5898/JHRI.2.1.

P. Bremner, A. Pipe, C. Melhuish, M. Fraser, and S. Subramanian. 2009. Conversational gestures in human–robot interaction. In *2009 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 1645–1649.

A. Bryman. 2017. Quantitative and qualitative research: Further reflections on their integration. In *Mixing Methods: Qualitative and Quantitative Research*. Routledge, 57–78.

J. K. Burgoon. 1991. Relational message interpretations of touch, conversational distance, and posture. *J. Nonverbal Behav.* 15, 4, 233–259. DOI: https://doi.org/10.1007/BF00986924.

J. K. Burgoon and L. Aho. 1982. Three field experiments on the effects of violations of conversational distance. *Commun. Monogr.* 49, 2, 71–88. DOI: https://doi.org/10.1080/03637758209376073.

G. Calbris. 1990. *The Semiotics of French Gestures*, Vol. 1900. Indiana University Press.

G. Calbris. 1995. Anticipation du geste sur la parole. *Dins Verbal/Non Verbal, Frères juneaux de la parole. Actes de la journée d'études ANEFLE.* Besançon, Université de Franche-Comte, 12–18.

G. Calbris. 2011. *Elements of Meaning in Gesture*, Vol. 5. John Benjamins Publishing.

G. Calbris, J. Montredon, and P. W. Zaü. 1986. *Des gestes et des mots pour le dire*. Clé International, Paris, 145.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

J. Cassell. 1998. A framework for gesture generation and interpretation. *Computer Vision in Human–Machine Interaction*. 191–215.

J. Cassell, H. H. Vilhjálmsson, and T. Bickmore. 2004. BEAT: The behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

G. Castellano, S. D. Villalba, and A. Camurri. 2007. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 71–82. DOI: https://doi.org/10.1007/978-3-540-74889-2_7.

N. E. Chafai, C. Pelachaud, and D. Pelé. 2007. A case study of gesture expressivity breaks. *Lang. Resour. Eval.* 41, 3–4, 341–365. DOI: https://doi.org/10.1007/s10579-007-9051-7.

J. Chandler, P. Mueller, and G. Paolacci. 2014. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behav. Res. Methods* 46, 1, 112–130. DOI: http://doi.org/10.3758/s13428-013-0365-7.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

D. Chi, M. Costa, L. Zhao, and N. Badler. 2000. The emote model for effort and shape. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. 173–182. DOI: https://doi.org/10.1145/344779.352172.

C.-C. Chiu and S. Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents*. Springer, 127–140.

C.-C. Chiu and S. Marsella. 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, 781–788.

M. Chu, A. Meyer, L. Foulkes, and S. Kita. 2014. Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *J. Exp. Psychol. Gen.* 143, 2, 694. DOI: http://dx.doi.org/10.1037/a0033861.

A. J. Cienki and J.-P. Koenig. 1998. Metaphoric gestures and some of their relations to verbal metaphoric expressions. *Discourse and Cognition: Bridging the Gap*. 189–204.

K. Cooperrider. 2014. Body-directed gestures: Pointing to the self and beyond. *J. Pragmat.* 71, 1–16. DOI: https://doi.org/10.1016/j.pragma.2014.07.003.

S. Corera and N. Krishnarajah. 2011. Capturing hand gesture movement: A survey on tools, techniques and logical considerations. In *Proceedings of Chi Sparks*.

A. B. de Marchena and I.-M. Eigsti. 2014. Context counts: The impact of social context on gesture rate in verbally fluent adolescents with autism spectrum disorder. *Gesture* 14, 3, 375–393. DOI: https://doi.org/10.1075/gest.14.3.05mar.

C. M. De Melo, L. Zheng, and J. Gratch. 2009. Expression of moral emotions in cooperating agents. In *International Workshop on Intelligent Virtual Agents*. Springer, 301–307.

A. De Santis, B. Siciliano, A. de Luca, and A. Bicchi. 2008. An atlas of physical human–robot interaction. *Mech. Mach. Theory* 43, 3, 253–270. DOI: https://doi.org/10.1016/j.mechmachtheory.2007.03.003.

P. DiMaggio. 1997. Culture and cognition. *Annu. Rev. Sociol*. 23, 1, 263–287. DOI: https://doi.org/10.1146/annurev.soc.23.1.263.

W. H. Dittrich, T. Troscianko, S. E. Lea, and D. Morgan. 1996. Perception of emotion from dynamic point-light displays represented in dance. *Perception* 25, 6, 727–738. DOI: https://doi.org/10.1068/p250727.

J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. 2010. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2399–2402.

D. Efron. 1941. *Gesture and Environment*. King's Crown Press. DOI: https://doi.org/10.1177/000271624222000197.

P. Ekman and W. V. Friesen. 1969a. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1, 88–106. DOI: https://doi.org/10.1080/00332747.1969.11023575.

P. Ekman and W. V. Friesen. 1969b. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal Communication, Interaction, and Gesture*. 57–106. DOI: https://doi.org/10.1515/semi.1969.1.1.49.

I. Embodied. Moxie. https://embodied.com/products/moxie-reservation.

C. Ennis, R. McDonnell, and C. O'Sullivan. 2010. Seeing is believing: Body motion dominates in multisensory conversations. *ACM Tran. Graph. (TOG)* 29, 4, 1–9. DOI: https://doi.org/10.1145/1778765.1778828.

F. Eyben, M. Wöllmer, and B. 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, 1–6. DOI: http://doi.org/10.1109/ACII.2009.5349350.

B. Farahi. 2016. Caress of the gaze: A gaze actuated garment. In *ACADIA 2016: Posthuman Frontiers, published in Proceedings of the 36th Annual Conference*. USA.

B. Farahi. 2018. Heart of the matter: Affective computing in fashion and architecture. In *ACADIA 2018: Recalibration: Imprecision and Infidelity, Published in proceedings of the 38th Annual Conference*. Mexico City, Mexico.

B. Farahi. 2019. Iridescence: Bio-inspired emotive matter. In *ACADIA 2019: Ubiquity and Autonomy, Published in Proceedings of the 39th Annual Conference*. Austin.

A. Feng, Y. Huang, M. Kallmann, and A. Shapiro. 2012. An analysis of motion blending techniques. In *International Conference on Motion in Games.* Springer, 232–243.

D. Feng, D. C. Jeong, N. C. Krämer, L. C. Miller, and S. Marsella. 2017. "Is it just me?" Evaluating attribution of negative feedback as a function of virtual instructor's gender and proxemics. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 810–818.

D. Feng, P. Sequeira, E. Carstensdottir, M. S. El-Nasr, and S. 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, 509–516.

Y. Ferstl and R. McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98. DOI: https://doi.org/10.1145/3267851.3267898.

Y. Ferstl, M. Neff, and R. McDonnell. 2020. Adversarial gesture generation with realistic gesture phasing. *Comput. Graph*. 89, 117–130. DOI: https://doi.org/10.1016/j.cag.2020.04.007.

M. Fridin and M. Belokopytov. 2014. Embodied robot versus virtual agent: Involvement of preschool children in motor task performance. *Int. J. Hum.-Comput. Int.* 30, 6, 459–469. DOI: https://doi.org/10.1080/10447318.2014.888500.

R. W. Gibbs Jr. 2008. *The Cambridge Handbook of Metaphor and Thought*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511816802.

S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506. DOI: https://doi.org/10.1109/CVPR.2019.00361.

S. Goldin-Meadow and M. W. Alibali. 2013. Gesture's role in speaking, learning, and creating language. *Annu. Rev. Psychol*. 64, 257–283. DOI: https://doi.org/10.1146/annurev-psych-113011-143802.

S. Goldin-Meadow, H. Nusbaum, S. D. Kelly, and S. Wagner. 2001. Explaining math: Gesturing lightens the load. *Psychol Sci.* 12, 6, 516–522. DOI: https://doi.org/10.1111/1467-9280.00395.

J. F. Gorostiza, R. Barber, A. M. Khamis, M. Malfaz, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and M. A. Salichs. 2006. Multimodal human–robot interaction framework for a personal robot. In *ROMAN 2006—The 15th IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 39–44. DOI: https://doi.org/10.1109/ROMAN.2006.314392.

D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier. 2009. Mechatronic design of NAO humanoid. In *2009 IEEE International Conference on Robotics and Automation*. IEEE, 769–774. DOI: https://doi.org/10.1109/ROBOT.2009.5152516.

J. Grady. 1997. *Foundations of Meaning: Primary Metaphors and Primary Scenes*. University of California, Berkeley.

H. Gunes and M. 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, 1148–1153.

A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. 2018. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2255–2264.

D. J. Gurney, K. J. Pine, and R. Wiseman. 2013. The gestural misinformation effect: Skewing eyewitness testimony through gesture. *Am J. Psychol.* 126, 3, 301–314. DOI: https://doi.org/10.5406/amerjpsyc.126.3.0301.

U. Hadar. 1989. Two types of gesture and their role in speech production. *J. Lang. Soc. Psychol.* 8, 3–4, 221–228. DOI: https://doi.org/10.1177/0261927X8983004.

L. M. Hiatt, A. M. Harrison, and J. G. Trafton. 2011. Accommodating human variability in human–robot teams through theory of mind. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

G. Hoffman and W. Ju. 2014. Designing robots with movement in mind. *J. Hum.-Rob. Interact.* 3, 1, 91–122. DOI: https://doi.org/10.5898/JHRI.3.1.Hoffman.

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

T. Holz, M. Dragone, and G. M. O'Hare. 2009. Where robots and virtual agents meet. *Int. J. Soc. Robot.* 1, 1, 83–93. DOI: http://dx.doi.org/10.1007/s12369-008-0002-2.

A. B. Hostetter and M. W. Alibali. 2008. Visible embodiment: Gestures as simulated action. *Psychon. Bull. Rev.* 15, 3, 495–514. DOI: http://dx.doi.org/10.3758/pbr.15.3.495.

C. J. Hutto and E. 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*.

B.-W. Hwang, S. Kim, and S.-W. Lee. 2006. A full-body gesture database for automatic gesture recognition. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 243–248. DOI: https://doi.org/10.1109/FGR.2006.8.

J. M. Iverson and S. Goldin-Meadow. 1997. What's communication got to do with it? Gesture in children blind from birth. *Dev. Psychol.* 33, 3, 453. DOI: https://doi.org/10.1037/0012-1649.33.3.453.

J. M. Iverson and S. Goldin-Meadow. 1998. Why people gesture when they speak. *Nature* 396, 6708, 228–228. DOI: https://doi.org/10.1038/24300.

J. M. Iverson and S. Goldin-Meadow. 2001. The resilience of gesture in talk: Gesture in blind speakers and listeners. *Dev. Sci.* 4, 4, 416–422. DOI: https://doi.org/10.1111/1467-7687.00183.

N. Jacobs and A. Garnham. 2007. The role of conversational hand gestures in a narrative task. *J. Mem. Lang.* 56, 2, 291–303. DOI: https://doi.org/10.1016/j.jml.2006.07.011.

A. Jamalian and B. Tversky. 2012. Gestures alter thinking about time. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

K. Jokinen, C. Navarretta, and P. Paggio. 2008. Distinguishing the communicative functions of gestures. In *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 38–49. DOI: https://doi.org/10.1007/978-3-540-85853-9_4.

H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. 2017. Panoptic Studio: A massively

multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1, 190–204.

C. Jost, V. André, B. Le Pévédic, A. Lemasson, M. Hausberger, and D. Duhaut. 2012. Ethological evaluation of human–robot interaction: Are children more efficient and motivated with computer, virtual agent or robots? In *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 1368–1373. DOI: https://doi.org/10.1109/ROBIO.2012.6491159.

S. Joty, G. Carenini, and R. T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Comput. Linguist.* 41, 3, 385–435. DOI: https://doi.org/10.1162/COLI_a_00226.

S. D. Kelly, D. J. Barr, R. B. Church, and K. Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *J Mem. Lang.* 40, 4, 577–592. DOI: https://doi.org/10.1006/JMLA.1999.2634.

A. Kendon. 1997. Gesture. *Annu. Rev. Anthropol.* 26, 1, 109–128. DOI: https://doi.org/10.1146/annurev.anthro.26.1.109.

A. Kendon. 2000. Language and gesture: Unity or duality. *Lang. Gesture* 2, 47–63. DOI: https://doi.org/10.1017/CBO9780511620850.004.

A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511807572.

P. Khooshabeh, C. McCall, S. Gandhe, J. Gratch, and J. Blascovich. 2011. Does it matter if a computer jokes. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 77–86. DOI: https://doi.org/10.1145/1979742.1979604.

M. Kipp. 2014. ANVIL: A universal video research tool. In *Handbook of Corpus Phonology.* 420–436.

M. Kipp and J.-C. 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, 1–8.

M. Kipp, M. Neff, and I. Albrecht. 2007. An annotation scheme for conversational gestures: How to economically capture timing and form. *Lang. Resour. Eval.* 41, 3–4, 325–339. DOI: https://doi.org/10.1007/s10579-007-9053-5.

S. Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Lang. Cogn. Process*. 24, 2, 145–167. DOI: https://doi.org/10.1080/01690960802586188.

S. Kita and A. 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. *J. Mem. Lang.* 48, 1, 16–32. DOI: https://doi.org/10.1016/S0749-596X(02)00505-3.

N. Kock. 2005. Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE Trans. Prof. Commun.* 48, 2, 117–130. DOI: https://doi.org/10.1109/TPC.2005.849649.

S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International Workshop on Intelligent Virtual Agents.* Springer, 205–217. DOI: https://doi.org/10.1007/11821830_17.

S. Kopp, K. Bergmann, and I. Wachsmuth. 2008. Multimodal communication from multimodal thinking—towards an integrated model of speech and gesture production. *Int. J. Semant. Comput.* 2, 01, 115–136. DOI: https://doi.org/10.1142/S1793351X08000361.

S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier. 2014. An architecture for fluid real-time conversational agents: Integrating incremental output generation and input processing. *J. Multimodal User Interfaces* 8, 1, 97–108. DOI: https://doi.org/10.1007/s12193-013-0130-3.

E. Krahmer and M. Swerts. 2007. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *J. Mem. Lang.* 57, 3, 396–414. DOI: https://doi.org/10.1016/j.jml.2007.06.005.

N. Krämer, S. Kopp, C. Becker-Asano, and N. Sommer. 2013. Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior. *Int. J. Hum. Comput. Stud.* 71, 3, 335–349. DOI: https://doi.org/10.1016/j.ijhcs.2012.09.006.

F. W. Kron, M. D. Fetters, M. W. Scerbo, C. B. White, M. L. Lypson, M. A. Padilla, G. A. Gliva-McConvey, L. A. Belfore II, T. West, A. M. Wallace, T. C. Guetterman, L. S. Schleicher, R. A. Kennedy, R. S. Mangrulkar, J. F. Cleary, S. C. Marsella, and D. M. Becker. 2017. Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial. *Patient Educ Couns*. 100, 4, 748–759. DOI: https://doi.org/10.1016/j.pec.2016.10.024.

G. Lakoff and M. Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Q. Le, J. Huang, and C. Pelachaud. 2012. A common gesture and speech production framework for virtual and physical agents. In *ACM International Conference on Multimodal Interaction*.

ICMI. 2012. Workshop on Speech and Gesture Production in Virtually and Physically Embodied Conversational Agents, October 26, 2012, SantaMonica, CA. ACM.

Q. A. Le and C. Pelachaud. 2011. Generating co-speech gestures for the humanoid robot NAO through BML. In *International Gesture Workshop*. Springer, 228–237. DOI: https://doi.org/10.1007/978-3-642-34182-3_21.

D. Y. Lee, M. R. Uhlemann, and R. F. Haase. 1985. Counselor verbal and nonverbal responses and perceived expertness, trustworthiness, and attractiveness. *J. Couns. Psychol.* 32, 2, 181. DOI: https://doi.org/10.1037/0022-0167.32.2.181.

K. M. Lee, Y. Jung, J. Kim, and S. R. Kim. 2006. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *Int. J. Hum. Comput. Stud.* 64, 10, 962–973. DOI: https://doi.org/10.1016/j.ijhcs.2006.05.002.

I. Leite, C. Martinho, and A. Paiva. 2013. Social robots for long-term interaction: A survey. *Int. J. Soc. Robot.* 5, 2, 291–308. DOI: https://doi.org/10.1007/s12369-013-0178-y.

T. Leonard and F. Cummins. 2011. The temporal relation between beat gestures and speech. *Lang. Cogn. Neurosci.* 26, 10, 1457–1471. DOI: https://doi.org/10.1080/01690965.2010.500218.

S. C. Levinson. 1996. Language and space. *Annu. Rev. Anthropol*. 25, 1, 353–382. DOI: https://doi.org/10.1146/annurev.anthro.25.1.353.

E. T. Levy and D. McNeill. 1992. Speech, gesture, and discourse. *Discourse Process.* 15, 3, 277–301. DOI: https://doi.org/10.1080/01638539209544813.

D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

M. Lhommet and S. C. Marsella. 2013. Gesture with meaning. In *International Conference on Intelligent Virtual Agents*. Springer, 303–312. DOI: https://doi.org/10.1007/978-3-642-40415-3_27.

M. Lhommet and S. Marsella. 2014. Metaphoric gestures: Towards grounded mental spaces. In *International Conference on Intelligent Virtual Agents*. September. http://www.ccs.neu.edu/marsella/publications/pdf/Lhommet_IVA2014.pdf.

M. Lhommet, Y. Xu, and S. Marsella. 2015. Cerebella: Automatic generation of nonverbal behavior for virtual humans. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 4303–4304.

J. Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int. J. Hum. Comput. Stud.* 77, 23–37. DOI: https://doi.org/10.1016/j.ijhcs.2015.01.001.

K. P. Lickiss and A. R. Wellens. 1978. Effects of visual accessibility and hand restraint on fluency of gesticulator and effectiveness of message. *Percept. Mot. Ski.* 46, 3, 925–926. DOI: https://doi.org/10.2466/pms.1978.46.3.925.

P. Luo, M. Kipp, and M. Neff. 2009. Augmenting gesture animation with motion capture data to provide full-body engagement. In *International Workshop on Intelligent Virtual Agents*. Springer, 405–417. DOI: https://doi.org/10.1007/978-3-642-04380-2_44.

R. Maatman, J. Gratch, and S. Marsella. 2005. Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*. Springer, 25–36. DOI: https://doi.org/10.1007/11550617_3.

F. Maricchiolo, A. Gnisci, M. Bonaiuto, and G. Ficca. 2009. Effects of different types of hand gestures in persuasive speech on receivers' evaluations. *Lang. Cogn. Neurosci.* 24, 2, 239–266. DOI: https://doi.org/10.1080/01690960802159929.

S. C. Marsella, S. M. Carnicke, J. Gratch, A. Okhmatovskaia, and A. Rizzo. 2006. An exploration of Delsarte's structural acting system. In *International Workshop on Intelligent Virtual Agents*. Springer, 80–92. DOI: https://doi.org/10.1007/11821830_7.

S. Marsella, J. Gratch, and P. Petta. 2010. Computational models of emotion. *A Blueprint for Affective Computing—A Sourcebook and Manual* 11, 1, 21–46.

S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, and A. Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics*

Symposium on Computer Animation, SCA '13. ACM, New York, NY, 25–35. ISBN: 978-1-4503-2132-7. DOI: http://doi.acm.org/10.1145/2485895.2485900.

C. McCall, D. P. Bunyan, J. N. Bailenson, J. Blascovich, and A. C. Beall. 2009. Leveraging collaborative virtual environment technology for inter-population research on persuasion in a classroom setting. *PRESENCE Teleop. Virt. Environ.* 18, 5, 361–369. DOI: https://doi.org/10.1162/pres.18.5.361.

D. McNeill. 1985. So you think gestures are nonverbal? *Psychol. Rev.* 92, 3, 350. DOI: https://doi.org/10.1037/0033-295X.92.3.350.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press.

D. McNeill. 2006. Gesture: A psycholinguistic approach. *The Encyclopedia of Language and Linguistics*. 58–66. DOI: https://doi.org/10.1016/B0-08-044854-2/00798-7.

D. McNeill, J. Cassell, and E. T. Levy. 1993. Abstract deixis. *Semiotica* 95, 1–2, 5–20. DOI: https://doi.org/10.1515/semi.1993.95.1-2.5.

D. Morris. 2015. *Bodytalk: A World Guide to Gestures*. Random House.

R. Morris, D. McDuff, and R. Calvo. 2014. Crowdsourcing techniques for affective computing. In *The Oxford Handbook of Affective Computing*. Oxford University Press, 384–394. DOI: https://doi.org/10.1093/oxfordhb/9780199942237.013.003.

O. Mubin and C. Bartneck. 2015. Do as I say: Exploring human response to a predictable and unpredictable robot. In *Proceedings of the 2015 British HCI Conference*. 110–116. DOI: https://doi.org/10.1145/2783446.2783582.

K. M. Murphy. 2003. Building meaning in interaction: Rethinking gesture classifications. *Crossroads of Language, Interaction, and Culture* 5, 29–47.

M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Trans. Graph. (TOG)* 27, 1, 1–24. DOI: https://doi.org/10.1145/1330511.1330516.

M. Neff, Y. Wang, R. Abbott, and M. Walker. 2010. Evaluating the effect of gesture and language on personality perception in conversational agents. In *International Conference on Intelligent Virtual Agents*. Springer, 222–235. DOI: https://doi.org/10.1007/978-3-642-15892-6_24.

V. Ng-Thow-Hing, P. Luo, and S. Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4617–4624. DOI: https://doi.org/10.1109/IROS.2010.5654322.

R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud. January. 2009. *Greta: An Interactive Expressive ECA System*, Vol. 2. 1399–1400. DOI: https://doi.org/10.1145/1558109.1558314.

S. Nishio, K. Ogawa, Y. Kanakogi, S. Itakura, and H. Ishiguro. 2018. Do robot appearance and speech affect people's attitude? Evaluation through the ultimatum game. In *Geminoid Studies*. Springer, 263–277. DOI: https://doi.org/10.1007/978-981-10-8702-8_16.

S. Nobe. 2000. Where do most spontaneous representational gestures actually occur with respect to speech. *Language and Gesture* 2, 186. https://doi.org/10.1017/CBO9780511620850.012.

M. A. Novack and S. Goldin-Meadow. 2017. Gesture as representational action: A paper about function. *Psychon. Bull. Rev.* 24, 3, 652–665. DOI: https://doi.org/10.3758/s13423-016-1145-z.

R. E. Núñez and E. Sweetser. 2006. With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cogn. Sci.* 30, 3, 401–450. DOI: https://doi.org/10.1207/s15516709cog0000_62.

M. Ochs, G. de Montcheuil, J.-M. Pergandi, J. Saubesty, C. Pelachaud, D. Mestre, and P. Blache. 2017. An architecture of virtual patient simulation platform to train doctors to break bad news. In *Conference on Computer Animation and Social Agents (CASA)*.

Ş. Özçalışkan and S. Goldin-Meadow. 2005. Gesture is at the cutting edge of early language development. *Cognition* 96, 3, B101–B113. DOI: https://doi.org/10.1016/j.cognition.2005.01.001.

T. Pedersen, S. Patwardhan, and J. Michelizzi. 2004. WordNet::Similarity—Measuring the relatedness of concepts. In *AAAI*, Vol. 4. 25–29.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Linguistic Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum Associates, Mahway, NJ. 71, 2001.

I. Poggi and C. Pelachaud. 2008. Persuasion and the expressivity of gestures in humans and machines. *Embodied Communication in Humans and Machines*. 391–424. DOI: https://doi.org/10.1093/acprof:oso/9780199231751.003.0017.

I. Poggi and L. Vincze. 2008. Gesture, gaze and persuasive strategies in political discourse. In *International LREC Workshop on Multimodal Corpora*. Springer, 73–92. DOI: https://doi.org/10.1007/978-3-642-04793-0_5.

I. Poggi, C. Pelachaud, F. de Rosis, V. Carofiglio, and B. De Carolis. 2005. Greta. A believable embodied conversational agent. In *Multimodal Intelligent Information Presentation*. Springer, 3–25. https://doi.org/10.1007/1-4020-3051-7_1.

F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. 2001. Perceiving affect from arm movement. *Cognition* 82, 2, B51–B61. DOI: https://doi.org/10.1016/s0010-0277(01)00147-0.

G. Radden. 2003. The metaphor time as space across languages. *Zeitschrift für interkulturellen Fremdsprachenunterricht*, 8, 2.

F. H. Rauscher, R. M. Krauss, and Y. Chen. 1996. Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychol. Sci.* 7, 4, 226–231. DOI: https://doi.org/10.1111/j.1467-9280.1996.tb00364.x.

D. Reidsma, I. de Kok, D. Neiberg, S. C. Pammi, B. van Straalen, K. Truong, and H. van Welbergen. 2011. Continuous interaction with a virtual human. *J. Multimodal User Interfaces* 4, 2, 97–118. DOI: https://doi.org/10.1007/s12193-011-0060-x.

L. Ren, A. Patrick, A. A. Efros, J. K. Hodgins, and J. M. Rehg. 2005. A data-driven approach to quantifying natural human motion. *ACM Trans. Graph. (TOG)* 24, 3, 1090–1097. DOI: https://doi.org/10.1145/1073204.1073316.

L. D. Riek. 2014. The social co-robotics problem space: Six key challenges. *Robotics Challenges and Vision (RCV2013)*.

L. D. Riek, P. C. Paul, and P. Robinson. 2010. When my robot smiles at me: Enabling human–robot rapport via real-time head gesture mimicry. *J. Multimodal User Interfaces* 3, 1–2, 99–108. DOI: https://doi.org/10.1007/s12193-009-0028-2.

S. Robotics. 2018. Pepper. https://www.softbankrobotics.com/emea/en/pepper.

H. Robotics. 2019. Sophia. https://www.hansonrobotics.com/sophia.

S. V. Rouse. 2015. A reliability analysis of Mechanical Turk data. *Comput. Hum. Behav.* 43, 304–307. DOI: https://doi.org/10.1016/j.chb.2014.11.004.

M. Salem, S. Kopp, I. Wachsmuth, and F. Joublin. 2010. Generating robot gesture using a virtual agent framework. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 3592–3597. DOI: https://doi.org/10.1109/iros.2010.5650572.

M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. 2012. Generation and evaluation of communicative robot gesture. *Int. J. Soc. Robot.* 4, 2, 201–217. DOI: https://doi.org/10.1007/s12369-011-0124-9.

M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.* 5, 3, 313–323. DOI: https://doi.org/10.1007/s12369-013-0196-9.

S. Satake, T. Kanda, D. F. Glas, M. Imai, H. Ishiguro, and N. Hagita. 2009. How to approach humans? Strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*. 109–116. DOI: https://doi.org/10.1145/1514095.1514117.

C. Saund, M. Roth, M. Chollet, and S. Marsella. 2019. Multiple metaphors in metaphoric gesturing. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 524–530. DOI: https://doi.org/10.1109/ACII.2019.8925435.

B. Scassellati. 2002. Theory of mind for a humanoid robot. *Auton. Robots* 12, 1, 13–24. DOI: https://doi.org/10.1023/A:1013298507114.

B. Schuller, A. Batliner, S. Steidl, and D. Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* 53, 9–10, 1062–1087. DOI: https://doi.org/10.1016/j.specom.2011.01.011.

S. Shigemi, A. Goswami, and P. Vadakkepat. 2019. ASIMO and humanoid robot research at Honda. In *Humanoid Robotics: A Reference.* Springer, 55–90.

C. L. Sidner, C. Lee, and N. Lesh. 2003. The role of dialog in human robot interaction. In *International Workshop on Language Understanding and Agents for Real World Interaction*.

M. Siegel, C. Breazeal, and M. I. Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2563–2568. DOI: https://doi.org/10.1109/IROS.2009.5354116.

M. Slater, A. Sadagic, M. Usoh, and R. Schroeder. 2000. Small-group behavior in a virtual and real environment: A comparative study. *Presence Teleop. Virt Environ.* 9, 1, 37–51. DOI: https://doi.org/10.1162/105474600566600.

M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler. 2004. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Trans. Graph. (TOG)* 23, 3, 506–513. DOI: https://doi.org/10.1145/1186562.1015753.

N. M. Sussman and H. M. Rosenfeld. 1982. Influence of culture, language, and sex on conversational distance. *J. Pers. Soc. Psychol.* 42, 1, 66–74. DOI: https://doi.org/10.1037/0022-3514.42.1.66.

W. R. Swartout, J. Gratch, R. W. Hill Jr, E. Hovy, S. Marsella, J. Rickel, and D. Traum. 2006. Toward virtual humans. *AI Magazine* 27, 2, 96–96. DOI: https://doi.org/10.1609/aimag.v27i2.1883.

A. Takeuchi and T. Naito. 1995. Situated facial displays: Towards social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 450–455. DOI: https://doi.org/10.1145/223904.223965.

K. Takeuchi, D. Hasegawa, S. Shirakawa, N. Kaneko, H. Sakuta, and K. 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*. 365–369. DOI: https://doi.org/10.1145/3125739.3132594.

L. Talmy. 1985. Grammatical categories and the lexicon. *Language Typology and Syntactic Description*, Vol. 3. 57–149.

S. Thellman, A. Silvervarg, A. Gulz, and T. Ziemke. 2016. Physical vs. virtual agent embodiment and effects on social interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 412–415.

X.-T. Truong and T.-D. Ngo. 2016. Dynamic social zone based mobile robot navigation for human comfortable safety in social environments. *Int. J. Soc. Robot.* 8, 5, 663–684. https://doi.org/10.1007/s12369-016-0352-0.

S. Turchyn, I. O. Moreno, C. P. Cánovas, F. F. Steen, M. Turner, J. Valenzuela, and S. Ray. 2018. Gesture annotation with a visual search engine for multimodal communication research. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

USC Institute for Creative Technologies. SmartBody. https://smartbody.ict.usc.edu/download2.

G. Van de Perre, H.-L. Cao, A. De Beir, P. G. Esteban, D. Lefeber, and B. Vanderborght. 2018. Generic method for generating blended gestures and affective functional behaviors for social robots. *Auton. Robots* 42, 3, 569–580. DOI: https://doi.org/10.1007/s10514-017-9650-0.

I. Wachsmuth and S. Kopp. 2001. Lifelike gesture synthesis and timing for conversational agents. In *International Gesture Workshop*. Springer, 120–133.

J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric. 2007. Embodiment and human–robot interaction: A task-based perspective. In *RO-MAN 2007—The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 872–877. DOI: https://doi.org/10.1109/ROMAN.2007.4415207.

A. Whiten and R. W. Byrne. 1988. *The Machiavellian Intelligence Hypotheses*. DOI: https://doi.org/10.1007/978-1-4419-1428-6_1048.

A. D. Wilson, A. F. Bobick, and J. Cassell. 1996. Recovering the temporal structure of natural gesture. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*. IEEE, 66–71. DOI: https://doi.org/10.1109/AFGR.1996.557245.

J. R. Wilson, N. Y. Lee, A. Saechao, S. Hershenson, M. Scheutz, and L. Tickle-Degnen. 2017. Hand gestures and verbal acknowledgments improve human–robot rapport. In *International Conference on Social Robotics*. Springer, 334–344.

C. Wolff. 2015. *A Psychology of Gesture*. Routledge.

Y. Xu, C. Pelachaud, and S. Marsella. 2014. Compound gesture generation: A model based on ideational units. In *International Conference on Intelligent Virtual Agents*. Springer, 477–491.

Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. 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, 4303–4309.