Movement Speaks for Itself : the Kinematic and Neural Dynamics of Communicative Action and Gesture

James P. Trujillo

ISBN: 978-94-6284-206-9
Cover Design: James Trujillo
Printing: Ipskamp Printing

Movement Speaks for Itself : the Kinematic and Neural Dynamics of Communicative Action and Gesture

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 13 februari 2020 om
14:30 uur precies

door

James P. Trujillo

geboren op 28 september, 1988

te Anaheim, Californië, Verenigde Staten

Promotoren

Prof. dr. Harold Bekkering

Prof. dr. Asli Özyürek

Copromotor

Dr. Irina Simanova

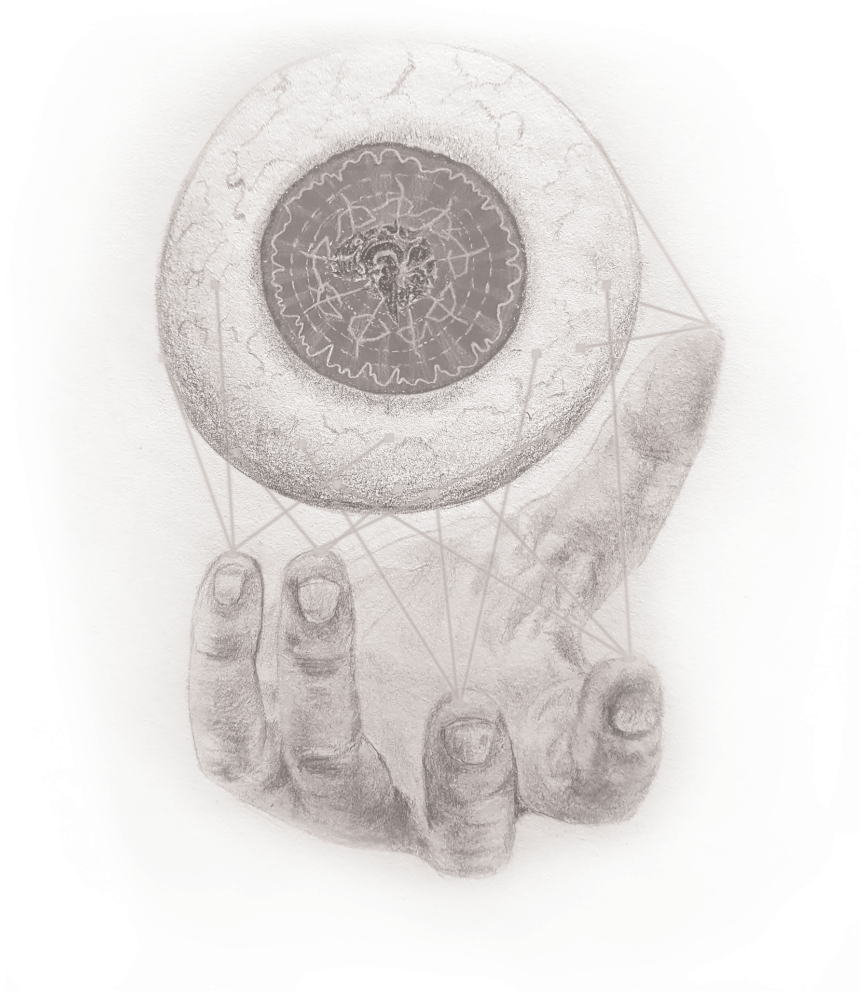
Manuscriptcommissie

Prof. dr. Ivan Toni

Prof. dr. Cristina Becchio
(Istituto Italiano di Tecnologia, Genova, Italië)

Prof. dr. Sotaro Kita
(University of Warwick, Verenigd Koninkrijk)

Contents



Chapter 1

General Introduction

CHAPTER 1

Humans are social creatures. Our lives are full of encounters with other humans in which we use various forms of communication in order to interact. When we think of communication we often think about language, whether spoken, written, or signed, but communication is much more than that. We also can very effectively communicate without saying a word. For example, imagine you are in a restaurant with some friends. Your friend raises her glass into the air and likely you are able to quickly recognize whether she is doing this to take a drink or to perform a toast with you. This is because you are able to quickly read her intention before she has completed the action, allowing you to respond quickly and appropriately. Similarly, if your friend does not have a glass she could still raise her hand *as if* she were raising a glass to toast, and you would likely understand this as well. This ability is part of what makes human social interaction work so effectively, allowing us to communicate efficiently, coordinate our actions with others (such as coordinating the raise of your own glass in the case of a toast) and to influence and learn from others.

Intention reading is possible because of what we refer to as social signaling. This refers to how we are constantly sending signals to those around us, allowing them to understand what our internal state is (e.g. if we are annoyed with the situation, or happy about it), or what our intention is so they can respond appropriately. Some of these signals, often also referred to as “body language” are well studied and recognized, such as orienting your body towards someone when you speak to them in order to show engagement in the interaction. Others are much more subtle, such as fine-grained differences in the way we perform an action. For example, we may reach out and grasp a cup with the intention to drink from it, or we may grasp the same cup with the intention to raise it up in a toast. In the second case, we are grasping it with a social intention, and our movements and eye-gaze act as a signal, allowing an observer to recognize what we intend to do, before we do it.

How exactly we utilize such complex signals, both in terms of producing them and understanding them, is the topic of my thesis. Specifically, I bring together action and gesture to understand how our intentions, in terms of action goals and social goals, shape our movements more generally, and how movement fits together with other bodily signals such as eye-gaze to facilitate communication. In this chapter I will provide some context for the studies described in the next chapters. I begin by discussing how social context shapes our behavior, followed by what we know

about communicative movements, and how the brain understands movements to be communicative. Then, I will discuss how these mechanisms of signaling may play a larger role in social interaction.

1.1 The Role of Social Context on Behavior

In everyday life we perform a variety of actions throughout the day. Many of these actions will be repeated many times, perhaps even within the same day, such as making a sandwich. Although the action itself is the same, the social context in which you produce the action will impact the way the action is performed in subtle but noticeable ways. When talking about social context, this can refer to many situations. For example, whether another person is present or not, whether you are currently interacting with them or not, and even whether this person is a child or an adult. In these different contexts, we are likely to have different intentional stances if our actions are relevant for that person. Think of the example of making a sandwich: we may perform this action just for ourselves, or use our action as a signal to request a response from the other person, or we may use gestures – communicative hand movements that are often paired with speech but that can also be used silently (See Box 1.1 for a definition of actions and gestures as discussed in this thesis).

Clear evidence of the impact of social context on behavior comes from research on adult-child interaction. When interacting with children, as compared to with other adults, adults tend to produce actions that are more eye-catching and may be more easily understood. For example when adults demonstrate to a child how to use a novel toy, they use more repetitions and more clearly segmented actions than when they demonstrate these toys to other adults (Brand, Baldwin, & Ashburn, 2002). The hand gestures we produce while speaking show a similar effect. Campisi and Özyürek found that, when describing how to use a coffee maker, adults who thought they were explaining this to a child produced larger, more complex gestures that were described as being more 'informative' when compared to the gestures produced for other adults (Campisi & Özyürek, 2013). This is similar to how adults use speech that is more informative in its content (Campisi & Özyürek, 2013) and acoustically more salient (Fernald, 1985; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989), when speaking to children. We additionally tend to make more direct eye-contact when interacting with children (Brand, Shallcross, Sabatos, & Massie, 2007), which is thought to maintain the continued interaction. We therefore have evidence from

Box 1.1 Actions and Gestures

Throughout the thesis I will refer to actions and gestures. In general, there are various definitions and forms of both of these movement types.

When I refer to *actions*, I am typically referring to what are known as "instrumental" or "object-directed" actions. That is, they are manual actions that involve the grasping and manipulation of physical objects. They correspond to the action hierarchy levels (see Figure 1 in section 1.2) of both "action" and "action sequence". As a general rule, the reader can consider actions to be the manual acts that involve manipulating objects with a direct goal, such as pouring coffee or opening a book.

Gestures refer to the communicative hand movements that we produce without manipulating any physical object. These movements form an integral part of the way we package and convey information for communication, supporting both the communicator as well as the addressee. From a neuroscientific perspective, gestures are considered to be generated by the same system as object-directed actions (Chu & Kita, 2015; Novack & Goldin-Meadow, 2017). While actions and gestures may be similar in their physical implementation, a major difference between them is that gestures often schematize information, focusing on the relevant aspects of the represented action (Kita, Alibali, & Chu, 2017). There are multiple types of gestures (see McNeill, 1992 or Kendon, 2004 for a more in-depth discussion), but throughout this thesis I will typically be referring to representational gestures.

Representational Gestures are hand movements that depict object features, such as tracing the outline of a shape, or simulating actions. I will mostly be talking about the action simulation variety of gestures. In these gestures, the person is acting out an action as if they are actually performing it, but without manipulating any objects.

A further distinction that should be made is that of *co-speech gestures* compared to *silent gestures*. Co-speech gestures are those that we produce alongside, such as to visually depict something we are talking about, or to provide some spatial information to complement our speech. I investigate co-speech gestures in **Chapter 5**. Silent gestures, on the other hand, are those that are produced without any accompanying speech. These gestures, also referred to as *pantomime gestures,* are sometimes used experimentally to investigate gesture production, as they allow researchers to separate the motor processes of gesture production from linguistic influences. I utilize silent gestures in **Chapters 2-4**.

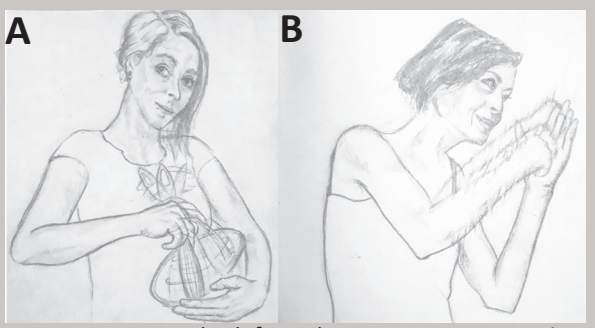


*Figure B.1.1On the left, **A** depicts a communicative action, demonstrating 'whisking'. On the right, **B** shows a pantomime gesture, depicting 'grating'.*

several communicative signals (i.e. actions, gestures, eye-gaze, and speech) that suggest that we try to make our communicative message more salient and more informative when we address children.

It may seem obvious that we behave differently when interacting with children compared to other adults. However, we adjust our actions in other contexts as well, taking into account how relevant our actions are to our partner. Sartori and colleagues showed this with an experiment in which participants reached out and lifted three colored objects placed on the table (Sartori, Becchio, Bara, & Castiello, 2009). The objects needed to be lifted in a certain order, which amounted to a code given to the participant. On some trials there was another person present and simply watching but not interacting, while on other trials this observer was blindfolded, and sometimes the observer was supposedly writing down the "code". The study showed that the presence of the observer, even when not interacting, changed the way the reaching and lifting actions were produced. The effect was even larger when the observer was trying to decipher the code. In other words, people took into account whether they had a partner, and whether that person was gaining anything from watching their actions (Sartori et al., 2009). While this study showed the effect of having an interactive partner, other social contexts, such as competition or cooperation (Manera, Becchio, Cavallo, Sartori, & Castiello, 2011), also shape the velocity and trajectory of reaching movements in a context-specific manner.

Similar results have been found in gestures. In one study, it was found that the trajectory and velocity of pointing gestures were different depending on whether or not the gesture was informative to an observer (Peeters, Chu, Holler, Hagoort, & Özyürek, 2015). A more recent study built on this result by asking participants to point at different targets, while another person observed either from the left or the right of the participant. In this study, participants adapted the trajectory of their movements based on the location of the observer (Winner et al., 2019). These two studies show that people take into account the presence and viewpoint of those around them. A similar idea has been tested in co-speech gestures. Özyürek (2002) showed that when people are describing motion events, such as someone going 'into' or 'out of' a location, their gestures are consistently oriented to move into or out of the shared space between speaker and addressee (Özyürek, 2002). In another study, Kelly and colleagues (S. D. Kelly, Byrne, & Holler, 2011) asked students to describe wilderness survival items to one of two audiences: one was another

group of students using the information for a dormitory orientation activity, and the other was a group of students who were actually preparing for a rugged camping trip. The study showed that the students who believed they were describing the items to the camping group used three times as many gestures, and spent three times as much time gesturing (S. D. Kelly et al., 2011). Although this study found no difference in speech quantity between the two scenarios, the previously discussed study by Campisi and Özyürek (2013) found an increase in speech quality (i.e. informativeness) when talking to other adults who were less knowledgeable about the task being described. Taken together, we see that people take into account the relevance of their actions and gestures, changing their behavior depending on the social context in which they are acting.

As we have seen, social context can shape the way we produce actions and gestures. However, our discussion of context is not complete without looking at both sides of the interaction. Some previous work suggests that these behavioral modulations can make the action or gesture more understandable, but this is only part of the story. One of the powerful attributes of humans is our ability to learn from others by focusing on relevant information. Of course, it is possible to learn from others by simply seeing their behavior(S. W. Kelly, Burton, Riedel, & Lynch, 2003). But such an approach would make it impossible to know what information about someone's behavior is relevant, or what aspects of the world around us we should pay attention to. Rather than simply observing, humans utilize ostensive cues to direct attention and learn from the parts of behavior that are most relevant. A theory known as natural pedagogy (Csibra & Gergely, 2009) suggests that from a young age we are sensitive to cues such as eye-gaze that signal to us that an upcoming behavior is relevant, or direct our attention to a particular object or direction(Senju & Csibra, 2008). These cues direct our attention to relevant information in others' behavior, allowing us to effectively engage with what is happening.

Natural pedagogy has primarily been researched using explicit communicative cues, such as making eye-contact or saying one's name (Senju & Csibra, 2008). Less is known about whether we can recognize this intention to communicate from more subtle cues, such as the changes in movement behavior. However, there is a body of work suggesting that clues to our intentions are embedded in our overt behavior and must be readable by an observer (Becchio, Manera, Sartori, Cavallo, & Castiello, 2012; Cavallo, Koul, Ansuini, Capozzi, & Becchio, 2016; Manera et al., 2011; Runeson

& Frykholm, 1983). This is important because the effects described above should not be seen as arising only from the context itself, but rather from the intentional stance that this context elicits. In other words, the context can be seen as a larger framing of the interaction, but ultimately it is the person producing the movements who shapes the movement qualities. This means that not only what we do, but the way we do it is its own complex communicative system, allowing our movements to "speak" for themselves. In the next section, I will discuss in more detail how our intentions shape the way that we move.

1.2 Communication in Movement

Intentions Shape the Way We Move

I have so far introduced the idea of actions being modulated, or shaped, by one's intentions. Now let us zoom in on what exactly this means. We often think of actions in terms of labels such as "reaching" and "grasping", or in terms of the even higher, sequence-level labels such as "drinking" or "pouring". This action hierarchy (Hamilton & Grafton, 2006; Ondobaka & Bekkering, 2012; Pacherie, 2008) can further be broken down into single ballistic movements of a body part that together form a coherent action. For instance, in order to grasp your coffee cup, you must move your arm in such a way as to bring your hand towards the cup while simultaneously extending your fingers to create the appropriate grasping shape, and finally you fold each of your fingers around the cup. Collectively, we would call this a "reach" and a "grasp", or more simply "grabbing" (see Figure 1). The velocity and trajectory of these movements are referred to as their kinematics (see Box 1.2.1 for an overview of how motion capture can be used to quantify movement kinematics). We can therefore move up the hierarchy, starting at movement kinematics and moving up to reaching and grasping, further up to drinking and pouring, and still higher to longer sequences of actions, such as preparing a meal. Higher levels are therefore made up of many lower level actions or movements. What makes this organization interesting is that higher levels in the hierarchy influence the way lower level actions are performed (Runeson & Frykholm, 1983). So, when we reach and grasp a glass in order to take a drink, the reach and grasp movements will be different from when the same actions are performed in order to raise the glass in a toast.

This hierarchy does not mean that a given action or action sequence is constrained

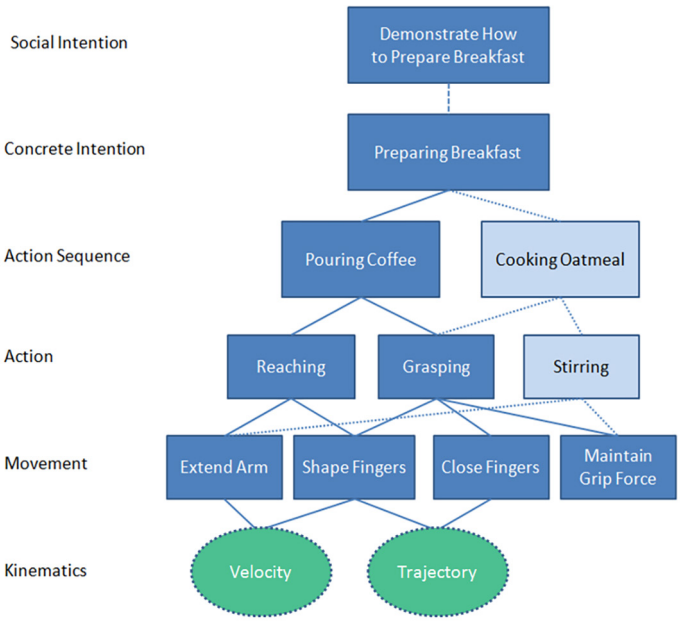


Figure1. A schematic example of an action hierarchy with the concrete end-goal (or intention) of preparing breakfast. The main focus here is the breakdown of 'pouring coffee' as one action sequence, but note that the intended 'preparation of breakfast' would entail many more action sequences, each with their own hierarchy of individual movements and kinematics. The focus of the diagram is on the dark blue boxes. The light blue boxes (e.g. "cooking oatmeal") show examples of additional actions or action sequences that may utilize similar lower-level movements, but for the sake of simplicity these are less fully defined in this graphic and also should not be taken as an exhaustive list. At the top of the hierarchy, one may have an additional social intention, such as demonstrating this action to someone else. While this hierarchy has typically been used to explain object-directed actions, I suggest that a similar hierarchy would hold for representational gestures.

to a particular set of movements or kinematics. Our ability to produce actions is highly flexible, allowing us to use objects in novel ways (van Elk, van Schie, & Bekkering, 2014), or to perform actions with different social intentions. These social intentions, like concrete intentions, affect the levels below them. As an example, social intentions, like a demonstration, also lead to differences in the movement kinematics. For instance, the trajectory (Quesque, Lewkowicz, Delevoye-Turrell, & Coello, 2013; Sartori et al., 2009) and velocity (Becchio, Sartori, & Castiello, 2010; Quesque et al., 2013) of communicatively intended reaching movements differ from that of non-communicative movements. Similarly in gestures, pointing gestures

Box 1.2.1 Motion Tracking and Naturalistic Data Collection

Typically, motion tracking utilizes markers that are placed on the body while their movements are captured either by a set of synchronized infrared cameras or by the emission of an electromagnetic field. By using multiple viewpoints, such as with the camera-based system, we are able to record movements in 3D. These techniques have long been used by movement scientists and have since been adopted into other domains to study kinematic differences in action production, variations in movement behavior in different social settings, etc.

Figure B.2.1. Photo of the Microsoft Kinect version 2, as used in the experiment described in Chapter 2 of this thesis.

While "markered" motion tracking has proven accurate and reliable, it comes with the disadvantage that movements may be restricted by the placement of markers, and people may be more aware that their movements are the subjects of investigation. A recent development in this area is the Microsoft Kinect (depicted in Figure B.2.1), which was originally developed as an input device for the Xbox gaming console, but has since been adopted for research purposes. Using an infrared emitter and sensor, the Kinect is able to see the environment in 3D with only one camera. Combined with vision-based human body detection algorithms, the Kinect provides 3D, markerless motion tracking. All studies described in Chapters 2–6 of this thesis use Microsoft Kinect motion tracking. This technology allows capturing complex movements in 3D, without any physical or psychological interference from markers being placed on a participant's body. It also allows isolating movements and transforming the data into *"stick-light figures*". In **Chapters 3-4** of this thesis I used "stick-light figures" as experimental stimuli, *to* study effects of kinematics in movement comprehension without other confounding effects such as background, facial expression, appearance of the actor, etc. (Figure B.2.2).

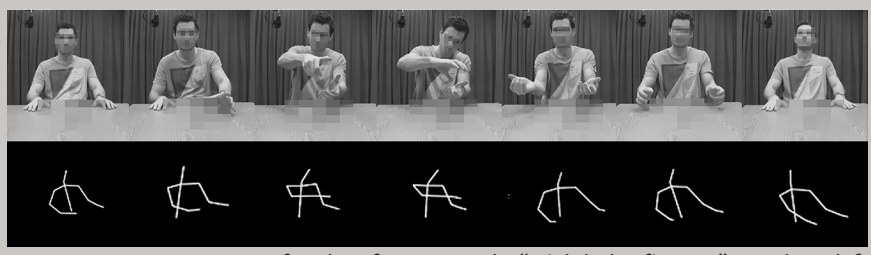


Figure B.2.2. Comparison of video frames with "stick-light figures" produced from simultaneously recorded motion tracking data.

made with a more-communicative intent showed a different velocity profile than those with a less-communicative intent (Peeters et al., 2015; Winner et al., 2019). Taken together, we see that communicative intentions shape actions and gestures at the kinematic level by varying the velocity and trajectory of the movements.

Although these earlier studies have looked at shorter segments of movement, such as reaching or pointing, a similar mechanism, and thus a similar underlying hierarchy, seems to be at play in more complex representational gestures. Representational gestures are those that utilize movements and hand-shapes to visually depict objects and actions (Kendon, 2004; McNeill, 1994). Typically, studies of these complex gestures have not used quantitative measures of kinematics, but the results still suggest that movement qualities are shaped by the intention to communicate. In the example by Campisi and Özyürek given in the previous section, the more communicative stance taken when interacting with a child led to increased gesture size as well as gesture complexity. These findings paint a picture of our communicative intentions being embedded, and thus potentially visible, in all of our movements. Indeed, a compelling theory developing in recent years is that the kinematic modulation by abstract intentions is a signal designed for an observer (Pezzulo & Dindo, 2013). In other words, we shape our movements to draw attention to relevant information. An interesting question left open by these studies is how action kinematics is linked to other articulators (e.g. eyes and lips) during communication. This would tell us if there is a general "communicative mode" that is effectively making any movement potentially communicative.

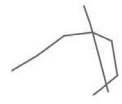
The extension of Pezzulo and Dindo's (2013) framework of communication to other articulators is particularly relevant when we realize the highly integrated nature of the body in general, and communicative behaviors more specifically. Consider the integration of speech and co-speech gesture. Work by Kita and Özyürek suggest that when we plan an utterance, speech and gesture together form an interface (Kita, 2000; Kita & Özürek, 2003) where gestures are not simply manual expressions of what we are trying to communicate, nor is speech a complete expression. Instead, the two communicative signals interact during the early planning phase to create a coherent, structured whole. Beyond this conceptual coupling, there is also a biomechanical coupling, such that effortful movements in gesture lead to acoustic changes in speech (Pouw, Harrison, & Dixon, 2019). Gestures also seem to be coupled with the

Box 1.2.2 Movement and Speech

While communicative intent seems to influence action and gesture kinematics, it is unclear how these intentions influence the other articulators, or perhaps even the dynamic relationship between them. For example, the model of speech gesture interface model proposed by Kita and Özyürek (2003) suggests that the modalities (e.g. speech and/or gesture) are selected based on what information one is intending to convey. After this, the actual speech and gestures are specified and produced. While this model describes the general process of generating multimodal utterances, Pezzulo and Dindo's communicative signaling framework (2013) suggest that any behavior can be adapted for communication. An interesting question is where this adaptation fits into the model. It could be that communicative intent simply places more effort into the articulators that have already been selected by the communication planner, as described by Kita and Özyürek. For example, we speak more clearly, exaggerate our gestures, and so on. This would fit with recent findings of biomechanical coupling between speech and gesture, which suggest that effort in one articulator leads to a similar peak in effort in the other (Pouw et al., 2019). Alternatively, communicative intention could be a part of the speech-gesture planning mechanism, affectively helping to select which modalities to utilize based on which one is likely to be effective given the context.

lips, as evidenced from the "echoing" of visual configurations or movements from the hands to the lips in sign languages (Woll, 2014; Woll & Sieratzki, 1998). Speech, in turn, is of course highly related to our lip movements. While this may seem trivial, lip movements play a major role in disambiguating what we are saying, even to the extent that when lip movements and speech do not match, observers "hear" something in between what the two signals were actually conveying (Mcgurk & Macdonald, 1976). See Box 1.2.2 for further discussion on speech-gesture coupling.

Beyond speech, gestures and lips, we also use our eye-gaze in coordination with gestures and actions during communication (Bavelas, Coates, & Johnson, 2002; Cañigueral & Hamilton, 2019; Senju & Johnson, 2009). We therefore see that communicative behavior consists of many articulators working together to express some information. Returning to the action hierarchy and Pezzulo and Dindo's (2013) model of communicative signaling, an important question is how communicative intentions fit into a larger model of communicative behavior, including movement kinematics, lips, voice and eyes.



In order to effectively communicative, we therefore need to orchestrate the various communicative signals in such a way as to influence the mental state of the addressee in a desired way. This requires us to take into account contextual factors, including the common ground between our knowledge and that of the addressee. It also requires us to be motivated to adjust our communicative strategies as needed, together with the cognitive capacity to deal with this complex task (van Rooij et al., 2011). In other words, the ability to communicate effectively is not related simply to language skills (Willems et al., 2010) or other classical psychometric measures on their own (Volman, Noordzij, & Toni, 2012). Instead, communication may itself be a relatively independent skill that guides the implementation of different communicative signals, such as speech, eye-gaze behavior, body language, gestures, or novel communicative methods. Therefore, understanding how communicative intentions fit into models of action, gesture, and speech production will be valuable to better understanding how humans are able to process and create complex social interaction. While these questions are quite large in their scope, a first step would be to understand if and how actions and gestures fit into a common framework of communicative kinematic modulation, and thus how people externalize communicative intentions. Specifically, I suggest that the kinematic markers of social intentions that have been found in reaching and pointing movements will also extend to the kinematics of complex actions and gestures.

Seeing Intentions in Movement

Successful communication depends not only on the communicator sending information, but it is also dependent on the intended receiver recognizing that what the communicator is doing is relevant. In other words, the communicator must make both their message and their intention to communicate clear to the addressee (Sperber & Wilson, 1986). As discussed in relation to natural pedagogy, this may occur if we hear our name, if someone makes direct-eye contact, or even if they are oriented towards us when speaking (Nagels, Kircher, Steines, & Straube, 2015). These are highly salient acts that signal an intention to engage in interaction. If kinematic modulation is indeed also a signal of communicative intent, then it should be recognizable as such to an observer.

In line with this idea, several studies have demonstrated that early kinematic differences can be utilized by observers to accurately predict the end-state of

an action before it has unfolded entirely (Cavallo et al., 2016; Sartori, Becchio, & Castiello, 2011; Stapel, Hunnius, & Bekkering, 2012). A similar picture is seen in abstract intentions, where people are able to discriminate between competitive and cooperative actions (Manera et al., 2011) as well as between actions with a social or personal intention (Lewkowicz, Quesque, Coello, & Delevoye-Turrell, 2015). These findings suggest that the kinematic modulation associated with abstract, communicative intentions is also visible to naïve observers.

An interesting question is how we are able to recognize that kinematic modulation is a communicative signal, rather than just variation in the way people move. One way to accomplish this is to take advantage of the consistency with which people typically perform an action. When producing an action, our motor system tunes the trajectories and velocities of the movements to be optimally efficient (Todorov, 2004). Simply put, we do not exert any more control or energy into the movements than what is necessary to achieve its goal. When we see others performing actions, we expect them to behave in a similarly efficient way (Gergely & Csibra, 2003; Hudson, McDonough, Edwards, & Bach, 2018). Communicatively intended actions are thus inefficient if we only consider a concrete end-goal intention. We as observers recognize this inefficiency.

Our ability to recognize inefficient actions follows from our natural inclination to learn from novel, relevant information in the environment (see the discussion on natural pedagogy in section 1.1; Csibra & Gergely, 2009). Studies on learning during development show that we form expectations about what is going to happen, effectively making predictions about what others are doing, and breaches of these expectations capture our attention. For example, novel information, such as the way an action is performed (Southgate, Chevallier, & Csibra, 2009), or the unexpectedness of the action given previous experience (e.g. using a different strategy than normal or performing an inefficient action; Liu & Spelke, 2017; Stahl & Feigenson, 2015), seem to trigger attention in children. The motoric inefficiency of communicative actions could therefore act as a signal to potential interactive partners, letting them know that there is relevant information for them.

While this theory has been tested in children for wholly irrational or unusual actions, it has not been tested in terms of kinematic modulations. Recent computational accounts highlight the flexibility of communicating by modulating one's movements,

as it allows any action or movement to potentially be communicative (Pezzulo, Donnarumma, & Dindo, 2013; Pezzulo, Donnarumma, et al., 2018). As any action can be modulated at the kinematic level, investigating how such a flexible yet subtle cue can be utilized to recognize the intention to communicate would help us to better understand the flexibility of human communication. Given our ability to recognize and utilize novel information, I expect that observers are able to use this kinematic modulation in order to infer a communicative intention.

1.3 How the Brain Infers Intentions from Movement

In the previous section I discussed how breaches of expectation can be perceived as a signal of one's intentions, allowing us to use our own experiences with actions to infer the underlying, higher level meaning of the act. This seems to make sense when considering how we learn from novel or unexpected events in our environment. Yet it is important to look at how the brain responds to and processes novel and unexpected information in order to understand the inner working of how we make these inferences about the intentions of others (see Box 1.3 for an overview how brain imaging can be utilized and how it is implemented in the current thesis). Understanding the neural implementation of this process opens a window into how the brain has evolved to deal with the complexities of social interaction.

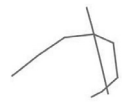
In order to understand the intentions underlying someone's actions, we must often first understand what they are doing. Typically, we can readily understand an action by recognizing the movements as something that we have seen before. One theory is that the brain accomplishes this by using part the motor system to "mirror" the movements of others. The aptly named Mirroring System allows us to infer the intended outcome of a series of movements (i.e. the "concrete intention", or semantic goal, of the action) by comparing the observed movements with our previous experience with performing or perceiving those same movements (Kilner, Friston, & Frith, 2007; Rizzolatti, Cattaneo, Fabbri-Destro, & Rozzi, 2014). The ability to use our own motor system to understand others' actions seems to develop early in life and allows us to not only understand what we have seen after the action is complete, but also to actively predict the outcome of an action as it is unfolding (Oztop, Wolpert, & Kawato, 2005) using kinematic cues such as velocity (Stapel, Hunnius, & Bekkering, 2015). While it should be noted that the actual "mirroring" properties

Box 1.3 Measuring Brain Function

One way to measure brain activity, which is used in this thesis (**Chapter 3**), is functional magnetic resonance imaging (fMRI). MRI scanners use a strong magnetic field to align the magnetic dipoles in hydrogen particles in the scanned area. Radio-frequency pulses are then used to produce a shift in this alignment. After the pulse, the particles relax back to their aligned positions. The density of hydrogen particles, which is primarily dependent on the tissue being measured, affects the amount of energy given off by the particles as they return to their initial position. This energy emission, in turn, is measured by a conductive coil around the participant's head. While activation of neurons is what we are specifically interested in, MRI captures the amount of oxygen in the blood, which similarly affects energy emission after a radio frequency pulse. Because neural activity requires energy, an increase in the amount of oxygenated blood to a particular brain region is indicative of an increase in neural activation. This response, known as the blood-oxygen level dependent (BOLD) response has proven to be a reliable proxy measure of brain activity, while the 3D images on which it is captured allow a much more detailed investigation of where the activation is occurring.

Typically, when discussing brain activation in fMRI studies, we are modeling some aspect of our stimulus, such as the moment a participant sees an image, as producing an increase in the BOLD response. We then look at the correlation between the expected BOLD response and the actual signal that we are seeing. As an extension of this, we can additionally model other parameters of the stimulus to further specify the model. To use a simple example, we could additionally assume that the brightness of an image influences the BOLD response. This would predict not only a response at each occurrence of the stimulus, but a response that scales with the brightness of the image. By testing this hypothesis in each voxel of the brain, we create a 3D map of the brain regions that respond to a particular stimulus or stimulus quality (e.g. brightness). Similarly, this can be used to find high-level neural architecture, such as conceptual knowledge about an object, regardless of whether we see a picture, hear the name, or see the written name of the object (Simanova, Hagoort, Oostenveld, & van Gerven, 2012). In **Chapter 3** of this Thesis I used this approach to identify brain regions that respond to the "communicativeness" of movement.

While brain activation is a good way to investigate the regions that respond to a particular stimulus or mental process, cognitive functioning is not achieved by separate brain areas. Connectivity, or the exchange of information between these areas, is a vital piece of the puzzle. By looking at the dynamics of how different regions respond at slightly different times or magnitudes, we can model which regions are communicating with one another, and even the direction of information exchange. For example, Dynamic Causal Modeling uses what we know about how neural activation translates into BOLD responses and how neural populations communicate with one another to assess how one brain region may influence another. By adding our experimental inputs (e.g. image brightness) into this model, we can determine how a particular stimulus affects the dynamics of information exchange between particular brain regions. I used this approach to identify the effect of the communicativeness of kinematics on the functional connectivity in **Chapter 3**.



of this system are heavily debated (Hickok, 2013; Vannuscorps & Caramazza, 2016; Wurm & Lingnau, 2015), it is sufficient for the purpose of this discussion that the collection of brain regions termed the Mirroring System do seem to be involved in processing the actions of others, although this may be in a more purely perceptual manner. While the Mirroring System allows us to understand typical actions that we have previously seen or experienced, it may not be able to account for our ability to understand unusual or irrational actions (Van Overwalle & Baetens, 2009). In the case of irrational or inefficient actions, the deviation from efficiency is unexpected. Since we like our environment to be predictable, we must rationalize the observation by making an inference about the person's abstract intentions or mental state.

We often make inferences about the mental state of other people, such as their beliefs, desires, and intentions. This is referred to as mentalizing, or having a "theory of mind" (Nichols & Stich, 2003; Premack & Woodruff, 1978), and actively doing so is associated with activation of a set of brain regions called the Mentalizing System (Frith & Frith, 2006). This network is crucial to our ability to predict and interact with other people as it allows us to re-evaluate their goals or intentions when their behavior is unexpected (Schiffer, Krause, & Schubotz, 2014). The system is also known to respond to unusual actions (Brass, Schmitt, Spengler, & Gergely, 2007). This is an important feature as it allows us to recognize events that may be informative, as novel or unusual events allow us to potentially learn new things about the world around us (Csibra & Gergely, 2009). Returning to the case of unusual actions, the Mentalizing System seems to work in concert with the Mirroring System (Van Overwalle & Baetens, 2009), allowing us to think about why the action was performed in the way that it was. Most previous studies have looked at intention recognition of wholly irrational actions, such as turning on a light switch with one's knee when the hands are free (Brass et al., 2007), but there is also evidence that these systems respond to the efficiency of movement trajectories (Marsh, Mullett, Ropar, & Hamilton, 2014).

Besides responding to unusual or inefficient actions, the Mentalizing System is also activated by overt social signals, such as making eye-contact (Schilbach et al., 2006) or hearing one's name being called (Kampe, Frith, & Frith, 2003). This shows the importance of the system in communication and social interaction, especially in the context of communicative movements. This suggests that this Mentalizing System is sensitive to the high-level properties of a stimulus, such as whether it is familiar or

efficient, and whether it is socially relevant. Such high-level processing is important for processing the relevance of what we are perceiving and could help select the most appropriate behavior based on this high-level interpretation (Wang & Hamilton, 2012). An interesting question is how these high-level properties may be related. Returning to how we learn from novel information, it could be that recognizing efficiency is also part of recognizing social relevance. For example, the Mentalizing System may be responding to the salience, or relevance, of a stimulus based on input from lower levels of processing. For inferring social intentions from movement kinematics, this would likely be done in concert with the Mirroring System.

Although both systems seem to be crucial for understanding the social intentions underlying actions, some studies suggest that the Mirroring and Mentalizing Systems are only concurrently engaged when there is an explicit need to reflect on the intentions underlying an action (Angela Ciaramidaro, Becchio, Colle, Bara, & Walter, 2013; de Lange, Spronk, Willems, Toni, & Bekkering, 2008; Van Overwalle & Baetens, 2009) or when processing very unusual information (Marsh et al., 2014). However, the abovementioned studies typically use actions that are unusual based on their context, such as observing someone lifting an object over an obstacle in one condition, and observing the same movement trajectory when the obstacle is no longer present (Marsh et al., 2014). While this research has provided insights into how expectations can shape our attributions of intention, it is less clear how these brain systems interact when intentions should be inferred from subtle changes in articulators, rather than contextual constraints. An interesting open question is thus whether this interplay between the mirroring and mentalizing systems can support communication by recognizing intentions in the kinematics of an action. If this is the case, then activation of these two systems should be directly related to the extent of communicative kinematic modulation in an observed action or gesture.

1.4 Clarifying Meaning in Movement

Thus far we have mainly discussed communicative signaling in terms of its high-level goal of signaling the intention to communicate. Successful communication requires more than just the recognition that what someone is doing is relevant to you. It also requires you to understand the information that is being transmitted. For example, let us return to the earlier example of your friend raising her glass at dinner. Now let us imagine that she only raises her hand as if she were raising her

glass. Her fingers are shaped as if she were grasping it, but the glass itself is absent. In effect, she is using a representational gesture to signal that you should make a toast. In section 1.2 we discussed how the kinematics of her movements would allow you to recognize that this movement was conveying some information to you (i.e. that it was communicatively intended). In order for this to be meaningful to us, we must also understand *what* she is communicating, which we call the semantic content of the gesture. While we know that gestures and actions made in different communicative contexts are qualitatively different, how kinematic differences affect semantic comprehension has received less attention in the literature.

On the role of kinematics in semantic comprehension, most research has looked at the effect of child-directed actions on learning. Actions produced for children have more repetitions and are more clearly segmented into individual action units (e.g. grasping, moving, lifting; Brand et al., 2002). Later studies showed that these child- or infant-directed actions promote imitation (Williamson & Brand, 2014) and learning, and are preferred by infants (Brand & Shallcross, 2008). One interpretation of these findings is that the increased segmentation of the action allows the individual parts to be more readily recognized, thus allowing the complete action to be recognized.

Complimentary to the child-directed action research, pointing gestures have been used in the field of robotics to understand how we can make robots more understandable or predictable. By testing different kinematic models of pointing gestures, researchers have found an interesting parallel with the communicative kinematic work being done with humans. Specifically, a pointing gesture that is optimized for efficiency (i.e. requiring as little movement as possible) is difficult for a human to interpret. When the kinematics instead are optimized to balance movement efficiency with some exaggeration of the trajectory, the gesture becomes easier to understand (Dragan & Srinivasa, 2014; Holladay, Dragan, & Srinivasa, 2014). More recently, people have also been shown to implement this same type of adjustment, where trajectories are exaggerated in specific ways that allow an observer to better recognize the target of the pointing gesture (Winner et al., 2019). Together, this suggests that communicative actions and gestures may be doing more than just signaling high-level intentions. By segmenting the act into smaller units and exaggerating the relevant features of those units, the deviation from typical kinematics signals the act as being communicatively intended, and the sequence of movements, whether it be action or gesture, becomes easier to understand.

This segmentation and exaggeration framework also fits well with the idea of action perception as hypothesis testing (Donnarumma, Costantini, Ambrosini, Friston, & Pezzulo, 2017). When viewing others, we are constantly trying to predict what we will see next in order to ultimately understand what the person's goal, or intention, is (Cuijpers, Schie, Koppen, Erlhagen, & Bekkering, 2006). We may have some idea about what kinds of actions are possible given the context, or even what we would be doing in this scenario, but this is not enough to know what this particular person is doing right now. However, we can help our predictions by testing hypotheses about what they might be doing as the action unfolds. This can be achieved by directing the eyes to parts of the visual scene that will inform our predictions (Donnarumma, Costantini, et al., 2017). The exaggerations in trajectory described above, for example, support this process. When the trajectory is exaggerated more to the left, we can use this information to predict the outcome even before the movement is complete. To take this one step further, when we recognize that something is intended communicatively, whether due to kinematic modulation, eye gaze, or something else, it draws our visual attention to what is happening. Whether kinematic modulation is able to fulfill this dual role, and how it may interact with other communicative articulators such as eye-gaze, has not previously been investigated. I suggest that the kinematic modulation arising from the intention to communicate is able to clarify meaning by exaggerating salient movement features. This would be a powerful function of communicative movement in cases when speech or gaze are not possible, such as noisy environments, or when eye-gaze is directed elsewhere.

1.5 Lending a Hand to Degraded Speech

Much of what we have discussed in this section has related to clarifying information for children, or programming robot movements to be clearer to us. Similarly, many studies on communicative signaling have used paradigms in which the interacting participants cannot verbally communicate with one another, forcing them to use visual signaling. While this may not seem directly applicable to the typical interactions between adults, we do not always have the luxury of clear communication. In fact, using visual signaling may be especially useful when verbal communication becomes more difficult, for example at a crowded cocktail party.

In many social gatherings, background noise can make it more difficult to understand what your partner is saying. A well-studied effect of such noise is called the Lombard

Effect. The Lombard Effect is the way that we speak louder, elongate our vowels, and increase pitch, which together increase the audibility of our speech (Lombard, 1911; Zollinger & Brumm, 2011). While the effect was originally found in speech, it has since been extended into "visual speech", such as mouth opening, lip movements, and eyebrow movement (Davis, Kim, Grauwinkel, & Mixdorff, 2006; Kim, Davis, Vignali, & Hill, 2005). Whether the effect also extends to co-speech gestures is not known. This is particularly important because listeners benefit not only from the changes in auditory and visual speech, but also from the speakers' gestures (Drijvers & Özyürek, 2017). This situation is interesting because it represents a relatively common social context in which we must communicate, and in which one of our main communicative signals is disrupted. An interesting hypothesis would be that the Lombard Effect would indeed extend to gestures, enhancing the legibility of the kinematics in a similar way to how speech is also made more audible. This can be seen as an extension of our framework of communicative intent, ensuring our gestures are understood regardless of the communicative context (e.g. adult-to-child, noisy environment).

Extending the idea of communicative kinematic modulation to noisy situations is interesting because it does not change the overall relevance of communication. Rather, noise degrades the speech signal for the addressee, but does not disrupt the speaker's ability to gesture, exaggerate lip movements, or use eye-gaze to signal attention. In particular, modulating lip movements and gestures could be a more effective way to compensate for noise, as opposed to exaggerating speech. However, given the biomechanical coupling between gesture and speech (Pouw et al., 2019), and also to lip movements (Woll, 2014; Woll & Sieratzki, 1998), it is also possible that speakers would simply respond with a general increase in communicative effort, which would lead to an overall exaggeration of speech, lips, and gestures. Studying multimodal communication in noise therefore provides a unique and useful way to investigate, at the level of articulators and their interactions, how communicative intention and context affect the way we express what we are trying to communicate.

In sum, in addition to the social context described in previous sections, environmental context, such as background noise, affects our communication. While the effects of a noisy environment on speech and lip movements have been studied, less is known about its effects on gesture and action kinematics. Kinematic modulation of communicative actions could be part of a broader function that allows us to select

and exaggerate relevant information from various articulators and thus enhance communication. If kinematic modulations represent a communicative mode that adapts our behavior to better convey information, then kinematic modulation should play a role when verbal communication becomes difficult.

1.6 State of the Art and Current Contribution

The way we make ourselves understood through non-conventionalized movement has largely been explored in two separate strains of research. On the one side, gesture researchers have explored the role of hand gestures as important components in human communication and shown how different social contexts can influence the way gestures are produced. On the other side, action researchers have explored how information embedded in fine-grained kinematics allows us to signal and understand different action intentions. In fact, both lines of research are investigating movements under different contexts and intentions, but from two different approaches. In this thesis, I try to bridge these different lines of research in order to provide a deeper understanding of communicative movements.

The influence of communicative intentions on movement kinematics, as well as the way observers read abstract intentions from kinematics, has primarily been investigated on simple movements, such as reaching to grasp or pointing. **In Chapter 2**, I extend this research to complex object-directed actions and representational gestures. I use motion tracking to investigate how a communicative intention shapes action and gesture kinematics and test whether these kinematic modulations are sufficient for reading communicative intentions in both actions and gestures.

Previous research has investigated how the brain infers intentions from contextually unusual actions. In **Chapter 3**, I ask how brain dynamics allow the recognition of communicative intentions from movement kinematics alone. I use functional magnetic resonance imaging (see Box 1.3) to measure brain activation and connectivity while participants classified the gestures of stick-light figures (see Box 1.1) as being communicative or not.

In **Chapter 4**, I turn to the semantic side of communicative movements and ask whether communicative kinematic modulation allows an observer to more easily identify the gesture. Previous work has shown that the kinematics of reaching

movements allow one to predict the upcoming action, while features such as size and "punctuality" of an action make it more intelligible to an observer. By selectively showing only segments of a gesture and decreasing the amount of visual information available to the observer, we test the specific role, and timing, of kinematics in supporting comprehension of a representational gesture.

In **Chapter 5**, I look at how we modulate and coordinate multiple bodily signals when interacting in a noisy environment. When faced with a noisy environment, speakers show exaggeration of both acoustic (e.g. intensity and pitch) and visual (e.g. lip movements) features (i.e. the Lombard Effect), and listeners benefit both from these audio and visual changes, but also from the speaker's gestures. Whether the speaker actually modulates their gestures in a similar way as their speech and lip movements has not been investigated. Furthermore, whether this modulation is part of a general increase in communicative effort or a strategic adaptation of the most relevant signals is also not understood. In the experiment described in this chapter we used a dyadic interaction task together with motion tracking and audio recordings to model how speech and gesture come together to support communication in noise.

In **Chapter 6**, I focus on the implications for using motion tracking to study the kinematics of meaningful movements such as actions and gestures. Although motion tracking has been applied to studying motor control and some high-level features of gestures, the many degrees of freedom for analysis have made it difficult to utilize for more complex, naturalistic movements. I provide a framework for quantifying kinematic features that are useful for understanding meaningful human movements, and discuss the implications and possible directions for this line of research in the future.

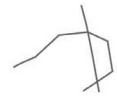
In **Chapter 7**, I bring together the results of all the experiments described in the previous chapters of this thesis and provide a discussion of the implications of this work for models of social interaction and communication, and in the fields of action and gesture research more generally.

Chapter 2

Communicative intent modulates production and comprehension of actions and gestures: A Kinect study

Chapter based on:

Trujillo, J.P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*, *180*, 38-51. https://doi.org/10.1016/j.cognition.2018.04.003



CHAPTER 2

Abstract

Actions may be used to directly act on the world around us, or as a means of communication. Effective communication requires the addressee to recognize the act as being communicative. Humans are sensitive to ostensive communicative cues, such as direct eye gaze (Csibra & Gergely, 2009). However, there may be additional cues present in the action or gesture itself. Here we investigate features that characterize the initiation of a communicative interaction in both production and comprehension.

We asked 40 participants to perform 31 pairs of object-directed actions and representational gestures in more- or less- communicative contexts. Data were collected using motion capture technology for kinematics and video recording for eye-gaze. With these data, we focused on two issues. First, if and how actions and gestures are systematically modulated when performed in a communicative context. Second, if observers exploit such kinematic information to classify an act as communicative.

Our study showed that during production the communicative context modulates space-time dimensions of kinematics and elicits an increase in addressee-directed eye-gaze. Naïve participants detected communicative intent in actions and gestures preferentially using eye-gaze information, only utilizing kinematic information when eye-gaze was unavailable.

Our study highlights the general communicative modulation of action and gesture kinematics during production but also shows that addressees only exploit this modulation to recognize communicative intention in the absence of eye-gaze. We discuss these findings in terms of distinctive but potentially overlapping functions of addressee directed eye-gaze and kinematic modulations within the wider context of human communication and learning.

Introduction

Our hands may be used in a variety of ways to interact with the world around us. Two such interactions are object-directed actions, in which the hands interact with a physical object (e.g., to open a jar), and representational gestures (Kendon, 2004; McNeill, 1994), in which the hands are used to simulate an interaction or visually represent a non-present object (hands move as if opening a jar). What is specific to humans is that both categories of movements can be recruited for the purpose of communication, allowing us to teach through demonstration (Campisi & Özyürek, 2013; Southgate et al., 2009) or convey the intention for an observer to act in response (Tomasello, 2010).

Characteristic of communicative acts is the accompanying addressee-directed eye-gaze (Brand et al., 2007). Humans in particular seem inherently sensitive to ostensive communicative cues, such as direct eye gaze and eyebrow raise (Csibra & Gergely, 2009). Direct eye-gaze is particularly powerful, displaying a willingness to interact (Cary, 1978), as well as altering cognitive processing and behavioral response (Senju & Johnson, 2009). For example, a recent study by Innocenti et al. investigated the impact of eye-gaze on a requesting gesture, e.g. reaching out and grasping an empty glass with the implied request to have it filled. The study showed that both the speed and size of a communicative gesture and addressee-directed eye-gaze affected kinematics of the response act. Therefore, the mere presence of direct eye-gaze induced a measurable effect on the response of the addressee (Innocenti, de Stefani, Bernardi, Campione, & Gentilucci, 2012).

For communication in general, there are at least two main requirements: the communicator must make his or her intention to communicate recognizable, and they must represent the semantic information they wish to be received by the observer (Sperber & Wilson, 1986). The first step in communicating using actions or gestures is thus for the communicator to make the action or gesture recognizable as being a communicative act. In doing so the communicator might use kinematic modulation (see, for example, Becchio, Cavallo, et al., 2012) as well as addressee-directed eye-gaze (Kampe et al., 2003; Schilbach et al., 2006). Secondly the communicator's cues need to be picked up by addressee in order to interpret actions or gestures as communicative. Here, again, both the kinematics of the manual acts and the ostensive cues, or the interaction of both, can play a role. In the present study, we

address the overall profile of communicative actions and gestures within the larger context of production and comprehension. We compare for the first time actions and gestures in communicative versus non communicative contexts to see if they are subject to similar kinematic modulations and are coupled by ostensive cues. We then investigate whether and how these cues are in turn interpreted by addressees. To quantify kinematic modulation effects, we use the Kinect device to obtain a non-intrusive, objective and precise measure of action and gesture.

The next few paragraphs summarize the current literature on the kinematic modulation and on the perception of actions and gestures in communicative context.

Production of communicative actions and gestures

At the basic motor control level, actions are thought to follow a principle of motor efficiency (Todorov & Jordan, 2002). In this framework, control of an action is a balance between reducing cost and achieving the goal of the action. While this framework explains action control in a neutral setting, there is evidence that other contextual or cognitive domains influence these dynamics. The intention to communicate affects the velocity of reach-to-grasp movements (Sartori et al., 2009), and can modulate the trajectory of such movements to make a target more predictable to a co-actor (Sacheli, Tidoni, Pavone, Aglioti, & Candidi, 2013). Furthermore, child-directed communicative actions are marked by several kinematic modulations, including an increased range-of-motion and punctuality (Brand et al., 2002). At the level of cognitive and neural implementation of motor control, this indicates a top-down influence on action production that is theorized to facilitate interactions by balancing the initial efficiency principle with the additional factor of disambiguating the end-goal for an observer (Pezzulo et al., 2013). In line with the account by Pezzulo and colleagues, we suggest that the kinematic modulation from a communicative context can be summarized as an optimization of space-time dimensions (Pezzulo et al., 2013). In this account, communicative modulation is an effort to present the optimal amount of visual information to disambiguate the act (optimization of space) within an efficient amount of time (optimization of time). We extend this framework by investigating specific kinematic cues, and testing the framework in gestures as well as actions

Although the motor efficiency/optimization principle does not specifically refer to gestures, they too are manual acts with a specific extrinsic goal. Often, this goal is to

change the internal state of an observer, but gestures may also be performed without communicative intention. For instance, in the context of co-thought gestures, one uses gestures while trying to solve complex visuospatial tasks (Chu & Kita, 2011). Additionally, clinicians often use pantomime production tasks as a clinical measure in aphasia (Goldenberg, Hartmann, & Schlott, 2003; Hermsdörfer, Li, Randerath, Goldenberg, & Johannsen, 2012). Gestures then are likely to also follow an initial efficiency principle which may further be modulated depending on the goal or intention. Like actions, gestures are also influenced by a communicative context. For example, when meant to be more informative to an observer, pointing gestures are made slower than when the gesture will not be used by an observer (Peeters, Holler, & Hagoort, 2013). Furthermore, during a demonstration or explanation, a gap in common knowledge between speaker and addressee leads to gestures that are larger (Bavelas, Gerwing, Sutton, & Prevost, 2008; Campisi & Özyürek, 2013), more complex or precise (Galati & Galati, 2015; Gerwing & Bavelas, 2004; Holler & Beattie, 2005) and are produced higher in space (Hilliard & Cook, 2016). Whether these kinematic modulations are comparable to those observed in actions in similar communicative settings, has not been assessed.

Perception of communicative actions and gestures

Although communicative intent driven modulation is present during the production of actions and gestures, as shown above, it is less clear whether and how this modulation is seen or used by observers. Studies show that children prefer actions marked by increased range of motion and exaggerated movement boundaries (Brand et al., 2002), which leads to increased visual attention in infants (Brand & Shallcross, 2008), and more frequent imitation of a demonstrated action in children (Williamson & Brand, 2014). In regard to intention recognition, a study on social actions by Manera et al., showed that observers are able to distinguish between cooperative and competitive actions using only the kinematics (point-light-displays; Manera et al., 2011). This suggests that kinematic modulation, at least in regard to child-directed actions and social context, is noticed by observers.

With regard to perception of the communicativeness of gestures, a recent study by Novack et al. shows that movements in the presence of objects are seen as representations of actions, while the *same* movements made in the absence of objects are described as being movement for its own sake (Novack, Wakefield, &

Goldin-Meadow, 2016). This suggests that even though kinematics clearly affects the way the action or gesture is perceived, observers rely strongly on situational constraints to understand the underlying intention. Further evidence comes from a study on body orientation and iconic gesture use (Nagels et al., 2015). Nagels and colleagues found that when a speaker is oriented toward an addressee and gestures during speech, the addressee feels more addressed, thereby indicating a better recognition of communicative intent. Interestingly, both the condition with the speaker orientated towards the addressee but not using iconic gestures as well as the condition with the speaker oriented away from the addressee but using iconic gesture were also rated as being more communicative than the condition in which the speaker faced away and did not use gestures (Nagels et al., 2015). These studies indicate that, at least for iconic gestures, both eye-gaze directed to the addressee and gestures can convey a communicative intent. It is important to note that although iconic gesture use contributed to the feeling of being addressed, the kinematics of gestures themselves were not modified in that study. Therefore, the question remains of how such a modulation will impact the perceived communicativeness of the gesture or the action.

Current study

The current study seeks to link previous findings on communicative manual acts by investigating the characteristic features that facilitate the initiation of a communicative interaction, taking into account both production and comprehension. Specifically, we ask if communicative intent modulates the kinematics of, and eye-gaze behavior accompanying both actions and gestures, and if observers use kinematic modulation and/or eye gaze to recognize the communicative intention of the action and gesture. Previous studies have shown that communicative intent may modulate different aspects of actions and/or gestures. However, these two modalities have not been investigated in a single design, utilizing the same communicative context and considering both production and comprehension. To address these questions, we used two experiments: one for production and one for comprehension.

In the first experiment, two groups of participants performed a set of everyday actions, as well as the corresponding representational gestures. One group of participants performed in a more communicative context, and the other in a less-communicative, or self-serving context. In order to provide a non-intrusive, naturalistic setting, we

did not specifically instruct participants to "be communicative", but used a subtle manipulation of the context in which they performed the task. We used high-definition video recordings for manual coding of eye-gaze behavior. Furthermore, we used the Microsoft Kinect to collect full-body 3D joint tracking data. Use of the Kinect allows tracking of the participants' 3-dimensional movements, allowing streamlined, quantitative coding of kinematic features. We chose this approach as opposed to the more traditionally used optical tracking as the Kinect does not require markers or calibration. This supports the naturalistic aspect of our experiment, while maintaining high quality motion capture performance (Chang et al., 2012; Fernández-Baena, Susín, & Lligadas, 2012). Although relatively new in the field of research, the Kinect has successfully been implemented for gesture (Biswas & Basu, 2011; Paraskevopoulos, Spyrou, & Sgouropoulos, 2016) and sign-language recognition (Pedersoli, Benini, Adami, & Leonardi, 2014) and was shown to be a reliable tool for measuring kinematics. In the second experiment, we showed a selection of single acts to a new set of participants in order to understand how these features are used by an addressee. These participants were asked to classify each act as either communicative or non-communicative. We then assessed which features contributed to an observer's context classification. In the third experiment, the same subset of videos was modified, to obscure the eye-gaze information. The clips were then shown to a group of naïve participants, replicating the Experiment II, to further distinguish relative contribution of the kinematic modulation and eye-gaze in the detection of the communicative intent.

In sum, this study aims to elucidate the profile of communicative action and gesture, and place this profile in the larger frame of production and recognition. We ask which kinematic features are modulated by communicative interactions on the production side, and how this modulation facilitates comprehension of the communicative intent.

Methods – Experiment I

Participants

Forty participants were included in this study, recruited from the Radboud University. Participants were selected on the criteria of being aged 18 – 35, right-handed, healthy and fluent in the Dutch language. Additionally, one confederate also participated in all experiments. The confederate was a 23 year old, female, native Dutch speaker.

The experimental procedure was in accordance with a local ethical committee.

Context Settings

Participants were divided into two groups: more communicative (n =20, 13 females, mean age = 23.6 years) and less communicative (n =20, 13 females, mean age = 23.8 years). For the more-communicative group, the confederate was introduced as having the task of watching the experiment through the camera placed in front of the participant and learning the participant's actions/gestures. In the less-communicative group, the confederate was introduced as having the task of watching the experiment through the camera and learning the general experimental set-up. Critically, this means that in both groups the confederate was considered to be watching and learning, but only in the communicative group was the confederate stated to be learning directly from the participant's manual acts. The paradigm therefore aimed to create a continuum of behavior, extending from less communicative, self-serving behavior, to highly communicative behavior that was highly oriented towards the addressee. This novel paradigm builds on designs using confederates to control feedback while eliciting an interactive setting (eg. Holler & Wilkin, 2011; Sartori et al., 2009). Crucially, our context manipulation aims to influence the intentional stance of the participant towards the addressee, similar to Peeters and colleagues (2013), while keeping all other (e.g. presence of confederate and instructions to participant) factors equal**.** Participants were pseudo-randomly assigned to groups, with consideration only being given to a relatively equal distribution of males and females to each group.

Items

The full set of actions/gestures contained 31 item sets, most of which consisted of two objects. Auditory instructions accompanied each item set and were recorded by a female, native Dutch speaker. Items were presented in random order for each participant and modality (action and gesture). All instructions were similarly constructed in a simplistic way as to indicate the object(s) and a verb (e.g. The participant may be given a pitcher of water and an empty glass, with the accompanying instructions "Giet het water in het glas", *pour the water into the glass*). A full list of the instructions used for these items can be found in appendix 1.

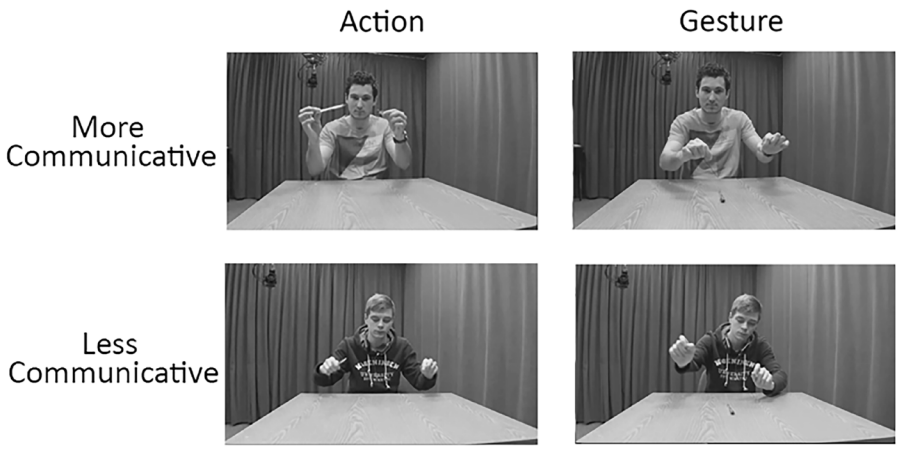


Figure 2. Overview of the experimental design. Each image depicts an example frame taken from a video of the corresponding factor. In each image, the action or pantomime being performed is 'remove the cap from the pen'. The x-axis displays modality (action vs. gesture as a within factor) and the y-axis depicts context (more communicative vs. less-communicative as a between factor).

Modality

Both groups executed the full list of items in each of two conditions, reflecting two modalities of movement: action and gesture. For the action condition, participants were simply instructed to follow the auditory instructions using the items on the table. In the gesture condition, participants were instructed to follow the instructions as if they were using the objects, but without actually touching them. The order of modalities was counterbalanced across subjects. An overview of the design with example frames taken from each factor (modality x context) can be seen in Figure 2.

Procedure

For both groups, we used the following procedure: the participant entered the experiment room and was briefly introduced to a confederate, as described above. After the brief introduction, the confederate moved to an adjoining room. The participant was then seated at a table with a camera hanging directly in front of the table, facing the participant at approximately eye-level. The participant was shown two areas marked on the table to designate the starting point for his/her hands and instructed on the experimental procedure. After asking both the participant and the confederate if they are ready to begin, the door separating the participant from the

confederate was closed. Each item began with (an) object(s) being placed in front of the participant in the middle of the table. After the experimenter was out of sight from the participant, and both hands were resting on the designated starting points, auditory instructions were played indicating what action/gesture should be executed. After the instructions were played, a short interval followed before a bell sound was played, indicating the participant may begin executing the action/gesture. Participants were told that they must not begin acting until they hear the bell sound, at which point the camera would begin recording. When the action/gesture was completed, the participant returned his/her hands to the indicated starting places. At the end of the first block (modality), the experimenter explained the instructions for the second block and again asked for verbal confirmation from the confederate if their task was still going well. After this, the door was again closed and the second block began. During both conditions, after the 10th and 20th item, the experimenter also briefly asked the confederate and the participant if their respective tasks were going well. This was done in order to enforce the idea that another participant was present throughout the experiment. At the end of the second block, the participant was debriefed regarding the purpose of the experiment and the presence of the confederate.

Data collection

In order to optimize and streamline analysis of kinematic features, we employed the Microsoft Kinect V2 to collect 3D joint tracking data. The Kinect utilizes single-camera motion tracking and allows automatic, markerless tracking of 25 joints on the human body. For the purpose of this study, we collected data from all 25 joints, although the hips and legs were not used for any analysis. For a graphic overview of the joints utilized in this study, see figure 3A. Although relatively new in the field of research, studies have shown that the Kinect offers hand and arm tracking performance with accuracy comparable to that of high performance optical motion tracking systems such as the OptiTrack (Chang et al., 2012). Data was collected at 30 frames per second (fps). Film data was collected at 25 fps by a camera hanging at approximately eye-level, directly in front of the participant.

Due to technical problems, Kinect data was not collected for seven recording acquisitions: for one less-communicative and one more-communicative participant no Kinect data was acquired, and for two less-communicative and one more-

communicative participant no Kinect data for the Action modality was acquired.

Data Processing

All kinematic analyses were carried out in MATLAB 2015a (The MathWorks, Inc., Natick, Massachusetts, United States) using in-house developed scripts. To account for the noise inherent in Kinect recordings, we first applied a Savitsky-Golay filter with a span of 15 and degree of 5.

The following kinematic features were calculated individually for each item: *Distance* was calculated as the total distance travelled by both hands in 3D space over the course of the item. *Peak velocity* was calculated as the greatest velocity achieved with the right (dominant) hand. *Maximum amplitude* refers to the maximum vertical height, as indexed by six categories (see supplementary Figure 1.1 for a visual representation of these categories), achieved by either hand in relation to the body. *Hold time* was calculated as the total time, in seconds, counting as a hold. Holds were defined as an event in which both hands and arms are still for at least 0.3 seconds. *Submovements* were calculated as the number of individual ballistic movements made, per hand, throughout the item. Our approach was based on the description given by Meyer and colleagues (Meyer, Abrams, Kornblum, Wright, & Smith, 1988). For a more detailed description of how the individual features were calculated, see Appendix 2.

In order to allow comparisons between items with relatively different kinematic profiles, we first standardized all kinematic features. Each feature was transformed into a z-score, per item, by subtracting the mean (n=40) for that item-feature and dividing by the same item-feature's standard deviation. This allowed us to keep any variability between subjects, while removing between-item variability.

We additionally calculated the overall duration of each item. The duration of the item was calculated as the total time between the beginning and end of the item. The beginning of the item was marked by the bell sound, which indicated the beginning of the trial for the participant, which occurred approximately 500ms before the participant began to move his or her hands from the starting points; the end of the item was defined as approximately 500ms after the participants' hands returned to the starting points, when the second bell sound was played. The 500ms windows before and after hand movements were approximate in nature due to the fact that

they are linked to the bell sound that was manually played by the experimenter. Participants tended to respond approximately 500ms after hearing the sound, but if the participant waited more than 1000ms or less than 250ms, this window was given a duration of 500ms. The bell was likewise played approximately 500ms after both hands were resting on the table, but the duration was set to the bell sound (which could vary due to a variable response by the experimenter) in order to only capture the time-frame within which participants believed they were visible to the confederate. We transformed the durations into z-scores, per item, using the same method as described for the kinematic features.

Eye gaze was manually coded on a frame-by-frame basis using the video annotation software ELAN (www.lat-mpi.eu/tools/elan/). Eye-gaze was coded by taking the amount of time between the beginning and end (as calculated for our duration measure) in which the participant looked directly at the camera, in milliseconds, and divided by the total duration of the item. This provided a general measure of the proportional gaze time, indicating the percent of the overall item duration in which eye-contact was made with the camera. Including the 500ms included before initial hand movement and after final hand movement was done in order to incorporate gaze cues immediately preceding or following an action, during the time in which participants thought they were being observed or recorded.

Data Analysis

In order to determine whether the two contexts could be differentiated on the basis of kinematic features, we performed a mixed-effects logistic regression. This was done in order to incorporate all of the data variance into our analyses. This analysis was performed using R (R Core Team, 2012) and lme4 (Bates, Mächler, Bolker, & Walker, 2014). We created six linear mixed-effects models, each with one of the features of interest (distance, maximum amplitude, submovements, hold-time, peak velocity, gaze) as the dependent variable, with context as a fixed-effect, and a random intercept for the item factor. To test the significance of these models, we used chi-square tests to compare the models of interest with a null model, thereby comparing whether the variable of interest, context, explains significantly more of the variance than the random-intercept-only model. In order to account for potential correlations between kinematic features and eye-gaze, as well as the increased type-I error rate associated with multiple comparisons, we used Simple Interactive

Statistical Analysis (http://www.quantitativeskills.com/sisa/calculations/bonfer.htm**)** to calculate an adjusted Bonferroni correction using the mean correlation between the six tested features (action r= 0.12; gesture r = 0.16), which led to a Bonferroni adjusted alpha value of to $p< 0.011$ for gestures and $p < 0.010$ for actions.

No statistical comparisons were performed between actions and gestures. This is due to the z-transformation of the kinematic values, which normalizes the data between items, but results in similar distributions for actions and gestures. Any difference in the mean of these two distributions is therefore due to an uneven distribution of data around the mean, rather than a difference of the mean itself.

Results – Experiment I

In the action modality, the communicative context was associated with an increased proportion of addressee-directed eye-gaze of 4% ± 0.53% of the total video duration ($\chi^2(1)= 54.61$, $p< 0.001$), as well as an increase of 0.21 ± 0.04 SDs in *distance* ($\chi^2 (1) = 26.94$, $p < 0.001$), an increase of 0.18 ± 0.06 SDs in *submovements* ($\chi^2 (1) = 10.10$, $p = 0.001$) and an increase of 0.16 ± 0.06 SDs in *maximum amplitude* ($\chi^2 (1) = 7.21$, $p = 0.007$) and near-significant increase of 0.11 ± 0.01 SDs in *peak velocity* ($\chi^2(1) = 5.99$, $p = 0.014$. *Hold-time* was not significantly different between the two contexts ($\chi^2(1) = 0.16$, $p = 0.691$). More communicative actions were found to be longer in overall duration when compared to less-communications actions (t (1159.79) = 2.79, $p = 0.005$).

In the gesture modality, the communicative context was estimated to increase the proportion of addressee-directed eye-gaze by 7% ±0.82% of the total video duration ($\chi^2 (1) = 61.01$ $p < 0.001$), as well as *distance* by 0.24 ± 0.05 SDs ($\chi^2 (1) = 19.57$,$p< 0.001$), *peak velocity* by 0.31 ± 0.06 *SDs* ($\chi^2 (1) = 30.97$, $p< 0.001$), *submovements by* 0.28 ± 0.06 SDs ($\chi^2 (1) = 23.36$, $p< 0.001$)and *maximum amplitude* by 0.36 ± 0.06 SDs ($\chi^2(1) = 37.43$, $p < 0.001$). *Hold-time was increased* by 0.12 0.06 SDs, which was not significant with the adjusted alpha threshold ($\chi^2 (1) = 4.42$, $p = 0.011$). More communicative gestures were found to be longer in duration when compared to less-communicative gestures (t (1160.69) = 3.93, $p< 0.001$). An illustrative example of the kinematic profile from sample cases of actions and gestures can be seen in Figure 3, and overview of the eye-gaze and kinematic results can be seen in Figure 4.

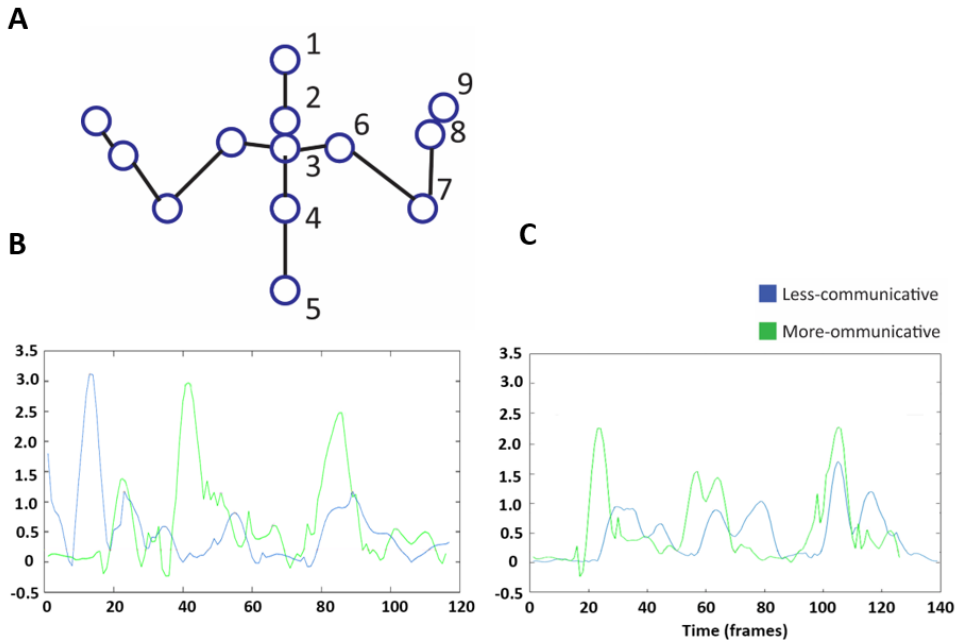


Figure 3. Illustration of the tracked skeleton (A) and comparison of velocity profiles (B,C). Panel A illustrates the joints tracked by the Kinect for analysis of kinematics. The circles represent each individual joint: 1. Top of head 2. Neck 3. Spine – upper 4. Spine – middle 5. Spine – lower 6. Shoulder 7. Elbow 8. Wrist 9. Hand. Panels B and C depict two representative velocity profiles (measured from the right hand), taken from the same item ("Place the apple in the bowl"), shown overlaid for comparison. Panel B depicts items from the Action modality, while panel C depicts items from the Gesture modality. The green line corresponds to a more-communicative act, while the blue line corresponds to a less-communicative act. The x-axis represents time, given in frames. The y-axis represents velocity, given in meters per second (m/s).

Conclusion and Discussion – Experiment I

The aim of our first experiment was to quantify the kinematics and eye-gaze behavior of actions and gestures produced in more or less communicative setting. We found that both modalities were modulated in regards to the overall size, number of submovements, and maximum amplitude, with gestures also showing an increase in peak velocity in the communicative context. Furthermore, both modalities elicited more addressee-directed eye-gaze in the communicative context. We also showed this to be the case for a variety of items.

At a motor control level, actions are performed in a manner that optimally balances the successful completion of the action with energy cost, fine control of the

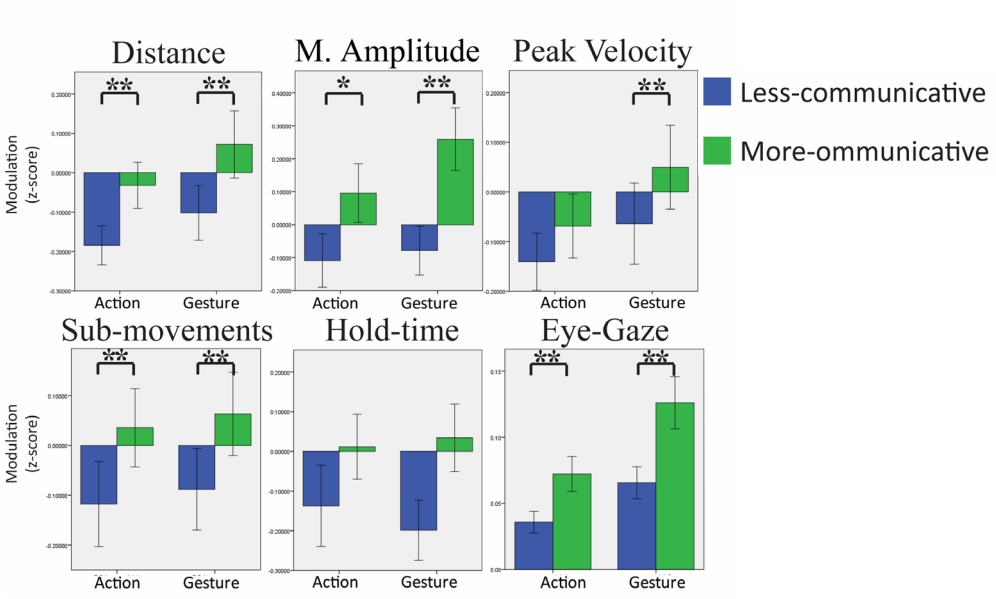


Figure 4. Comparison of more-communicative and less-communicative kinematic features and eye-gaze. Features are displayed in separate plots. Action and gesture are separated on the x-axis. For kinematic features, the y-axis displays the standardized value (positive values therefore indicate higher-than-average features, while negative values indicate lower-than-average features); for eye-gaze, the y-axis represents proportional addressee-directed eye-gaze. Blue bars depict less-communicative average values, while green bars represent more-communicative average values. M. Amplitude = Maximum Amplitude. * $p < 0.05$; ** $p < 0.001$.

movement (Todorov & Jordan, 2002), and environmental constraints (Gergely & Csibra, 2003). Although this explains action control in a neutral setting, previous studies have shown an effect of social context on action kinematics (Becchio, Sartori, Bulgheroni, & Castiello, 2008; Sartori et al., 2009). In these studies, the velocity of movements is differentially modulated dependent on whether or not the actor is attempting to communicate, or whether the action is being performed in a competitive or a cooperative setting. Our findings confirm and expand upon these studies by showing that multiple aspects of movement kinematics are modulated by a communicative context across a wide selection of manual acts. The results indicate a top-down, or context-driven modulation of the motor control system (Friston, 2011).We additionally show that a similar pattern of kinematic modulation is seen both for object-directed actions as well as for the corresponding representational gesture.

Although highly similar, gestures differed from actions in that gestures also had a faster peak-velocity and a subtle increase in hold-time. These features may be more subtle, or they may result from the additional presence of objects during action production, which provides an extra constraint. These two features fit well with the idea of communicative acts being produced with more punctuation, with the difference between modalities suggesting that this may not always be possible when acting with an object. While we cannot test the two modalities against each other, visual inspection of the data (see Figure 4) suggest that modulation may be more pronounced in gestures compared to actions, with vertical amplitude showing the greatest modulation.

We suggested that the communicative context enhances communication efficiency by optimizing space-time dimensions. We found that more-communicative acts covered more visual space and involved more submovements than less-communicative acts, although this was at the cost of requiring more time to produce. The increase in size may optimize the overall amount of information available (i.e. Providing more visual sampling of that movement within the same time-frame), while the increase in submovements may indicate a more detailed representation within the presented information. The fact that these increases are produced at the cost of affecting the overall duration provides support for computational accounts of modulations occurring as an optimization of space-time dimensions (Pezzulo et al., 2013). In other words, the amount of utilized visual space increases, but this is balanced against how much time the overall act requires to produce. This is in line with the rather minimal difference in standardized durations (more communicative actions were 0.15 standard deviations larger than less-communicative actions, while more communicative gestures were 0.22 standard deviations larger than less-communicative gestures). Our finding of a heightened peak-velocity in the gesture modality is also mirrored in a study by Vesper and Richardson, where a cooperative context elicits increased size and peak-velocity during a joint-tapping task (Vesper & Richardson, 2014).This finding can also be interpreted as an optimization of space-time parameters, with the larger movement providing more information and the faster peak-velocity reducing the overall time to produce the act. Although we do not specifically investigate differences between individual manual acts, our study provides experimental evidence that this kinematic optimization may be a signature of more communicative acts in general, regardless of what the specific act is.

Communicative acts are inherently designed for a second person with whom the actor wishes to interact. Although movement kinematics are modulated by the communicative context, it must still be determined what the effect of this modulation is on the observer. For example, although end-goal intentions also modulate the initial phases of an action, a study by Naish and colleagues showed that this information cannot be read by an observer (Naish, Reader, Houston-Price, Bremner, & Holmes, 2013). The role kinematic modulation plays must still be investigated in order to understand their importance in communicative signaling relative to eye-gaze, which is a well-known cue in social interaction (de C Hamilton, 2016),

The aim of our second experiment was therefore to determine if any of the aforementioned features of communicative manual acts are as important for signaling the intention to communicate as addressee-directed eye-gaze. To this end, we used a selection of the videos produced in our first experiment and asked a new set of participants to classify each video as communicative or non-communicative.

Methods – Experiment II

Participants

Twenty participants were included in this study, recruited from the Radboud University. Participants were selected on the criteria of being aged 18 – 35, right-handed, healthy, native Dutch-speakers, and without having participated in the previous experiment. The experimental procedure was in accordance with a local ethical committee.

Materials

Eighty videos (of the 2480) recorded from experiment I were selected for inclusion in this experiment. To provide a representative sampling of each of the two groups, all individual items from all subjects included in the previous experiment were ranked according to eye-gaze and overall kinematics (z- scores). The two groups were ordered such that items in the more communicative context with high communicative context with low eye-gaze and kinematic values were ranked higher than those with low values. This placed all items on a continuum that ranks how representative their features are of their respective groups. This was done due to the observation that, due to the subtle manipulation of context in Experiment I, there was considerable overlap of behavior in the lower ends of each spectrum (i.e. Some participants in

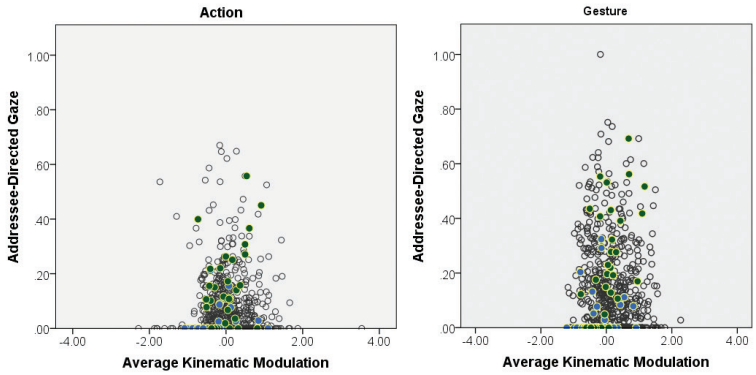


Figure 5. Selection of items used in Experiment II. The left plot shows Action items. The right plot shows Gesture items. In both plots, the x-axis represents the mean modulation of the five kinematic features from Experiment I (distance, maximum amplitude, hold-time, sub-movements, and peak velocity). The y-axis represents proportional addressee-directed eye-gaze. Filled blue circles depict the selected less-communicative items, while filled green circles depict the selected more-communicative items, and empty black circles depict the remaining non-selected items.

the more communicative context showed behavior more similar to those of the less communicative context, and vice-versa). Due to the necessarily restricted number of videos to be included in this experiment, we chose to include items which represented a spectrum of eye-gaze and kinematic features representative of their respective context. It should be noted that although this method allowed a more clear separation of the contexts, our further selection procedure (described below) ensured that items were included across a wide range of this ranked continuum. Included items were therefore not the extreme ends only, as shown in Figure 5.

After creating the ranked continuum of items, inclusion moved from highest to lowest ranked items. Each of the 31 items, as defined in Methods I – Items, was included a minimum of two times and maximum of three times across the entire selection, while ensuring that each item also appeared at least once as an action and once as a gesture, and at least once in more-communicative context and once in the less-communicative context. This was done to ensure an equal representation of each item across modalities and contexts. One action and one gesture video was included from each participant in Experiment I. This ensured that when watching the videos participants of Experiment II would be less likely to learn the context of any given actor (Experiment I participant).

Procedure

Before beginning the experiment, participants were given a brief description of the task in order to inform them of the nature of the stimuli. This ensured that participants knew to expect both actions and gestures, and that this was not relevant for their task. Participants were seated in front of a 24" Benq XL2420Z monitor with a standard keyboard for responses. Stimuli were presented at a frame rate of 29 frames per second, with a display size of 1280x720. During the experiment, participants would first see a fixation cross for a period 1000 ms with a jitter of 250 ms. One of the item videos was then displayed on the screen, after which the first question appeared: "Was the action performed for the actor self or for you?" Participants could respond with the 0 (self) or 1 (you) keys on the keyboard. Actions classified as being performed for the actor self were considered non-communicative, while those classified as being performed for "you" (in this case, the participant) were considered communicative. Immediately after answering, participants received the next question prompt: "How certain are you about your decision?" Participants could then respond with the 0 – 5 number keys, representing a range from "very uncertain" (0) to "very certain" (5), as was also indicated on the screen. After 40 items, participants were informed via the computer screen that they were halfway through the experiment, and were allowed to take a short break if needed. Probe trials were presented every 7 – 9 trials, in which participants were additionally asked what had made them more or less certain about their judgment. For this question, free response typed answers were recorded. These trials were not used for statistical analysis. Context judgments were recorded for each trial, as well as the accuracy of the response.

Data Analysis

Overall performance reflected the accuracy of classifying less-communicative videos as being performed for the actor self, and more-communicative videos being performed for the participant. Before any analyses were performed, we removed outliers in two steps. First, we determined whether there were any participants with outlying performance accuracy, reflected by mean accuracy of less than 2.5 SDs below the mean. After removing any outlying participants, we then calculated mean RT across all participants and excluded any single trials where RT was less than 2.5 SDs below the mean**.** In order to determine the overall accuracy of performance, a

one-sample t-test with test-value = 50 was performed to test if accuracy was greater than chance. Chi-square tests were used to determine if accuracy was equal in both modalities, as well as in both contexts (i.e. To test whether context judgment was more difficult for actions or gestures, or for discriminating one context over the other).

To assess the contribution of eye-gaze and kinematic features to the judgment of communicative context, we performed a two-step linear mixed-effects logistic regression with context judgment as the dependent variable. Before building the models or differentiating between action and gesture, we tested all predictor variables (eye gaze and kinematic features) for multicollinearity by calculating the variance inflation factor (VIF) using the methodology of Zuur and colleagues (Zuur, Ieno, & Elphick, 2010). Predictors with a VIF greater than three were excluded from all subsequent models.

Statistical models were assessed for actions and gestures in order to test for differences in relevant predictor variables, and utilized the modulation values described in Methods – Experiment I, *Data Processing*. We included both correct and incorrect judgments in our statistical model as we were most interested in the perceived context. In the first step of the regression we included eye-gaze as the predictor variable, as eye-gaze is recognized in the literature as a highly salient cue for communication (Csibra & Gergely, 2009). In the second step of the regression model we included all kinematic features that were not previously excluded due to multicollinearity, thereby ensuring the models for action and gesture were alike. We used a likelihood ratio test to compare the two steps of the model, thereby assessing the additional contribution of kinematics to the prediction of communicative context, over and beyond the (expected) contribution of eye-gaze. The contributions of individual predictors (i.e. eye gaze and individual kinematic features) are additionally reported in order to show the relative weight of each predictor in the complete model. Random intercepts were included for actor and item at each step of the model.

Certainty was assessed in two domains: first, the effect of modality and context was determined using Welch's t-tests, as implemented by R. This approach corrects for (potential) inequalities of variance, thereby providing a more robust comparison of the means. Second, the contribution of eye-gaze and kinematic features on an

observer's context judgment was determined using a linear mixed-effects regression. Following the same block procedure as described for the logistic regression we included certainty as the dependent variable, with eye-gaze in a first predictive step of the model and kinematic features (modulation values) in the second step. In order to test the significance of eye-gaze, we again used a likelihood ratio test comparing the model that included eye-gaze as a predictor against the model that only contained the random effects. For these models, random intercepts were again included for actor and item. We additionally modeled random slopes for judgment together with each predictor variable at both steps of the model. This was done because we predict that kinematic modulation and direct eye-gaze are positively associated with judging an act to be communicative, therefore the predictor variables should be positively associated with certainty when the video was judged to be communicative, but negatively associated with certainty when the video was judged to be less-communicative.

Results – Experiment II

One participant was excluded due to outlying classification accuracy, and an additional 43 trials were excluded due to slow RT. Analysis of multicollinearity revealed a VIF of 3.12 for Distance, leading us to discard this feature from all subsequent analyses. After removing Distance, the VIF of all remaining predictors was found to be less than two.

Overall performance in classifying context was 60.86%, which was significantly greater than the 50% chance level, $t(18) = 8.68$, $p < 0.001$. Performance was significantly better in recognizing less-communicative (67% accuracy) compared to more-communicative (57% accuracy) contexts, $t(35.97) = 2.49$, $p = 0.017$. We found only marginally higher accuracy in classifying gestures (M = 62.48%, SD = 0.06) compared to actions (M = 59.20%, SD = 0.08), $t(34.34) = -1.428$, $p = 0.16$.

Eye-gaze was a strong predictor for context judgment in both actions (parameter estimate = 7.87, error = 1.78, z = 4.41, $p < 0.001$) and gestures (parameter estimate = 8.48, error = 1.09, z = 7.72, $p < 0.001$). Adding kinematics did not contribute to the model for actions ($\chi^2(4) = 4.15$, $p = 0.39$) or gestures ($\chi^2(4) = 0.56$, $p = 0.97$). An overview of the model results can be seen in Table 1, including the parameter estimate, the standard error of the estimate, and the associated Z-score of each predictor in the full model. We report here the statistics for eye-gaze from the first

Table 1. Effect of eye-gaze and kinematics on context judgments

	Action			Gesture		
Model Parameter	***Parameter Estimate***	***Std. Error***	***Z***	***Parameter estimate***	***Std. Error***	***Z***
Eye-gaze	7.69	1.62	4.73**	8.31	1.37	6.07**
Max. Amplitude	0.35	0 .21	1.72	0.09	0.19	0.46
Hold-time	0.16	0.19	0.83	0.06	0.13	0.47
Submovements	0.06	0.19	0.31	-0.16	0.23	-0.67
Peak Velocity	0.15	0.33	0.44	0.11	0.13	0.47

*$p < 0.05$, ** $p < 0.01$

step of the model, and the statistics of the kinematics from the second step.

Certainty in the less-communicative context judgments (M = 3.53, SD = 0.69) was not significantly different than certainty in the more-communicative context (M = 3.64, SD = 0.50), $t(32.90) =$, $p = 0.588$. Certainty when judging actions (M = 3.65, SD = 0.56) was not significantly different compared to when judging gestures (M = 3.52, SD = 0.65), $t(35.36) = 0.65$, $p = .529$. In both actions and gestures, eye-gaze showed a linear relation with certainty (action: $\chi^2(3) = 8.17$, $p = 0.043$; gesture: $\chi^2(3) = 17.80$, $p < 0.001$), with increased direct eye-gaze changing certainty by 0.16 ± 1.65. This change was positive or negative depending on whether the video was judged to be communicative or non-communicative (see Supplementary Figure 2.2). Including kinematics did not significantly improve this model for actions ($\chi^2(16) = 6.86$, $p = 0.976$) or gestures ($\chi^2(16) = 2.97$, $p = 0.999$).

Conclusion and Discussion – Experiment II

A communicative context is dependent upon interaction, and thus recognition of the communicative intention by the addressee. We therefore sought with our second experiment to examine the role of communicative acts from the standpoint of the addressee. The optimality principle of motor control (Todorov & Jordan, 2002), together with that of contextual efficiency (Gergely & Csibra, 2003), suggests a dynamic (i.e. variable), yet effectively constrained system of action production. We suggested that a deviation from these efficiency principles would be noticeable by an observer, and thereby used as a signal of intention.

Contrary to our initial hypothesis we found that kinematics do not contribute to an

observer's recognition of communicative intent. Instead, observers rely much more on addressee-directed eye-gaze. Our second experiment therefore lends additional evidence to the idea that eye-gaze cues may be the most important indicator of communicative intent for the addressee (Csibra & Gergely, 2009). Although the suggestion that intention can be read from kinematics (Ansuini, Cavallo, Bertone, & Becchio, 2014; Becchio, Manera, et al., 2012) finds support in the literature, it may be that eye-gaze is such an important cue for recognizing intentions that it overrides kinematic information when both are available. Rather than an interaction, the two cues may alternatively be seen as a hierarchy with regard to cue importance. To test this assumption, we conducted the third experiment to determine whether detecting of intentions from kinematics could be limited to a particular modality (actions or gestures), or to situations where eye-gaze information is unavailable.

Methods - Experiment III

Participants

Twenty naïve participants were included in this study, recruited from the Radboud University. Participants were selected on the criteria of being aged 18 – 35, right-handed, healthy, native Dutch-speakers, and without having participated in either of the previous experiments. The experimental procedure was in accordance with a local ethical committee.

Materials

The same selection of videos was used as in Experiment II, but with the faces of the actors obscured in order to remove the possibility of using eye-gaze information. In order to obscure the faces, we utilized the Mosaic feature in Adobe Premiere Pro to create a pixilated oval (pixel size = 80 x 80) which covered the entire face in each of the videos.

Procedure and Data Analysis

Experimental procedure and data analysis were carried out exactly as in Experiment II. Despite the fact that the faces were blurred and the eye-gaze information was therefore obscured, we included eye-gaze as the first step in each of our models to ensure comparability between the models in the Experiments II and III.

Table 2. Effect of (non-visible) eye-gaze and kinematic modulation on context judgments

		Action		*Gesture*		
Model Parameter	***Parameter Estimate***	***Std. Error***	***Z***	***Parameter estimate***	***Std. Error***	***Z***
Eye-gaze	0.05	0.99	0.05	0.13	0.84	0.16
Max. Amplitude	0.19	0.13	1.50	0.31	0.12	2.59**
Hold-time	-0.14	0.12	-1.13	0.47	0.08	0.57
Submovements	0.15	0.12	1.23	0.17	0.15	1.14
Peak Velocity	-0.16	0.20	-0.77	0.01	0.12	0.07

*p < 0.05, ** p < 0.01

Results – Experiment III

Due to a technical issue, 40 trials from one participant were lost from the initial dataset. One participant was excluded due to outlying performance accuracy and an additional 26 trials were removed due to outlying RT. Multicollinearity tests revealed distance to have a VIF of 3.21, leading us to exclude it from further analyses. After removing distance, the remaining predictors had VIFs of less than two.

Overall accuracy of context judgments was 52.47%, which was significantly above chance level, $t(18) = 2.99$, $p = 0.008$. We found no difference in accuracy when judging communicative (M = 51.61%, SD = 0.05) compared to less communicative (M = 53.18%, SD = 0.06) videos, $t(36.49) = 0.82$, $p = 0.419$. We similarly found no difference when judging actions (M = 52.51%, SD = 0.06) compared to gestures (M = 52.44%, SD = 0.05), $t(35.94) = 0.04$, $p = 0.967$.

In actions, direct eye-gaze was not associated with context judgment (z = 0.05, $p = 0.962$), while kinematics contributed to a near-significant increase in the model fit, $\chi^2(4) = 9.42$, $p = 0.051$. In gestures, direct eye-gaze was associated with context judgment (z = 2.09, $p = 0.035$), and kinematics contributed to a significant improvement to the model, $\chi^2(4) = 10.57$, $p = 0.032$. An overview of the parameter estimates, standard error, and z-scores for each predictor in the full model can be seen in Table 2.

When judging actions, direct eye-gaze did not influence certainty ($\chi^2(3) = 5.09$, $p = 0.165$), nor did kinematics ($\chi^2(16) = 7.42$, $p = 0.964$). In gestures, we similarly found no association between direct eye-gaze and certainty ($\chi^2(3) = 6.01$, $p = 0.111$), nor

did kinematic modulation significantly predict certainty ($\chi^2(16) = 8.22$, $p = 0.942$).

Conclusion and Discussion – Experiment III

The results of this study show a marginally better-than-chance recognition of more- compared to less-communicative actions and gestures, and also indicate that both modalities (actions and gestures) and contexts (more- and less-communicative) are recognized with similar levels of accuracy. We further show that while eye-gaze was not associated with context judgments in either modality, increased kinematic modulation was predictive of gestures being judged as more-communicative. Specifically, increasing maximum amplitude of a gesture leads to it being perceived as more communicative.

The lack of association between eye-gaze and context judgment in actions was expected, as eye-gaze information is not available to the participants in this experiment. That we found this association in gestures may be due to the generally increased direct eye-gaze in the more communicative setting, which could naturally lead to this association arising even when the information is not available. As the association is no longer present in the full model, this result suggests that kinematics contribute more to the model than eye-gaze. That the action modality did not show this effect may be due to the relatively low accuracy overall, which would obscure any natural association. This is also evident in the lack of association between kinematics and context judgment. This suggests that judging the action videos may have been more based on chance, rather than using specific kinematic or gaze features. In gestures on the other hand, we see a strong relation between increased maximum amplitude and a higher rate of being perceived as more-communicative. That this effect is present in the gesture modality, despite low accuracy, also suggests that participants were more receptive to the kinematic modulation in gestures, and more readily interpreted them as communicative. Although speculative, this would be in line with theories by Goldin-Meadow and colleagues suggesting that gestures have a special role in communication (Goldin-Meadow, 2017), and as such may be more likely to be interpreted as intended for someone besides the actor (Novack et al., 2016).

These results highlight the difficulty of recognizing communicative context from kinematics alone. However, the results also indicate that, at least in gestures, kinematic modulation may play a role in guiding this recognition process. This result

is intriguing given that the kinematic modulation in the present stimuli set was highly subtle, with a large overlap between the less- and more-communicative contexts. Future studies will therefore be needed to explore the influence of kinematic modulation on the recognition of communicative intent.

General Discussion

In this study we set out to characterize the initiation of a communicative interaction in both production and comprehension. To do this, we first used motion-tracking and automatic feature calculation to quantify spatial and temporal kinematic features and accompanying eye-gaze behavior of communicative actions and gestures (production), and then assessed the contribution of kinematic modulation and addressee-directed eye-gaze to the judgment of communicative context by addressees (comprehension). Overall, our results show that space-time dimensions of both action and gesture kinematics are modulated by a communicative context. Addressee-directed eye-gaze is also increased in the communicative context and is the best determinant of an observer's classification of an act as being communicative, although kinematic modulation plays a role when eye-gaze information is unavailable.

Results from our first experiment showed that in a more communicative context both actions and gestures are made larger, with greater vertical amplitude and with a more complex movement pattern when compared to a less-communicative context. Additionally, we find increased addressee-directed eye-gaze in the more-communicative context. This finding is in agreement with previous studies showing increased addressee-directed gaze in more communicative contexts, and further supports the notion that this effect is not simply reliant on the participant being watched (as was true in both the more- and less-communicative contexts of our experiment), but that it is directly related to the communicativeness of the context. Our finding of kinematic modulation is in line with research on infant-directed gestures. Infant directed actions show evidence for 'motionese', a form of kinematic modulation which is argued to help sustain attention in infants as well as to make action intentions more legible (Brand et al., 2002). Specifically, this kinematic modulation includes a greater range of motion as well as increased 'punctuality', a qualitative measure of fluid versus segmented movement. While range of motion can be seen as a parallel of the distance measure in our study, punctuality may also reflect our quantification of submovements and holds. We similarly found more

submovements and, at least in gestures, a trend-level increase in communicative holds, which may reflect the more segmented movement profile described by Brand and colleagues. This similarity provides support for our results, as motionese can be seen as an exaggeration of communicative gestures in general. Our finding of kinematic modulation may therefore be a functionally similar exaggeration. For communication with adults, we exaggerate the kinematics of our movements; for communication with children, we exaggerate kinematics even more. In addition to showing that this exaggeration occurs in both actions and gestures, we additionally expand the fundamental framework in which these modulations can be seen by proposing that kinematic modulation is an extension of motor efficiency that optimizes the space-time dimension of communicative acts. This work therefore bridges earlier behavioral studies (Brand et al., 2002; Campisi & Özyürek, 2013) with computational models (Pezzulo et al., 2013) using modern motion tracking and automatic feature quantification to define specific kinematic features relating to the spatial and temporal characteristics of actions and gestures.

Results from our second experiment showed that addressee-directed eye-gaze remains the most salient cue for recognizing an act as being communicative. While previous studies have suggested that a communicative intention can be read from kinematics (Ansuini, Santello, Massaccesi, & Castiello, 2006; Becchio, Manera, et al., 2012), our study suggests that kinematics are not a primary source of information for this classification.

Our third experiment attempted to disentangle eye-gaze from kinematics by occluding facial information. Results from this experiment showed that, at least in gestures, spatial information can act as a cue to communicative intent. Although the correlation between kinematic features and intention recognition did not hold for actions, we speculate that this may be related to the magnitude of the effect. Upon visual inspection of the production data from Experiment I, vertical amplitude is the most strongly modulated kinematic feature, and this appears more pronounced in gesture than actions. Similarly, vertical amplitude in gestures is the only feature that is found to be a significant predictor of intention recognition in Experiment III. As eye-gaze is known to have a strong impact on attention and cognitive processing (Calder et al., 2002), these results suggest that kinematics are simply lower in a hierarchy for intention recognition. The dominance of eye-gaze as a signal for communicative intention does not mean kinematic modulation is entirely useless to the addressee,

as it can also be used as a cue for intention when more primary social cues are obscured. However, the primary role of kinematic modulation may lie elsewhere in the communicative interaction.

Communication requires both the recognition of the intention to communicate as well as comprehension of the semantic content being conveyed. We suggest that kinematic modulation occurs in order to enhance the saliency or legibility of the semantic content being communicated (i.e. the specific movements or their meanings). In this view, eye-gaze signals the intention to communicate, while the kinematics are modulated in order to make the message more easily understood. While speculative, this theory is in line with the interpretation of kinematic modulation in motionese as enhancing action legibility (Brand et al., 2002). In this view, larger, more punctuated actions are thought to make the semantic content more legible. Although legibility was not directly tested by Brand et al., later studies showed that mothers begin exaggerating their action kinematics when infants are capable of learning the action (Fukuyama et al., 2015), infants prefer watching actions featuring motionese (Brand & Shallcross, 2008), and children are more likely to reproduce these actions (Williamson & Brand, 2014). Furthermore, studies in joint actions in adults also reveals actions that direct the attention of the addressee to a certain object using “an exaggerated manner or conspicuous timing” (H. H. Clark, 2005) which may be analogous to spatial and temporal modulation of kinematics. Robotics research, which combines theory-based robotic production of gestures or actions with validation through human comprehension experiments, supports the notion that exaggeration of kinematics improves semantic interpretation of a manual act (Dragan & Srinivasa, 2014; Holladay et al., 2014). This theory has also been explored in the framework of computational modeling, where movement trajectories are modulated to disambiguate the end-goal (Pezzulo et al., 2013). Together, these findings suggest that kinematic modulation may play a role in learning and communication when semantic content needs to be made clear. By modulating the kinematics to be optimally unambiguous, the communicator is thus able to optimize the space-time dimensions of the interaction.

On the other hand, eye-contact is a strong social cue (Calder et al., 2002; Senju & Csibra, 2008) that initiates a pedagogical stance even early in life (Csibra & Gergely, 2009; Senju & Csibra, 2008; Williamson & Brand, 2014). Although cognitively separate from the processing of action semantics (Rizzolatti et al., 2014), this

initiation of interaction may therefore be necessary to prepare the addressee to benefit from kinematic modulation. We speculate that kinematic modulation likely serves another purpose in human communication, i.e., to enhance the saliency or legibility of the semantic content being communicated, but can also serve as a cue for intention recognition when more primary cues, such as eye-gaze, are not available. Future studies are, however, needed in order to bring further light to this hypothesis.

Strengths and Limitations

Our study provides novel insights into the kinematics of communicative actions and gestures. Using robust motion-tracking technology we were able to automatically quantify several kinematic features, which relate to different spatial and temporal components of the act's kinematic profile. This lends precision to our results and may provide a framework for future studies examining kinematic features of actions or gestures. Furthermore, the naturalistic elicitation of more- and less- communicative contexts provides ecological validity to our results, in that participants performed ordinary, everyday acts, such as pouring water or slicing bread, without the use of physical markers being placed on the body. Our study is also the first to examine actions and gestures within the same framework of communicative contexts and manual acts, providing a novel investigation of the similarities and differences between the two modalities. Especially in regard to using the same manual acts in both communicative contexts, we are able to attribute kinematic differences to the context itself, while avoiding differences due to different motor end goals intentions (van Elk et al., 2014). Finally, the relatively large sample size (n = 40) and variety of action/gesture pairs used (n = 31) provides evidence for the external validity of our findings.

While the naturalistic setting of our study provides ecological validity, we recognize that this comes at the cost of some control over experimental variables. As participants were never specifically asked to be communicative, we rely on the assumption that the subtle manipulation of instructions elicited genuinely communicative behavior. Given the significant performance in context judgment in the second experiment, however, we believe that our context distinction is valid. Lastly, our study was limited in its ability to directly compare actions and gestures statistically due to the methodology used. While this methodology allowed investigation of many different

acts, and thus allows generalization of these findings to other acts, it also hindered us from making between-modality comparisons. The difference in significant results between actions and gestures, however, allows some conclusions to be drawn regarding the differences in kinematic modulation. Finally, the subtle elicitation of the more communicative context may have led to kinematic differences between the two contexts that are difficult to entirely separate.

Conclusion

In summary, we examined the features characterizing the initiation of a communicative interaction, examining both the production and comprehension of actions and gestures. We found that a communicative context elicits kinematic modulation of both actions and gestures, together with an increase in addressee-directed eye-gaze. While eye-gaze strongly contributes to the recognition of communicative contexts, kinematic modulation only serves this purpose in gestures when eye-gaze information is unavailable. We suggest that eye-gaze is primarily responsible for initiating the interaction, while kinematics may contribute to enhancing the legibility of the movement, potentially facilitating transmission of the semantic content of the communicative act.

Acknowledgements

We are grateful to Samantha Aarts for her role as confederate in the first experiment and to Ksenija Slivac for assistance in data collection during the second experiment. This research was supported by the Language in Interaction Gravitation Grant. The authors declare no conflict of interest in this study.

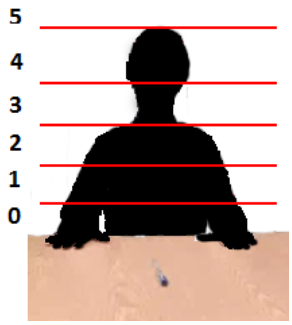
Appendix 2.1. Production instructions

Original (Dutch)	***English***
doe de apple in de kom	Place the apple in the bowl
borstel je haar met de borstel	Brush your hair with the brush
veeg het papier af	Brush off the paper
kreukel het papier	Crumple the paper
snij het brood met de mes	Cut the bread with the knife
knip het papier doormidden	Cut the paper in half
wis de figuur met de gom	Erase the figure with the eraser
vouw het papier doormidden	Fold the paper in half
sla de spijkers met de hammer	Hammer the nails with the hammer
meet het papier met het meetlint	Measure the paper with the measuring tape
open het potje	Open the jar
open het slot met de sleutel	Open the lock with the key
pel de banaan	Peel the banana
doe het dopje op de pen	Put the pencap on the pen
giet het water in het glas	Pour the water in the glass
doe de hoed op	Put on the hat
doe de ring aan	Put on the ring
verwijder het kurkje van de fles	Remove the cork from the bottle
verwijder het dopje van de pen	Remove the pencap from the pen
schrob het bureau met de spons	Scrub the desk with the sponge
schud de kaarten door elkaar	Shuffle the cards
pers de citroen uit	Squeeze the lemon
stapel de blokken op elkaar	Stack the blocks on top of each other
stempel het papier	Stamp the paper
niet de papieren samen	Staple the papers together
dompel het theezakje in het water	Steep the teabag in the water
roer de thee met de lepel	Stir the tea with the spoon
doe de zonnebril op	Put on the sunglasses
scheur het papier doormidden	Tear the paper in half
gooi de dobbelstenen	Roll the dice
schrijf je naam op het papier met de pen	Write your name on the paper with the pen

Appendix 2.2. Calculation of kinematic features

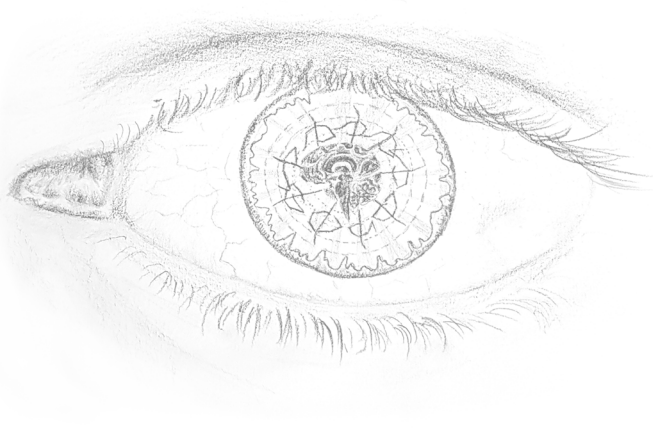
Spaces for the Vertical Amplitude feature were dynamically defined in equal distances between the midline of the torso, base of the neck, and top of the head at each frame of acquisition. This yielded a total of 5 heights that were dependent on the height of the participant and their current body position. For a visual depiction of the spaces defined, see Supplementary figure 2.1.

Submovements were defined by using the velocity profile of a given hand. Following the description by Meyer and colleagues (Meyer et al., 1988), submovements were operationalized as movements that exceed a given velocity threshold, with the beginning and end marked by either the crossing of a near-zero velocity threshold (going from static to moving) or showing a secondary acceleration (reversal from deceleration to acceleration). We used a standard peak analysis to determine the total number of peaks within the velocity profile of each hand that can be considered submovements. For our study, we assigned a minimum velocity threshold of 0.2 meters per second, a minimum distance between peaks of 8 frames, and a minimum peak height and prominence of 0.2 meters.



Supplementary figure 2.1. Visual representation of *Vertical Amplitude* feature, as calculated in reference to a participant's skeleton using the Kinect. Red lines indicate the cut-off points (approximated for illustration), with the numbers on the left indicating the value assigned to the space between the upper and lower lines. Note that 0 is bounded by the table, while 5 has no upper bound and is therefore bounded by the participant's maximum arm extension.

Supplementary Figure 2.2. Judgment-specific slopes for correlation between direct eye-gaze and certainty. The left panel shows the fit lines for the action modality, while the right panel shows the fit lines for the gesture modality. Blue lines depict judgment as being less-communicative (intended for actor), while green lines indicate judgment as being more-communicative (intended for viewer).



Chapter 3

Seeing the unexpected:

how brains read communicative intent through kinematics

Chapter based on:

Trujillo, J.P., Simanova, I., Özyürek, A., & Bekkering, H. (2019). Seeing the unexpected: How brains read communicative intent through kinematics. *Cerebral Cortex*, in press.

Abstract

Social interaction requires us to recognize subtle cues in behavior, such as kinematic differences in actions and gestures produced with different social intentions. Neuroscientific studies indicate the putative mirror neuron (pMNS) in the premotor cortex and mentalizing systems (MS) in the medial prefrontal cortex support inferences about contextually unusual actions. However, little is known regarding the brain dynamics of these systems when viewing communicatively exaggerated kinematics.

In an event-related fMRI experiment 28 participants viewed stick-light videos of pantomime gestures, recorded in a previous study, which contained varying degrees of communicative exaggeration. Participants made either Social or Non-Social classifications of the videos. Using participant responses and pantomime kinematics we modeled the probability of each video being classified as communicative. Inter-region connectivity and activity was modulated by kinematic exaggeration, depending on the task.

In the Social Task, communicativeness of the gesture increased activation of several pMNS and MS regions and modulated top-down coupling from the MS to the pMNS, but engagement of the pMNS and MS was not found in the Non-Social task. Our results suggest that expectation violations can be a key cue for inferring communicative intention, extending previous findings from wholly unexpected actions to more subtle social signaling.

Introduction

In order to successfully interact with others, it is important to understand their social and communicative intentions. The human brain is remarkable in its ability to attribute goals and intentions to actions, allowing us to interpret not only what a person is doing (i.e. the concrete intention) but why they are doing it (i.e. the abstract intention; Van Overwalle, 2009). For example, as a customer lifts a glass the waiter can predict whether the customer is going to drink from the glass or uses this act as a request to have another drink. In this example, the social or communicative intention of the actor must be quickly read from their motor behavior (Blakemore, &Decety, 2001). An interesting question is how the brain picks up on the subtle, socially relevant modulation of the motor act to accomplish this abstract intention reading.

Previous research suggests that humans modulate the kinematics of their movements based on high-level, abstract intentions (Becchio, Manera, et al., 2012; Pezzulo et al., 2013). For example, when an object-directed action is produced with a communicative intention, the kinematic profile of the action is quantitatively different from when the same action is produced without or with a different degree of communicative intention (Campisi & Özyürek, 2013; Sartori et al., 2009). In a previous behavioral study we quantified the differences in kinematics of motor acts produced in a more- compared to less-communicative context. We found that in actions and gestures both spatial and temporal kinematic features were modulated, becoming more exaggerated in the more-communicative context (Trujillo, Simanova, Bekkering, & Özyürek, 2018). Furthermore, we found that observers were able to read this communicative intent from the actors' movement kinematics (Trujillo et al., 2018). These results are well in line with previous suggestions that humans are able to use differences in kinematic profiles in order to infer an underlying intention (Becchio, Manera, et al., 2012).

The ability to read intentions from movement kinematics has been shown both for concrete end-state intentions, e.g. grasp to drink versus grasp to pour (Becchio, Koul, Ansuini, Bertone, & Cavallo, 2018; Cavallo et al., 2016) as well as for more abstract social intentions, e.g. engaging in a social task, (Manera et al., 2011; Trujillo et al., 2018). It has been suggested that the end-state intentions may be read by directly mapping the kinematics onto actions in our own motor repertoire (Blakemore &

Decety., 2001; Cavallo et al., 2016; Rizzolatti et al., 2014). While direct mapping could work for concrete (action end-state) intentions, it is less clear how we read more abstract (i.e. high-level) social intentions that may not have a direct mapping. Abstract intentions are more difficult due to the necessity of having a mapping of all potential socially modulated forms of every action.

A potential solution is to infer intentions based on whether the action follows a typical, expected kinematic pattern or not. This follows from literature describing how we ascribe high-level intentions to movements that are otherwise unusual or implausible, given the context, as a way to rationalize them (Brass et al., 2007; Csibra & Gergely, 2007; Gergely & Csibra, 2003). For example, when we see someone activating a light switch with their knee, we may rationalize this as being due to their hands being occupied by a heavy stack of books (Brass et al., 2007). In this way we explain away the unusual movement as being due to the observable context. In the case of communicatively intended acts, the exaggerated kinematics would be inconsistent with how an observer expects the action to be produced according to previous experience, resulting in the observer attributing a more abstract intention to the actor. This is consistent with the theory of sensorimotor communication (Pezzulo et al., 2013), which suggests that movements can be made communicative by deviating from the most optimal way of performing the action. This also fits with previous results showing that kinematically inefficient movements are seen as unexpected (Hudson et al., 2018). This framework would predict that we do not understand by mapping the observed kinematics to our own motor system, but rather actively infer a hidden intention that would explain the unusual movement.

In the brain, processing abstract intentions typically involves the mentalizing system (Angela Ciaramidaro et al., 2013; Frith & Frith, 2006; Kampe et al., 2003; Spunt, Satpute, & Lieberman, 2011). At the same time, a meta-analysis by van Overwalle & Baetens suggests that the brain likely utilizes the motor system to understand what the observed action is together with the mentalizing system to process the intention (Van Overwalle & Baetens, 2009). This is especially important when considering the case of communicative kinematic modulation. If we are to read the underlying intention from kinematic modulation alone, we must first recognize that the action is being performed in an unusual or exaggerated fashion. Recognizing the act as unusual likely involves the putative mirror neuron system (pMNS; Newman-Norlund, Van Schie, Van Hoek, Cuijpers, & Bekkering, 2009) attempting to match the observed

action with one already in the observer's motor repertoire (Kilner et al., 2007). The exaggerated kinematics would therefore elicit a breach of expectation, resulting in the recruitment of the mentalizing system (MS) to process the underlying intention that generated the unusual behavior (Brass et al., 2007; de Lange et al., 2008; Schiffer et al., 2014). The recruitment of the pMNS and MS in response to unusual movements and the reading of intentions has been shown previously, utilizing movements that are unusual given their end-goal (e.g. using one's knee to activate a light switch) and context (e.g. whether one's hands are free). Distinctly unusual kinematics, specifically in terms of movement trajectory, have also been shown to recruit pMNS and MS regions (Marsh, & Hamilton, 2011; Marsh et al., 2014). This suggests that observers are sensitive to the rationality or efficiency of movement, and unexpected kinematics may lead to intention inferences. However, these studies did not explicitly test whether brain response scales with unexpectedness or inefficiency of the movement kinematics.

Here, we specifically investigate the question of whether a difference in the intention to communicate can be recognized from the kinematics provided. As kinematic modulation is a relatively subtle intentional signal based purely in movement, testing the recruitment of the pMNS and MS in recognizing abstract intention provides a direct test of this model of intention reading.

Processing of abstract intentions in the pMNS and MS is likely achieved via an interaction between the two systems. This is because the two systems are often not activated concurrently. Instead, studies of intention recognition often show activation of either the pMNS or the MS, but not both for the same task, suggesting that information likely flows from one to the other when both are needed. The results from van Overwalle & Baetens (2009) seems to suggest that this process would be bottom-up, with the pMNS influencing the MS when breaches of movement expectation are encountered. In this framework, expectations originate in the premotor cortex, and the MS is recruited to resolve these breaches of expectation. An alternative account is the predictive coding framework (Kilner et al., 2007). This framework suggests that high-level expectations, originating in this case in the MS, might influence lower-level expectations, such as movement expectations (Ondobaka, De Lange, Wittmann, Frith, & Bekkering, 2015). Although the theoretical framework of predictive coding computationally predicts bidirectional influence (i.e. top-down and bottom-up), experimental work seems to primarily find top-down

modulation (Chambon et al., 2017; Chennu et al., 2016). This is particularly the case when participants are actively attending to the unexpected stimulus (Chennu et al., 2016). This would argue for a stronger top-down influence, with the MS primarily influencing the pMNS. This account is supported by findings from studies of perceptual breaches of expectation, where unexpected changes in auditory stimuli (Chennu et al., 2016) as well as the processing of more abstract intentions (Chambon et al., 2017), result in modulation of top-down connectivity strength. It is therefore necessary to investigate directional connectivity in order to understand how the two systems interact when reading abstract (e.g. communicative) intentions from movement.

An important aspect of previous studies on intention recognition is the role of context. For example, in the study by Brass and colleagues (Brass et al., 2007) the unusual action of turning on a light switch was informed by the presence of a stack of folders that the actor was holding. The act itself was of course unusual due to the effector used (i.e. the knee, rather than the hand) to complete the action. Similarly, intention may be largely inferred from the combination of action and object. For example, picking up an apple and extending it towards the viewer is likely to be seen as communicatively or socially intended, whereas picking up a book and opening it directly in front of one's self is seen as privately or personally intended (Ciaramidaro et al., 2007). In order to understand how kinematics can inform intention recognition we must therefore disentangle subtle, communicatively intended kinematic modulation from other visual contextual cues.

Finally, it is important to address the effect of exogenous cues on intention recognition. While it is clear that observers can read even abstract intentions from movement kinematics, this inference on the underlying intention is not likely to be actively made under all circumstances (de Lange et al., 2008; Spunt & Lieberman, 2013). Instead, intention inferences may only be made when it is task-relevant. However, it is possible that the brain responds in a similar way even when the intention is not being attended. Therefore, testing whether activation and connectivity changes are dependent on the presence of explicit task instructions would indicate whether the brain responds implicitly to communicative cues in movement kinematics.

Current study

This study aims to determine the neural systems and mechanisms underlying the

recognition of communicative intention at the level of movement kinematics. Particularly, we test whether 1) communicative kinematic modulation results in activation of the pMNS and MS and 2) determine whether there is evidence for a top-down or bottom-up interaction between the systems. We additionally will determine whether there is evidence for implicit processing of abstract intentions from kinematic modulation alone. We further build on previous studies by investigating whether this neural mechanism of intention inference also holds for more complex movement sequences such as representational gestures (i.e. movements that visually simulate a manual action).

We will address these issues using two forced-choice gesture viewing tasks during functional magnetic resonance imaging (fMRI). In the two tasks, participants viewed stick-light figures created in a previous study where we measured the kinematics of more- and less-communicative gestures (Trujillo et al., 2018). In one task, the Social Task, participants were asked after each video if they believe the action being depicted in the video was intended for the actor or the viewer (representing more-communicative and less –communicative intentions). In the other task, the (Non Social) Handedness Task, participants saw the same videos but were asked to decide whether the action being depicted was performed with the left hand or the right hand. Using participant responses, we calculated the average perceived communicativeness of the kinematic modulation in each of the videos. By correlating this value with fMRI BOLD response, we calculated the extent to which brain activation increases with increasingly communicative kinematics. We therefore use kinematics to provide an extension of the abstract intention inference model beyond the perception of purely categorical, contextually embedded stimuli. We further specify the model by assessing whether communicative kinematic modulation affects top-down or bottom-up information flow between the systems (effective connectivity analysis). Finally, as a secondary analysis, we use the Handedness Task to determine whether the neural response to communicative kinematics is dependent on task instruction (Secondary Task Analysis).

Methods

Participants

Twenty-eight participants took part in this study, recruited from the Radboud University. Participants were recruited with the criteria of being between the ages

of 18 and 35, right handed, with correct or corrected-to-normal vision, native speakers of Dutch, with no history of psychiatric or communication impairments. One participant was excluded due to an error in the projection of stimuli, resulting in a difference in size in the projection. One additional participant did not complete the first task due to discomfort in the scanner. This led to a total sample size of 26 participants (11 male) with a mean age of 25.10 years. The procedure was approved by a local ethics committee.

Materials

a. *Kinematic feature quantification*

The current study used the same kinematic features quantified in Trujillo et al., 2018. We used a toolkit for markerless automatic analysis of kinematic features, developed earlier in our group (Trujillo, Vaitonyte, Simanova, & Özyürek, 2019). The following briefly describes the feature quantification procedure: All features were measured within the time frame between the beginning (hands start to move) and the ending (hands no longer moving) of the gesture. This was the same method used by Trujillo et al., (2018), allowing us to more faithfully replicate behavioral findings, and ensuring the kinematic features represent the movement in the entirety of the video. Motion-tracking data from the Kinect provided measures for our kinematic features: *Distance* was calculated as the total distance travelled by both hands in 3D space over the course of the item. *Vertical amplitude* was calculated on the basis of the highest space used by either hand in relation to the body. *Peak velocity* was calculated as the greatest velocity achieved with the dominant hand. *Hold time* was calculated as the total time, in seconds, counting as a hold. Holds were defined as an event in which both hands and arms are still for at least 0.3 seconds. *Submovements* were calculated as the number of individual ballistic movements made, per hand, throughout the item. Ballistic movements were calculated using a peak analysis, similar to the description of submovements given by Meyer and colleagues (Meyer et al., 1988). In line with the Trujillo et al. (2018) study, our peak analysis used a velocity threshold of 0.2m/s, between-peak distance of 8 frames, and minimum peak height and prominence of 0.2m. To account for the inherent differences in the kinematics of the various items performed, z-scores were calculated for each feature/item combination across all actors including both conditions. This standardized score represents the modulation of that feature, as it quantifies how much greater

or smaller the feature was when compared to the average of that feature across all of the actors. This means that high z-score values for a video indicate that the kinematics were significantly larger than what is typical for that action. For a more detailed description of these quantifications, see Trujillo et al., 2018.

b. *Stimuli*

We included 120 videos recorded in a previous study (Trujillo et al., 2018). In this previous study, 40 participants performed 31 different representational (pantomime) gestures. Twenty performed the gestures in a less-communicative context, while the other twenty performed them in a more-communicative context. Motion capture data of participants (henceforth actors) in this previous experiment were captured using Microsoft Kinect while the actors were seated at a table. The gestures were pantomime versions of object-directed actions, such as cutting paper with scissors or peeling a banana. For each act, actors began with their hands placed on designating starting points on the table, marked with tape. Target objects were placed on the table (e.g. scissors and a sheet of paper for 'cutting paper with scissors') but actors were instructed beforehand not to actually touch the objects. After placing the object(s) on the table, the experiment moved out of view and recorded instructions were played in Dutch (e.g. 'knip het papier doormidden met de schaar' ['cut the paper with the scissors']). Immediately following the instructions, a bell sound was played, indicating that the actor could start performing the gesture. Once the act was complete, the hands returned to the starting points, after which another bell sound indicated the end of the trial. The more-communicative context was elicited by introducing a confederate who sat in an adjacent room and was said to be watching through the video camera and learning from the participant. In this way, an implied communicative context was created. The same procedure was applied to the less-communicative context, except the confederate was said to be learning the experimental set-up. The less-communicative context was therefore exactly matched, including the presence of an observer, but only differed in that there was no implied interaction.

In order to provide a representative sample of the videos we first ranked all videos according to the overall kinematic modulation (z-scores derived from the kinematic features described in section *b*) and the communicative context (more- or less-communicative). This placed all of the videos on a continuum from low kinematic

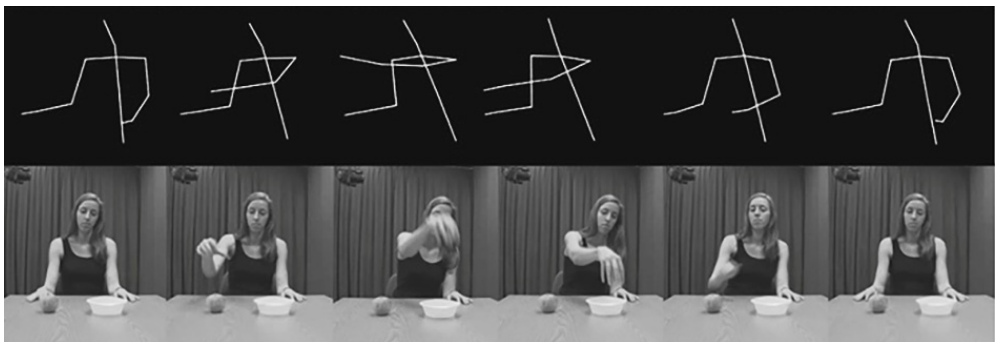


Figure 6. Still frames of a stick-light figure and a comparison with the corresponding video images. The lower panel depicts a series of still frames from one of the videos recorded in (Trujillo et al., 2018) at various stages of action completion. The upper panel depicts the corresponding stick-light figure derived from the kinematics of this action. Note that the images in the upper panel represent what was seen by participants, who had no exposure to the video images. Figure adapted with permission from (Trujillo, Vaitonyte, et al., 2019).

modulation, as was typical of the less-communicative videos, up to high kinematic modulation, as seen in the more-communicative videos. We then selected 60 more-communicative videos, favoring high z-scores, and 60 less-communicative videos, favoring low z-scores, on the basis of keeping the two contexts matched in all raw kinematic (i.e. non-modulation) values as well as overall duration, while also keeping the modulation values of all kinematic features significantly different. This was done using standard t-tests on the raw and modulation values. Therefore, the more-communicative videos were primarily characterized by high positive z-scores, and less-communicative videos were characterized by high negative (e.g. slower, smaller than typical) z-scores. Once a suitable selection was made, the selected videos were transformed into stick-light figures based on the Kinect motion capture data (see Figure 6 for still frames). This ensured that the visual information being processed while viewing the videos was identical besides the movements, or kinematics, of the act.

c. *Physical Setup and Briefing*

Participants were informed that they would be viewing short videos of actions being depicted by 'stick figures', which were created from the motion capture data of real participants in a previous experiment. They were informed that half of the participants performed the actions for themselves, and the other half performed them explicitly for someone else. We informed the participants that in their first task

they should try to guess if each action was performed for the actor or for the viewer, and in the second task they should try to determine if the actions were performed more with the left hand or the right hand. The Social task was always given first, followed by the Handedness task. The ordering was fixed to ensure that the stimuli were novel during the Social Task.

Participants were positioned in the supine position in the scanner with an adjustable mirror attached to the headcoil. Through the mirror participants were able to see a projection screen outside the scanner. Participants were given an MRI compatible response box which they were instructed to operate using the index finger of their right hand to press a button on the right, and the index finger of their left hand to press a button on the left. Button locations corresponded to response options given on the screen, which always include two options: one on the left of the screen, and one on the right of the screen. The resolution of the projector was 1024 x 768 pixels, with a projection size of 454 x 340mm, and 755mm distance between the participant and the mirror. Video size on the projection was adjusted such that the stick figures in the videos were seen at a size of 60 x 60 pixels. This ensured that the entire figure fell on the fovea, reducing eye movements during image acquisition. Stimuli were presented using an in-house developed PsychoPy (Peirce et al., 2019) script.

d. *Tasks*

Social Task

The Social Task was designed to explicitly elicit intention recognition by attending to the movements. In this task, participants first saw a Dutch action verb that served as a linguistic prime for the upcoming video. This was provided to ensure participants understood the gesture that they were seeing. Next, there was a 3.5 second (s) fixation cross, with a 1.5s jitter. Participants were then presented with the stick-light gesture. Average duration for these videos was 6.34 seconds. After the video completed, participants were then visually presented with the question of whether the action was intended for the actor or the viewer. The two options were presented on random sides of the screen and participants responded by pressing either the left or right button of the response box. No feedback was given regarding the accuracy of the response. The order of videos was randomized for each participant.

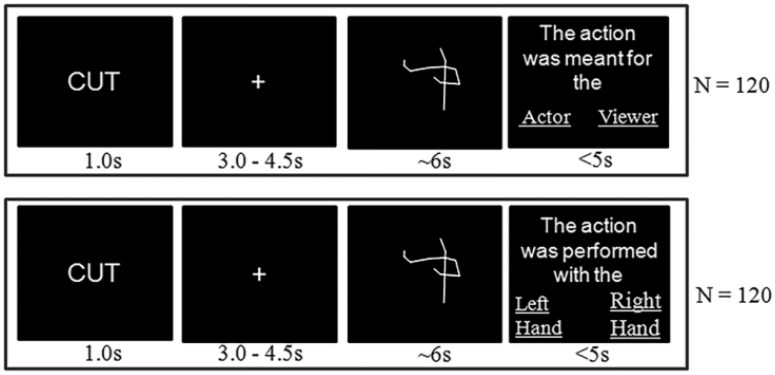


Figure 7. Overview of trial progression. The upper panel depicts the Social Task, while the lower panel depicts the Non Social Handedness task. Participants first saw a single prime word, followed by a fixation cross of variable length, then the video, and finally the task-specific response screen.

Handedness Task

The Handedness Task was designed so that participants would attend to the movements without any social or communicative implication, allowing us to test for evidence of automatic processing of intention. This task followed the same procedure, with a new randomized order of stimuli. However, in this task participants were asked whether the action was performed with the left hand or the right hand. See figure 7 for a schematic timeline of one trial.

Behavioral Data

Data Preparation & Implementation

Response time (RT) and intention classification were utilized for analyses. Data were first checked for outliers at the participant level in terms of RT, with outliers considered to be more than 2.5 standard deviations above the group mean. This led to a removal of 73 individual trials in the Social Task and a removal of 76 trials in the Handedness Task. All preparatory procedures and statistical tests were carried out separately for the Social and Handedness tasks. All testing of behavioral data was performed using the R statistical program (R Development Core Team, 2007). Mixed effects modeling utilized the lme4 package (Bates et al., 2014) and p-values were estimated using the Satterthwaite approximation of denominator degrees of freedom, as implemented in the lmerTest package (Kuznetsova, 2016).

Statistical Analyses

I. Social Task

Statistical analyses were carried out in order to assess whether kinematic modulation was correlated with intention classification. Note that we did not test whether classification decisions matched the context labels from the previous study (Trujillo et al., 2018). This is because the primary interest of the study was the spectrum of kinematic modulation, rather than the initial categories which are also highly variable.

We used linear mixed-effects modeling to determine the correlation between kinematic features and intention classification. Kinematic modulation values were entered into the model as fixed effects with the classification decision (communicative – for the viewer, or non-communicative – for the actor) as the dependent variable. In the first model, participant was additionally included as a random intercept variable, allowing us to control for individual variation between participants. We used a χ^2 test to determine if this model better explained the data than a null model in which only participant variation was given as an explanatory (independent) variable. Next, we compared our initial model with a more complex model that additionally included actor and action as random intercepts. This model was again tested against the null and initial models to determine which provided the best explanation of the data using χ^2 tests. Only fixed effects results from the winning model are interpreted. To reduce the risk of Type I error, we used the Simple Interactive Statistical Analysis tool (http://www.quantitativeskills.com/sisa/calculations/bonfer.htm) to calculate an adjusted alpha threshold based on the mean correlation between all of the tested kinematic features, as well as the number of tests (i.e. number of variables in the mixed model). Our four variables (vertical amplitude, peak velocity, submovements, hold-time) showed an average correlation of 0.063, leading to a Bonferroni corrected alpha threshold of 0.013.

II. Handedness Task

Statistical analyses were carried out in order to assess whether participants were attending to the movement kinematics. This ensures that our fMRI results reflect only a difference in the task, rather than the stimuli, which participants should be attending to similarly in both the Social and Handedness Tasks.

We used linear mixed-effects modeling following the same procedure described for the Social Task. The only difference was that we included peak velocity and submovements for the left hand and excluded vertical amplitude and hold-time. This was done due to vertical amplitude and hold-time being features that were quantified from both hands. Therefore, we included the single hand features for both right and left in order to test the hypothesis that participants classified the handedness of the videos according to hand-specific features. In other words, we assume that right-handed classifications will be made based on submovements and/or peak velocity of the right hand if participants are attending to the kinematics.

We again calculated an adjusted alpha threshold based on the mean correlation of the tested kinematic features and the number of tests (again four). The four variables in this model set (right peak velocity, right submovements, left peak velocity, left submovements) showed a mean correlation of 0.138, leading to a Bonferroni corrected alpha threshold of 0.015.

Calculation of 'Communicativeness' Metric

In order to test our hypothesis that the communicative quality of movement kinematics would be correlated with hemodynamic response in the mirroring and mentalizing systems, we used the behavioral data to calculate a metric of how communicative each video was. In order to calculate this communicativeness value, we first calculated a new mixed effects model with intent classification as the dependent variable, vertical amplitude, hold-time, peak velocity, submovements, and response time as fixed effects predictors, and actor, action, and participant and random intercepts. Response time was included in this model as a measure of certainty, allowing us to not only capture the effect of the kinematics on the final classification decision of the participants, but also how quickly the participants made this decision. Finally, we used this model to calculate the mean predicted probability of judging each video as communicative. As the predicted probability serves as a measure of how likely a new participant would be to judge a video as communicative, this is taken to represent a quantification of video communicativeness. The process of calculating the predicted probability was carried out in a leave-one-out manner, where the values were calculated separately for each individual participant, based only on the rest of the participants' response data. For example, to calculate the communicative values that would be used to model participant 5's brain response,

we used the response data from participants 1-4 and 6-26 to calculate a mean value for each video. Participant 5's data are thus not included in the calculation of her own fMRI regressors. This was repeated for each participant. This was done to prevent over-fitting the data. In the end, each participant had a unique set of communicativeness values assigned to the videos, with one value per video. The c*ommunicativeness* metric therefore provided a single value for each video that described, based on participant responses and the underlying kinematic modulation values, the probability that the video would be classified as being communicatively intended when viewed by a new, naïve participant. These values were then used to model the fMRI data at the first (subject) level.

Brain Imaging

a. *fMRI Data Acquisition*

Anatomical and task-related MRI images were acquired on a 3-Tesla Siemens Magnetom Skyra MR scanner (Erlangen, Germany) with a 32-channel head coil at the Donders Institute for Brain, Cognition and Behaviour in Nijmegen, the Netherlands. Structural images (1 x 1 x 1 mm^3) were acquired using a T1-weighted magnetization-prepared rapid gradient echo-sequence with repetition time (TR) = 2300ms, echo-time (TE) = 3.03ms, flip angle = 8°, field of view (FOV) = 256 x 256 x 192 mm^3. Two behavioral tasks (described below) were carried out by participants while T2*-weighted dual-echo EPI BOLD-fMRI images were acquired using an interleaved ascending slice acquisition sequence (slides = 40, TR = 730ms, TE = 37.8ms, flip angle = 90°, voxel size = 3 x 3 x 3, slice gap = 0.34mm, FOV = 212 x 212mm^2).

b. *fMRI Analysis – General Linear Model*

All analyses were performed using SPM12 (Statistical Parametric Mapping; Wellcome Department London, UK, http://www.fil.ion.ucl.ac.uk/spm). All functional data were preprocessed following the same pipeline: functional and structural images were realigned coregistered, spatial normalization with the Montreal Neurological Institute (MNI) template, and spatial smoothing using an 8mm full width half maximum kernel. After preprocessing, we checked motion parameters in the task-related acquisitions to ensure that no participants moved more than 3° in rotation or 3mm in translation.

We created an event-related design matrix for within-subject first-level analysis,

wherein we modeled the video-viewing period, response, and fixation as separate regressors. Communicativeness of the videos was added as a parametric modulator, with the values convolved with the video viewing events in a separate regressor. Finally, the six motion parameters were added as regressors of no-interest. Our primary first-level contrast was *communicativeness* over baseline, which effectively modeled a linear correlation between the BOLD signal and the *communicativeness* score. The two tasks were modeled in separate design matrices, with no direct comparisons between the two. This is because the Handedness Task was only used to test whether brain activation or connectivity is related to kinematic modulation when the task does not require a communicative intent decision.

Contrast images from the first-level analysis were used in the second (group) level analysis, using whole-brain voxel-wise t-tests. Contrast maps were thresholded at $p < 0.001$, uncorrected, with cluster threshold set as $k > 10$.

c. *fMRI Analysis – Dynamic Causal Modeling*

I. *General overview*

We used Dynamic Causal Modeling (DCM; Friston, Harrison, & Penny, 2003) in order to quantify how the mentalizing and mirroring system interact during intention understanding. DCM allows the researcher to define a subset of brain regions and their connections and model how the activity of the regions or strength of the connections is dependent upon an experimental manipulation. After building and estimating a set of potential causal models, a model selection analysis is performed in order to find the model that represents the best fit to the data. In order to keep the models relatively simple and balanced, we opted to only model two regions: one from the mentalizing, and one from the mirroring system. We based our initial selection criteria on the meta-analysis of intention understanding by van Overwalle & Baetens (Van Overwalle & Baetens, 2009), which lists the posterior superior temporal sulcus (pSTS), anterior inferior parietal sulcus (aIPS), and premotor cortex (PMC) as the primary mirroring system regions, and the temporal parietal junction (TPJ) and medial prefrontal cortex (mPFC) as the primary mentalizing regions. As the TPJ, aIPS, and pSTS show some degree of overlap, we chose not to use these regions, and therefore selected the PMC as the representative mirroring region and the mPFC as the representative mentalizing region to contrast the two networks in a neuroanatomically optimal manner.

II. *Regions of Interest*

We defined the location of these group-level regions of interest around the peak-voxel coordinates of our second-level *communicativeness* contrast from the Social Task. Functional regions were defined from the coordinates based on the definitions by Lacadie and colleagues (Lacadie, Fulbright, Arora, Constable, & Papademetris, 2007). Note that the same coordinates were used in our DCM analysis of the Handedness Task in order to ensure a direct comparison of the results, and that this analysis is carried out regardless of GLM results of the Handedness Task as this was an a priori planned analysis in order to compare against the Social Task. The PMC was located at x = 24, y = -10, z = 53, while the mPFC was located at x = -9, y = 38, z = 23. The coordinates were used as starting points to locate subject specific regions. This was done using SPM12's volume of interest (VOI) utility which takes a starting coordinate and moves it, per participant, to the nearest peak voxel within a 5mm range. This method takes individual variation in functional neuroanatomy into account and increase sensitivity of subsequent analyses. Each newly assigned peak was manually checked to ensure that it still was in the designated region. Mean time courses were extracted from a 10mm sphere surrounding the peak coordinate, using the *communicativeness* contrast and a liberal threshold of $p < 0.100$ to ensure a robust estimate of the time series.

III. *Model Space*

We created an initial model comprised of the PMC and mPFC with bidirectional intrinsic connections. The video viewing event (video onset, with length equal to video duration) was modeled as a possible direct, or driving, influence on regional activity, while the *communicativeness* regressor (as explained under the subsection *Calculation of 'Communicativeness' Metric*) was defined as a possible modulating influence on the strength of inter-region connections. By varying the presence of the driving and modulation influences on the two regions and connections, we created fourteen models that included all possible combinations of these influences, including one fully parameterized model that had both driving influences and both modulations, as well as one 'null' model that had no influence from the task. See Supplementary Figure 3.1 for a schematic overview of all of these models.

IV. *Model Selection*

Bayesian model selection (BMS) was used to test the probability of our data given each of the models. As our participants are relatively homogeneous (i.e. no group based inferences) we utilized a fixed effects approach. A posterior probability of > 0.95 was taken to be strong evidence in favor of a particular model.

Results

Behavioral Results – Social Task

For the Social Task we tested whether higher kinematic modulation values predicted classification of an act as being communicative. In line with our hypothesis, our mixed-effects regression model containing the kinematic features as fixed effects predictors was a better fit to the data than the null model that did not contain kinematics, $\chi^2(4)= 51.629$, $p < 0.001$. Adding actor and action as random intercepts further improved model fit, $\chi^2(2)= 18.605$, $p < 0.001$. All results at the kinematic feature level are therefore based on the full model, including all kinematic modulation values as fixed effects as well as participant, actor, and action as random intercepts. In terms of kinematic features, we found that increased vertical amplitude ($z = 4.113$, $p < 0.001$) and hold-time ($z = 3.243$, $p = 0.001$) were significantly predictive of classifying an act as communicative. Increased number of submovements showed a near significant relation to intent classification ($z = 2.432$, $p = 0.015$), while peak velocity was not related to communicative intent classification ($z = 0.924$, $p = 0.356$). Results therefore confirm that intention classification was related to kinematic modulation.

Behavioral Results – Handedness Task

For the Handedness Task we tested whether higher kinematic modulation values of a particular hand predicted classification of an act being performed more with that same hand. This was to ensure participants were attending to the kinematics in this task. We found that the model containing kinematic modulation values was a better fit to the data than the null model, $\chi^2(4)= 83.291$, $p < 0.001$. Adding actor and action to the model further improved model fit, $\chi^2(2)= 368.57$, $p < 0.001$. All results at the kinematic feature level are therefore based on the full model, including all kinematic modulation values as fixed effects as well as participant, actor, and action as random intercepts. In terms of kinematic features, we found that submovements of the right hand were predictive of classifying an act as being more right-handed (z

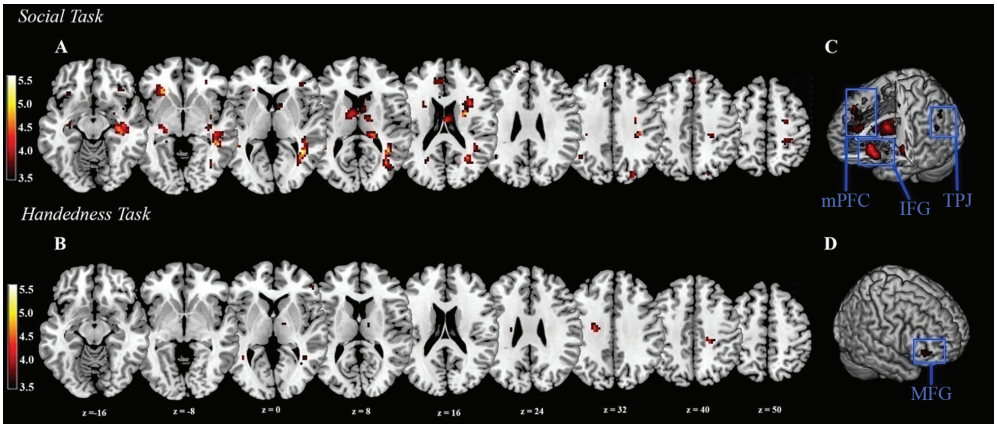


Figure 8. Overview of GLM results. The top panels (A & C) depict slices from the Social Task, while the bottom panels (B & D) depict the Handedness Task. Red areas indicate significant ($p < 0.001$) correlation between BOLD response and video communicativeness. The red color bars show the corresponding T values. Panels A and B provide a slice by slice overview of the two tasks, while panels C and D provide a 3D rendering of the same data, with significant areas of interest highlighted (mPFC = medial prefrontal cortex; IFG = inferior frontal gyrus; TPJ = temporoparietal junction; MFG = middle frontal gyrus)

= 5.143, $p < 0.001$). We found no association between handedness classification and submovements of the left hand (z = -1.676, $p = 0.094$), peak velocity of the right hand (z = 1.817, $p = 0.069$), or peak velocity of the left hand (z = 1.643, $p = 0.100$). Results therefore confirm that participants attended to kinematic modulation also during the Handedness Task, while further suggesting that the right hand was attended to primarily.

Whole-brain results – Social Task

Whole-brain results reflect BOLD correlation with video *communicativeness*. Results of the whole-brain analysis of the Social Task show primarily regions associated with the pMNS, such as the right premotor cortex and right inferior parietal lobe, as well as regions associated with the MS, such as the left medial prefrontal cortex and left temporoparietal junction. We additionally found activation in the left inferior frontal gyrus, left caudate nucleus, right hippocampus, and several areas of the cerebellum. Table 3 provides an overview of peak coordinates, given in MNI space, with statistics and cluster sizes. All regions were significant at $p < 0.001$. Figure 8 provides a graphical overview of these results.

Table 3. Significant activation correlated with communicativeness across tasks

L/R	BA	Region	T	Z	k	x	y	z
			Social Task					
R		Hippocampus	6.02	4.69	474	30	-19	-10
L		Caudate Nucleus	5.59	4.46	438	-9	-1	14
L	32	mPFC	5.26	4.28	362	-9	38	23
L	47	IFG	5.23	4.26	130	-24	29	-1
L		Hippocampus	5.06	4.16	55	-27	-16	-7
L	39	TPJ	4.49	3.81	23	-54	-49	29
R	46	IPL	4.31	3.69	36	39	35	5
R	7		4.12	3.57	18	27	-79	38
R	40		3.99	3.47	52	57	-34	38
R		Cerebellum	3.94	3.44	11	9	-28	-40
R	6	Premotor Cortex	3.86	3.338	11	24	-10	53
R		Cerebellum	3.82	3.36	16	3	-76	41
L			3.78	3.33	18	-24	-76	-25
R	6	Premotor Cortex	3.74	3.3	11	21	11	47
			Handedness Task					
R	46	MFG	4.16	3.56	17	51	41	2

BA = Brodmann area; k = cluster size; mPFC = medial prefrontal cortex; IFG = inferior frontal gyrus; TPJ = temporoparietal junction; IPL = inferior parietal lobe; MFG = middle frontal gyrus.

Whole-brain results – Handedness Task

Results of the whole-brain analysis of the Handedness Task show only the middle frontal gyrus being correlated with c*ommunicativeness*. See Table 3 for peak coordinates and statistics. Figure 8 provides a graphical overview of these results.

Connectivity results – Social Task

In the Social Task, we found strong evidence (exceedance probability = 1.00) for a model with no driving effects of video-viewing on the premotor or mPFC but modulation of the top-down (mPFC → premotor) connection. See Figure 9 for a schematic overview of the winning model and the exceedance probability.

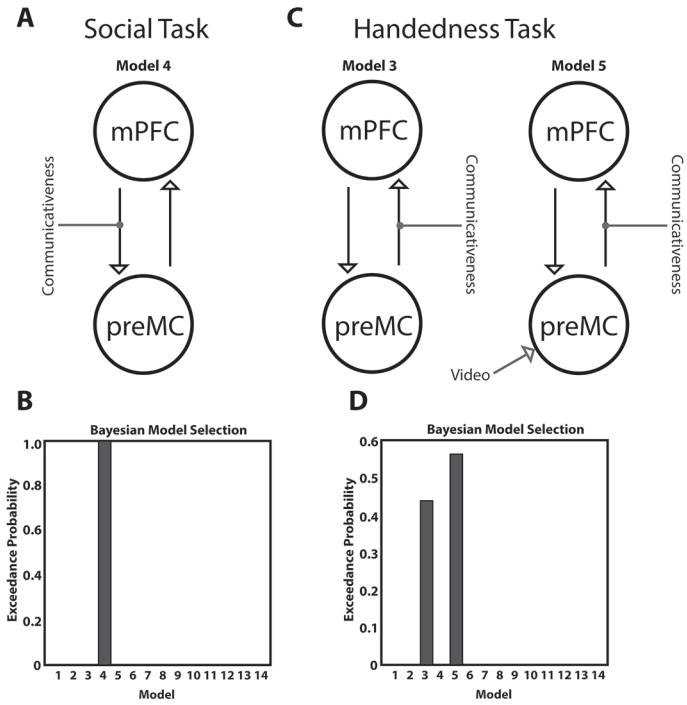


Figure 9. Overview of winning DCM models. **A** depicts the winning model for the Social Task, **B** presents the exceedance probability. In all models, circles depict the individual regions, while arrows depict the intrinsic, directional coupling between them. Video viewing is modeled as a driving input to the regions. Communicativeness is modeled as a modulator of coupling strength. **C** depicts the two high probability models for the Handedness Task, and **D** presents the exceedance probabilities for these models. mPFC = medial prefrontal cortex; preMC = premotor cortex.

Connectivity results – Handedness Task

In the Handedness Task, we did not find evidence above our defined probability threshold. However, two models together showed an exceedance probability of 1.00. The model with the highest evidence (exceedance probability = 0.561) showed driving influence of video viewing on the premotor cortex and modulation by communicativeness of the videos on the bottom-up (premotor → mPFC) connection. The second model (exceedance probability = 0.439) showed no driving effects but modulation by communicativeness of the bottom-up connection. Together, this can be taken as strong evidence in support of modulation of the bottom-up connection, with weaker support for the driving effect on the premotor cortex. See Figure 9 for a schematic overview of the two models and the exceedance probabilities associated with them.

Discussion

General overview of findings

This study set out to test the brain activation and connectivity during the recognition of communicative intentions from kinematic modulation. We found that 1) participants recognize communicative intent based on spatial and temporal kinematic features if explicitly asked to classify intentionality, 2) the perceived communicativeness of the videos correlates with activation of the mentalizing and mirroring system when this is task-relevant, 3) top-down connectivity between these systems is altered by communicativeness in the Social Task, while bottom-up connectivity is modulated in the Non-social Task.

Behavioral results

Our behavioral results show that our participants were able to utilize kinematic modulation in their intention classifications. This result is a direct replication of earlier work from our group that showed that increased vertical amplitude was perceived as communicative (Trujillo et al., 2018). The current study replicated this finding while extending it in two important ways. First, we additionally found hold-time to be predictive of communicative intent classification. Second, our use of stick-light figures, rather than real videos, shows that intention recognition can occur even from highly reduced stimuli. Together, these results support the hypothesis that communicative intent can be read purely from movement kinematics (Becchio, Manera, et al., 2012; Cavallo et al., 2016), and that both spatial and temporal features are important signals of intention.

We found that the exaggeration of submovements of the right hand was associated with perceiving an act as right-handed. This finding indicates that participants also attended to kinematic modulation in the Handedness Task, although the specific features were different from the Social Task. Given this finding, we are able to compare brain activation and connectivity results between the two tasks, as the primary difference is whether participants were basing judgments of communicative intentionality or handedness on the perceived kinematic modulation.

Brain activation in response to communicative kinematics

In the Social Task, we found activation of areas associated with the mentalizing

system, such as the mPFC and left temporoparietal junction (TPJ), as well as several areas associated with the mirroring system such as the inferior parietal lobe and premotor cortex. Our results largely replicate the meta-analytic findings by van Overwalle and Baetens regarding brain activation while reading intentions from unusual or unexpected actions, experimental findings of brain activation in response to unexpected or unusual motions (Marsh et al., 2011, 2014; Van Overwalle & Baetens, 2009), as well as implicit intention recognition tasks using object-directed actions (Ciaramidaro et al., 2013). Similar to previous reports on violations of movement expectations, we found the right premotor cortex (Koelewijn, van Schie, Bekkering, Oostenveld, & Jensen, 2008; Manthey, Schubotz, & Von Cramon, 2003; Van Overwalle & Baetens, 2009), mPFC (Schiffer et al., 2014; Van Overwalle & Baetens, 2009) and left TPJ (Ciaramidaro et al., 2013) responding to increasingly communicative movements. One major distinction between our findings and those of the meta-analysis is that we found the left TPJ, whereas van Overwalle and Baetens found the right TPJ. This can be explained by the left TPJ being primarily responsible for the processing of communicative intentions (Becchio, Manera, et al., 2012; Ciaramidaro et al., 2013; Van Overwalle & Baetens, 2009), whereas the right TPJ is involved in the processing of many other types of intentions as well (Ciaramidaro et al., 2013; Van Overwalle & Baetens, 2009). These results are therefore directly in line with the idea that inferring abstract intentions is based on breaches of expectation originating in the MS, while expanding these previous findings by specifically showing that the brain responds similarly to subtle breaches at the kinematic level.

Besides the a priori predicted mentalizing and mirroring areas, we also found activation of the hippocampus and caudate nucleus to be correlated with communicative kinematics. Activation of both of these regions is directly in line with our theoretical framework. For example, previous work shows the caudate nucleus responding to expectation violations in a human movement observation paradigm (Schiffer & Schubotz, 2011) as well as more generally in response to less familiar action sequences (Diersch et al., 2013). The hippocampus has similarly been linked to processing less familiar actions (Diersch et al., 2013) and is furthermore involved in signaling the presence of novel information (Lisman & Grace, 2005) such as unfamiliar actions (Caligiore, Pezzulo, Miall, & Baldassarre, 2013). These findings suggest that the caudate nucleus and hippocampus play an important role in processing unexpected movement kinematics in order to infer communicative intentions.

In the Handedness Task, we did not find any activation in our a priori defined regions of interest. This means that the regions found in the Social Task only respond when communicativeness is task-relevant. This finding is contrary to studies that used implicit viewing tasks and still found significant activation. However, a major difference in our study is that while we used kinematic variations of the same overall action, previous studies typically use categorically different actions, such as lifting up an apple to take a bite compared to lifting it up to pass to the observer (Ciaramidaro et al., 2013). Thus, while the brain may respond robustly to categorically distinct socially intended actions, response to subtle kinematic differences may itself also be much more subtle in the absence of explicit attention to the underlying intention. On the other hand, we are not the first to report a task-dependent response to the intentionality of observed actions. Our finding is in agreement with an earlier study by de Lange and colleagues who similarly found activation of the mentalizing system in response to unusual actions, but only when explicitly attending to the intention (de Lange et al., 2008). De Lange et al. additionally found that an area of the mirroring system remained active in response to unusual actions even when not explicitly attending to the intention. Similarly, we found the middle frontal gyrus, which may also be involved in the pMNS (Molenberghs, Mattingley, Cunnington, & Mattingley, 2011). Similarly, Spunt and Lieberman (2013) found that cognitive load, in the form of a competing memory task, extinguished activation of MS regions during abstract intention inference. Overall, we suggest that robust activation of the MS and pMNS in response to communicative kinematic modulation only occurs when the observer is actively attending to this aspect of the movement. Future studies will be needed in order to determine whether kinematic modulation will naturally draw attention in the absence of explicit task instructions, given that our control task may have inadvertently drawn attention away from this feature of the stimuli, rather than simply making it less task-relevant.

Effective connectivity

In the experiment, participants had to infer intentionality of the observed actions, i.e. decide if the action was performed “for the actor” or “for the viewer”. The model-driven connectivity analysis showed that the kinematic modulation affected top-down coupling strength between mPFC and PreMC and not vice versa. Our findings therefore provide evidence for a hierarchical system utilizing top-down expectations and bottom-up detection of kinematic deviations. This suggested mechanism allows

us to draw a parallel with perceptual studies that empirically test the effect of unexpected stimuli on brain dynamics. Specifically, recent studies using DCM show that while attending to auditory stimuli, unexpected omissions or mismatches of the stimulus result in changes to top-down connections between relevant brain regions (Auksztulewicz & Friston, 2015; Chennu et al., 2016). More generally, these findings are also directly in line with models of top-down control in social cognition (Hillebrandt, Blakemore, & Roiser, 2013; Wang & Hamilton, 2012).

Our finding fits well with experimental evidence of expectations shaping the dynamics of higher and lower-level cognitive systems when processing concrete (i.e. end-goal) intentions. For example, in a recent study Jacquet and colleagues measured corticospinal excitability to show that when viewing and identifying the end-goal of an action, changes to expectations regarding end-goal intentions results in a tuning of the motor system (Jacquet et al., 2016). Interestingly, and in line with our study, these expectations could be based on observed kinematics and whether or not they were optimal for goal completion. While Jacquet et al. only looked at the motor system, a later study by Chambon and colleagues investigated the use of sensory evidence versus prior expectations to recognize concrete intentions while measuring whole-brain activation (Chambon et al., 2017). Chambon et al. found that top-down connections within the mentalizing system are modulated by an increasing reliance on prior expectations, which occurs when sensory evidence becomes less available or reliable (Chambon et al., 2017). Similarly, Ondobaka and colleagues found that the posterior cingulated cortex, another region of the mentalizing system, has a top-down affect on the action observation network during the processing of movement expectations of others (Ondobaka et al., 2015). While the specific regions in this study are different from our results, this may be due to the difference in the types of movement goals, or intentions, being processed. Ondobaka et al. conclude that their result shows support for a hierarchical account of action goal understanding with high-level midline (mentalizing) regions processing expected goals (or intentions) and lower level action observation, or mirroring, regions processing the movements. However, this study did not directly show changes in connectivity between higher and lower levels. Our results therefore provide an interesting extension to these previous findings, showing evidence for the importance of top-down connections when observing other's actions –including gesture.

In the Handedness Task, we see the pattern of connectivity modulation reversed.

Increased communicativeness of the videos results in more modulation of the bottom-up coupling strength. This is in line with the study of coupling strength changes in response to unexpected auditory stimuli. In that study, top-down coupling changes were associated with an unexpected stimulus when this stimulus was the focus of attention. When the stimulus was not the focus of attention, the top-down coupling effect was still present, but paired with a bottom-up coupling change as well (Chennu et al., 2016). However, the DCM results from the Handedness Task should be interpreted with caution, as the GLM analysis of this task did not reveal significant activation of these regions at our specified threshold. Additionally, the fixed task order and different cognitive demands of the two tasks makes it difficult to determine whether these connectivity differences are due to that lack of explicit attention to the communicative intent, or to some other factor. We will therefore keep our discussion of these results to a minimum.

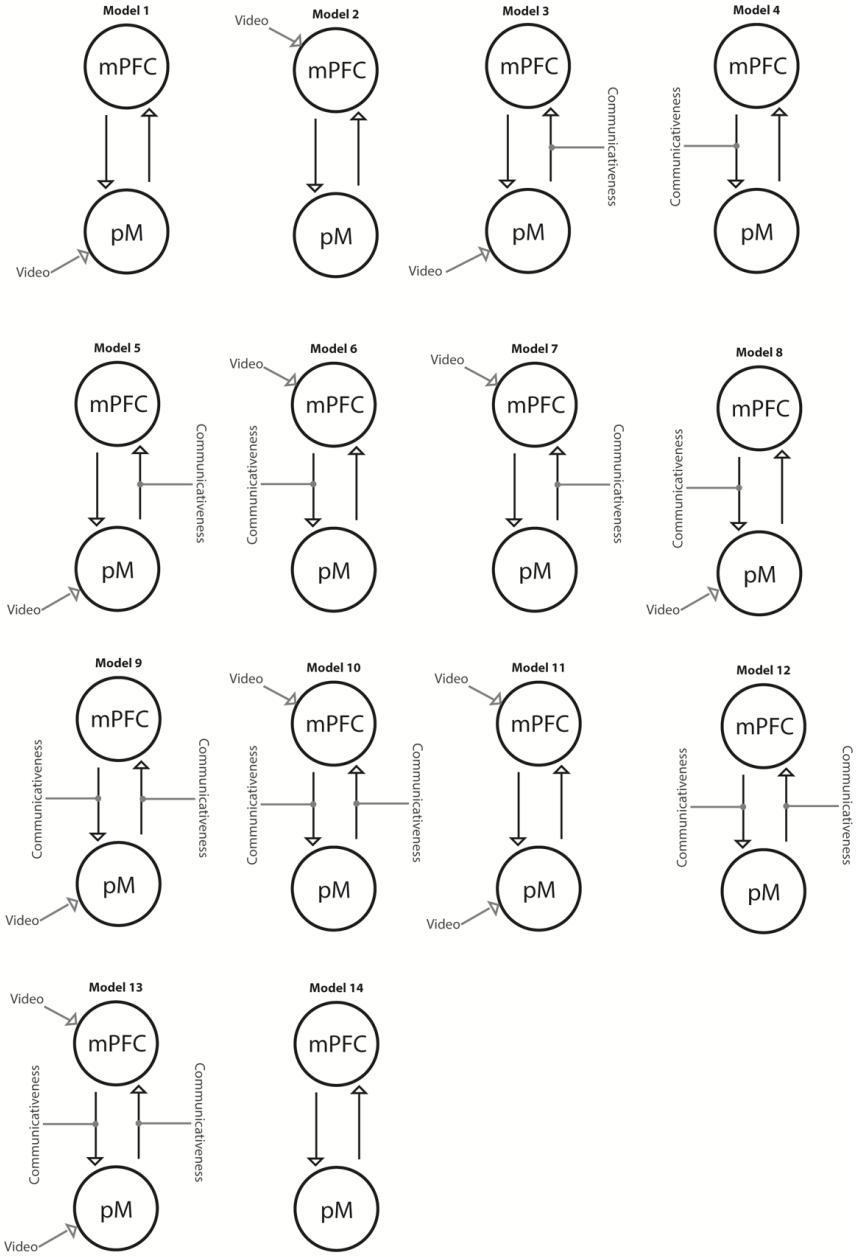
Overall, these results suggest that unexpected events result in top-down changes in connectivity at multiple levels of the brain. The detection of unexpected kinematics allows the recognition of communicative intentions.

Conclusions

In sum, we found that communicative intent can be read from isolated and subtle kinematic cues, and that this recognition process is reflected in activation and (top-down) changes in connectivity of the mirroring and mentalizing systems. These results shine new light on how motor and social brain networks work together to process statistical irregularities in behavior to understand or "read" the complex dynamics of socially and communicatively relevant actions. Most directly, it highlights expectation violations as a key cue for inferring communicative intention, linking studies of movement, communication, and low-level perception. In particular, we show that even subtle kinematic differences in an otherwise typical motor act can be used to infer intention. This has theoretical implications for understanding the fundamental neurobiological mechanisms underlying perceptual inferences and communicative behavior as well as the evolutionary origins of communicative signaling. Practical implications extend to understanding human and human-machine interactions and providing a novel neuroscientific basis to investigate clinical conditions in which movement or social skills are impaired (e.g. Autism Spectrum Disorder).

Acknowledgments

This research was supported by the NWO Language in Interaction Gravitation Grant (024.001.006). The authors declare no conflict of interest in this study.



Supplementary Figure 3.1.Schematic overview of all DCMs in the model set. In all models, circles depict the individual regions, while arrows depict the intrinsic, directional coupling between them. Video viewing is modeled as a driving input to the regions, while communicativeness is modeled as a modulator of coupling strength.

Chapter 4

The communicative advantage:

how kinematic signaling supports semantic comprehension

Chapter based on:

Trujillo, J.P., Simanova, I., Bekkering, H., & Özyürek, A. (2019). The communicative advantage: how kinematic signaling supports semantic comprehension. *Psychological Research*, in press. https://doi.org/10.1007/s00426-019-01198-y

Abstract

Humans are unique in their ability to communicate information through representational gestures which visually simulate an action (moving hands as if opening a jar). Previous research indicates that the intention to communicate modulates the kinematics (e.g. velocity, size) of such gestures. If and how this modulation influences addressees' comprehension of gestures have not been investigated. Here we ask whether communicative kinematic modulation enhances semantic comprehension (i.e. identification) of gestures. We additionally investigate whether any comprehension advantage is due to enhanced early identification or late identification.

Participants (n=20) watched videos of representational gestures produced in a more- (n=60) or less-communicative (n=60) context and performed a forced-choice recognition task. We tested the isolated role of kinematics by removing visibility of actor's faces in Experiment I, and by reducing the stimuli to stick-light figures in Experiment II. Three video lengths were used to disentangle early identification from late identification. Accuracy and response-time quantified main effects. Kinematic modulation was tested for correlations with task performance.

We found higher gesture identification performance in more- compared to less-communicative gestures. However, early identification was only enhanced within a full visual context, while late identification occurred even when viewing isolated kinematics. Additionally, temporally segmented acts with more post-stroke holds were associated with higher accuracy.

Our results demonstrate that communicative signaling, interacting with other visual cues, generally supports gesture identification, while kinematic modulation specifically enhances late identification in the absence of other cues. Results provide insights into mutual understanding processes as well as creating artificial communicative agents.

Introduction

Human communication is multimodal, utilizing various signals to convey meaning and interact with others. Indeed, humans may be uniquely adapted for knowledge transfer, with the ability to signal the intention to interact as well as to manifest the knowledge that s/he wishes to communicate (Csibra & Gergely, 2006). This communicative signaling system is powerful in that the signals are dynamically adapted for the context in which they are used. For example, representational gestures (Kendon, 2004; McNeill, 1994), show systematic modulations dependent upon the communicative or social context in which they occur (Campisi & Özyürek, 2013; Galati & Galati, 2015; Gerwing & Bavelas, 2004; Holler & Beattie, 2005). Although these gestures are an important aspect of human communication, it is currently unclear how the addressee benefits from this communicative modulation. The current study aims to investigate for the first time whether and how kinematic signaling enhances identification of representational gestures.

There is growing evidence that adults modulate their action and gesture kinematics when communicating with other adults, depending on the communicative context. For example, adults adapt to addressees' knowledge by producing gestures that are larger (Bavelas et al., 2008; Campisi & Özyürek, 2013), more complex (Gerwing & Bavelas, 2004; Holler & Beattie, 2005), and higher in space (Hilliard & Cook, 2016) when conveying novel information. Instrumental actions intended to teach show similar kinematic modulation, including spatial (McEllin, Knoblich, & Sebanz, 2018; Vesper & Richardson, 2014) and temporal (McEllin et al., 2018) exaggeration. Evidence from our own lab corroborates these findings of spatial and temporal modulation in the production of both actions and gestures. In our recent work, we quantified the spatial and temporal modulation of actions and pantomime gestures (used without speech) in a more- relative to a less-communicative context (Trujillo, et al., 2018). We showed that spatial and temporal features of actions and pantomime gestures are adapted to the communicative context in which they are produced.

A computational account by Pezzulo and colleagues suggests that modulation makes meaningful acts communicative by disambiguating the relevant information, effectively making the intended movement goal clear to the observer (Pezzulo et al., 2013). This framework focuses on actions, but could be extended to gestures. One recent experimental study directly assessed how kinematic modulation affects

gesture comprehension. By combining computationally-based robotic production of gestures with validation through human comprehension experiments, Holladay et al. showed that spatial exaggeration of kinematics allows observers to more easily recognize the target of pointing gestures (Holladay et al., 2014). Similarly, Gielnak and Thomaz showed that when robot co-speech gestures are kinematically exaggerated, the content of an interaction with that robot is better remembered (Gielniak & Thomaz, 2012). Another study used an action-based leader-follower task to show that task leaders not only systematically modulate task-relevant kinematic parameters, but these modulations are linked to better performance of the followers (Vesper, Schmitz, & Knoblich, 2017).

These previous studies suggest that the kinematic modulation of communicative movements (e.g. actions and gestures) serves to clarify relevant information for the addressee. However, it remains unclear whether this also holds for more complex human movements, such as pantomime gestures. This question is important for our understanding of human communication given that complex representations form an important part of the communicative message (S. D. Kelly, Ozyurek, & Maris, 2010; Özyürek, 2014).

The mechanism by which kinematic modulation might support semantic comprehension, or identification, of complex movements remains unclear. Several studies suggest disambiguation of the ongoing act, either through temporal segmentation of relevant parts (Blokpoel, van Kesteren, Stolk, Haselager, Toni & van Rooij, 2012; Brand, Baldwin, & Ashburn, 2002), or spatial exaggeration of relevant features (Brand et al., 2002) as the mechanism. In the case of disambiguation, the "semantic core" (Kendon, 1986), or meaningful part of the movement, is made easier to understand as it unfolds. However, there is also evidence suggesting that early kinematic cues provide sufficient information to inform accurate prediction of whole actions before they are seen in their entirety (Cavallo et al., 2016; Manera et al., 2011). One study, for example, used videos of a person walking, and at a pause in the video participants were asked whether the actress in the video would continue to walk, or start to crawl. The authors showed that whole-body kinematics could support predictions about the outcome of an ongoing action (Stapel et al., 2012). However, another study showed videos of a person reaching out and grasping a bottle, and asked the participants to predict the next sequence in the action (e.g. to drink, to move, to offer) and found that they were unable to use such early cues

for accurate identification in this more complex, open-ended situation (Naish et al., 2013). Furthermore, identification of pantomime gestures has previously been reported to be quite low when no contextual (i.e. object) information is provided (Osiurak, Jarry, Baltenneck, Boudin, & Le Gall, 2012). Given these inconsistencies in the literature, an open question remains: are early kinematic cues sufficient to inform early representational gesture identification, or does kinematic modulation primarily aid gesture identification as the movements unfold (i.e. late identification)?

Finally, in order to understand how kinematic modulation might support gesture identification, it is important to consider other factors that might influence the semantic comprehension of an observer. In a natural environment, movements such as gestures are accompanied by additional communicative signals, such as facial expression and eye-gaze, and/or finger kinematics relevant in the execution of the gestures. Humans are particularly sensitive to the presence of human faces, which naturally draw attention (Cerf, Harel, Einhäuser, & Koch, 2007; Hershler & Hochstein, 2005; Theeuwes & Van der Stigchel, 2006). This effect is most prominent in the presence of mutual gaze (Farroni, Csibra, Simion, & Johnson, 2002; Holler et al., 2015), but also occurs in averted gaze compared to non-face objects (Hershler & Hochstein, 2005). Hand-shape information can also provide clues as to the object one is manipulating (Ansuini, Cavallo, Koul, D'Ausilio, Taverna & Becchio, 2016), and more generally the kinematics of the hand and fingers together provide early cues to upcoming actions (Becchio et al., 2018; Cavallo et al., 2016), which together may allow the act to be more easily identified. In order to understand the role of kinematic modulation in communication, the complexity of the visual scene must also be taken into account.

In sum, previous studies show kinematic modulation occurring as a communicative cue in actions and gestures. While research suggests that this modulation serves to enhance comprehension, this has not been assessed directly in terms of semantic comprehension of complex movements, such as representational gestures. Furthermore, it is currently unclear if improved comprehension would be driven by early action identification or by late identification of semantics, and which kinematic features provide this advantage.

The current study addresses these questions. In two experiments, naïve participants perform a recognition task of naturalistic pantomime gestures recorded in our

previous study (Trujillo et al., 2018). In the first experiment they see the original videos with the face of the actor either visible or blurred, to control for eye-gaze effects. In the second experiment the same videos are reduced to stick-light figures, reconstructed from Kinect motion tracking data. The stick figure videos allow us to test the contribution of specific kinematic features, because only the movements are visible, but not the face or hand-shape. In both experiments we additionally manipulate video length to test whether any communicative benefit is driven more by early identification (resulting in differences only in the initial fragment), or late identification (resulting in differences in the medium and full fragments). Experiment II provides an additional exploratory test of the contribution of specific kinematic features to gesture identification.

We hypothesize that kinematic modulation serves to enhance semantic legibility. As early kinematic information is less reliable for open-ended action prediction (Naish et al., 2013) and pantomime gestures may generally be difficult to identify without context (Osiurak et al., 2012), we expect better recognition scores for the communicative gestures in the medium fragments and full fragments compared to initial fragments. We furthermore predict that performance will correlate with stronger kinematic modulation. Additionally, we expect performance to be lower overall with stick-light figures, compared to the full videos due to decreased visual information, but with a similar pattern (i.e. better performance in medium and full fragments compared to initial). For our exploratory test, we expect that exaggeration of both spatial and temporal kinematic features will contribute to better gesture identification.

Experiment I – Full visual context

Our first experiment, with actual videos of the gestures, was designed to test whether 1) kinematic modulations leads to improved semantic comprehension in an addressee, 2) if the advantage is better explained by early identification or late identification of the gestures, and 3) whether the effect is altered by removing a salient part of the visual context, the actor's face.

Methods

Participants

Twenty participants were included in this study, (mean age = 28; 16 female), recruited from the Radboud University. Participants were selected on the criteria of being aged 18 – 35, right-handed and fluent in the Dutch language, with no history of psychiatric disorders or communication impairments. The procedure was approved by a local ethics committee and informed consent was obtained from all individual participants in this study.

Materials

Each participant performed the recognition task with 60 videos of pantomimes that differed in their context (more or less communicative), video duration (short, medium and full), and face visibility (face visible versus blurred). Detailed description of the video recordings, selection and manipulation follows below.

a. *Video recording procedure*

Stimuli were derived from a previous experiment (Trujillo, et al., 2018). In this previous experiment, participants (henceforth, actors) were filmed while seated at a table, with a camera hanging in front of the table. Motion-tracking data was acquired using Microsoft Kinect system hanging slightly to the left of the camera. Each actor performed a set of 31 gestures, either in a more-communicative or a less-communicative setting (described below). Gestures consisted of simple object-directed acts, such as cutting paper with scissors or pouring water into a cup. Target objects were placed on the table (e.g. scissors and a sheet of paper for the item 'cut the paper with the scissors') but actors were instructed to perform as if they were acting on the objects, without actually touching them. For each item, actors began with their hands placed on designated starting points on the table (marked with tape). After placing the target object(s) on the table, the experimenter moved out of view from the participant and the camera, and recorded instructions were played. Immediately following the instructions, a bell sound was played, which indicated that the participant could begin with the pantomime. Once the act was completed, actors returned their hands to the indicated starting points, which elicited another bell sound, and waited for the next item. For this study, videos began at the first bell sound, and ended at the second bell sounded. In the more-

communicative context we introduced a confederate who sat in an adjacent room and was said to be watching through the video camera and learning the gestures from the participant. In this way, an implied communicative context was created. In the less-communicative context, the same confederate was said to be learning the experimental set-up. The less-communicative context was therefore exactly matched, including the presence of an observer, but only differed in that there was no implied interaction. Despite the subtle task manipulation, our previous study (Trujillo, et al., 2018) showed robust differences in kinematics between the gestures produced in the more-communicative versus the less-communicative context.

b. *Kinematic feature quantification*

For the current study, we used the same kinematic features that were quantified in our earlier study (Trujillo et al., 2018). We used a toolkit for markerless automatic analysis of kinematic features, developed earlier in our group (Trujillo, Vaitonyte, Simanova, & Özyürek, 2019). The following briefly describes the feature quantification procedure: All features were measured within the time frame between the beginning and the ending bell sound. Motion-tracking data from the Kinect provided measures for our kinematic features, and all raw motion tracking data was smoothed using the Savitsky-Golay filter with a span of 15 and degree of 5. As described in our previous work (Trujillo et al., 2018), this smoothing protocol was used as it brought the Kinect data closely in line with simultaneously recorded optical motion tracking data in a separate pilot session. The following features were calculated from the smoothed data: *Distance* was calculated as the total distance travelled by both hands in 3D space over the course of the item. *Vertical amplitude* was calculated on the basis of the highest space used by either hand in relation to the body. *Peak velocity* was calculated as the greatest velocity achieved with the right (dominant) hand. *Hold time* was calculated as the total time, in seconds, counting as a hold. Holds were defined as an event in which both hands and arms are still for at least 0.3 seconds. *Submovements* were calculated as the number of individual ballistic movements made, per hand, throughout the item. To account for the inherent differences in the kinematics of the various items performed, z-scores were calculated for each feature/item combination across all actors including both conditions. This standardized score represents the modulation of that feature, as it quantifies how much greater or smaller the feature was when compared to the average of that feature across all of the actors. (Addressee-directed) eye-gaze was coded in ELAN as the proportion

of the total duration of the video in which the participant is looking directly into the camera. For a more detailed description of these quantifications, see Trujillo et al. (2018). Also note that the kinematic features calculated using this protocol are in line with the same features manually annotated from the video recordings (Trujillo, Vaitonyte, et al., 2019). This supports our assumption that the features calculated from the motion tracking data represent qualities that are visible in the videos.

c. *Inclusion and randomization*

Our stimuli set included 120 videos (of the 2480) recorded in our previous study (Trujillo, et al., 2018). Our selection procedure (See Appendix 4.1) ensured that our stimulus set in the present experiment included an equal number of more- and less-communicative videos. Each of the 31 gesture items from the original set was included a minimum of three times and maximum of four times across the entire selection, performed by different actors, while ensuring that each item also appeared at least once in the more-communicative context and once in the less-communicative context. Three videos from each actor in the previous study were included. Appendix 2.1 provides the full list of gesture items. Supplementary Figure 4.1 illustrates the range of kinematics, gaze, and video durations included across the two groups in the current study with respect to the original dataset from Trujillo et al., (2018). We ensured that the stimulus set for the present study matched the original dataset in terms of context-specific differences in the kinematics and eye-gaze, ensuring that the current stimulus set is a representative sample of the data shown in Trujillo et al., (2018). These results are provided in 4.1.

d. *Video segmentation*

In order to test whether kinematic modulation primarily influences early or late identification (question 2), we divided the videos into segments of different length. Based on previous literature (Kendon, 1986; Kita, van Gijn, & van der Hulst, 1998), we defined segments as following: *Wait* covered the approximate 500ms after the bell was played, but before the participant started to move. *Reach to grasp* covered the time during which the participant reached towards, and subsequently grasped the target object. In the case of multiple objects, this segment ended after both objects were grasped. *Prepare* captured any movements unrelated to the initial reach to grasp, but was not part of the main semantic aspect of the pantomime. *Main movement* covered any movements directly related to the semantic core of

	Phase 1		Phase 2		Phase 3	
	Reach-to-grasp	**Prepare**	**Main Movement**	**Auxiliary**	**Return Object**	**Retract**
Open jar	Right hands extends to jar	Right hand lifts jar. Left hand grasps lid	Twisting hands to depict unscrewing the lid	Hands moved apart to show separating lid from jar	Hands return to object starting positions	Hands returned to indicated starting position
Cut paper	Right hand extends to scissors, left hand to paper	Both hands lifted, configured to start cutting paper	Cutting motion depicted with right hand	Hands spread apart to show that the cutting is complete	Hands return to object starting positions	Hands returned to indicated starting position

Table 4. Movement phase examples

the item. *Auxiliary* captured any additional movements not directly related to the semantic core. *Return object* captured the movement of the hands back to the objects starting position, depicting the object being replaced to its original location. *Retract* covered the movement of the hands back to the indicated starting position of the hands, until the end of the video. Note that the "prepare", and "auxiliary" segments were optional, and only coded when such movements were present. All other segments were present in all videos. Phases were delineated based on this segmentation. *Phase 0* covered the "wait" segment. *Phase 1* covered "reach to grasp" and "prepare". *Phase 2* covered the "main movement" and "auxiliary". *Phase 3* covered "return object" and "retract". See Table 4 and Figure 10 for examples of how these phases map onto specific parts of the movement. After defining the segments for each video, we divided the videos into three lengths, referred to as initial fragments (M = 3.27±1.52s), medium fragments (M = 4.62±2.19s), and full videos (M = 5.59±2.53s). Initial fragments consisted of only phase 0 and phase 1, medium fragments consisted of phases 0-2, and full videos contained all of the phases. An overview of these segments and phases can be seen in figure 9. We performed ANOVAs on each of the fragment lengths to ensure video durations of

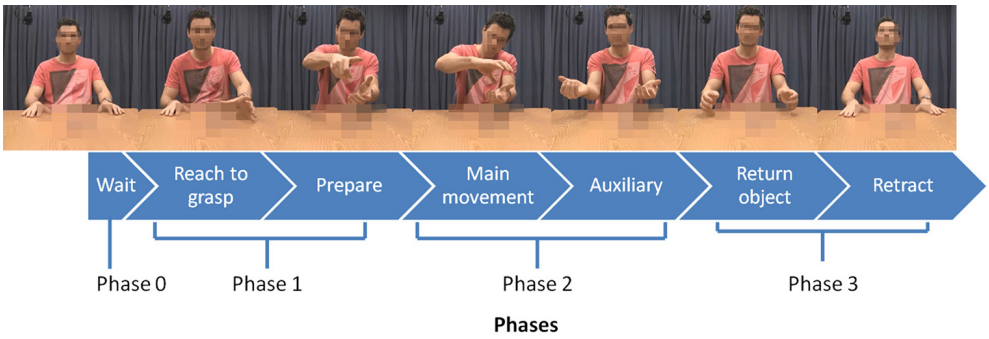


Figure 10. Overview of video segmentation and phases. Along the top, representative still frames are shown throughout one video (item: "open jar"). The individual blue blocks indicate individual segments. Below this, phase division is depicted.

the same fragment length did not differ significantly across cells (see Supplementary Table 1 for statistics). This resulted in initial fragments only providing initial handshape and arm/hand/finger configuration information, medium fragments providing all relevant semantic information, and full videos providing additional eye-gaze (when present) and additional time for processing the information.

e. *Blurring*

In all videos, a Gaussian blur was applied to the object, which was otherwise visible in the video. This ensured that the object could not be used to infer the action. To determine whether the face in general, in particular the gaze direction, has an effect on pantomime recognition, we also applied a Gaussian blurt to the face in half of the videos. Blurring the faces in this way allowed us to manipulate the amount of available visual information, providing a first test for how kinematic modulation affects gesture identification in a less complete visual context (question 3). This was balanced so that each actor had at least one video with a visible face and one with a blurred face.

Task

Before beginning the experiment, participants received a brief description of the task in order to inform them of the nature of the stimuli. This ensured that the participants knew to expect incomplete videos in some trials. Participants were seated in front of a 24" Benq XL2420Z monitor with a standard keyboard for responses. Stimuli were presented at a frame rate of 29 frames per second, with a display size of 1280x720.

Fragment Length		Context			
		Face Visibility		Face Visibility	
	Context	More-Communicative	More-Communicative	Less-Communicative	Less-Communicative
	Face	Visible	Blurred	Visible	Face Blurred
	Fragment	Initial	Initial	Initial	Initial
	Mean Duration	4.49s	5.03s	4.50s	4.03s
	Context	More-Communicative	More-Communicative	Less-Communicative	Less-Communicative
	Face	Visible	Blurred	Visible	Blurred
	Fragment	Medium	Medium	Medium	Medium
	Mean Duration	4.72s	4.43s	4.34s	4.57s
	Context	More-Communicative	More-Communicative	Less-Communicative	Less-Communicative
	Face	Visible	Blurred	Visible	Blurred
	Fragment	Full	Full	Full	Full
	Mean Duration	4.73s	4.34s	4.29s	4.61s

Table 5. Overview of analysis cells for Experiment I. There are 10 videos in each of the cells.

During the experiment, participants would first see a fixation cross for a period of 1000 ms with a jitter of 250 ms. One of the item videos was then displayed on the screen, after which the question appeared: “What was the action being depicted?” Two possible answers were presented on the screen, one on the left, and one on the right. Answers consisted of one verb and one noun that captured the action (e.g. The correct answer to the item “pour the water into the cup” was “pour water”). Correct answers were randomly assigned to one of the two sides. The second option was always one of the possible answers from the total set. Therefore, all options were presented equally often as the correct answer and as the wrong (distractor) option. Participants could respond with the 0 (left option) or 1 (right option) keys on

the keyboard. Accuracy and response time (RT) were recorded for each video.

Analysis

Main effects analyses: communicative context, fragment length, and visual context. Both RT and accuracy of identification judgments were calculated for each of 12 cells (Table 5): Fragment Length (initial fragment vs. medium fragment vs. full video) x Face (blurred vs. visible) x Context (more-communicative vs. less-communicative) in order to test 1) whether more-communicative gestures were identified faster or with higher accuracy (main effect of context), 2) performance was higher in only initial fragments (providing evidence for early identification theory) or only in medium fragments (providing evidence for late identification), as well as 3) whether face visibility impacted performance, which informs us whether there is an effect of visual information availability on the identification performance. Separate repeated-measures analyses of variance (RM-ANOVA) were run for accuracy and RT in order to test for the presence of main and interactional effects. We used Mauchly's test of Sphericity on each factor and interaction in our model and applied the Greenhouse-Geisser correction where appropriate.

Results – Experiment I

We used RM-ANOVA to test for a significant main effect of communicative context, fragment length, or face visibility on performance. In terms of accuracy, results of the fragment length x face visibility x communicative context RM-ANOVA showed a significant main effect of communicative context, $F(1,19) = 2.912$, $p = 0.029$, as well as a main effect of fragment length, $F(2,38) = 53.583$, $p < 0.001$, but no main effect of face visibility, $F(1,19) = 0.050$, $p = 0.825$. Planned comparisons revealed higher accuracy in the more-communicative context for initial fragments (More-communicative mean = 87.13%, less-communicative mean = 81.17%; $t(18) = 3.025$, $p = 0.007$) , but there was no difference between contexts in the medium fragments (More-communicative context mean = 97.37%, less-communicative mean = 96.49%; $t(18) = 0.785$, $p = 0.443$) or full videos (more-communicative mean = 97.37%, less-communicative mean = 97.22%; $t(18) = 0.128$, $p = 0.899$). In sum, performance was high overall on more-communicative compared to less-communicative videos, with specifically more-communicative initial fragments showing higher performance than less-communicative initial fragments. Accuracy, regardless of communicative context, was additionally higher in medium and full fragments compared to initial.

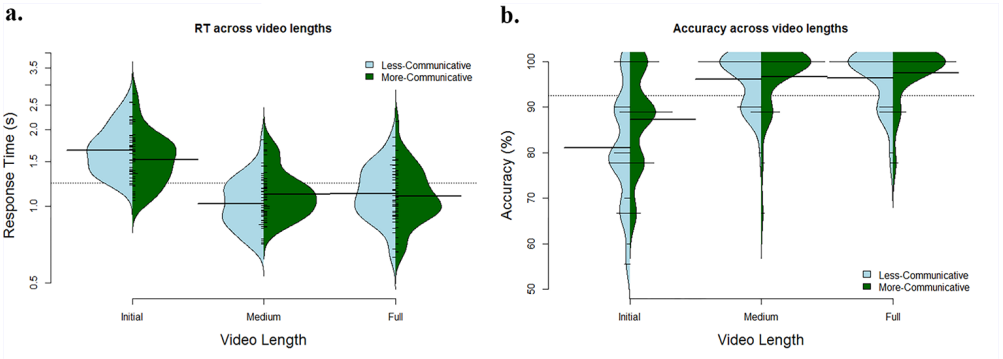


Figure 11. Overview of semantic judgment performance over context and fragment length, combined for face visibility. Bean plots depict the distribution (kernel density estimation) of the data. The dotted lines indicate the overall performance mean, the larger solid bars indicate the mean per video length and communicative context, shorter bars indicate mean values per participant, and the filled curve depicts the overall distribution of scores. Panel A shows mean accuracy across the three video lengths. Panel B shows RT across the three video lengths. In all panels, fragment length is depicted along the x-axis, the y-axis shows mean performance (in panel, mean accuracy; in panel, mean RT in seconds), while blue (left) plots depict the less-communicative context and green (right) plots the more-communicative context.

See Figure 11A for an overview of these results.

In terms of RT, results of the fragment length x face x context RM-ANOVA revealed a significant main effect of communicative context, $F(1,19) = 5.699$, $p = 0.028$, and of fragment length, $F(2,38) = 192.489$, $p < 0.001$, but not of face visibility, $F(1,19) = 3.725$, $p = 0.069$. Planned contrasts revealed faster RT in more-communicative compared to less-communicative initial fragments (More-communicative mean = 1.446; less-communicative mean = 1.583s), $t(19) = 3.824$, $p = 0.001$ but faster RT for less- compared to more-communicative medium fragments (more-communicative mean = 1.094s; less-communicative mean = 1.029s), $t(19) = 3.479$, $p = 0.003$, but no difference between more- and less-communicative full videos (more-communicative mean = 1.094; less-communicative mean = 1.129), $t(19) = 1.237$, $p = 0.231$. We also found faster RT for medium fragments (M = 1.093) compared to initial fragments (M = 1.630), $t(19) = 12.538$, $p < 0.001$, as well as for medium fragments compared to full videos (M = 1.142), $t(19) = 2.326$, $p = 0.031$. In sum, RT was similar in both the

more- and less-communicative contexts, but faster responses were seen in medium fragments compared to initial and full fragments. See Figure 11B for an overview of these results.

Discussion – Experiment I

In our first experiment, we sought to determine how communicative modulation affects identification of pantomime gesture semantics. We found that pantomime gestures produced in a more-communicative context were better recognized when compared to those produced in a less-communicative context. Specifically, more-communicative initial fragments were recognized more accurately and faster than less-communicative initial fragments.

The higher accuracy in recognizing more- compared to less-communicative initial fragments suggests that at least some of the relevant information is available even in the earliest stages of the act, and that communicative modulation enhances this information. Since the face visibility did not contribute significantly to better performance, we suggest that improved comprehension may come from fine-grained kinematic cues, such as hand-shape and finger kinematics. As objects are known to have specific action and hand-shape affordances (Grèzes & Decety, 2002; Tucker & Ellis, 2001), hand-shape can also provide clues as to the object being grasped, and thus also the upcoming action (Ansuini et al., 2016; van Elk et al., 2014). These results are therefore in line with the early prediction results described for action chains (Becchio, Manera, et al., 2012; Cavallo et al., 2016). Our results may also be explained by immediate comprehension. In other words, the visual information provided by the shape and configuration of the hands may be sufficiently clear to activate the semantic representation of the action without any prediction of the upcoming movements. Although we cannot determine the exact cognitive mechanism, we can conclude that communicative modulation supports comprehension through early action identification.

We found no evidence for higher accuracy in more- compared to less-communicative medium fragments, nor for full videos. It seems that the overall accuracy in medium and full fragments does not allow a difference to be found between the contexts. In both more- and less-communicative medium fragments, accuracy was above 96%, suggesting that ceiling-level performance may have already been reached. This indicates that even if communicative modulation supports late identification, general

task difficulty was not high enough in our task to allow us to find any difference. Surprisingly, faster RT was found for less- compared to more-communicative medium fragments. This unexpected result may reflect a trade-off between kinematic modulation, which is thought to be informative, and direct eye-gaze, which serves a communicative function but may not lead to faster responses. Along this line, Holler and colleagues (2012) argue that direct eye-gaze leads to a feeling of being addressed, which in turn forces the addressee to split their attention between the eyes and hands of the speaker. If this interpretation is correct, we would expect that although responses are faster for the less-communicative videos, accuracy should still be higher in the more-communicative videos. In order to draw any conclusions about how communicative modulation affects late identification, we suggest that it is necessary to increase task difficulty.

In sum, our results show that communicatively produced gestures are more easily recognized than less communicative gestures, and that this effect is explained by early action identification. This result is in line with the research on child-directed actions (Brand et al., 2002), as well as the more recent developments regarding early action identification based on kinematic cues (Ansuini et al., 2014; Cavallo et al., 2016).

Experiment II – Isolated Kinematic Context

Although this first experiment shows evidence for a supporting role of kinematic modulation in semantic comprehension of gestures, it remains unclear whether the effect remains when only gross kinematics are observed, and facial, including attentional cueing to the hands, and finger kinematics, including hand-shape, are completely removed. Removing additional visual contextual information would therefore help to disentangle the effects of gross (i.e. posture and hands) kinematic modulation from other (potentially communicative) visual information. For example, while extensive research has looked at the early phase of action identification from hand and finger kinematics (Ansuini et al., 2016; Becchio et al., 2018; Cavallo et al., 2016), the higher level dynamics of the hands and arms, which we call gross kinematics, have not been well studied. This is particularly relevant as these high-level kinematic features are similar to the qualities described in gesture research. Thus, in Experiment II we replicate Experiment I, but reduce the stimuli to present a visually simplistic scene consisting of only lines representing the limbs of the actor's

body. If kinematic modulation is driving the communicative advantage seen in our first experiment, we can expect the same effect pattern as seen in Experiment I. If other features of the visible scene, such as finger kinematics, provided the necessary cues for semantic comprehension then the effect on early identification should no longer be present. Due to the visual information being highly restricted, we expect task difficulty to be increased. In this way, we are able to determine if kinematic modulation supports early action identification in the absence of other early cues such as hand-shape, and whether it supports ongoing semantic disambiguation when gesture recognition is more difficult. Overall, this experiment will build on our findings from Experiment I by providing a specific test of how kinematic modulation affects semantic comprehension when isolated from other contextual information. Additionally, it will test which specific kinematic features contribute to supporting semantic comprehension.

Methods – Experiment II

Participants

Twenty participants were included in this study (mean age = 24; 16 female), recruited from the Radboud University. Participants were selected on the criteria of being aged 18 – 35, right-handed, fluent in the Dutch language, without any history of psychiatric impairments or communication disorders, and not having participated in the previous experiment. The procedure was approved by a local ethics committee and informed consent was obtained from all individual participants in this study.

Materials

We used the same video materials as in the Experiment I, but this time the videos were reduced to stick-light-figures. Motion-tracking data was used to reconstruct the movements of the upper-body joints (Trujillo, Vaitonyte, et al., 2019). Videos consisted of these reconstructions, using x,y,z coordinates acquired at 30 frames per second of these joints (see figure 12 for an illustration of the joints utilized). Note that no joints pertaining to the fingers were visually represented. This ensured that hand-shape was not a feature that could be identified by an observer. These points were depicted with lines drawn between the individual points to create a light stick-figure, representing the participants' kinematic skeleton. Skeletons were centered in space on the screen, with the viewing angle adjusted to reflect an azimuth of 20°

		Context	
	Context	More-Communicative	Less-Communicative
	Fragment	Initial	Initial
	Mean Duration	4.22s	4.24s
Fragment Length	*Context*	More-Communicative	Less-Communicative
	Fragment	Medium	Medium
	Mean Duraiton	4.68s	4.73s
	Context	More-Communicative	Less-Communicative
	Fragment	Full	Full
	Mean Duration	4.59s	4.51s

Table 6. Overview of analysis cells for Experiment II. There are 20 videos in each of the cells.

and an elevation of 45° in reference to the center of the skeleton.

Analysis

a. Main effects analyses: communicative context, fragment length, and visual context

To determine if there was an overall effect of communicative context on accuracy or RT, and to again test for evidence of either the early identification or late identification hypothesis, we used two separate 3 (Fragment Length) x 2 (Context) one-way ANOVAs (Table 6). When appropriate, independent samples t-tests were used to determine where these differences occurred across the 3 video lengths. When a non-normal distribution was detected, results are reported after a Greenhouse-Geisser correction.

b. Feature-level regression analysis: exploratory test of kinematic modulation values

Given that Experiment II aims to test the specific contribution of kinematic modulation on semantic comprehension, we additionally performed an exploratory linear mixed-effects analysis using the kinematic modulation values that characterize the stimulus

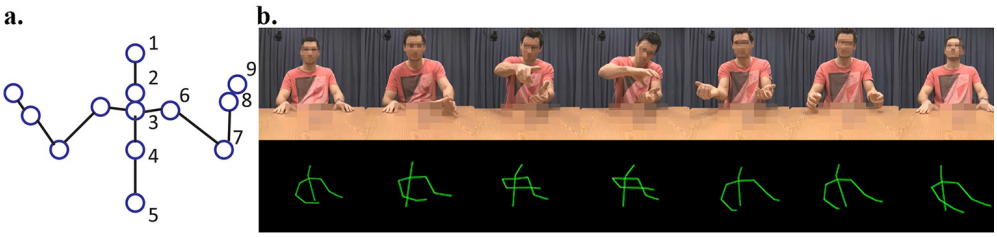


Figure 12. Illustration of materials used for Experiment II. **a**. Diagram of joints represented in the videos of Experiment II: 1. Top of head, 2. Bottom of head, 3. Top of spine, 4. Middle of spine, 5. Lower spine, 6. Shoulder, 7. Elbow, 8. Wrist, 9. Center of hand. Note that numbers 6-9 are present for both the left and right arms. **b**. Still frames from an actual stimulus video, depicting the visual information made available to the participants, underneath the corresponding actual video frames (not shown to participants) for comparison.

videos. This was done to assess the relation between specific kinematic features and semantic judgment performance. Kinematic modulation values were available from our previous study, where these stimulus videos were created (Trujillo et al., 2018), and were meant to quantify kinematic features in the semantic core of the action. We therefore chose to perform this additional analysis in Experiment II as a follow-up assessment of the significant difference between more- and less-communicative medium fragments.

We performed linear regression analyses between the set of kinematic features and RT and a logistic regression between the set of kinematic features and accuracy. Regression analyses were performed on the medium fragments as this is where a statistically significant difference was found between more- and less-communicative videos. Statistical analyses utilized mixed effects models implemented in the R statistical program (R Development Core Team, 2007) using the lme4 package (Bates et al., 2014). P-values were estimated using the Satterthwaite approximation for denominator degrees of freedom, as implemented in the lmerTest package (Kuznetsova, 2016). Our regression models first factored out video duration and subsequently tested the three main components of kinematic modulation that have been identified in previous research: range of motion (Bavelas et al., 2008; Hilliard & Cook, 2016) (here quantified as vertical space utilized), velocity of movements, and punctuality (Brand et al., 2002) (here quantified as the number of submovements and the amount of holds between them). Kinematic features were defined as main effects, while a random intercept was added for participant. For a detailed

description of how the model was defined, see Appendix 4.2. To reduce the risk of Type I error, we used the Simple Interactive Statistical Analysis tool (http://www.quantitativeskills.com/sisa/calculations/bonfer.htm) to calculate an adjusted alpha threshold based on the mean correlation between all of the tested features (regardless of whether they are in the final model or not), as well as the number of tests (i.e. number of variables remaining in the final mixed model). Our six variables (duration, vertical amplitude, peak velocity, submovements, hold-time) showed an average correlation of 0.154, leading to a corrected threshold of $p = 0.019$.

Results – Experiment II

a. Main effects analyses: communicative context, fragment length

Our first RM-ANOVA tested whether accuracy was affected by the communicative context, or the fragment length of the videos. We found a significant main effect of communicative context on accuracy, $F(1,19) = 5.108$, $p = 0.036$, as well as a main effect of fragment length, $F(2,38) = 10.962$, $p<0.001$. Planned comparisons revealed no difference between accuracy of more-communicative and less-communicative initial fragments (more-communicative mean = 59.58%, less-communicative mean = 56.76%), $t(19) = -0.646$, $p = 0.526$, or in full videos (more-communicative mean = 64.87%, less-communicative mean = 62.76%), $t(19) = 0.492$, $p = 0.628$. We found significantly higher accuracy in more-communicative medium fragments (M = 75.69%) compared to less-communicative medium fragments (M = 66.11%) videos, $t(19) = 2.99$, $p = 0.007$. We found no fragment length by communicative context interaction, $F(2,36) = 0.659$, $p = 0.523$.

Our second RM-ANOVA tested whether RT was affected by communicative context or fragment length. We found a significant main effect of fragment length on RT, $F(2,38) = 7.263$, $p = 0.003$, but no main effect of communicative context, $F(1,19) = 2.12$, $p = 0.162$. We additionally found a video length x context interaction, $F(2,38) = 3.87$, $p = 0.031$. Planned comparisons revealed significantly faster RT in medium fragments (M = 1.817s) compared to initial fragments (M = 1.953s), $t(19) = 3.982$, $p = 0.001$, but no difference between medium fragments and full videos (M = 1.872s), $t(19) = 1.339$, $p = 0.196$. See figure 13 for an overview of these results. In sum, communicative context did not affect RT, but responses were faster in medium compared to initial fragments.

b. Feature-level regression analysis: exploratory test of kinematic modulation

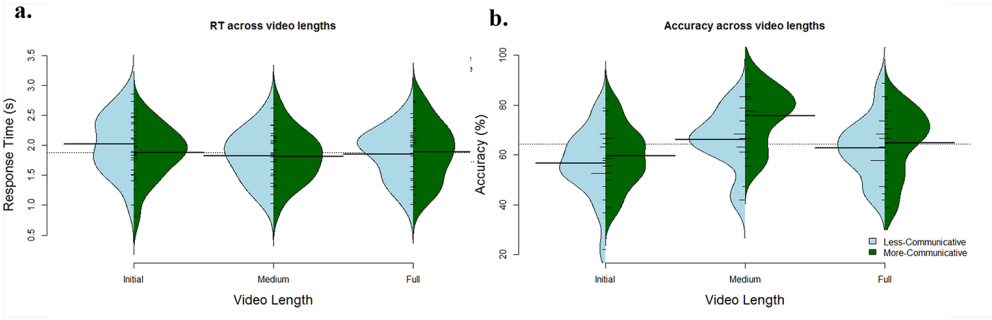


Figure 13. Overview of semantic judgment performance over context and fragment length in Experiment II. Bean plots depict the distribution (kernel density estimation) of the data. The dotted lines indicate the overall performance mean, the largest solid bars indicate the group mean per video length and context, and shorter bars indicate individual participant means. Panel A shows mean accuracy across the three video lengths. Panel B shows RT across the three video lengths. In all panels, fragment length is depicted along the x-axis, the y-axis shows mean performance (in panel, mean accuracy; in panel, mean RT in seconds), while blue (left) plots depict the less-communicative context and green (right) plots the more-communicative context.

values

To test which specific kinematic features, if any, affected accuracy, we used mixed models to assess whether accuracy on each video could be explained by the kinematic features of that video. We found kinematic modulation of punctuality (hold-time and submovements) to explain performance accuracy better than the null model, χ^2 (5) = 16.064, $p < 0.001$. Specifically, we found kinematic modulation of punctuality (hold-time and submovements) to explain performance associated with higher accuracy (b = 0.377, z = 3.962,$p < 0.001$), although submovements were not (z = -0.085, p = 0.932). We found no correlation between duration and accuracy (z = -1.151, p = 0.249) in our kinematic model. Response time was not significantly explained by any of the kinematic feature sets. Duration, as assessed in the null model, was also not related to response time (t = –1.768, p = 0.077). In sum, kinematic modulation of hold-time was specifically related to higher performance accuracy.

Discussion – Experiment II

Experiment II was designed to test the isolated contribution of kinematics to semantic comprehension and further differentiate between early identification

versus late identification. We found that more-communicative videos were still recognized with overall higher accuracy than less-communicative videos even in the absence of contextual cues such as hand-shape, finger kinematics, or actor's face.

Higher accuracy in recognizing more-communicative compared to less-communicative medium fragments suggests that the advantage given by kinematic modulation predominantly affects identification of the pantomime after it has unfolded. The unfolding of the final phase of the pantomime may provide enough extra time for the overall act to be processed completely and the pantomime to be recognized accurately regardless of modulation. This finding is therefore in line with the hypothesis that kinematic modulation mainly contributes to ongoing semantic disambiguation. We further explored the contribution of specific kinematic features to semantic comprehension in the absence of further visual context such as hand-shape or facial cues. We found that temporal kinematic modulation (i.e. increasing segmentation of the act) was an important factor influencing semantic comprehension. Specifically, increasing hold-time positively impacted accuracy. Our results suggest that although the effect may be subtle in production, this feature plays an important role in clarifying semantic content through temporal unfolding of the gesture.

General Discussion

This study aimed to determine the role of kinematic modulation in the semantic comprehension of (pantomime) gestures. First, we asked whether kinematic modulation influences semantic comprehension of gestures and found that more-communicatively produced gestures are recognized better than less-communicatively produced gestures (Experiments I & II). Second, by utilizing different video fragment lengths, we tested the underlying mechanism of this communicative advantage. We found evidence for enhanced early identification when provided with a more complete visual scene, including the hand shape (Experiment I), but enhanced late identification when provided with only gross kinematics (Experiment II). Finally, we show in Experiment II that increased post-stroke hold-time has the strongest effect on the communicative gesture comprehension advantage.

When provided with a wealth of visual cues, as in Experiment I, participants gained a communicative advantage even in the early stages of movement. This finding fits nicely with the idea that the end goal of an action, or perhaps the upcoming

movements themselves, can be predicted by utilizing early kinematics together with visual contextual information (Cavallo et al., 2016; Iacoboni et al., 2005; Stapel et al., 2012). Our results from the Experiment II suggest that kinematic modulation of gross hand movements alone is not sufficient for this effect as when the visual stimulus was degraded this advantage was removed. It should be noted that we cannot conclude that kinematic information is insufficient, but rather that the gross hand kinematics that are typically used to assess gestures are insufficient. This is particularly relevant given the evidence that hand and finger kinematics inform early manual action identification (Becchio et al., 2018; Cavallo et al., 2016; Manera et al., 2011). We therefore conclude that both kinematic and non-kinematic cues play a role in early gesture recognition, while modulated arm and hand kinematics provide cues to identify the act as it unfolds, even in the absence of other visual cues.

Our conclusion regarding the role of temporal modulation, and more specifically the increased hold-time, as supporting semantic comprehension matches well with the factor 'punctuality', as defined by Brand and colleagues (Brand et al., 2002) in their study of child-directed action. Punctuality of actions refers to movement segments with clear beginning and end points, allowing the individual movements to be clear to an observer (Blokpoel et al., 2012). Exaggerating the velocity changes between movements and increasing hold-time (Vesper et al., 2017) can make the final body configuration more salient by allowing longer viewing time of this configuration for the addressee.

Our findings have several important implications. By combining naturalistic motion-tracking production data with a semantic judgment task in naïve observers, our study provides new insights and support for models of effective human-machine interactions. Specifically, our results expand and contrast the robotics literature that demonstrate spatial modulation as a method of defining more legible acts (Dragan, Lee, & Srinivasa, 2013; Dragan & Srinivasa, 2014; Holladay et al., 2014). Our findings suggest that while spatial modulation may be effective for single-movement gestures such as pointing, temporal modulation has a larger role in this clarification effect in more complex acts.

We additionally build on studies of gesture comprehension, showing the importance of kinematic cues in successful semantic uptake and bringing new insights to previous findings. For instance, our findings provide a mechanistic understanding

of larger scale, qualitative features, such as informativeness (Campisi & Özyürek, 2013). Differences in the informativeness of complex gestures may be understood by looking at the underlying kinematic differences and how these relate to the comprehension of such gestures. As an example, gestures are understood through the individual movements that comprise them, rather than static hand configurations (Kendon, 2004; McNeill, 1994). Increasing the number of clearly defined movements consequently increases the amount of visual information available to an observer, which could lead to the perception of increased informativeness.

Our work has further implications for clinical practice, where it can be applied to areas such as communication disorders. Research has shown that people with aphasia use gestures, including pantomimes, to supplement the semantic content of their speech (deBeer, Carragher, van Nispen, de Ruiter, Hogrefe & Rose, 2015; Rose, Mok, & Sekine, 2017). Knowledge of which features contribute to semantically recognizable gestures could therefore be applied to developing therapies for more effective pantomime use and understanding.

Summary

Our study is the first to systematically test and provide a partial account of how the kinematic modulation that arises from a more-communicative context can support efficient identification of a manual act. We found that communicatively produced acts are more easily understood early on due to kinematic and non-kinematic cues. While comprehension is dependent on how much of the visual scene is available, communicative kinematic modulation alone leads to improved recognition of pantomime gestures even in a highly reduced visual scene. Particularly, temporal kinematic modulation leads to improved late identification of the act in the absence of other cues.

Acknowledgements

The authors are grateful to Ksenija Slivac for her contribution to stimulus preparation and data collection in Experiment I, as well as Muqing Li for her contribution to data collection and analyses in Experiment II. We additionally thank Louis ten Bosch for his insights and discussions regarding methodology. This research was supported by the NWO Language in Interaction Gravitation Grant. The authors declare no conflict of interest in this study.

Appendix 4.1.

Item Selection Procedure. To provide a representative sampling of each of the two groups, all individual items from all subjects included in the previous study were ranked according to eye-gaze and overall kinematic modulation (z-scores derived from the kinematic features described in the section *b*). The two groups were ordered such that items with high values for addressee-directed eye-gaze and kinematic modulation were ranked higher than those with low values. This placed all items on a continuum that ranked their communicativeness. This was done due to the observation that, due to the subtle manipulation of context in Experiment I of Trujillo et al. 2018, there was considerable overlap of kinematic modulation in the middle of the spectrum (i.e. Some actors in the more-communicative context showed modulation more similar to those of the less-communicative context, and vice-versa). We chose to include items which represented a range of eye-gaze and kinematic features representative of their respective communicative context. This method allowed a more clear separation of the contexts, while our further selection procedure (described below) ensured that items were included across a wide range of this ranked continuum.

After creating the ranked continuum of items, inclusion moved from highest to lowest ranked items. Each of the 31 items, as described in Appendix 2.1, was included a minimum of three times and maximum of four times across the entire selection, performed by different actors, while ensuring that each item also appeared at least once in more-communicative context and once in the less-communicative context. Three videos from each actor were included. This ensured an even representation of the data from our previous study. Supplementary Figure 4.1 illustrates the range of kinematics, gaze, and video durations included across the two groups in the current study with respect to the original dataset.

We ensured that the current stimulus set was representative of the original data by repeating the same mixed-model analyses described in Trujillo et al. (2018). In line with the original dataset, we found higher values in communicative compared to non-communicative Vertical Amplitude (Communicative = 0.160±0.99; Non-Communicative = -0.449±0.809; $\chi^2(4)= 12.263$, $p < 0.001$), Submovements (Communicative = 0.161±789; Non-Communicative = -0.661±585; $\chi^2(4)=32.821$, $p < 0.001$), Peak Velocity (Communicative = 0.181±1.08; Non-Communicative = -0.683±0.649; $\chi^2(4)=23.965$, $p = 0.001$), and direct eye-gaze (Communicative = 0.235±0.220; Non-Communicative = 0.013±0.041; $\chi^2(4)=44.703$, $p < 0.001$). Also in line with the original data, we found a less robust difference in Hold-time (Communicative = 0.107±1.159; Non-Communicative = -0.448±0.892; $\chi^2(4)= 7.917$, $p = 0.005$). Duration was also longer in Communicative (M = 7.237±1.754) compared to Non-Communicative (M = 6.132±1.235) videos.

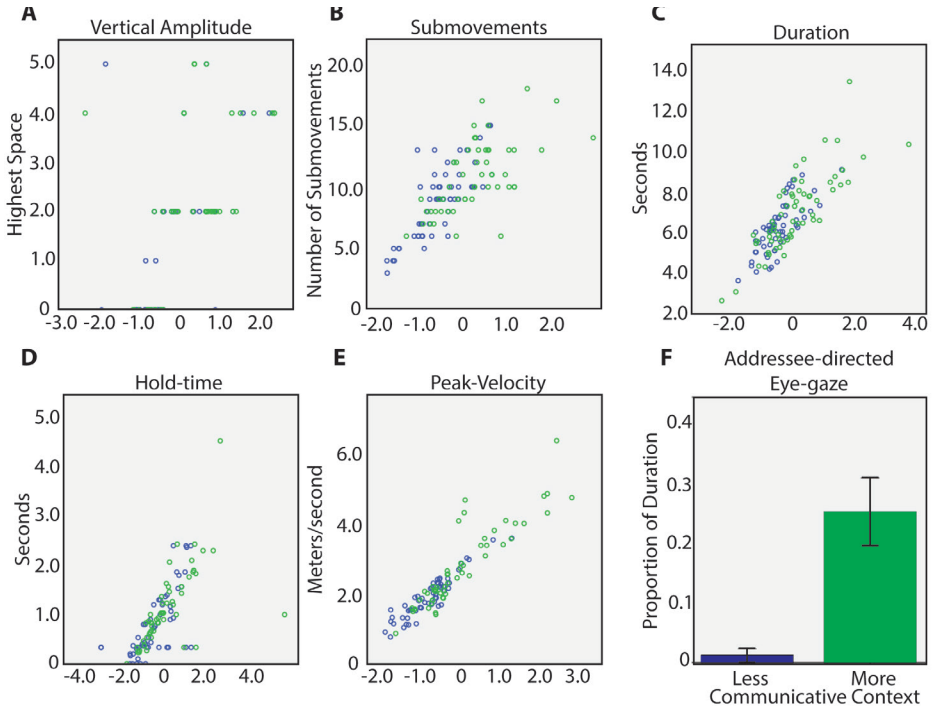
Appendix 4.2.

Mixed Effects Modeling Procedure. The order in which the predictor variables were entered into the mixed effects model was determined based on the a priori hypothesized contribution of the three components: range of motion has been found to be increased in adult-child interactions (Brand et al., 2002; Fukuyama et al., 2015); peak velocity was found to be increased in a communicative context in at least one study (Trujillo et al., 2018); punctuality was previously not found to be changed in child-adult interactions by (Brand et al., 2002), but was found to be increased in a communicative context by (Trujillo et al., 2018).

As more-communicative videos were, on average, longer than less-communicative videos, we included video duration (ms) in our regression models. This allowed us to test the contribution of kinematic features after taking into account total duration, ensuring that any effect of kinematics is not explained by duration alone. We report the video duration correlation from the best-fit model if this model is a better fit to the data than the null model. If the null model is a better fit, then we report the video duration correlation from the null model. Duration was fitted before the kinematic variables in order to ensure that any significant contribution of kinematic modulation to the model fit was over and above that of duration. In other words, our models were set up to specifically test the contribution of kinematic modulation after taking into account video duration and inter-individual differences.

Typically, when utilizing mixed-effects models the researcher must first find the model that is the best-fit for the data before making inferences on the model parameters. The best-fit model was determined by first defining a 'null' model that only included duration and as fixed effect and participant as random intercept. We used a series of log-likelihood ratio tests to determine if each kinematic feature term (described above: range of motion, velocity, punctuality) contributed significantly to the model fit. For example, if a comparison between a model that includes peak velocity, with a model that does not include this effect term yields a non-significant result, then we do not include this kinematic feature in the model. If the comparison yields as a significant result, we keep this kinematic feature and compare this model with a new model that contains the next non-tested kinematic feature. In a step-wise fashion we thus test the contribution of each of the kinematic features. We report effects from the final, best-fit model, if it is still a better fit than the null model.

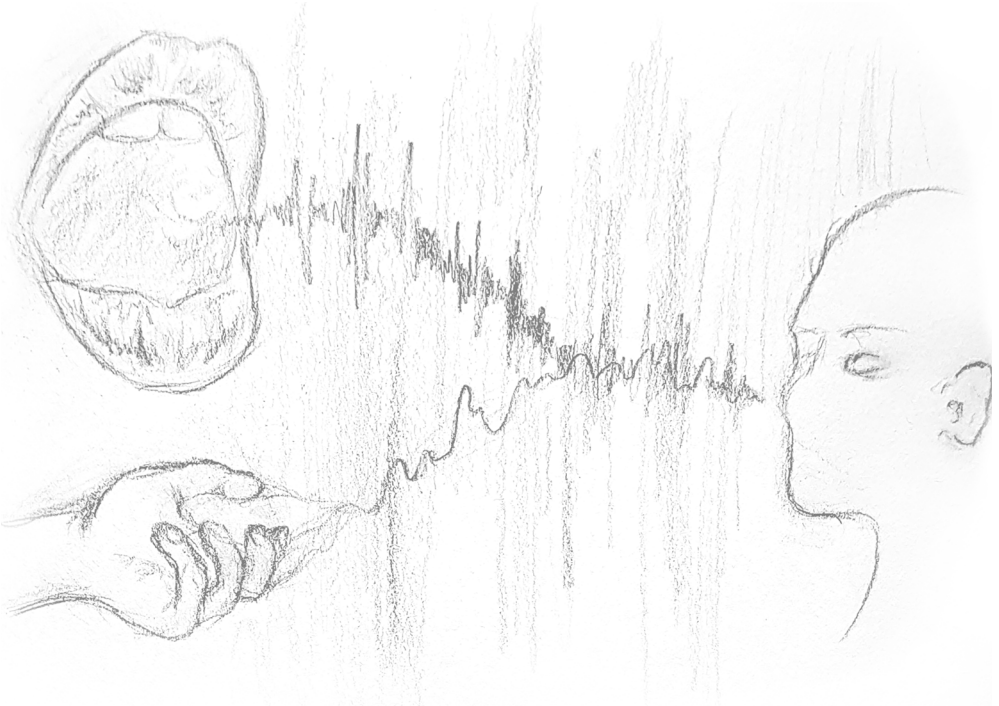
Supplementary Material



Supplementary Figure 4.1. **A-E**. Overview of raw and modulation values for kinematics and duration of included videos. In all scatter plots the y-axis depicts raw values, while the x-axis depicts modulation (z-score) values. Blue circles are less-communicative videos, green circles are more-communicative videos. **F**. Comparison between more-communicative and less-communicative selections of the proportion of the total duration during which addressee-directed eye-gaze was detected.

Supplementary Table 4.1. Comparison of video durations across conditions for Experiment I

	df	*F*	*p*
		Initial	
Face visibility	1	0.001	0.977
Communicative Context	1	0.202	0.656
Residual	34		
		Medium	
Face visibility	1	0.642	0.429
Communicative Context	1	3.404	0.074
Residual	34		
		Final	
Face visibility	1	2.361	0.133
Communicative Context	1	3.129	0.086
Residual	34		



Chapter 5

Kinecting Speech, Noise, and Gesture: Evidence for a Multimodal Lombard Effect

CHAPTER 5

Abstract

In many natural face-to-face interactions, we are challenged with communicating in non-ideal settings, such as noisy environments. Typically, we are able to successfully communicate despite interference from noise. This is partially due to communicative adaptations made by the speaker. The classic example of such adaptation is the Lombard Effect, which refers to involuntary changes in speech intensity and pitch in a noisy environment. Until now however, there is no research on how co-speech gesture is adapted to such situations when there are changes in noise, and whether and how speech production is different when paired with gestures.

Here, we present results from a dyadic communication task carried out at the Lowlands music festival. In the task, participants wore headphones with varying levels of noise. One participant, called the Producer, communicated action verbs to the Addressee. We use quantitative motion capture methods to assess kinematic features of both visible speech and gesture, and acoustic analysis of the speech signal.

Results show that 1) increasing levels of noise are associated with an increase in speech intensity and the kinematics of gestures, and 2) while in moderate noise these modulations occur either as increased speech acoustics paired with decreased gesture kinematics, or vice-versa, in severe noise increased speech acoustics and gesture kinematics go hand-in-hand. This demonstrates that the Lombard response to noise is not constrained to speech, but is a truly multimodal, communicative adaptation.

Introduction

When communicating in natural face-to-face interactions, we often find ourselves in noisy situations, such as a cocktail party or a crowded restaurant. In these cases, our interactional partner may have trouble understanding what we are saying due to our speech being degraded by the background noise. Previous research has shown that in such noisy environments speakers modulate auditory and visual (e.g. lip movements) features of their speech (Davis et al., 2006; Kim et al., 2005), and that these modulations help a listener to understand the degraded speech signal (Davis et al., 2006). Speech is often considered the main communicative signal, but it is complimented and heavily integrated with signals from the face and body. Multimodal communication, using our facial expressions (Ekman & Rosenberg, 1997) and our hand gestures (Kendon, 2004), can be helpful in noisy situations when verbal communication fails. Recent work has additionally shown that iconic co-speech gestures also help listeners to understand degraded speech (Drijvers & Özyürek, 2017). However, given the integrated role of gestures in speech and communication more generally (Kita & Özyürek, 2003), an interesting question is whether these iconic gestures are also modulated by the presence of noise. If this is the case, then this would be evidence for a truly multimodal, communicative adaptation to noise.

When speaking in noise, there is an automatic modulation of speech that is known as the Lombard effect (Lombard, 1911). This modulation can generally be seen as in increase in vocal effort, but specifically includes an increase in speech intensity (i.e. loudness), a shift in the fundamental frequency (F0; perceived as pitch), elongation of vowels, and increased speech rate. While this effect is partially reflexive (Pick, Siegel, Fox, Garber, & Kearney, 1989), it is also further modulated by communicative setting, with Lombard effects being enhanced when the speaker has a partner (Garnier, Henrich, & Dubois, 2010; Junqua, Fincke, & Field, 1999; Lane & Tranel, 1971). Importantly, these modulations also make the speech signal easier for listeners to understand (Cooke, Lecumberri, & Barker, 2008; Pittman & Wiley, 2001). This suggests that modulation of the speech signal in response to noise is at least partially a communicative adaptation designed for the listener.

Beyond the speech signal itself, our lip movements while speaking also provide information for our listener, allowing them to better understand speech in noise (Drijvers & Özyürek, 2017; Ma, Zhou, Ross, Foxe, & Parra, 2009; Macleod &

Summerfield, 2009; Sumby & Pollack, 1954). In line with the view of speech modulation as a communicatively intended adaptation, we can adapt not only the auditory features of speech, but also visible aspects of speech, such as lip movements (Davis et al., 2006; Kim et al., 2005). These modulations, which are positively correlated with speech intensity (Davis et al., 2006) are also related to better speech perception in noise (Kim, Sironic, & Davis, 2011). Furthermore, while these visible speech modulations are present even in non-interactive settings such as reading aloud, the modulation is greatest when there is an interactive partner (Fitzpatrick et al., 2011a; Garnier, Ménard, & Alexandre, 2018). This communicatively intended increase in the modulation is also related to better speech comprehension performance when compared to non-communicative Lombard speech comprehension (Fitzpatrick, Kim, & Davis, 2011b). Taken together with the speech acoustic modulations, there appears to be a communicatively intended audio-visual speech modulation in response to noise. This modulation of the two signals is furthermore utilized to increase intelligibility when speech is degraded by noise.

Our communicative message is conveyed not through audio-visual speech alone, but also through co-speech representational gestures (McNeill, 1994; Özyürek, 2014). Representational gestures are the hand-movements that visually represent objects, actions, events, or spatial relations through the movements and configurations of the hands (Kendon, 2004; McNeill, 1994). When paired with speech, gestures contribute to the overall perceived meaning of an utterance (Beattie & Shovelton, 2002; Holler, Shovelton, & Beattie, 2009; Özyürek, 2014). Similar to speech, gestures can also naturally occur in the absence of a communicative context (Chu & Kita, 2011). Also similar to audio-visual speech, different communicative contexts lead to kinematic differences in gestures. For example, compared to gestures produced for an adult, gestures produced for children are larger and more precise (Campisi & Özyürek, 2013). This extends to gestures made in the absence of speech. Our previous study found that silent gestures performed with the intention to communicate were modulated in terms of their spatial and temporal kinematics when compared to the same gestures performed with no incentive to communicate (Trujillo et al., 2018). Similarly, the temporal and spatial characteristics of pointing gestures are adapted to the presence (Peeters et al., 2015) and viewpoint (Winner et al., 2019) of an addressee. Together with studies of audio-visual speech signals, it therefore seems that each communicative signal can be adapted to better suit the communicative context in which it is produced.

These modulations in audio-visual speech and gesture have typically been discussed as evidence for signal-specific effects of the (communicative) context in which they are produced. An alternative explanation for these findings is that individuals put more effort into their communication whenever this is necessary. This would follow from more recent experiments showing that peaks in gesture effort (e.g. peak velocity) lead to peaks in F0, simply due to the biomechanical coupling of the two articulators (Pouw et al., 2019). A similar phenomenon is borne out in sign language in what is called echo phonology, where mouth movements "echo" the temporal and movement characteristics of the hand movements (Woll, 2014; Woll & Sieratzki, 1998). Importantly, these mouth actions are obligatory for correctly producing the sign, but they do not carry any meaning in isolation from the hand movements (Crasborn, Van Der Kooij, Waters, Woll, & Mesch, 2008). This provides further support for a neurobiological coupling between the hands and mouth. While these accounts suggest a lower-level speech-gesture coupling, several higher-level cognitive accounts would also predict some degree of coupling. The information packaging hypothesis (Kita, 2000) and the interface hypothesis (Kita & Özürek, 2003; Ozyürek, 2010), for example, both suggest that speech and gesture interface during the conceptual planning phase of production. In the information packaging hypothesis, the spatio-temporal nature of gestures allows the packaging of information into units that can be conveyed in speech (Kita, 2000). In the interface hypothesis, the linguistic structure of speech constrains how gestures are planned, and thus potentially how these packages of information can be formed (Kita & Özürek, 2003; Ozyürek, 2010). In other words, gestures allow information to be broken down into workable chunks, and linguistic structure provides some constraint to how these chunks can be organized (Kita et al., 2007). Adaptation to communicative context may therefore occur at the level of this multimodal 'message generator'. Importantly, this perspective assumes that speech and gesture are linked, but this link is dynamic, rather than fixed. Understanding how multimodal utterances are shaped in response to communicatively challenging situations would help to elucidate how speech and gesture interact, and how this dynamic is further shaped by the environment (e.g. communicative context). However, it is currently not known if and how gestures are modulated in response to noise, or how speech produced in noise is the same when it is produced together with gestures as compared to without gesture.

When considering audio-visual speech and gesture produced in communicatively challenging situations, there is also the question of whether both audiovisual speech

and gesture would be modulated. In other words, we could put our effort into both speech and gesture, or strategically put more effort into one or the other. This question has also been asked in regard to cognitive constraints of speech-gesture production. Rather than any asymmetry between the two, De Ruiter and colleagues suggested that speech and gesture parallel one another in terms of the effort being afforded to them (de Ruiter et al., 2012). This finding is in line with Pouw and colleagues' description of biomechanical effort driving speech-gesture synchrony (Pouw et al., 2019), as well as the hypothesis put forward by So, Kita, and Goldin-Meadow (2009) that gesture and speech go hand-in-hand. In their framework, more speech should result in more gestures, and less speech should be paired with fewer gestures (So, Kita, & Goldin-Meadow, 2009).

In contrast to the idea of gestures and speech largely paralleling one in quantity and content, other studies have found that when communication becomes more difficult due to increased cognitive load or conceptual difficulty, there is an increase in the quantity of co-speech gestures, while the number of words produced remains the same or even decreases (Hoetjes & Carro, 2017; Hostetter & Alibali, 2004; Melinger & Kita, 2007). However, these findings are based on situations in which communicative difficulty is manipulated in terms of cognitive load or conceptual difficulty, thus affecting the producer alone. There was no disruption of the actual communicative signals, only in the ease with which the producer could actually conceptualize or produce the relevant information. The presence of external noise likely does not make the information more difficult to externalize, but rather forces him or her to adapt the communicative signals in order to overcome the noise.

Similar to the shift towards gestures found by Hoetjes and colleagues (Hoetjes & Carro, 2017; Hoetjes, Krahmer, & Swerts, 2015), Fitzpatrick and colleagues found a shift from auditory modulation to visible speech (i.e. lip/mouth movement) modulation specifically in face-to-face interaction (Fitzpatrick et al., 2011a). In other words, speech acoustics were modulated more when the speaker's face was not visible to the addressee, but when their face was visible to the addressee there was less speech acoustic modulation but more modulation of lip movements. This suggests that speakers may selectively modulate either the auditory or visual modality, depending on which is more useful. An interesting question is whether noise leads to a modulation of any modality that is currently being used, or whether this noise modulation only occurs selectively in one modality, or even one signal, at

a time. The latter would suggest that communicative modulation may be a focused strategic adaptation, whereas the former would be evidence for a more general, potentially effort-based adaptation that affects any communicative articulator in use. Specifically, we could expect gestures to be more modulated than speech, as the visual signal is always useful, and this may be a more salient visual signal than the lips. Speech and lips, on the other hand, may be more strongly modulated when gestures are not present. Given that comprehension of moderately degraded speech benefits more from gestures than severely degraded speech (Drijvers & Özyürek, 2017), such a strategic modulation of specific signals may also depend on the amount of noise, and whether the presence of gestures are sufficient to disambiguate the speech. For example, it may be that if speech is not useful, such as in severe noise, people strategically shift to using gestures only, or gesture and visible speech. This is particularly relevant because it would tell us more about the underlying mechanisms of communicative adaptation. In other words, it would help to explain how auditory and visual components of speech are coupled with gesture during communicative adaptation to noise.

In the current study, we used a live dyadic interaction task to test whether communication in noise leads to a general adaptation of both audiovisual speech and gesture, or to a more strategic adaptation of one or the other. Specifically, we use audio recording and markerless motion tracking to assess the influence of noise on speech acoustics (i.e. auditory speech), face kinematics (visible speech), and gesture kinematics, as well as their interaction with one another.

We kept the communicative context the same throughout the experiment by having participants try to communicate action verbs to another participant. This relatively unconstrained task, along with the non-traditional lab environment (i.e. at a music festival), allowed a more naturalistic, ecologically valid elicitation of communicative behavior in noise. We used multi-talker babble, played through headphones worn by both participants, to induce a Lombard effect in the participants. While listeners had a constant (moderate) noise level, speakers had either a clear condition, moderate noise or high noise. The three noise levels were used due to the finding that gestures are most beneficial to comprehension of degraded speech at a moderate noise level (Drijvers & Özyürek, 2017). We hypothesize that multimodal adaptation to noise is strategically and communicatively motivated, and thus expect that acoustic modulation in response to moderate noise will be lower when speech is produced

together with gestures, given that gestures are predicted to be most beneficial in this moderate condition. As gestures are generally less helpful in high noise levels (Drijvers & Özyürek, 2017), we expect both audio-visual speech and gesture to be most strongly modulated in this condition. Given the addressee can always see the producer, we expect modulation of lip movements to be independent of whether the utterance is speech-only or speech and gesture (i.e. multimodal).

While participants attempted to communicate these words through the noise, we captured several features of their audio-visual speech and gestures. We investigated speech acoustic measures that have previously been most strongly linked to Lombard speech, namely the intensity and F0. We calculated face kinematics that we expected to represent main aspects of visible speech based on previous research, such as mouth opening distance (Fitzpatrick et al., 2011a; Garnier et al., 2018), along with total lip movement and lip velocity, which we see as representative of the more general lip and jaw movement parameters captured by the principle component analysis of Davis and colleagues (Davis et al., 2006; Kim et al., 2005). For gesture kinematics, we looked at the features that have previously been linked to communicatively intended modulation, such as velocity, holds, and size (Trujillo et al., 2018).

To test whether audiovisual speech and gestures are modulated as a general increase in effort or as a strategic modulation of one or the other, we run an additional test on any communicative feature (e.g. auditory speech, visible speech, or gesture features) that is found to be modulated by noise. For these features, we test whether the presence or absence of the other modality (i.e. presence of speech when testing gesture kinematics, or presence of gestures when testing auditory or visible speech) interacts with the main effect of noise. In other words, we ask whether noise modulation of any specific feature is part of a general increase in effort, or a strategic use of either the visual or auditory modality.

In sum, this study aims to further elucidate the functional cooperation between speech and gesture, and how the demands of a communicative context can shape the acoustics and kinematics of a communicative utterance. Second, we asked whether noise leads to a general increase in communicative effort (i.e. modulation of all signals) or a strategic modulation of the most useful signal (i.e. gestures in multimodal utterances, or speech acoustics and lips in speech-only utterances). We

predicted that increased noise would lead to an increase in speech intensity and F0, as well as an increase in total lip movement, lip velocity, and mouth opening, together with an exaggeration of the spatial and temporal kinematics of gestures. In terms of the distribution of effort, we predicted that visible speech would be more strongly modulated when produced without co-occurring gestures in the moderate noise condition, indicating a shift to the visual modality when this is still beneficial. To this end, we ask how the presence of noise influences the modulation and interplay speech acoustics, face kinematics, and gesture kinematics.

Methods

Data was collected at *Lowlands Science*, a science-outreach focused event taking place at the three-day Lowlands music festival in Biddinghuizen, The Netherlands on August 17-19, 2018. The festival is attended by approximately 55,000 people. Festival goers can freely enter the Lowlands Science terrain to participate in a number of socio-psychological experiments. Experiments are advertised simply by a short name, in the case of the current study this name was “Praten in 3D” [Talking in 3D]. All participants were tested between noon and 8pm across three consecutive days. We obtained ethics approval from the Faculty of Arts of the Radboud University Nijmegen prior to the festival. Participants were required to give consent to the use of their data for scientific research prior to participating, with the option to give consent for use of images and videos in publications and/or popular media. Participants did not receive any financial compensation. Prior to participating, we collected information regarding participants’ age, gender, hand-preference, number of alcoholic beverages consumed on the day of the experiment, and whether any drugs were used on that day.

Participants

In total we tested 91 pairs of participants, resulting in an initial sample size of 182 participants. In an initial screening of the data we excluded participants for whom there was incomplete data or technical problems during acquisition, those who appeared intoxicated, and those who appeared to have memorized the list of words before participating. For the purpose of this study we additionally limited our analyses to the first participant (the Producer) in each pair (see subsection Paradigm for an explanation). This led to a sample size of 58 participants included in the current study. Of these participants, there were 32 females, 20 males, and

six with missing gender information. Mean age of our sample was 27.75±7.9 years. All were native speakers of Dutch. Fifty-four were right handed. For the four left-handed participants, we focused on left-handed kinematics, rather than right handed kinematics. In terms of alcohol consumption, 24 had no alcohol on testing day, 24 had between one and three beverages, six had between four and six beverages, and four had six or more beverages. Four participants indicated have used drugs in the past 24 hours. The set-up can be seen in figure 14.

Set-Up

Participants took part in the experiment in pairs, with one starting as the "Producer" and the other starting as the "Addressee". The two participants were separated by a one-way screen that reduced visibility of the Addressee to the Producer, but allowed the Addressee to see the Producer. Both participants wore noise-cancelling headphones. The addressee always heard 4-talker babble[1], while sound in the Producer's headphone varied randomly from round (i.e. word) to round between clear (no noise), 4-talker babble, and 8-talker babble. Noise volume for both participants was manually adjusted by the experimenter to achieve the highest volume that participants could tolerate without being painful. Producers were recorded using two Microsoft Kinects and one video camera, all positioned approximately one meter away from the Producer, positioned at head-height directly next to the one-way screen. One Kinect was used to track whole body motion, while the second was used to track the face[2]. The addressee was recorded using a video camera, positioned one meter away, directly next to the one-way screen. The Addressee additionally wore eye-tracking glasses. However, given the focus of this study was on the Producer, we will not discuss data or results from the Addressee.

1 Multi-talker babble is pre-recorded audio in which speech from multiple speakers are overlaid on top of each other, thus simulating noise similar to that of a noisy cocktail party or restaurant. This type of noise was utilized as it may have a stronger effect on speech production when compared to white noise (Kim et al., 2005).

2 The Microsoft Kinect was used to capture both face and gesture kinematics. Although there is currently no research using the Kinect for face tracking, we utilized the Kinect in order to determine whether this system could detect meaningful changes in facial (i.e. lip) kinematics. The potential advantage of the Kinect over standard video-based approaches is that its use of depth images allows us to capture a participant's face in 3D while only using the single face-tracking Kinect. This novel methodological approach therefore has further implications for the further development of markerless face tracking in future studies.

Task

The task of the Producer was to communicate a verb to the Addressee. The word was presented by one of the experimenters using large white flashcards with the word printed in black. Producers could speak, gesture, or use any combination of movement and speech, as long as they remained standing at the point they started. This prevented participants from approaching the screen, moving around it, or otherwise moving out of view of the cameras and Kinects. Each round ended either when the Addressee correctly verbally identified the word, or when the experimenter determined that too much time had passed. The latter typically occurred after approximately 30 seconds. Feedback was given as a "thumbs-up" or "thumbs-down" gesture by the experimenter, signaling that the Producer could stop and wait for the next word. Sound level for the Producer was controlled via a button-press that was given at the end of each round.

Data and Processing

Video was recorded at 25 frames per second (fps) with audio sampling at 44,100Hz. The Kinect tracked the face and body at 30fps. Kinect data for the body consisted of 25 tracked joints. Tracking for the face consisted of 1,324 points.

Kinect data was used to calculate a set of kinematic features describing movements of the face and hands (see subsection Feature Calculation). All motion tracking data was smoothed using a Savitsky-Golay filter with a span of 15 and degree of 5 to correct for artefacts in the tracking. All data smoothing and kinematic features were calculated using MATLAB 2015a (The MathWorks, Inc., Natick, Massachusetts, United States) using a modified version of our kinematic feature extraction toolkit (Trujillo, Vaitonyte, et al., 2019). Audio data from the video recordings were used to calculate acoustic features of the speech signal. Due to the relatively unconstrained nature of the task we chose to focus on the first communicative attempt (see subsection Annotation of Communicative Attempts) within each round. This ensures that we capture the initial response to the specific noise condition as well as the first multimodal utterance, before it is affected by repetitions or changes in communicative strategy. Therefore, all features are calculated for single attempts, rather than single gestures or entire rounds. Similarly, the first time a target word is spoken during an attempt, this utterance is used to calculate acoustic features. In each feature we removed all outlying data points that were more than 1.5 times the

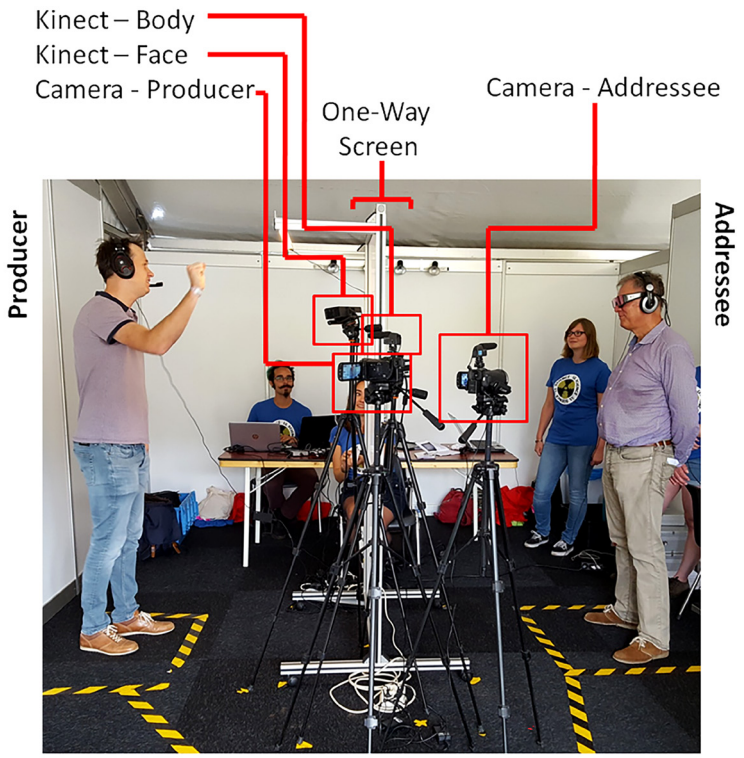


Figure 14. Overview of the physical set-up of the experiment. The producer can be seen on the left side, while the Addressee can be seen on the right side. A one-way screen separates them, allowing the addressee to see the producer, but obscuring the Producer's view of the addressee. Two Kinects (one for face tracking and one for body tracking) are directed at the Producer. One video camera is facing the Producer, while a second video camera faces the Addressee. The yellow and black markings on the floor indicate the area in which the participants must remain throughout the experiment.

interquartile range away from the median.

Annotation of Communicative Attempts

Communicative attempts were defined as being a single attempt to communicate the target word. The target word is the action verb that the addressee must identify in that round. This could be unimodal or multimodal, and could contain multiple gestures or speech utterances. In order to determine the timing of these individual communicative attempts, videos were manually annotated. Attempts were distinguished from one another based on temporal proximity and communicative strategy.

Individual gestures were identified and annotated based on the framework by Kita et al. (1997).For the purpose of finding communicative attempts, only representational gestures (Kendon, 2004; McNeill, 1994) were used. This excluded other types of gestures, such as interactive gestures (Bavelas et al., 1992) that were used to motivate the addressee to keep guessing. For speech, we only used speech utterances containing the target word. For the purpose of defining communicative attempts we therefore did not consider speech that was motivational rather than informative (e.g. “not quite..”, “come on”), or speech that was not directed to the addressee(e.g. “hmm, okay”). Speech utterances were annotated based on Clayman’s “Turn Constructional Units” (2013), with one “unit” being annotated as one utterance.

Unimodal utterances could be gestures with no temporally overlapping speech, or speech utterances with no temporally overlapping gesture. In the case of speech or gesture immediately preceding the other, these cases were considered to be one attempt when there were five or fewer video frames (approximately 200ms) between the two. This is based on the thresholds at which asynchronous speech and gesture are most effectively integrated during comprehension (Habets et al., 2011). In the case of multiple gestures, these are considered to be part of one attempt when there are two or fewer frames between them, or if there is no full retraction. In the case of three or more frames between gestures, or a complete retraction, even if this occurs in fewer than three video frames, these gestures are considered to be part of two separate communicative attempts. Two speech utterances were considered part of separate attempts if there was at least five frames (approximately 200ms) between them. This was based on the ending of a speech utterance marking the possibility for a response or feedback from the listener (Clayman, 2013), which minimally requires 200ms (Fry, 1975). These rules can be summarized as follows:

Unimodal: Gesture

No overlapping speech

No speech within 5 frames

No complete retraction gesture and speech onset

Multi-gesture: onset must be less than 3 frames from previous gesture

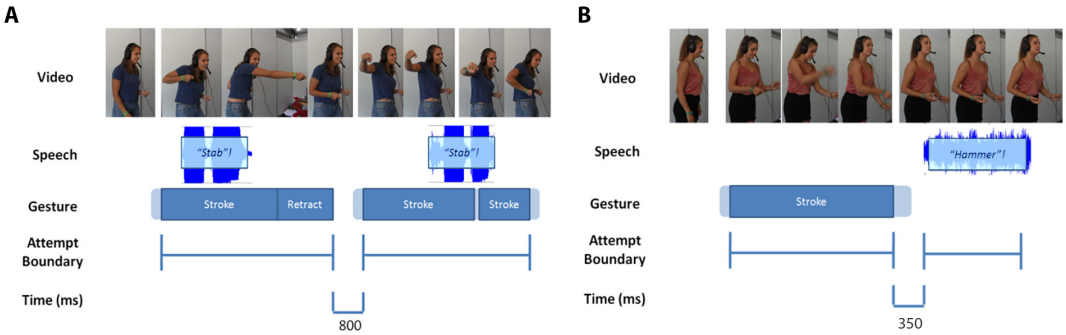


Figure 15. Examples of attempt coding, and unimodal versus multimodal attempts. Both examples depict rounds with two communicative attempts. In both panels, the top panel, *Video*, provides Still frames from the corresponding video. *Speech,* depicts the coded speech, overlaid on the speech waveform. *Gesture* shows the individual gesture strokes and, when present, retractions. *Attempt Boundary* shows how the two attempts were defined. *Time* shows the (rounded, for simplification) number of milliseconds between the two attempts. In **A**, the first attempt is considered multimodal, due to the temporally overlapping speech and gesture. The attempt is finished after the gesture, as there is full retraction. The second attempt is also multimodal, but this time with two gesture strokes as well. The two strokes belong to the same attempt both because they occur with the same speech utterance, but also due to the close temporal proximity, as visualized by the overlap in the opaque blue bar behind the Gesture coding. In **B**, both attempts are unimodal. The first attempt is gesture only. After the gesture stroke, there is no retraction, but it is followed by a speech utterance. As there are more than 200ms between the two, they are counted as two attempts.

Unimodal: Speech

No overlapping gesture

No gesture (with incomplete retraction) within 5 frames

Multi-speech: onset must be less than 5 frames from previous speech utterance

Multimodal

Speech and gesture temporally overlap, or

Speech and gesture no more than 5 frames apart

In case of gesture preceding speech, there must be no complete retraction of the gesture

Using this set of rules, the onset and completion time of the first two communicative attempts of each round were identified. Communicative attempts could therefore be speech-only, gesture-only, or multimodal (speech and gesture with temporal overlap). See Figure 15 for a visual example of how these rules could identify unimodal or multimodal attempts. Additionally, the onset and completion time of speech acts within each attempt were identified. This allowed us to focus our extraction and analysis of kinematic and acoustic features to the time frame of a single communicative attempt. This was done to reduce the effect of any feedback the producer may have been able to receive from the receiver, as well as any effect of repetition or general strategic changes across time while ensuring our window of analysis was relevant to our research questions. For the purpose of this study, we only utilize data from the first communicative attempt in each round. Table 7 provides an overview of the distribution of modality use, within the first attempt, across the noise levels.

Feature Calculation

i. Speech Acoustics

For speech, we calculated maximum intensity (i.e. loudness) and F0. Both features were calculated at the word level. *Intensity* was found by calculating a time-smoothed sound pressure level, in decibels (dB), of the audio waveform and taking the maximum value. We chose the maximum intensity value in order to match the use of peak velocity in both the gesture and face kinematics. *F0* was calculated using PRAAT (Boersma & Weenink, 2019). See Figure 16, panel I for a graphical overview.

ii. Face Kinematics

For the face tracking data, we only calculated features in attempts that included speech. We calculated the *Maximum Mouth Opening* by taking the distance between all tracked points corresponding to the inner area of the mouth and finding the maximum value per communicative attempt. Finally, we took the peak velocity achieved by the center point of the bottom lip in order to give the *Peak Lip Velocity*. Note that we took the peak velocity in order to correspond with the hand kinematic measures. These features were based on previously established visual features of the Lombard Effect as described by Heracleous et al. (2013) and Kim et al. (2005). See Figure 16, panel II for a graphical overview.

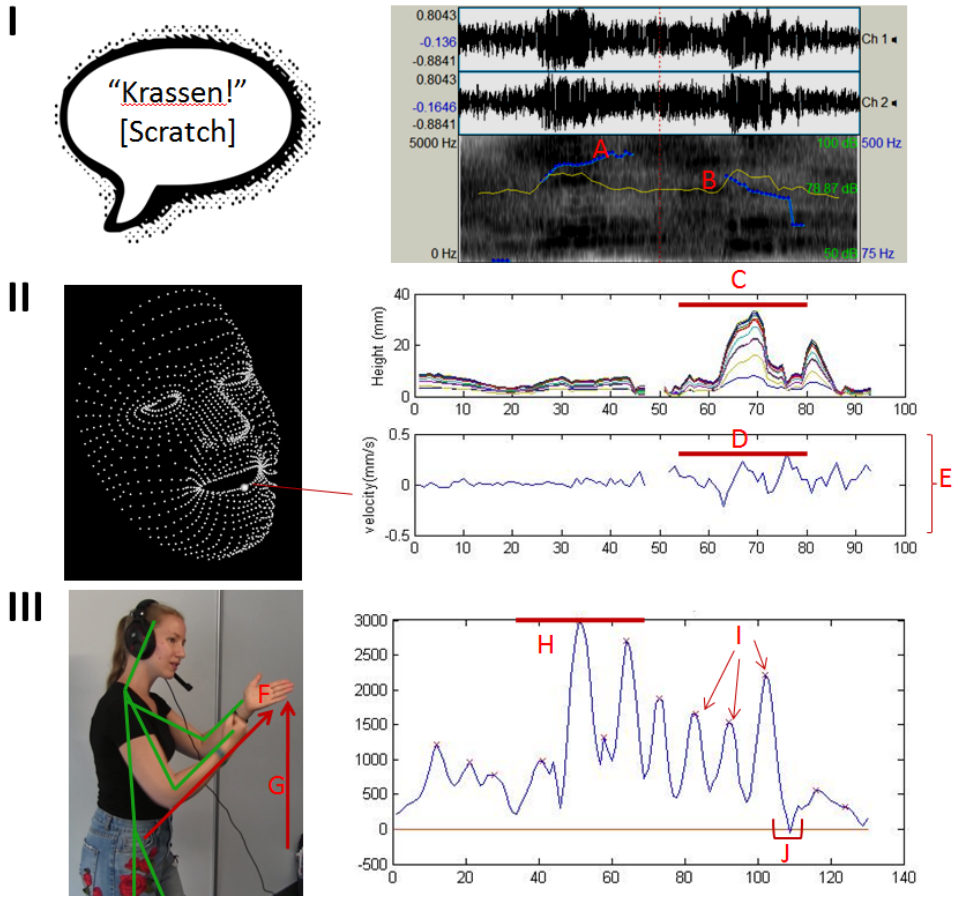


Figure 16. Graphical overview of all communicative features calculated. Panel **I** depicts the speech acoustics, with upper right plot (taken from PRAAT) showing the speech waveform together with the intensity and pitch envelopes. Panel **II** depicts the face kinematics. The tracked points of the face are displayed on the left, with emphasis on the middle lower lip point that was used for the peak velocity calculation. The right plots of Panel **II** show the mouth opening and lip movement through time. Panel **III** depicts the gesture kinematics. The graphic on the left is a still frame from the same communicative utterance from which the kinematic plots are derived, with an overlay of the Kinect tracking lines. The right bottom plot shows the velocity profile of the right hand. **A**. F0 is given as the blue line. **B.** Speech intensity is represented by the yellow line. **C.** Mouth opening is the highest value within one attempt by any two pair of points. **D**. Peak lip velocity was the highest velocity achieved by the lower lip. **E**. The overall amount of movement of the lower lip, taken as the average per second. **F**. Maximum distance of the hand from its starting point. **G**. Vertical amplitude of the hands. **H**. Peak velocity of the hand. **I**. Number of submovements, visible here as individual peaks. **J**. Holdtime, seen here as the amount of time spent below the velocity threshold.

iii. Gesture Kinematics

For the body tracking data, we calculated kinematic features using the toolbox developed in Trujillo, Vaitonyte et al. (2019). In sum, we calculated *peak velocity* of the dominant hand as the highest velocity achieved during the attempt, *submovements* as the number of individual movements made by the dominant hand, *hold-time* as the total amount of time during which the hands were still, and *vertical amplitude* as the maximum height achieved by either hand in relation to the body. We additionally calculated *maximum distance* as the maximum distance away from the body achieved by the dominant hand. This feature was added in order to include an additional purely spatial kinematic feature. See Figure 16, panel III for a graphical overview.

Analysis

All statistical analyses were carried out using the R statistical program (R Core Team, 2019). Before proceeding with statistical analyses, we tested all dependent variables (kinematic and acoustic features) for multicollinearity by calculating the variance inflation factor as described by Zuur and colleagues (Zuur et al., 2010). Predictors with a variance inflation factor greater than three were excluded from all subsequent analyses. Before running statistical tests, we excluded values that were more than 1.5 times the interquartile ratio from the median value. This was done for each separate feature.

We used the linear mixed-effects models to calculate the influence of noise on each of our dependent variables. Mixed-effects models were implemented using the lme4 package (Bates et al., 2014). We created nine linear mixed-effects models, each with one of the features of interest (submovements, maximum amplitude, hold-time, peak velocity, maximum distance, maximum mouth opening, lip movement, lip velocity, speech intensity, speech F0) as the dependent variable, with noise level as a fixed-effect, and a random intercept for each participant. To test the significance of these models, we used chi-square tests to compare the models of interest with a null model, thereby comparing whether the variable of interest, noise level, explains significantly more variance than the random-intercept-only model. As different words may lead to differences in kinematic or acoustic features, we first tested whether a null model containing both participant and word as random intercepts was a better fit to the data than a model with only participant as random intercept.

The better fitting model was used as the null model against which the kinematic and acoustic models were tested. The kinematic and acoustic models utilized the same random intercept structure as the best-fit null model against which it was tested.

When a model of interest (i.e. kinematic or acoustic) was a better fit than the null model, we additionally used the MultComp package (Hothorn, Bretz, & Westfall, 2008) to calculate pairwise comparisons between noise levels.

In order to account for potential correlations between kinematic features, and between acoustic features, as well as the increased type-I error rate associated with multiple comparisons, we used Simple Interactive Statistical Analysis (http://www.quantitativeskills.com/sisa/calculations/bonfer.htm) to calculate an adjusted Bonferroni correction using the mean correlation between the tested features. This mean correlation, and thus the adjusted alpha threshold, was calculated separately for body kinematics, face kinematics, and acoustic features. This is due to the fact that each of these signals represents a separate family of tests. In other words, we are not testing whether noise has an effect on specific features, but whether noise effects speech acoustics, face kinematics, or body kinematics. Body kinematics showed a mean correlation of 0.135, leading to a Bonferroni corrected alpha threshold of 0.0124. Face kinematics showed a mean correlation of 0.286, leading to a Bonferroni corrected alpha threshold of 0.031. Speech acoustics showed a mean correlation of -0.066, leading to a Bonferroni corrected alpha threshold of 0.026.

Results

Effect of Noise on Speech Acoustics

Speech acoustic analyses were based on data from 243 Speech-only attempts and 327 Multimodal attempts. Noise level was strongly associated with speech amplitude, $\chi^2(2) = 18.11$, $p < 0.001$. Specifically, 8-talker babble was associated with an increase of 0.35±0.08 dB compared to no noise ($z = 4.27$, $p < 0.001$), while 4-talker babble was weakly associated with an increase of 0.18±0.09 dB compared to no noise ($z = 2.09$, $p = 0.09$), and no significant difference was fond between 8-talker and 4- talker babble ($z = 1.99$, $p = 0.114$). Noise level was not significantly associated with F0, $\chi^2(2) = 0.278$, $p = 0.870$.See Figure 17 for an overview of these distributions.

	Modality		
	Speech Only	*Gesture Only*	*Multimodal*
8-Talker Babble	94	244	99
4-Talker Babble	71	252	88
Clear	78	236	140
Total	243	732	327

Table 7. Overview of modality usage across noise levels.

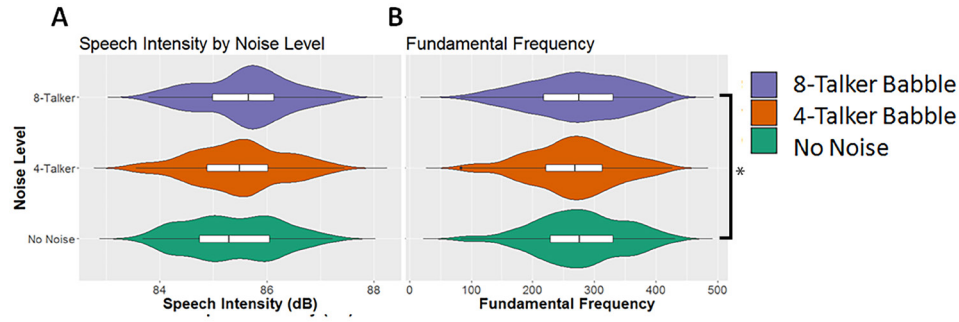


Figure 17. Overview of speech acoustics across noise levels. Panel **A** depicts Speech Intensity, **B** depicts Fundamental F0. In each panel, the y-axis shows the three noise levels in ascending order, while the x-axis shows the raw kinematic values. Violins represent the kernel probability of the data at each point. Within each violin a boxplot shows the median (middle bar) and first and third quartiles (box hinges). The whiskers on the boxplots show the range up to 1.5 times the interquartile range. Additionally, data points beyond the whiskers are depicted as black circles. * $P< 0.001$.

Effect of Noise on Face Kinematics

Speech acoustic analyses were based on data from 243 Speech-only attempts and 327 Multimodal attempts. Noise level was not significantly associated with peak lip velocity, $\chi^2(2) = 3.45$, $p = 0.178$, nor between noise level and maximum mouth opening, $\chi^2(2) = 3.69$, $p = 0.158$, or noise level and mean lip movement, $\chi^2(2) = 0.97$, $p = 0.615$. See Figure 18 for an overview of these distributions.

Effect of Noise on Gesture Kinematics

Data was based on 732 Gesture-only and 327 Multimodal attempts (see Table 7). Noise level was associated with the number of submovements, $\chi^2(2) = 10.47$, $p = 0.005$. Specifically, 8-talker babble was associated with an increase of 0.57±0.2 submovements compared to the clear condition ($z = 3.22$, $p = 0.004$). We observed

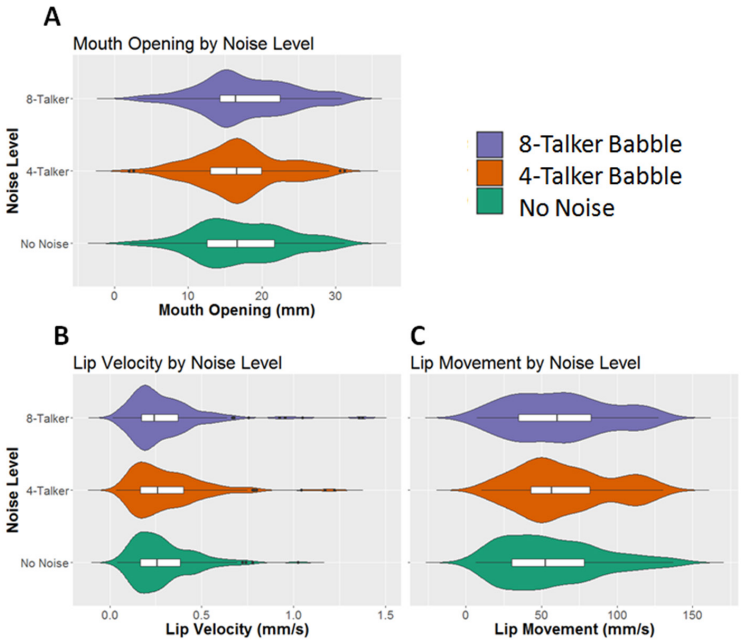


Figure 18. Overview of face kinematics across noise levels. Panel **A** depicts Max Mouth Opening, **B** depicts Peak Lip Velocity. In each panel, the y-axis shows the three noise levels in ascending order, while the x-axis shows the raw kinematic values. Violins represent the kernel probability of the data at each point. Within each violin a boxplot shows the median (middle bar) and first and third quartiles (box hinges). The whiskers on the boxplots show the range up to 1.5 times the interquartile range. Additional data points beyond the whiskers are shown as black circles.

no significant difference in submovements between 8-talker babble and 4-talker babble ($z = 1.20$, $p = 0.453$), nor between 4-talker babble and the clear condition ($z = 1.90$, $p = 0.138$). There was a marginally significant effect of noise on maximum distance, $\chi^2(2) = 6.16$, $p = 0.046$. See Figure 19 for an overview of these distributions. Noise level showed no association with peak velocity, $\chi^2(2) = 0.01$, $p = 0.998$, nor with hold-time, $\chi^2(2) = 3.85$, $p = 0.146$, or vertical amplitude, $\chi^2(2) = 4.44$, $p = 0.217$.

Interaction Between Speech and Gesture

In order to test whether signal modulation occurs in both unimodal and multimodal utterances (i.e. general effort hypothesis) or only in unimodal utterances (i.e. trade-off hypothesis), we assessed whether there was an interaction effect between noise level and modality (i.e. unimodal or multimodal). For submovements, we found no interaction between noise and modality, $\chi^2(3) = 0.973$, $p = 0.808$. For maximum

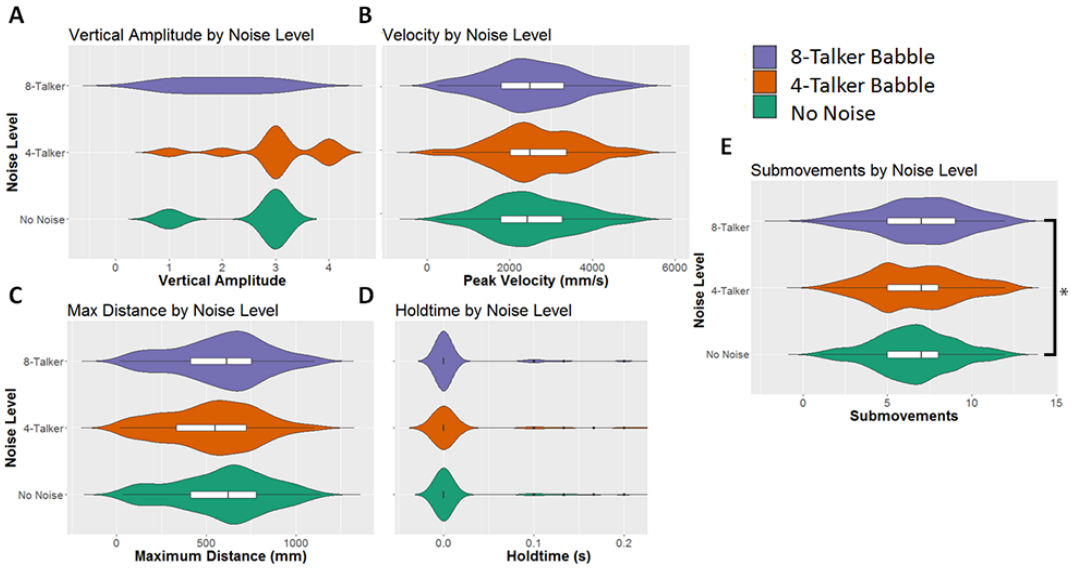


Figure 19. Overview of body kinematics across noise levels. Panel **A** depicts Vertical Amplitude, **B** depicts Peak Velocity, **C** depicts Max Distance, **D** depicts Holdtime, and **E** depicts Submovements. In each panel, the y-axis shows the three noise levels in ascending order, while the x-axis shows the raw kinematic values. Violins represent the kernel probability of the data at each point. Within each violin a boxplot shows the median (middle bar) and first and third quartiles (box hinges). The whiskers on the boxplots show the range up to 1.5 times the interquartile range. Additionally, data points beyond the whiskers are depicted as black circles. Note that Panel A does not contain box plots due to the bimodal distribution of the data.* $p < 0.001$

speech amplitude, we found a significant interaction between noise and modality, $\chi^2(3) = 9.44$, $p = 0.024$. In the model including multimodality, there is a significant increase in speech amplitude in both 8-talker babble compared to no noise ($z = 4.10$, $p < 0.001$) as well as in 4-talker babble compared to no noise ($z = 3.44$, $p = 0.002$). Additionally, multimodality (i.e. the co-presence of gesture) is associated with a lower speech amplitude in the 4-talker babble ($t = 3.06$). This means that 8-talker babble (compared to no noise) is associated with higher speech amplitude regardless of the co-occurrence of gesture, while 4-talker babble (compared to no noise) is associated with higher speech amplitude only when the speech does not occur together with gesture. In other words, speech seems to be strategically modulated based on the presence or absence of gestures, while gesture is modulated any time it is used (see Figure 20).

Exploratory Analysis: Relation between modulations of different visual articulators

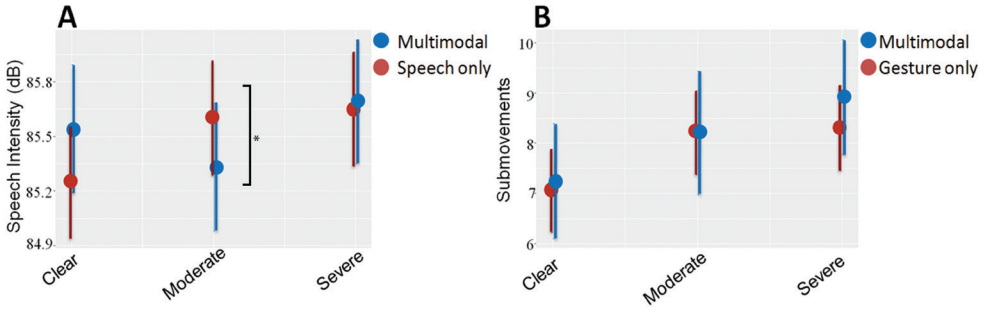


Figure 20. Speech intensity and gesture submovements across noise levels, as produced in unimodal or multimodal utterances. **A** depicts speech intensity (y-axis) plotted against the three noise conditions (x-axis), with blue lines representing multimodal (speech+gesture) utterances and red lines depicting unimodal (speech only) utterances. **B** depicts submovements (y-axis) plotted against the three noise conditions (x-axis), with blue lines representing multimodal (speech+gesture) utterances and red lines depicting unimodal (gesture only) utterances. In both plots circles represent the mean of the distribution, while line lengths extend to +/- one standard deviation. * t > 3.00.

As an additional exploratory test of how speech and gesture go together, we tested whether visible speech parameters could be explained by the extent of kinematic modulation. Although tests of the speech-gesture tradeoff hypothesis have not found evidence for a systematic shift from one modality to the other, research in audio-visual Lombard effects suggests that not everyone modulates visible speech parameters (Garnier et al., 2018). Furthermore, there may instead be a shift from auditory speech modulation to visible speech modulation that is specific to face-to-face settings (Fitzpatrick et al., 2011a). When gestures are available as an additional communicative signal, it may be that there is a similar shift towards modulation of gestures, as these would be more salient than lip movements. If this is the case, we could expect to see visible speech parameters being correlated with gestural kinematic modulation, rather than noise, simply due to a general modulation of the visual signals. The lack of a significant correlation would suggest that the facial kinematic features are either unrelated to Lombard speech or not accurately captured by Kinect. We therefore conducted the additional exploratory test. A positive correlation between the face and gesture kinematics in the absence of a main effect of noise on face kinematics provides some evidence for the effort-based hypothesis, rather than strategic modulation. For each of the three face kinematic parameters we set up the same linear mixed effects model as used in our

main analyses, including the same random effects structure, with the addition of submovements as a fixed effect. This model was again tested against the null model.

We found a highly significant correlation between gesture submovements and peak velocity of the lip, $\chi^2(3) = 34.138$, $p < 0.001$. Specifically, an increase of one submovement was related to an increase in 0.006±0.001 mm/s in the peak lip velocity. Similarly, we found a strong positive correlation between gesture submovements and maximum mouth opening, $\chi^2(3) = 61.96$, $p < 0.001$. Specifically, an increase of one submovement was associated with an increase of 0.354±0.05 mm in mouth opening height. We found no relation between gesture submovements and total lip movement, $\chi^2(3) = 2.265$, $p = 0.519$. In sum, although face kinematics are not systematically modulated by noise, the peak lip velocity and maximum mouth opening covary with gesture submovements, which is itself systematically modulated by noise.

Discussion

The aim of this study was to determine if and how noise modulates multimodal communicative features. Our primary interest was in determining if noise influences both audio-visual speech features as well as gesture kinematics. Second, we aimed to determine whether there is an interaction between the three communicative signals, indicating a strategic shift towards the visual modality, or whether there is evidence for speech and gesture paralleling one another as a form of general increase in effort. To address this question, we utilized markerless motion tracking and audio-video recording during a live communication between pairs of participants. We extracted kinematic features from the body (i.e. gesture kinematics) and visible speech (i.e. face kinematics) as well as speech acoustic features. Our results show that increasing noise leads to a modulation of gesture kinematics as well as speech acoustics. Specifically, we found increased noise was associated with an increase in speech intensity and an increase in gesture submovements. Furthermore, we found that in moderate noise there is a less prominent speech acoustic modulation when gestures are also present, indicating a shift towards the visual modality. In severe noise there was no interaction between speech acoustics and gesture kinematics, suggesting that the two signals indeed parallel one another in this condition.

As our primary research aim was to investigate whether the Lombard effect extends into gesture kinematics, it was important to first establish the classic Lombard

effect in speech. The most universal finding in this domain is an increase in vocal intensity in response to noise (Zollinger & Brumm, 2011), which we replicated in the present study. Vocal intensity is also thought to be the primary modulation, with other acoustic effects being secondary to intensity (Garnier & Henrich, 2014). The second acoustic feature frequently shown to be modulated by noise is F0. Changes in F0 are believed to occur automatically and not specifically in response to an addressee. We therefore initially expected that noise would also affect F0, although we did not find this effect. However, we believe the lack of effect can easily be explained by the overall high vocal intensities observed across all noise levels. Although F0 typically increases together with intensity (Titze & Sundberg, 1992), F0 levels saturate at high values (Rostolland, 1982), meaning that shouted or loud speech tends to show little variation in F0, even if there is variation in intensity. Given the overall loud environment, and the fact that participants were aware that the addressee was experiencing noise throughout the experiment, producers likely used a raised voice throughout the experiment. This is also evidenced in the mean vocal intensity values, as seen in figure 17 and supplementary table 5.1, which were similar to values reported for shouted speech (Raitio et al., 2013; Zhang & Hansen, 2007). Participants were able to further modulate vocal intensity in response to the different noise levels, but because the F0 values were already saturated it was not possible to detect any differences between the noise levels. Although it is not possible to determine whether such a saturation effect was due to the generally noisy environment in which the experiment took place or due to an adaptation to the addressee's listening condition, our results demonstrate a Lombard effect even in the presence of generally increased vocal effort.

Regarding the relation between speech intensity and modality, we found that severe noise was associated with an increase in speech intensity regardless of whether the speech was paired with gesture. This is in line with the theory of speech and gesture being modulated in parallel (de Ruiter et al., 2012; So et al., 2009). However, in moderate noise this increase was only present when the speech was not paired with gesture. In other words, speech intensity was lower in multimodal compared to unimodal (speech only) communicative attempts. This is directly in line with the hypothesis that there is a shift towards the visually prominent gesture signal when it is present. When the utterance is unimodal, only utilizing speech, then speech intensity is further modulated. This is supported by the idea that speakers flexibly draw on gestures to clarify speech when needed (Holler & Beattie, 2011), and the

previous findings of a shift of effort from the vocal signal to more visual signals when communication becomes difficult (Fitzpatrick et al., 2011a; Hoetjes & Carro, 2017; Hostetter & Alibali, 2004; Melinger & Kita, 2007). In the context of noise, gestures provide the most benefit to understanding degraded speech at a moderate, rather than high level of noise (Drijvers & Özyürek, 2017). As gestures are highly beneficial for clarifying speech at this noise level it would not be necessary to put more effort into the speech signal. If produced without accompanying gestures, then speech modulation would be required in order to increase the signal to noise ratio of the speech signal (Garnier & Henrich, 2014). In contrast, at high noise levels gestures still improve speech comprehension, but the effect is less pronounced (Drijvers & Özyürek, 2017). In order to overcome this, both kinematic and acoustic modulations must be used. To summarize these findings, there is a shift towards the visual modality (i.e. gesture, in this case) in the moderate noise condition in which gestures maximally benefit speech comprehension, whereas in severe noise gestures are less beneficial to the speech signal, and thus there is a general increase in effort in both gesture and speech.

For face kinematics, we did not find any evidence for a systematic effect of noise on the extent of mouth opening, lip velocity, or total movement of the lips. This is inconsistent with previous reports of jaw motion being modulated in Lombard speech (Davis et al., 2006; Kim et al., 2005). This is particularly interesting considering that increases in speech intensity typically show increased velocity (Huber & Chandrasekaran, 2006; Schulman, 1989). As these previous studies have used principle component analysis to show the overall effect of Lombard speech on mouth movement, it may be that our kinematic features were too specific. Previous studies have also suggested that mouth opening is increased in noise compared to clear conditions (Davis et al., 2006; Garnier et al., 2010). Our data were not in line with this prediction. One reason for this could be the overall high intensity values for speech throughout the experiment. This could mean that, much like F0, mouth opening was already so high that no observable increases were made when noise was introduced. This is in line with the fact that mean mouth opening values in our study, across all noise conditions, are similar to the values reported for loud speech by Schulman (1989). Overall, these findings are not consistent with an account of noise modulating face kinematics, at least in the specific context of our experiment.

In addition to modulation of speech and face kinematics, we additionally found that

gesture submovements are increased in noise compared to no noise. This finding of kinematic modulation in response to noise is directly in line with the idea that movement kinematics are adapted to the communicative context in which they are produced. In other words, the way we produce a gesture is dependent on the relevance of the information to our addressee (Campisi & Özyürek, 2013; Kelly et al., 2011; Trujillo, Simanova, Bekkering, & Özyürek, 2018). Submovements are specifically related to the amount of segmented visual information being presented through the hands. This can be either through the number of gesture strokes being produced, or the extent to which a complex gesture expression is segmented into clearly defined individual movements (Trujillo, Vaitonyte, Simanova, & Özyürek, 2019). However, a purely segmentation-based explanation may be expected to also produce a difference in hold-time, as this feature represents the clear punctuation between movements. Therefore, it is more likely that the increased submovements are related to an increase in the number of individual movements, either due to repetitions or due to increased complexity of the visual representation.

In line with the interpretation of increased gesture submovements representing a communicatively relevant increase in complexity or repetition, increasing complexity has previously been shown in gestures designed to be more informative (Campisi & Özyürek, 2013), and movement repetitions have similarly been suggested to increase the salience of a movement (Brand et al., 2002; Blokpoel et al., 2012; De Ruiter et al., 2010) and thus also be inherently communicative. In the present study we cannot draw conclusions about whether this increase in submovements relates specifically to more clear segmentation of similarly complex movements or whether the increase represents more movements being produced. While previous work has shown the importance and modulation of additional kinematic features in communicative manual movements (McEllin, Knoblich, & Sebanz, 2018; Trujillo, Simanova et al., 2019; Trujillo, et al., 2018; Vesper, van der Wel, Knoblich, & Sebanz, 2011), the current study did not manipulate the type of communication (e.g. leader-follower, demonstration versus coordination). However, our results provide the first evidence that gesture kinematics are modulated in response to noise. Specifically, this modulation involves the overall amount of visual information produced, rather than specific temporal or spatial components of the movements.

Our findings provide a nuanced answer to the question of whether communicative difficulty leads to an overall increase in effort afforded to both speech and gesture, or

whether there is a shift towards the visual modality. We find that the contributions of the specific signals (i.e. speech, lips, gesture) can be dynamically adapted to the needs of the current situation. Specifically, in moderate noise conditions speakers can either raise the intensity of their voice to overcome the noise, or they can take advantage of the supporting role of gestures to clarify their speech without putting extra effort into the auditory signal. In high noise conditions, gestures alone cannot compensate for the noise, requiring speakers to modulate the intensity of their voice and the amount of visual information represented in their gestures. More generally, we see that participants largely favored gesture-only utterances, at least for their first communicative attempt. This also suggests a general shift towards the visual modality, although it is possible that this is due to the social context in which the experiment took place, or from learning this strategy from other participants. Future studies will be needed to disentangle these effects.

Overall, our results fit well with the interface model of speech-gesture production proposed by Kita and Özyürek (2003). In their hierarchical model, modality selection occurs first, before the exact content of either speech or gesture occurs. This selection, as well as the generation of the communicative expressions (i.e. signals) are influenced by environmental factors. Our results suggest that the amount of noise in the environment indeed influences which modalities are used, and this selection modulates how the signals are produced at the kinematic or acoustic level. Additionally, the information packaging hypothesis suggests that gestures package information, thus influencing speech (Kita, 2000), while the interface hypothesis supposes that gestures are planned according to linguistic constraints on how gestures are shaped (Kita & Özürek, 2003). We build on these frameworks by suggesting that speech and gesture also interact at the kinematic/acoustic level, depending on environmental constraints. These findings are therefore important for understanding how speech and gesture are dynamically adapted to different communicative environments, and thus how they are coupled in one multimodal communicative system (See for an overview Wagner, Malisz, & Kopp, 2014).

In our exploratory analysis we additionally found that mouth opening as well as lip velocity were positively correlated with gesture submovements. This is intriguing because submovements were themselves modulated by noise levels. There are two potential explanations for this finding. The first is that mouth movements were indeed not systematically modulated by noise in the context of this experiment.

Instead, lip velocity and mouth opening follow the overall increase in effort put into the visual modality, which is most prominently expressed in gesture kinematics. An alternative explanation is that the facial kinematic features were either uninformative or otherwise not measured with sufficient accuracy in order to be informative. Given the robust relation between two of the kinematic features and submovements, we believe that the first explanation is more likely. These results could provide a more nuanced view of how communicative effort and communicative strategy can create complex adaptations to noise. When gestures are available, they are the most visually salient signal that can be used. Therefore, visual speech is not specifically modulated. However, the effort put into gesture kinematics carries over to the face, albeit in a less specific or less pronounced manner, due to the (neurobiological) coupling between hand and mouth (Woll, 2014; Woll & Sieratzki, 1998). This is also well in line with the idea of effort in gesture leading to changes in another signal (Pouw et al., 2019). However, given that this was an exploratory analysis, we suggest caution when interpreting these results. These results do suggest that the dynamic relation between modalities should be carefully assessed in future research.

The setting of our experiment provided several major advantages and disadvantages. First, the setting of a music festival meant that noise levels were relatively high throughout the experiment. This could be considered a disadvantage at the noise saturation likely contributed to the lack of several expected effects in response to our own noise manipulation. However, this gave us the opportunity to investigate the Lombard effect beyond the point of saturation. We therefore believe this setting was a very ecologically valid environment in which to test for a communicative, multimodal Lombard effect. Studies on the Lombard effect typically carefully control noise levels, leading to a comparison between a nearly entirely noise-free situation and one with some level of disruptive noise. In many real life situations, this distinction is not so clear. In fact, if we take the example of a cocktail party that is often used to explain Lombard effects, people are likely to be experiencing ambient noise throughout the party, but an increase in noise when they join a crowded room to interact with other guests. Similarly, at a music festival, individuals will experience noise throughout the festival, but a focal increase when near a stage or in a group of other individuals. In these cases, it is interesting to see that speech, visible speech, and gestures are still modulated beyond their already increased baseline levels. Particularly interesting is that even in this situation of highly saturated noise levels, there is still evidence for a gestural Lombard effect. This is especially relevant

when considering the highly heterogenous group that was tested and the relatively unconstrained task that was used. Given the high variance in our data, this could be considered a disadvantage, although we again argue that this contributes to the ecological validity of our findings. However, we suggest that future research should aim replicate these effects in more experimentally controlled conditions in order to determine whether these findings hold true across lower noise levels. Additionally, future research should investigate how these modulations interact with one another in time, and how they contribute to listener comprehension. Finally, to the best of our knowledge this is the first study to utilize markerless face tracking using the Kinect, which can be a useful tool for studying facial kinematics. The face kinematic features utilized in the present study could also be useful for capturing more specific aspects of the visual Lombard effect in speech. Although the features calculated here were relatively simply, this proof of concept for the method opens the door to calculating additional features and taking advantage of the high resolution of the Kinect face tracking.

Taken together, the present study shows that noise leads to a modulation of not only auditory speech signals, but also of gesture kinematics. Results demonstrate that gestures are the more prominently adapted visual signal when compared to visible speech. Secondly, we find that speakers may modulate speech and gesture strategically, based on which signal, or signals, are most effective given the level of noise. This suggests that noise-induced adaptation is a strategic response that likely occurs at the level of communicative planning. Kinematic modulation can therefore be seen as a very general communicative adaptation that can be used to signal intentions, clarify information, and dynamically compensate for communicatively challenging situations.

Acknowledgments

The authors are very grateful to Annika Brand, Eelke Spaak, Hedwig Trujillo-van der Meer, Julia Egger, Kimberley Mulder, Marlijn ter Bekke, Sybrine Bultena, and Yvonne van den Hoeven for their contribution to data collection, as well as Emma Berensen, Clarissa de Vries, and Charlotte Jonker for their contribution to the manual annotation of the videos. This research was supported by the NWO Language in Interaction Gravitation Grant. The authors declare no conflict of interest in this study.

Supplementary Material

	No Noise	4-talker babble	8-talker babble
		Speech Intensity (M±SD)	
Male	85.45±1.0	85.57±0.8	85.80±0.9
Female	85.03±1.4	85.22±1.1	85.64±1.1
		F0 (M±SD)	
Male	245.74±57.5	251.45±62.4	241.02±56.4
Female	301.78±76.2	287.16±77.8	299.67±82.5

Supplementary Table 5.1. Overview of speech acoustics (speech intensity and F0) across noise conditions and gender

Chapter 6

Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research

Chapter based on:

Trujillo, J.P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods, 51*(2). 769-777. https://doi.org/10.3758/s13428-018-1086-8

CHAPTER 6

Abstract

Action, gesture and sign represent unique aspects of human communication that use form and movement to convey meaning. Researchers typically use manual coding of video data to characterize naturalistic, meaningful movements at various levels of description, but the availability of markerless motion tracking technology allows quantification of the kinematic features of gestures, or any meaningful human movement. We present a novel protocol for extracting a set of kinematic features from movements recorded with Microsoft Kinect. Our protocol captures spatial and temporal features, such as the height, velocity, submovements/strokes, and holds. This approach is based on studies of communicative actions and gestures and attempts to capture features that are consistently implicated as important kinematic aspects of communication.

We provide open-source code for the protocol, a description of how the features are calculated, a validation of these features as quantified by our protocol compared to manual coders, and a discussion of how the protocol can be applied. The protocol effectively quantifies kinematic features that are important in production (e.g. characterizing different contexts) as well as in comprehension (e.g. used by addressees to understand intent and semantics) of manual acts.

The protocol can also be integrated with qualitative analysis, allowing fast and objective demarcation of movement units, providing accurate coding of even complex movements. This can be useful to clinicians, researchers studying multimodal communication as well as human-robot interactions. By making this protocol available we hope to provide a tool that can be applied to understanding meaningful movement characteristics in human communication.

Introduction

Human communication is intrinsically multimodal, consisting of not only speech but also visible communicative signals. Gesture, sign and communicative actions (e.g. joint-actions, demonstrations) are well-studied examples of communicative manual acts that can convey meaning in the presence or absence of co-occurring speech. A plethora of research in the last decade has shown that each of these modalities, while unique in certain ways, effectively utilizes movement and configuration to convey meaning and contribute to successful communication.

Among an array of visual bodily cues that people resort to when conveying meaning, gestures stand out as a unique attribute of the human communication system. A wealth of research has shown that gestures (we use the term 'gestures' here to refer to movements of hands and arms that are used to depict objects, ideas, events and experiences (Kendon, 2004; McNeill, 1994)) form an important aspect of communication. The study of gesture has opened a new window into human language, cognition and interaction, (e.g., McNeill, 1994; Kendon, 2004; for a recent collection see Church, Alibali, & Kelly, 2017) with important clinical applications, such as using the production and comprehension of pantomimes to assess disorders such as apraxia (Goldenberg et al., 2003; Gonzalez Rothi, Heilman, & Watson, 1985), Autism spectrum disorder (Anzulewicz, Sobota, & Delafield-Butt, 2016), or Parkinson's disease (Humphries, Holler, Crawford, Herrera, & Poliakoff, 2016).

Traditionally, researchers who study gesture recur to the analysis of video data. The video data are analyzed manually on the basis of pre-determined coding schemes, relying on such annotation tools as ANVIL (Kipp, 2001) or ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). It has recently become possible to employ more automatic ways to analyze multimodal data. The description of movement can now be carried out using motion capture, which is a technology allowing an automatic extraction and characterization of movement parameters (e.g. space, trajectory, distance, velocity). There is a host of motion capture techniques available, including the more well-known technologies, such as OptiTrack, Leap Motion, and the Microsoft Kinect. The Kinect is of particular interest due to the fact that it is inexpensive, portable and markerless, which increases ecological validity while providing accurate depth sensing (Wasenmüller & Stricker, 2017). The Kinect is a sensor consisting of two cameras (i.e. infrared and depth) that track human

skeletons in space, rendering a 3-dimensional structure of movement based on joint positions.

Since its release, the Kinect has been tested and applied to a multitude of research fields, including medical (R. A. Clark et al., 2015; Galna et al., 2014), robotics (Hussein, Ali, Elmisery, & Mostafa, 2014), augmented reality (Bostanci, Kanwal, & Clark, 2015), and multimodality of communication (Trujillo et al., 2018). Being a low-cost and non-invasive motion tracking system, the Kinect could indeed be applied to the study of gesture more widely. While the Kinect cannot fully replace manual coding, it can advance the analysis of movement in several ways. First, manual coding is extremely time-consuming, and requires more than one coder in order to calculate inter-coder reliability. A substantial amount of time is spent on training the coders as well as on carrying out the actual gesture coding. Time spent on coding can be reduced by allowing motion-capture data to provide a first-pass of the data, identifying individual gesture units on which the manual coders can perform further analysis. Inter-coder reliability would also be increased, as motion-capture data provides an objective demarcation of the gestural units, allowing the coders to work from the same framework. Second, the manual analysis is constrained by the reliance on 2-dimensional video data whereas the Kinect captures movement in 3-dimensional space. This can be especially advantage when analyzing complex movements, such as pantomimes. Third, the Kinect provides the opportunity to analyze movement quantitatively, which, depending on the research question(s), can be combined with a qualitative or categorical approach to gesture coding.

Here, we provide a Kinematic Feature Extraction protocol (available at: https://github.com/jptrujillo/kinematic_feature_extract) that quantifies several kinematic aspects of movements. We selected kinematic features in which researchers have shown interest in previous studies, and which we believe can be quantified for a variety of gestures or acts, including complex pantomimes. As the code is available open-source, it will additionally be possible to build off of our framework to add features that are of interest to the specific studies in which it is used.

Studies in the action and gesture domains have consistently noted the importance of size (Brand et al., 2002; Campisi & Özyürek, 2013; Gerwing & Bavelas, 2004), punctuality (Brand et al., 2002) and the use of holds (Gullberg & Kita, 2009), as

well as the velocity of movements (Manera et al., 2011; Sartori et al., 2009). We operationalize size here as being a cumulative utilization of space, and therefore include a measure of *distance*, which quantifies the accumulated distance traveled by the hands during the analyzed act. This feature will therefore capture both larger movements as well as the accumulation of many smaller movements. Punctuality was previously defined as having movements that are well marked in their beginning and end, a feature that is thought to help clearly segment the overall act for an observer (Brand et al., 2002). This fits well with work on motor control that shows that movements tend to be organized into smaller submovements. These are apparent as sharp changes in velocity, which result from changes in trajectory (e.g. reaching to grasp an object may consist of at least two submovements: an initial movement towards the object, and an additional corrective movement to ensure the hand is correctly aligned to grasp it; see, for example, the work by Meyer and colleagues, 1988). More punctual movements may therefore be seen as having more clearly defined submovements. This feature can also be seen as analogous to the gestural stroke (Kendon, 2004), allowing one to quantify the number of strokes produced. We operationalized the feature as *submovements*, which captures the number of submovements, or strokes, performed with each hand during a given act, as well as two *hold* features. Holds were defined as moments in which the hands and arms were completely still, representing a pause between submovements. These can also be seen as analogous to Kendon's pre- or post-stroke holds (Kendon, 2004). Our code calculates both *hold-time* (defined as the total amount of holding time in an act) as well as *hold-count* (the number of individual holds performed). While holds can be seen as quantifying the punctuality of an act, sub-movements and holds can together help to identify the key movement phases, as defined by Kita and colleagues (Kita et al., 1998), that are often studied by gesture researchers. Velocity has recently been shown in several studies as important in understanding different intentions underlying an act (Peeters et al., 2013; Sartori et al., 2009). We include *peak* velocity of each hand to capture the fastest recorded velocity during an act. This will quantify only the fastest movement, and therefore would capture fast preparatory movements while being insensitive to holds or the inclusion of slower movements later in the act. The height at which a gesture is performed has long been of interest for gesture researchers (Gullberg & Kita, 2009; McNeill, 1994). We therefore include a measure of *vertical amplitude,* which quantifies the peak height of the hands in relation to the body of the gesturer.

In addition to presenting code for quantifying these features, we validate these new methods with respect to the established methods to provide a proof-of-concept. Some recent work has shown that Kinect tracking is a valid alternative to optical tracking (Fernández-Baena et al., 2012) for clinical sciences (see, for a review, Da Gama, Fallavollita, Teichrieb, & Navab, 2015), as well as several projects developing gesture recognition algorithms for the Kinect (Biswas & Basu, 2011; Paraskevopoulos et al., 2016). We therefore compare the kinematic analysis of gestures carried out using the our script and Kinect data with the results obtained from manually coding the same kinematic gesture features in the ELAN annotation tool.

In sum, the following paragraphs address two primary goals: 1) provide a basic Kinematic Feature Extraction code that can be used with Kinect, providing a platform for developing more extensive feature extraction protocols, and 2) to contrast the automatic feature analysis (Kinect) described in Trujillo et al. (2018) with the manual analysis (human coders) of gestures by means of seeing whether, and to what extent, the two methods, the automatic and the manual, correlate.

Feature Extraction Method

Platform

MATLAB 2015a (The MathWorks, Inc., Natick, Massachusetts, United States) was used to develop all scripts. Files saved in the C3D file format are converted to text format, after which the script imports the data and proceeds with the data processing and feature extraction.

Data Processing

Taking the raw data, all points are smoothed using a Savitsky-Golay filter with a span of 15 and degree of 5. This accounts for the typical jitter and motion artifacts that can occur in raw Kinect data. If available, the data will be segmented into individual acts. This step requires the user to provide an additional input with onsets and offsets for each act. If this input is given, the output file will provide kinematic feature data for each individual act. If no onset/offset information is provided, the data file is treated as one act, and only one value for each feature is calculated (e.g. the total number of holds in the data file).

Kinematic features

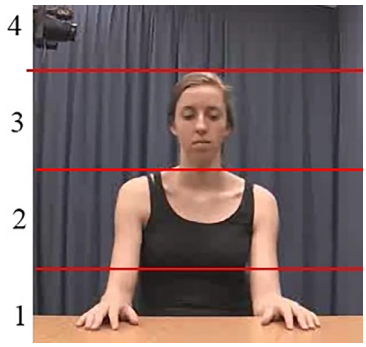


Figure 21. Visual representation of *Vertical Amplitude* feature, as calculated in reference to a participant's skeleton using the Kinect. Red lines indicate the cut-off points (approximated for illustration), with the numbers on the left indicating the value assigned to the space between the upper and lower lines. Note that 1 is bounded by the table, while 4 has no upper bound and is therefore bounded by the participant's maximum arm extension.

Vertical amplitude was defined as the highest point in relation to a participant's body reached with the right dominant hand during an act. Vertical amplitude was divided into four different categories, from the lowest, which was denoted by the hand not reaching above the midline of the torso, to the highest – above the top of the head. This was calculated by comparing the hands to the spine, neck and head at each frame of the recording (figure 21).

Peak velocity was defined as the fastest movement, reached with the right dominant hand. This was given as an absolute value in meters per second in our previous manuscript (Trujillo et al., 2018), but was binned into seven categories by placing all peak velocity values in the current data set onto a spectrum and subsequently dividing them into seven bins, evenly distributed across the included data.

Sub-movements were defined as smaller movement segments, which were made throughout the representational gesture item. This feature is based on the work of Meyer and colleagues, who described sub-movements as the individual ballistic movements that make up a given action (Meyer et al., 1988). In short, each item was divided into a number of basic movements, characterized by an initial increase in velocity followed by a decrease in velocity at the points of connection of the movement segments. Sub-movements can be comparable to gesture strokes, which are the most semantically meaningful gesture part (Kita et al., 1998). Sub-movements were operationalized exceeding a velocity threshold of $0.2m^2$, with the

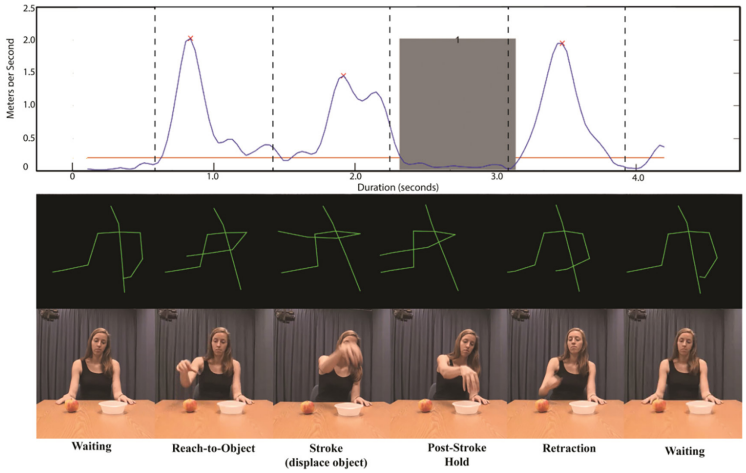


Figure 22. Graphical comparison of velocity profile (data collected with Microsoft Kinect) generated by the protocol with the corresponding video data. The depicted gesture was produced under the instruction to "place the apple in the bowl". The **upper plot** depicts an actual output image generated by the protocol, with the addition of vertical dashed lines, which are included to show the match between the kinematic and video data. The y-axis depicts velocity in meters per second, while the x-axis depicts time in seconds. The horizontal red bar is the cut-off used to separate sub-movements from other movement noise (either measurement error or slow, non-meaningful movements). The grey rectangle denotes a single hold, with the number printed between the bars indicating the number, or index, of the hold (e.g. if there are 4 separate holds in a dataset, then they will be numbered 1-4). The red X's indicate the peak of a counted sub-movement. The **middle plot** shows a series of still frames, depicting the primary movement phases of the gesture as captured by the Kinect. To match the corresponding video frames, the lines only depict the torso, arms, and head. The **lower plot** shows a series of still frames depicting the same phases as seen in the corresponding video. Below, a label is given for each depicted movement phase.

beginning and end marked by either the crossing of a near-zero velocity threshold (i.e. changing from static to moving) or showing a reversal from deceleration to acceleration. We used a standard peak analysis to determine the total number of peaks within the velocity profile of each hand that were at least 8 frames from the next nearest peak and with a minimum height of 0.2 meters. **Hold-counts** were defined as an absence of movement in both arms and hands, for at least 300ms. This number was utilized in (Trujillo et al., 2018) due to it being the approximate minimum time length that naïve observers consistently identify as a cessation of movement. This was operationalized as sets of frames where the hand, thumb,

elbow, and shoulder of both arms all show less than 0.01 meters of movement for at least 300ms (i.e. a minimum of 9 consecutive frames). Figure 21 provides a graphical representation of how the vertical amplitude feature is calculated against the producer's body. Figure 22 provides an example of visualization output from the protocol, matched to corresponding video frames from the same gesture.

Output

The code generates a .mat file containing all of the calculated kinematic features, with individual acts or moments separated per row in the table. If the data is not segmented by acts (see *data processing*) then the one row is a summary of the data-file. Additionally, a .fig plot is generated, one for each act, of the velocity profile of each hand, with submovements and holds indicated. For an example of such a plot see the top plot in Figure 22. This plot can be useful in providing a visualization of the collected data and calculated features, but can also be used to help guide the coding of gesture phases for further analysis. Using the save_skeleton.mat file, an additional video file can be generated of any act. This video has a black background with green lines that depict the connections between each of the measured joints. Example frames from such a file can be seen in the middle plot of Figure 22. These 'skeleton videos' can be used together with the standard recorded video to provide additional viewing angles to assist gesture coding, or as experimental stimuli. These implementations are further discussed below in the section titled Applications.

Validation Method

Materials

The materials in the present study consisted of a subset of videos from a production experiment from the Trujillo and colleagues' study (2018), in which 3D joint tracking data were collected by employing the Microsoft Kinect V2. Although the data was collected from all 25 joints of the human body that the Kinect's sensor is able to capture, the hips and legs were not used for any analysis. Data was collected at 30 frames per second (fps). Film data was collected at 25fps by a camera hanging at approximately eye-level, directly in front of the participant. In the Trujillo et al. study the kinematic features that were calculated were the following: distance, vertical amplitude, peak velocity, sub-movements, hold-time and hold-count. In the current study, it was chosen to analyze and compare across the two methods four kinematic

features: vertical amplitude, peak velocity, sub-movements and hold-count. The rationale for selecting these particular kinematic features was that they were the most amenable to hand-coding, in that it was possible to create meaningful categories for each of these features that could be captured with a naked eye. The video data used for the analysis contained only representational gestures, meaning that no videos showing actions were used for annotations. Manual data coding was carried out in the video annotation software ELAN (www.lat-mpi.eu/tools/elan/). The initial set of videos contained 120 video clips that were annotated by two human coders, however, due to the data loss in the Kinect the comparison between the manual and automatic coding is based on 111 videos.

Validation Procedure

First, Coder 1 annotated 111 videos by marking the four kinematic features in each video for each representational gesture (i.e. item). Descriptions of how the coder defined each feature are given below. Second, Coder 2, who first received training on how to code the data by Coder 1, annotated the same 111 videos. During the coding process, both coders were naïve to the kinematic values extracted by our script.

Manual coding of Kinematic features

As with the scripted analysis, **vertical amplitude** was calculated by comparing the hands to the spine, neck and head at each frame of the recording, using the same categories as the automatic coding.

Manual coders assigned **Peak velocity** values to different velocities in the range between 1 and 7. This was done after first viewing all of the videos and finding the peak movement, and then annotating each video as belong to one of the seven categories. A value of 1 therefore indicated that the fastest movement in the act was among the slowest in the dataset, while a value of 7 represented a movement that was among the fastest in the dataset.

For manual coders, **sub-movements** were defined as the number of movements that can be segmented based on an observable transition from deceleration to acceleration.

Coders defined **holds** as pauses in movement where both hands were still in a clearly

distinguishable manner for at least 300ms.

Statistical Comparison of Coding

Analyses consisted of two steps. The first step assumed calculating Spearman's *rho* in order to see the degree of association between the two human coders for each kinematic feature, and assessing inter-coder reliability for two features in particular. That is, Cohen's kappa was computed for *vertical amplitude* and *peak velocity* only because these features were quantified on set scales. Given that sub-movements and *hold-counts* could take on any value of 0 or greater, assessing inter-coder reliability was not possible.

The second step included comparing the Kinect features with the manual coding of Coder 1 (the second author) for which Spearman's *rho* was used in order to determine whether the two methods were correlated. Throughout the results section, corrected *p-values* are reported (Bonferroni correction was applied).

Validation Results

Human Coders

For *vertical amplitude* the correlation was $r_s(111)= .82$, $p < .001$ while for *sub-movements* it was $r_s(111)= .74$, $p < .001$. *Peak velocity* and *hold-counts* produced correlations of $r_s(111)= .70$ $p < .001$ and $r_s(111)= .60$, $p < .001$, respectively. The inter-coder reliability for *vertical amplitude* was $\kappa = .63$ while for *peak velocity* it was $\kappa = .40$. For an overview of all results, see Supplementary tables 1-4.

Manual-Automatic Coding

Vertical amplitude and *sub-movements* produced correlations of $r_s(111)= .83$, $p < .001$ and $r_s(111)= .41$, $p < .001$, whereas the correlations for *peak velocity* and *hold-counts* were $r_s(111)= .114$, $p = .233$ and $r_s(111)= .33$, $p < .001$, respectively.

Discussion

The Kinematic Feature Extraction toolkit presented here can be used to quantify spatial and temporal features of meaningful movements, including complex pantomimes. Together with markerless tracking technology such as the Microsoft Kinect, it provides a valuable tool for quantifying kinematic features that are

important for research in the production of communicative manual acts.

In order to validate this method, we compared automatically extracted kinematic features, based on Kinect data, with manually coded kinematic features, based on video data. Results of this validation process show that the Kinect can robustly measure both spatial and temporal kinematics of pantomimes, with automatically extracted features (i.e. *vertical amplitude*, *sub-movements*, *and hold counts*) largely similar to manually coded features. While the *peak velocity* showed very poor overlap between manual and automatic coding, inter-coder reliability in the manual coding for this feature was also lower. This suggests that the proposed method of automatic extraction may measure this feature more robustly.

Human Coders

The gesture coding between two manual coders resulted in high correlations for the kinematic features of *vertical amplitude*, *sub-movements* and *peak velocity* whereas the correlation for *hold-count* was slightly lower in comparison to other three features. While coding of peak velocity was highly correlated between the coders, there was somewhat lower reliability, as indicated by the lower kappa score. This suggests that while manual coders were consistent in ranking the videos (i.e. providing larger numbers for videos with faster movements), there was less reliability for selecting the exact same category. Due to the more subjective nature of this feature, it is not surprising that reliability is somewhat lower. However, overall high correlations between coders indicate that the coding of these features was carried out in a consistent and replicable manner.

Manual-Automatic Coding

Overall good agreement was seen in *vertical amplitude* and number of *sub-movements*. As *vertical amplitude* was relatively straightforward to define, with a clear reference point (participant body) against which to compare the height of the hands, this result was very much expected. *Sub-movements* also showed high overlap. The high correlation between human and automatic coding suggests that our automatic approach captures individual sub-movements, at least on the coarse level in which a human observer may also segment an act into individual movements. This is important because this shows that the automatic coding captures the primary movement boundaries in a similar way to human coders. As sub-movements can be

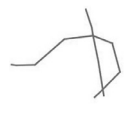
seen as analogous to gesture strokes, this provides some validation of the process as an objective and automatic way to code these gesture units.

When coding *hold-counts*, we find a significant positive correlation, although the fit of the model is lower than that of *vertical amplitude* or *sub-movements*. Closer inspection of the data revealed that in some cases it was difficult for the manual coders to accurately delineate the beginning and end of individual holds due to the presence of small movements, or a series of very brief holds. In this case, we suggest that the holds are likely to be more accurately counted by the automatic approach, as there is a clear cutoff point for movement and duration.

Although *peak velocity* did not show strong correspondence between automatic and manual coding, we suggest that this may have been due to differences in which movements were coded as being the fastest. When qualitatively comparing the automatic and manual analyses, it was noticed that manual coders would reliably capture larger movement segments within a given gesture but fail to extract very fast but short movements. The association between the two methods for *peak velocity* relied on the assumption that overall the same sub-movements were extracted by the Kinect and the human coder, which generally was true, however, when this was not the case, the fastest sub-movement recorded by the Kinect would be a different sub-movement labeled as the fastest by the human coder. In other words, the outcome of movement segmentation mattered for both the *sub-movements* and *peak velocity*. These results suggest that velocity is a very difficult metric to code visually due to it being mathematically very precise and therefore may be made more accessible by using more robust measures, such as the Kinect. In sum, the somewhat lower overlap between the automatic and manual method for *peak velocity* and *hold-counts* does not undermine the robustness of the obtained results. On the contrary, it indicates that the Kinect can be an effective means to code kinematic features that provide significant challenges for accurate manual coding. Using a mathematical approach with strict criteria therefore allows fine-grained and accurate quantification of these features.

Implementation

Our approach was recently applied in a study by Trujillo and colleagues (Trujillo et al., 2018) in which participants performed 31 object-directed actions (e.g. brushing hair, folding a paper, etc.) and the corresponding representational gestures (i.e. enacting



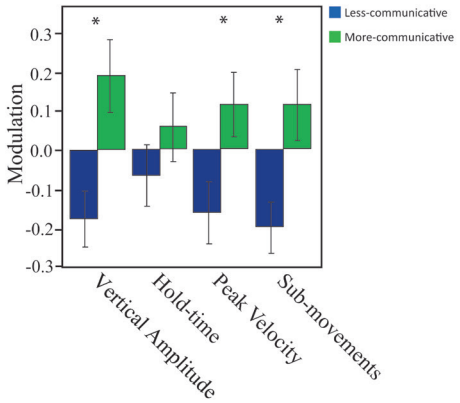


Figure 23. Kinematic modulation data in more- and less-communicative gestures, reproduced with permission from data from Trujillo and colleagues (Trujillo et al., 2018). Kinematic features are displayed along the x-axis, while modulation values (deviation from sample mean) are displayed along the y-axis. Blue bars depict the less-communicative context, while green bars depict the more-communicative context. * $p < 0.001$.

the same actions without the object being present) in two settings. The difference between these two settings was that in the first setting, participants were induced to believe someone was observing their actions and gestures with the aim to learn from them (i.e. more communicative context), whereas in the second setting, although they also believed they were being observed, the participants assumed they were performing actions and gestures for themselves (i.e. less communicative context). The key finding of the production experiment in Trujillo and colleagues' study was that both actions and gestures were kinematically modulated with respect to the context in which they were performed, with sub-movements and vertical amplitude being increased in both actions and gestures in the more- compared to less-communicative context. Peak velocity was additionally increased in more- compared to less-communicative gestures (Figure 23). The comprehension experiment in the same study showed that the kinematic modulations of gestures were reliably perceived and utilized by the addressees, in that naïve observers used the increased vertical amplitude to infer whether the actor performed the gesture for themselves or for the viewer. A follow-up study using the same production data additionally showed that these increases in sub-movements, peak velocity, and holds improve comprehension of the semantic content of the act (Trujillo, Simanova, et al., 2019). Together, these findings show that our toolkit can quantify kinematic features that are important characteristics of the communicative context of a manual act, and that

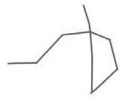
these same features are used by addressees to understand intention and semantic content.

Limitations

While this validation study shows promise for the quantification of kinematic features in action and gesture research, it should be noted that the features extracted and validated here only measure the qualities of movement in a given act. We therefore do not expect this methodology to replace manual coding, particularly in the case of qualitative classification of gestures. The feature extraction is also meant to capture a type of summary information of a given manual act. That is to say, this does not generate online or continuous coding of all movement, but is meant to be applied to a single act, or set of movements which one wishes to characterize. While the current protocol utilizes pre-defined start and end points to define what constitutes a single act, or time frame of analysis, this could be modified to be used together with automatic segmentation or gesture defining tools (see, for example, work by de Beugher et al. (Beugher, Brône, & Goedemé, 2018)).

Applications

Using the Microsoft Kinect to capture gesture production and automatically extract kinematic features can be an important tool for researchers interested in meaningful movements. Previous research has shown that velocity of pointing gestures may be modulated by the communicative context in which they are performed (Peeters et al., 2013), and the size (Bavelas et al., 2008; Campisi & Özyürek, 2013) or height (Hilliard & Cook, 2016) of gestures may also be modulated by the common-ground in knowledge between the speaker and addressee. Furthermore, velocity and size of communicative gestures has also been shown to effect the response of interactional partners (Innocenti et al., 2012), as well as signal communicative intention (Trujillo et al., 2018) and clarify the semantics of the act (Trujillo, Simanova, et al., 2019). Studies on communicative actions may also benefit from this tool. When compared to interacting with other adults, child-directed (Brand et al., 2002) as well as robot-directed actions (Vollmer et al., 2009) are modulated by distinct kinematic features. Similar features may also be useful in differentiating between various adult interactive contexts, such as demonstration and joint action coordination (McEllin et al., 2018). Clinicians may also benefit from such analysis, as pantomime production is often used when assessing aphasia (Goldenberg, Hermsdörfer, Glindemann, Rorden, &



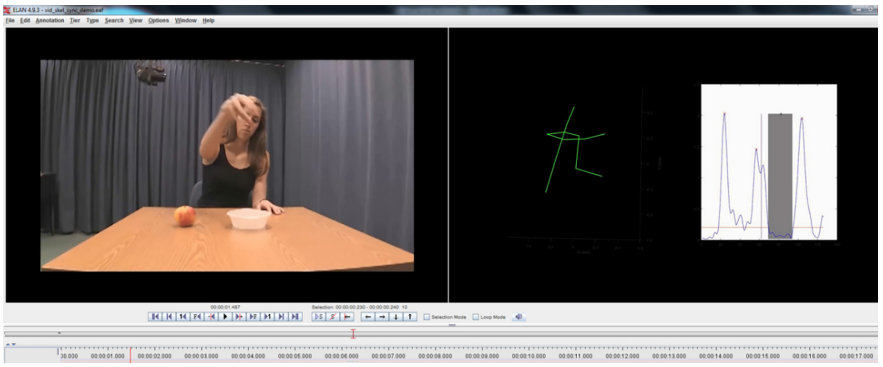


Figure 24. Example of video and kinematic pairing in ELAN. On the left, the standard video recording is being played, while on the right a skeleton of the motion capture data as well as the velocity profile of the right hand are played simultaneously. Note the horizontal bar on the velocity profile, which moves from left to right as the video plays, allowing a coder to see to which part of the plot the current video frame corresponds.

Karnath, 2007; Hermsdörfer et al., 2012). An additional advantage to this approach is that the Kinect does not require reflective markers or other physical components attached to the participant, allowing a somewhat more ecological approach where the participant may be less aware of the fact that their movements are being recorded. In the case of clinical applications, this markerless aspect allows the tool to be implemented without providing any additional discomfort to the patient.

Aside from the direct quantification of specific features, the velocity profile that is provided as output (see Figure 22) can also be used side by side with video data in order to assist in the manual coding of strokes and holds. While the gestural units themselves are accurately defined in time by the Kinect code, the manual coder can more easily code the qualitative or categorical features of these units. For example, by finding the onset of a velocity peak that has been marked as a submovement by the toolkit, one can easily and precisely find the onsets (and offsets) of strokes. Similarly, the onsets and offsets of holds are made more precise by finding the onsets and offsets as defined by the toolkit. In figure 23, we give an example of a video paired with a Kinect-acquired velocity profile video which can be used to find onsets and offsets of relevant gesture phases.

Finally, Kinect data can be used to supplement video data thanks to its 3-dimensional nature. While gesture data in the lab is often acquired with multiple cameras capturing distinct angles, fieldwork may make such multi-camera setups more difficult. In this case, standard video data may be used as the primary source for coding data, but the

Kinect acquisition would additionally provide the velocity profile output to support coding of gesture phases, as well as any number of angles of visualization to reduce ambiguities that may come from typical 2D data and limited angles of acquisition. As an example of this, Figure 24 depicts the Kinect acquisition playing alongside the video recording, where the movements can be seen at a slightly rotated viewing angle.

Summary

Our novel kinematic feature extraction protocol provides a robust measure of spatial and temporal kinematics, with extracted features being representative of what human observers can reliably code, while additionally allowing access to features that human coders have difficulty quantifying. Overall, we believe this methodology can be a useful tool for gesture researchers, clinicians, and others interested in quantifying the kinematics of meaningful human movement.

Acknowledgments

The authors would like to thank Judith Holler for helpful critiques on the applications of our methodology, and Harold Bekkering for his contribution to defining the kinematic features. This research was supported by the NWO Language in Interaction Gravitation Grant. The authors declare no conflict of interest in this study.

Supplementary table 1. Inter-rater agreement for Vertical amplitude

		Coder 2			
	value	1	2	3	4
Coder 1	1	**9**	19		
	2	1	**45**	1	
	3		5	**23**	1
	4				**7**

Supplementary table 2. Inter-rater agreement for Hold-count

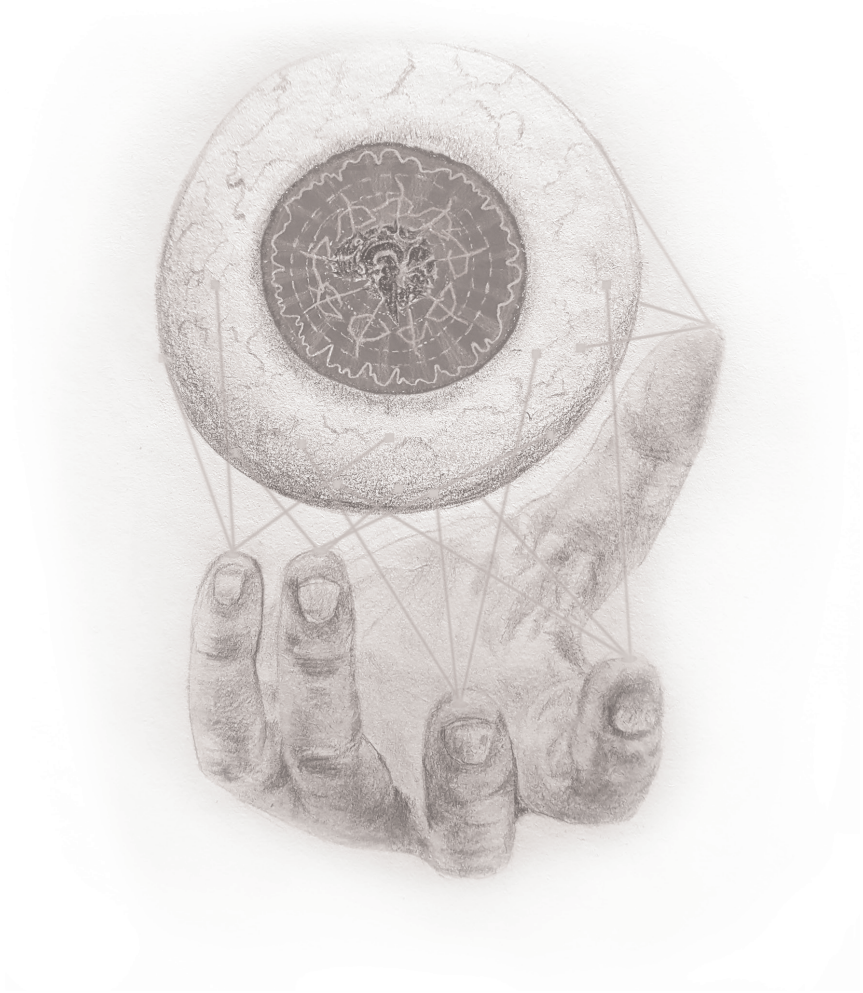
		Coder 2				
	value	0	1	2	3	4
Coder 1	0	42	2	1	1	
	1	6	29	10	1	1
	2	3	1	7		1
	3	3	2			
	4			1		

Supplementary table 3. Inter-rater agreement for Peak velocity

		Coder 2						
	value	2	3	4	5	6	7	8
Coder 1	2	2	1	1	1			
	3		13	4	4	1		
	4		7	22	3	3	1	
	5		2	5	5	4	1	
	6		1	1	2	3	1	
	7			1		3	4	2
	8					3	2	8

Supplementary table 4. Inter-rater agreement for Sub-movements

Coder 1 \ Coder 2	2	3	4	5	6	7	8	9	10	11	12	13	15
3	1	5	1				1						
4		2	5										
5			8	7	3								
6		1		2	5	2		1	1				
7	1	1	2	1	3	1	2	1	2				
8			1	1	2	1	6	3					
9		1		1	5	2	1	3					
10							1	1	1				
11							1	2	1	3			
12							1		2	2			
					1	1	1				3		
												1	
													1



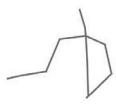
Chapter 7

General Discussion

As social creatures, humans rely heavily on the ability to understand what others are doing and why. Similarly, we make ourselves understandable to others. This is the glue that allows our complex social structure to function. Besides conventionalized communicative behaviors such as speaking or giving a “thumbs up”, we also regularly act on objects, or simulate doing so using hand gestures, in order to demonstrate how to do something or to instruct someone to act. The way that we perform these actions and gestures changes depending on whether we are doing them for ourselves or as a demonstration, and furthermore depending on for whom we are demonstrating (e.g. a child or an adult). Specifically, different social contexts lead to changes in the kinematics (e.g. velocity, size, complexity) of our movements. If our actions and gestures are shaped by the context in which they are produced, this means that information about our intentions, both in terms of what we are trying to convey and why, is externalized in our behavior.

While the idea of intentions being visible in our movements is not a new one, little research has thus far investigated the role of these kinematic changes in communicative actions and gestures. This thesis brings together ideas from studies on intentional actions, interactional gestures, development, and human brain imaging in order to demonstrate how communicative intentions shape our kinematics in an informative way, and how the brain extracts this information when we are observing such actions and gestures. At a general level, my hypothesis was that people in a more communicative setting would exaggerate their movements, and this exaggeration would not only make the act easier to understand, but would also reveal the underlying communicative intention. As I took this to be a global communicative strategy in order to more effectively convey information, I hypothesized that this finding would additionally extend to noisy scenarios, where the same kinematic exaggeration would be used to compensate for speech being a less reliable signal.

In addition to general theoretical advances, this thesis was novel in several ways. First, we applied motion tracking techniques to capture relatively unconstrained actions and their corresponding pantomime gestures. This allowed us to quantify the movement kinematics of a variety of actions and gestures, making our results generalizable beyond a single movement sequence. By using these same data and videos of actions and gestures in comprehension experiments with new participants, we were able to test not only what intention information is present in the kinematics



of these movements, but also how this information is used by observers in order to understand the meaning or to infer the intention of the person performing the action or gesture.

In this thesis I have demonstrated that the intention to communicate affects the kinematics of our actions and gestures, and that these kinematic differences both enhance the comprehensibility of what we are doing and act as a signal that what we are doing is intended for our addressee. Our addressee recognizes this intention because the exaggeration is unexpected based on previous experience. This leads them to infer that our intention was to use the action or gesture communicatively. Using brain imaging, I have shown that this process of intention inference is supported by a similar neural mechanism as is used to rationalize unusual or inefficient behavior observed in others. Extending this model of communicative kinematic exaggeration into noisy, co-speech gestures, I demonstrated a similar effect. Specifically, I found that increased noise leads to an increase in the visual information conveyed in the gesture. In the remainder of this chapter I will first outline the main findings described in the previous chapters and then discuss these findings within the broader literature, describing the implication of this work for current theory. Finally, I will speculate on how future research can build on these findings to better understand how movement kinematics fit into the bigger picture of human social interaction.

7.1. Summary of Main Findings

Chapters 2-3 focused on the signaling and recognition of communicative intentions. In **Chapter 2,** I found that a communicative intention leads to kinematic exaggeration in both actions and gestures. Specifically, the velocity, size, and segmentation (i.e. distinctive separation of constituent movements) were increased. This was paired with an increase in direct eye-gaze towards the viewer. When asked to classify these actions or gestures as being communicatively intended or not, observers primarily relied on the presence of direct eye-gaze. When this information was experimentally removed they used the kinematic modulation, more frequently classifying exaggerated gestures as being communicatively intended. In particular, gestures produced higher in space were seen as more communicative.

The videos and kinematic data from Chapter 2 were used as stimuli for the MRI experiment carried out in **Chapter 3.** In this experiment, I replicated the finding of kinematic exaggeration being perceived as communicative, and additionally found that regions of the mentalizing network and mirroring network are activated in response to this kinematic modulation. This effect occurred at the level of kinematics, such that increases in communicative kinematic exaggeration were directly correlated with brain response in these regions. Furthermore, I found that top-down connectivity from the mentalizing network to the mirroring network is also associated with kinematic exaggeration. In other words, communicative kinematic modulation directly changes the strength of the mentalizing network's influence on the mirroring network.

In **Chapter 4** I again used the videos and kinematic data from Chapter 2, but here to investigate how kinematics influence the intelligibility of a gesture. Overall, I found that the more communicatively intended gestures were recognized better than less communicative gestures. This effect seemed to come from several factors. First, exaggeration of temporal features, specifically a decrease in velocity and an increase in segmentation of a gesture, led to it being better understood by an observer. By varying the amount of the complete gesture that participants actually saw, we also found that this advantage for communicative gestures was driven in the early stages of the movement by the type of kinematic features that we have discussed thus far, but also by other cues, which we speculated to be hand and finger kinematics. When we experimentally removed visibility of the fingers, communicative gestures were still better recognized than less communicative gestures, but there was no longer an advantage in the early stages of the movement.

In **Chapter 5** I quantified speech acoustics, face kinematics, and gesture kinematics to investigate how a multimodal utterance is affected by a noisy environment. I replicated previous findings of increased speech intensity in response to noise, and additionally found an increase in gesture submovements in response to increased noise. I found no evidence for a modulation of face kinematics specifically in response to noise, but instead found that face kinematics such as mouth opening and lip velocity were strongly correlated with submovements, suggesting that face kinematics do not respond specifically to noise, but follow the overall communicative strategy of modulating visual features (e.g. gesture kinematics and face kinematics). Finally, I found evidence that while high noise conditions induce a modulation of

both speech and gesture, moderate noise levels may be compensated by either strongly modulated speech without co-speech gesture, or by less modulated speech that is accompanied by (kinematically modulated) co-speech gesture.

In **Chapter 6,** I present a methodological validation of the kinematic features discussed in previous chapters. I show that the kinematic features quantified using motion capture are valid parallels to the features that have been manually annotated based on video data alone in previous studies of communicative actions and gestures. I additionally discuss how this approach can support studies of gestures and actions, and can provide new possibilities for future research in communicative behavior.

7.2 Discussion

7.2.1 Communication in Movement

Intentions Shape the Way We Move

The first core question of this thesis was how communicative intentions are expressed in a variety of actions and gestures. In other words, I wanted to know whether there is a particular set of kinematic features that is modulated by the intention to communicate, signaling our intentions through more than just our eyes and speech. This was addressed in **Chapter 2,** where I found that co-occurring eye-gaze as well as spatial and temporal kinematic features are modulated across a variety of actions and gestures, depending on our intention to communicate.

Our intentions are typically thought of as private, internal things, unless we make them public through speech or through the things and people that we look at. However, in the 1980s there was already an idea that our intentions are visible even at the lowest (i.e. kinematic) level of our actions (Runeson & Frykholm, 1983). To return to the idea of action hierarchies discussed in **Chapter 1,** the intentions, or goals, of an action (i.e. upper levels of the hierarchy) influence the movement qualities (i.e. kinematics - the lower levels of the hierarchy). Since then, several studies have shown that concrete intentions, such as what we intend to do with an object, influence the kinematics of the movements used to reach out and grasp the object (Becchio et al., 2018; Cavallo et al., 2016; Naish et al., 2013). Similarly, abstract social intentions, such as the intention to communicate, influence the kinematics of reach-to-grasp movements (Quesque et al., 2013; Sartori et al., 2009), as well as the

qualitative features of co-speech gestures (Campisi & Özyürek, 2013). Our results build on these findings by showing a set of kinematic features that are modulated across a variety of different actions and gestures, and are at a high enough level of description to correspond to previous qualitative findings in social interactions. By this I mean that the kinematics describe things like overall size and segmentation of a movement, rather than lower level kinematics such as the orientation of one joint compared to another, or the specific trajectory of one single movement. This comparability with qualitative features is an important feature because it allows us to objectively quantify these movements while still building upon established phenomena in various research disciplines. Our findings provide evidence that both actions and gestures are modified in a similar way when we intend to use them communicatively. This general strategy corresponds well with findings related to leader-follower roles in joint action tasks (Candidi, Curioni, Donnarumma, Sacheli, & Pezzulo, 2015; Vesper et al., 2017), as well as child-directed actions (Brand et al., 2002; Fukuyama et al., 2015) and gestures (Campisi & Özyürek, 2013).

Although communicative movements are kinematically exaggerated, this exaggeration is still a goal directed modulation that is constrained by cost and effect. The findings I present in **Chapter 2** suggest that communicatively intended gestures are larger and more segmented. Increasing the size of a movement could increase its salience to an observer, while segmentation allows the observer to more easily identify the individual parts that comprise the whole gesture. While both functions seem to be useful for communication, this level of description is perhaps too broad. Further increasing segmentation would lead to arbitrarily many individual movements making up a whole action or gesture, while increasing size could make the action or gesture unidentifiable due to its extreme exaggeration from typicality, or lead to all movements appearing equally salient.

Following the theoretical framework of sensorimotor communication provided by Pezzulo and colleagues (Pezzulo et al., 2013), I suggest that relevant aspects of the movement are exaggerated. This would signal to an observer which aspects are important, allowing them pick up the essential information in these aspects. The increased segmentation likely differentiates movement components that are relevant at a goal level. For example, if we are demonstrating how to open a jar, we might use holds (i.e. pauses in movement) between grasping and turning the lid, and again between turning the lid the final time and removing it from the jar (see Figure

10). In this way, we emphasize each turn as its own movement within the complete action. Similarly, we could exaggerate the trajectory of our hand movement when removing the lid in order to emphasize its removal from the jar itself.

Our motor system must generate actions (or gestures) that accomplish a particular goal without being overly costly in their planning or production. While communicative actions at first seem to deviate from this efficient planning by being exaggerated, these communicative acts should instead be thought of as having an additional goal. The concrete goal is the completion of the action, or its direct, physical consequence. The additional goal is the abstract outcome, such as conveying some information to someone or influencing them into acting on an object. Exaggerating the kinematics of the action or gesture therefore allows this communicative goal to be completed, but this exaggeration must accomplish this communicative goal, and it should not be more extreme than what is necessary. This is similar to Grice's conversational maxim of quantity, which suggests that when we speak, we only say as much as is necessary to communicate the message that we wish to convey (Grice, Cole, & Morgan, 1975). Our motor system may act in a similar way, flexibly adapting movements to efficiently convey relevant information to our addressee.

The flexible adaptation of our movement is likely part of a larger hierarchy of communicative behavior. Focusing on the production of actions, the context we are in as well as our intentions play a role in generating an appropriate sequence of actions or movements that is best suited to our goal (Hamilton & Grafton, 2006; van Elk et al., 2014). But the eyes are an important signal to our intentions, and thus a complete model of communicative behavior should include eye-gaze behavior as being a part of the solution to achieve our goal. This idea forms a logical extension of the model of speech-gesture production proposed by Kita and Özyürek (2003) that postulates a communicative planner. This communicative planner selects the information that will be conveyed, and whether it will be expressed as gesture or speech. If we expand this model, we can imagine a larger, similarly hierarchical model of communicative behavior. Given the importance of eye-gaze, for example, our intention to communicate generates addressee-directed eye-gaze, selects the appropriate action or gesture, and co-occurring speech. The relative importance and implementation of each articulator, or signal, is likely modulated by the context and by the addressee's needs. For example, when an addressee knows very little about the topic (Campisi & Özyürek, 2013), or expresses a lack of understanding (Holler

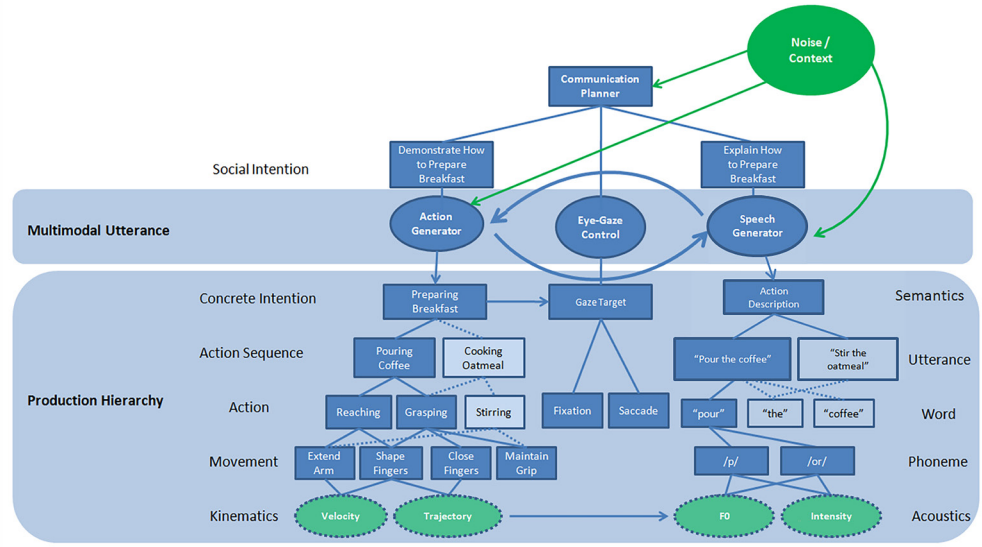


Figure 25. Graphical representation of multimodal production, under the influence of social intentions and communicative context. In this model, the top of the hierarchy is the Communicative Planner, which is where the modalities and semantic information are selected, similar to in the Interface Hypothesis Model (Kita &Özyürek, 2003). Here, I add eye-gaze as part of the complete multimodal utterance. The Production Hierarchy part of the model describes how different levels of detail in the production of speech, gesture, and eye-gaze behavior are all influenced by the upper levels. Namely, one's social (e.g. communicative) intention, the context in which the communication is occurring, and the configuration of the other communicative signals. The arrow between kinematics and acoustics is to show the biomechanical coupling between these levels. Lip movements are not listed separately here, but would entail their own action hierarchy under the Speech Generator (see Hickok, 2012 for a more detailed discussion of speech production). Note that for simplicity, the model does not show kinematics of eye-gaze, also because it is not clear whether they are influenced by communicative intention or context. Eye "actions", such as saccades and fixations are included, as addressee-directed eye-gaze is influenced by intentions.

& Wilkin, 2011), gestures may be modulated to become more informative. Such an intentional hierarchy should take the context, including addressee, into account to select the appropriate implementation of a range of communicative articulators including the hands, eyes, lips, speech, and body orientation. See Figure 25 for my visualization of such a model.

Seeing Intentions in Movement

The fact that we perform communicative actions and gestures in a kinematically distinct manner compared to non- or less-communicative actions and gestures provides a basis for observers to recognize this intention (Ansuini, Cavallo, Bertone, & Becchio, 2015; Becchio, Manera, et al., 2012; Cavallo et al., 2016). Results from **Chapter 2** support this idea, showing that naive observers are able to use the information embedded in kinematics in order to classify communicative intentions, at least in gestures. Importantly, by showing this effect in a variety of gestures, we demonstrate that the use of kinematic cues for intention recognition is not limited to single actions, but is likely a general mechanism.

The findings from **Chapter 2** specifically show that maximum vertical amplitude is the strongest cue that observers use for intention recognition. Vertical amplitude refers to the height of the hands in relation to the body. When observing a gesture, our results therefore suggest that movements produced higher in space than usual are seen as a signal of communicative intention. In **Chapter 1** I discussed the theory of natural pedagogy (Csibra & Gergely, 2009) as a potential explanation for how unusual or inefficient actions would trigger attention and could thus be used to signal the intention to communicate.

Our findings are similar to the work of Cavallo and colleagues (Cavallo et al., 2016) who showed that observers rely strongly on spatial kinematic features of reaching movements in order to decode concrete action intentions, such as discriminating between reaching to grasp a bottle in order to pour from it or to drink from it. Of these spatial features, vertical amplitude of the wrist during the reaching movement was most informative. In relation to the findings of **Chapter 2**, it is quite possible that observers recognize the exaggerated trajectories of communicative movements as being inefficient, or at least not corresponding to the how the complete action or gesture is typically performed. This leads them to rationalize that the actor must have an additional goal, or intention, that produced this spatially exaggerated action or gesture.

Although observers are able to recognize the underlying communicative intention of an action or gesture from the kinematics, this is only part of the picture. In fact, participants were more reliant on the eye-gaze behavior of the person performing the action or gesture, as direct eye-contact is a strong signal of the intention to

communicate or interact (Cañigueral & Hamilton, 2019; Csibra & Gergely, 2009). This suggests that just as there is a hierarchical structure in action production, there is likely also a hierarchy that describes the relative importance of different features or articulators. Direct eye-contact and hearing one's own name, for instance, are considered highly salient cues (Kampe et al., 2003) that likely prepare us to interact (Wang & Hamilton, 2012). When these cues are unavailable, we can rely on the kinematics. Body orientation also leads to the feeling of being addressed (Nagels et al., 2015), making this an even more complex set of cues. Following from Donnarumma and colleagues' theory of action perception as a form of continuous hypothesis testing (Donnarumma, Costantini, et al., 2017), observers may be able to take all of these cues, together with context, into account in order to determine which cues to focus on. This would form a sort of 'attentional hierarchy' that directs attention to the most salient information currently available. In this way they can best predict, or understand, why a person is doing what they are doing. An interesting avenue for future research is to further bridge these different lines of evidence in order to see how they all fit together into one hierarchy of potential information sources.

As the main cue that we found in **Chapter 2** is vertical height, it is important to consider an alternative hypothesis that could explain why observers preferentially use vertical amplitude. The hands are being brought closer to the actor's eyes, thus potentially making the movement more salient to the observer. Whether observers use the overall unexpectedness of the movement or its proximity to a highly salient area in the visual scene is not possible to discern in the experiment presented in Chapter 2. However, results from **Chapter** 3 additionally show that movement holds were also used as a cue to intention. These results provide evidence that communicatively intended movements are indeed made more salient through their kinematics. I discuss further evidence for the mechanism by which kinematics achieve this signaling process in the following section.

7.2.2 How the Brain Infers Intentions from Movement

In order to understand how humans are able to flexibly make use of kinematic information in order to make inferences about another person's intentions, I used fMRI (see **Chapter 1**, box 1.3) during an intention recognition task, as described in **Chapter 3.** I showed that activity in the mirror and mentalizing systems linearly correlates with the amount of communicative kinematic modulation in an observed

gesture. This finding is particularly interesting in light of our hypothesis that observers use the unexpectedness or unusualness of an observed movement to infer the underlying intention. The brain regions found in this experiment were directly in line with a meta-analysis of fMRI studies of intention processing during action observation (Van Overwalle & Baetens, 2009). As discussed in **Chapter 1,** these brain regions respond to wholly irrational actions, such as turning on a light switch with one's knee (Brass et al., 2007), but also to inefficient movement trajectories (Marsh et al., 2014).

The results presented in **Chapter 3** provide the first evidence that the brain uses the efficiency or unexpectedness of a gesture in order to attribute a communicative intention to the act. In terms of the specific kinematic cues that are being used, we found the same spatial feature as in **Chapter 2**, but also a temporal cue, namely the use of holds (i.e. pauses between movements). An important difference between the tasks used in **Chapters 2 and 3** is that **Chapter 3** used stick-light figures, a form highly reduced visual representation of the actor (see **Chapter 1**, Box 1.1) whereas **Chapter 2** used real videos. This is evidence that with less visual information, such as hand and finger kinematics or target object, observers rely on more kinematic cues to infer a communicative intention. This provides additional support for my hypothesis that intention recognition is based on the unexpectedness of the movement kinematics. However, this has another important implication. This fits well with the idea of a hierarchy of potential information sources that an observer can use in order to understand what a person is doing. The fact that we found an additional cue being used in **Chapter 3** suggests that the relative importance of kinematic information was higher in this visually simplified version of the videos. To further expand on this, we can imagine an attentional hierarchy in which there are cues that immediately signal the intention to communication, such as hearing one's name, or making direct eye-contact. However, if these very explicit cues are not present, we can still take advantage of more subtle cues, such as kinematics.

Finding that activation of the mirror and mentalizing systems correlates with kinematic exaggeration is interesting because it contributes to our understanding of intention recognition in general. First, rather than looking at distinct categories, such as 'efficient' and 'inefficient', we show that the brain responds to subtle changes in movement kinematics. We also show that this response is directly in line with how the brain responds to wholly unusual (Brass et al., 2007), unexpected (de Lange et

al., 2008), and inefficient (Marsh et al., 2014) actions. This provides evidence that the way we recognize communicative intentions is by recognizing that the action or gesture is not performed in a typical (i.e. non-communicative) manner. This is an important feature of our perception, and I believe it is related to theory of natural pedagogy. By keeping track of even the kinematics of how actions are typically performed, we are better able to focus on novel, potentially useful information, or simply use the regularities to understand the reason (i.e. the intention) for what someone is doing.

An additional finding from **Chapter 3** is that communicative kinematic modulation changes the influence of the mentalizing system on the mirroring system. This is a particularly interesting result because it matches well with a model of social mimicry described by Wang and Hamilton (2012), who suggested that top-down connectivity between these systems is a response to social stimuli that prepares us to respond appropriately (Wang & Hamilton, 2012). The finding of a similar pattern in response to communicatively intended movements suggests that this dynamic between the two systems may reflect a general mechanism to recognize communicatively intended behavior and prepare us to respond appropriately.

The most important contribution of the intention recognition tasks that were used in **Chapters 2 and 3** is showing evidence for a general mechanism by which the brain can recognize socially relevant behavior in others. As discussed above, this is not a task-specific mechanism, but rather a general mechanism that utilizes our ability to keep track of statistical regularities in the environment and in the behavior of others. Salient, relevant information from the continuous stream of sensory information with which we are constantly confronted allows us to make inferences about another person's intentions, thus making them more predictable. This predictability, in turn, allows us to more effectively engage in social interaction.

7.2.3 Clarifying Meaning in Movement

I showed in **Chapter 4** that temporal segmentation of a communicative gesture increases identification accuracy by naïve observers. This supports the theory that I put forth in **Chapter 1** suggesting that increased segmentation in communicatively intended actions and gestures could make it more easily identified.

The role of segmentation in making gestures more easily identified builds nicely

upon earlier work that showed a similar effect in object-directed actions (Brand et al., 2002). Overall, the results of **Chapters 2-4** show that kinematic modulation serves multiple purposes. On the one hand, it signals the underlying intention to the observer. In other words, it signals to the observer that what is happening is relevant to them. On the other hand, it also ensures that the movements themselves are easily identified. Together, the action or gesture is made both relevant and easily identified. This is particularly useful given that our expectations about how actions are performed can directly influence how we perceive these actions (Hudson et al., 2018). This means that our expectations shape our sensory experience. Drawing attention to particular aspects of a communicative signal, and exaggerating the qualities of this signal, could therefore help an addressee to get to the correct interpretation. This follows from the 'action perception as hypothesis testing' account (Donnarumma, Dindo, & Pezzulo, 2017) discussed in section 7.2.1. Simply put, the exaggerations break the action or gesture down into perceptually salient pieces. Observers' attention is attracted to these exaggerations due to their communicative quality, while the trajectory or timing of these movements is enhanced, making it easier for an observer to understand what is happening, and which aspects are relevant. In other words, the exaggerations tell an observer where to look, and what information is important, effectively supporting their 'hypothesis testing' (i.e. their prediction of what we are doing and why).

The multilayered role of kinematic modulation also fits well with previous theoretical accounts of communication. One classic framework of communication suggests that in order for communication to be successful, the speaker or actor must convey both the intention to communicate as well as the information that he or she wishes to communicate (Sperber & Wilson, 1986). We can of course learn from others' behavior incidentally by observing them, assuming the behavior is relevant or salient (S. W. Kelly et al., 2003). In order for true communication to take place, however, one must establish a communicative context. This is necessary because we cannot learn from everything we see. Instead, our sensory systems are actively tuned to novel or salient information (Pezzulo, Rigoli, & Friston, 2018), allowing us to deal with the constant stream of information without filtering out everything. While most accounts have focused on the need for explicit ostensive cues, such as eye-gaze, to establish a communicative context, I have shown that kinematics can fulfill both roles.

A further implication for this model of communicative modulation is that of communicative ability. The ability to communicate effectively is a highly complex skill that draws on many cognitive abilities (Boer, Toni, & Willems, 2013; Willems et al., 2010). Coordinating all of the communicative articulators in just the right way for a given context and addressee requires the brain to put together a highly complex, multifaceted signal that must be coordinated across the different articulators and across time. Similarly, for an addressee to actually understand this message, the bundle of signals, including contextual information, must be unified and decoded (i.e. Holler et al., 2015; S. D. Kelly, Healey, Özyürek, & Holler, 2015). The results I presented in **Chapter 4** suggest that kinematic modulation is one factor that contributes to better understanding in addressees (i.e., easier decoding of the message). However, not everyone performs actions the same way, and some people's movement kinematics make the actions easier or more difficult to understand than others (Koul, Cavallo, Ansuini, & Becchio, 2016). Given the complex nature of communication and communicative ability, it would be interesting to investigate whether our baseline kinematics are related to our ability to modulate these same kinematics, and whether communicative skill can be predicted from the extent of communicative modulation. Considering our hierarchical model of communicative behavior, one interesting hypothesis would be that successful communication can be predicted based on the degree of influences exerted by each upper level (e.g. concrete intention, social intentions, or context) on subsequently lower levels. Such a test could be important for understanding what makes communication successful.

7.2.4 Lending a Hand to Degraded Speech

In **Chapter 5** I presented the first evidence for a gestural Lombard Effect, showing that gesture kinematics are modulated by a noisy environment. This finding extends the framework from speech research in which changes in auditory and visual (i.e. lip movements) speech signals occur in response to noise by showing that this effect is not constrained to speech, but that this response is multimodal, with gesture kinematics also being adapted to the communicative context. While recent work has highlighted the joint contribution of visible speech and gesture to clarifying speech in noise (Drijvers & Özyürek, 2017), the results of **Chapter 5** demonstrate from the production side that both speech and gesture are modulated as part of a joint strategic response to noise. This is an important advance in understanding how multimodal communication is dynamically adapted to the interactional environment.

The finding that gesture kinematics are modulated by noise is also relevant when considering a general framework with which behavior can be made more communicative. This is directly in line with the theory of sensorimotor communication, put forward by Pezzulo and colleagues (2013), which suggests that any motor behavior can be modulated to more effectively signal information to an observer through the use of exaggeration. When considering **Chapters 2-4**, kinematic modulation can signal the intention to communicate and increase the perceptual clarity of the action or gesture. This signals that the action or gesture is relevant for the observer while simultaneously making the action or gesture more easily identifiable. In a noisy situation, such as in **Chapter 5,** the entire communicative expression is modulated in an attempt to make the meaning more clear. This puts the idea of communicative kinematic modulation into a larger framework in which meaningful movements, such as actions and gestures, are flexibly and dynamically adapted to the communicative situation. I have provided a somewhat simplified visualization of such a framework for signaling intention and meaning, and how addressees process this information, in Figure 26.

One of the main questions brought up in the introduction was how these various articulators fit together under the influence of a communicative intention. With biomechanical coupling (Pouw et al., 2019), neurobiological coupling (Woll, 2014), and cognitive coupling (Kita & Özyürek, 2003), it is fascinating to consider how the brain coordinates effective communication. Such a question, in its entirety, is beyond the scope of this thesis. However, I can at least provide some evidence for the question of whether our intention to communicate modulates behavior via a general increase in effort in all articulators, or whether it leads to a strategic shift of effort into particular articulators. Our findings from **Chapter 5** suggest that there is a strategic shift to the visual modality in moderately noisy conditions in which speech is less effective, but gesture can still help disambiguate the speech. However, we also see that lip movements seem more related to gestures than to the noise level itself, which suggests that the physical or neurobiological coupling between articulators is still an important part of the effects that we are investigating. Finally, in severe noise, when gestures may be less effective in disambiguating speech on their own, there was a general increase in both speech and effort. This shows that people not only strategically modulate certain signals, but they also take into account whether their addressee is likely to benefit from the gestures alone, or whether the speech signal is so degraded (e.g. in severe noise) that both speech and gesture must be

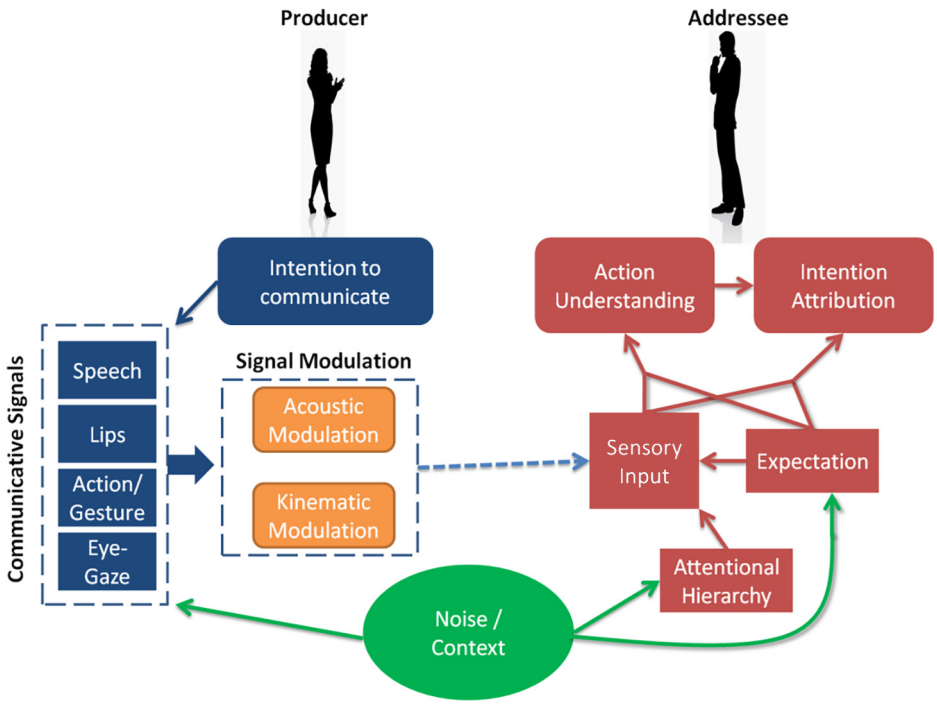


Figure 26. Graphical overview of how communicative intentions and context shape multimodal utterances, and how addressees make use of this modulation to inform action understanding and intention attribution. To summarize the previous sections of this chapter, both a producer's communicative intentions and the context in which the utterance is produced can lead to modulations of the communicative signals (e.g. speech, lip movements, actions/ gestures, eye-gaze). From the addressee side, the communicative context can influence their expectations (e.g. Brass et al., 2007) and their 'attentional hierarchy', which is the relative importance of the individual perceptual signals (e.g. visual information when noise is degrading speech). In turn, the attentional hierarchy and one's expectations will shape how sensory input is taken in. This sensory information, along with one's prior expectations, support action understanding. In the case of intention attribution, understanding the action itself is the first step, together with the integration of how one expected the action to unfold and how it was perceived to actually unfold.

modulated. Moving forward with the investigation of how these articulators are dynamically utilized for communication, it is important to consider not only what the specific context affords in terms of signaling, but also how the hands, eyes, and mouth are related to one another, and thus systematically affect one another.

7.2.5 Towards More Quantitative Studies of Communicative Movement

Given the prominent role of movement kinematics in communication, it is important to be able to study these complex features in an objective and replicable manner. While a strong study design is surely the starting point for useful research, strong methodology is equally important. In **Chapter 6** I showed that markerless motion tracking can provide a useful way to investigate the kinematics of meaningful movements such as actions and gestures. Specifically, I have provided a set of kinematic features that represent important temporal and spatial aspects of these movements that are sensitive to differences in social or communicative intentions. Beyond simply showing that such analysis can yield useful features for analysis, I have shown that manual annotation of the same features shows a strong correspondence. This demonstrates that these features, combined with low-cost markerless motion tracking technology such as the Microsoft Kinect, can be useful for capturing important spatial and temporal features of actions and gestures. The results and methodology presented in **Chapter 6** therefore provide a way to further advance the quantitative study of meaningful human movement.

7.3 Methodological Contributions of the Thesis

7.3.1 Ecological Validity

One of the primary highlights of the work I have presented in this thesis is the use of relatively naturalistic production experiments. Many of the theoretical developments on which I have based my work come from the domain of action production, where kinematic analysis is constrained to the very simple action of reaching to grasp an object, or where participants are explicitly instructed to do something communicative. These paradigms are obviously very useful and have provided many important breakthroughs in understanding how intentions shape the way we behave and are perceived by others. However, I believe the work I have presented provides an equally useful contribution. In **Chapter 2** I show how communicative intention shapes actions and gestures, but importantly I did this without ever explicitly telling participants to ‘try to be communicative’. In fact, the paradigm was carefully designed to ensure participants were not aware of the social or communicative manipulation, and they only believed that their task was to

accurately produce actions and gestures. Furthermore, the actions and gesture that they performed were highly varied, and analysis of the kinematics was focused on the entire sequence of movements.

The paradigm itself is an important contribution because it shows that even the very subtle manipulation in how we instructed participants had a quantifiable effect on their behavior. This is an important reminder of how crucial it is for instructions to be very consistent, and for social experiments to be very carefully planned. Furthermore, the use of markerless motion tracking ensured that participants were less aware of what we were interested in studying when compared to using markered tracking that required placing reflective markers on their body before starting. Creating a paradigm with such a subtle manipulation was important for this work because it shows that our effects are less likely to be based on what a person thinks they should do, but rather on how they actually respond to a situation. In other words, instructing someone to 'be communicative' could create an artificial behavior. Although I suggest that this paradigm has 'ecological validity', this is of course a relative term. Gesture studies often have less constrained tasks that allow speech, dialogues, or even full interactions. The paradigm I used in Chapter 2 is still very experimentally controlled in comparisons with these studies. In my experiments, I sought to find a useful middle ground for bringing cognitive and motor theories of action production and perception together with more communication-focused theories from gesture and language research.

In **Chapter 5**, I expanded the single-person paradigm to include a real addressee and further allowed any form of communication, making the experiment even more naturalistic than the one described in **Chapter 2**. This set-up provided a more ecological valid test of the visual aspects of the Lombard effect, including how it relates to speech. Typically, participants in studies of noise communication sit and speak while being otherwise isolated. This is of course quite different from most real-world noisy scenarios. These earlier studies have provided a good starting point, but I believe the experiment in Chapter 5 has taken this line of research even further by showing how the different communicative articulators interact and work together, and how people behave in a truly noisy environment.

To further bridge these two lines of research, I focused my analyses on the entire action or gesture. This means that I quantified the kinematics of the whole

expression. As things like average velocity or trajectory at specific points in time become very difficult to interpret for such complex, movement sequences, I also quantify the kinematics at a higher level of description. Rather than trajectory and location, I investigated features such as overall size and segmentation, which I refer to as *gross kinematics* in order to distinguish the level of detail from more fine-grained approaches. The advantage to this approach is that it also brings the level of description closer to that of gesture research and more qualitative action research, such as many child-directed action studies. This lends some quantitative validation to more qualitative findings, and hopefully shows how these two domains of research can benefit from one another. I believe the compromise between experimental control and ecological validity has been successful. However, in section 7.4 *Future Directions*, I expand on how I believe this line of research could continue to move forward.

7.3.2 Technological Innovation

The second major contribution of the thesis, beyond the empirical findings, is the technical implementation. In **Chapter 2** I used markerless motion tracking, which was quite a novel method for studying these complex movements, and for extracting kinematic features. The use of markerless tracking was useful for ecological validity, as discussed above, but also provided a very useful tool for future research. Whereas manual annotation requires trained coders to work with every new piece of data acquired, quantifying motion tracking data can be done automatically. In order to ensure that what I was calculating with the Kinect was actually valid and useful, I tested my scripts against the manual coding of two human annotators. The results, described in **Chapter 6**, show that this type of analysis can capture similar features to what humans code, but it can do so automatically, effectively streamlining the process and providing an objective and repeatable calculation of features. The scripts that I wrote for this quantification were reformatted to be easily implemented by future researchers on their own data, and were released, open source along with the accompanying paper (Trujillo, Vaitonyte, et al., 2019). In this way I hope to show the utility of this method while also making it more easily accessible.

Besides working with the Kinect body data, in **Chapter 5** I also implemented novel face-tracking data using the Kinect. This approach, to my knowledge, has not previously been used. This innovation is important because it allows a 3D capture

of a participant's face, which is an important articulator in human social interaction (Ekman & Rosenberg, 1997). The work that I present in **Chapter 5** is therefore a proof-of-concept that markerless, 3D face tracking can be used to capture visible speech features. Future research should compare the accuracy of this type of measurement with more commonly used, video-based face tracking. Such 3D face tracking could be very useful not only for empirical studies, but also for virtual reality settings where a participant's face is rendered onto a virtual avatar. This could provide more experimental control while allowing multiple people to interact in a virtual environment, or for studies where a participant would see his or her own face in the virtual environment (see Pan & Hamilton, 2018 for an overview of such research). As we move towards more multi-person research designs, the use of markerless tracking could prove to be a very powerful tool for capturing movement, whether for direct analysis, virtual rendering, or both.

7.4 Future Directions

7.4.1 Expanding Our Understanding of Social Context

The current thesis primarily utilized gestures and actions that were performed in the absence of speech. This allowed us to control extraneous effects such as speech production and discourse planning. However, communication is often multimodal, utilizing speech and context to convey meaning.

Our reliance on communicative actions and gestures in the absence of speech provided an important control, but also necessarily limits the scope in which we can interpret our results. In **Chapter 5** we go beyond silent movements, but these multimodal utterances were limited to conveying only a single word. It is quite likely that strategies are adapted to the number and effectiveness of communicative articulators. This follows from research showing that people use specific patterns of visual representation (e.g. acting out an action, or depicting an object) in both sign and gesture, depending on certain semantic qualities of the word (e.g. Ortega & Ozyürek, 2016). To extend this idea to gesture kinematics, if an action is more easily described verbally, then gestures may be less kinematically exaggerated. On the other hand, if the visual 'description' of an action is more useful, for example in the case of teaching complex action sequences, then kinematics may become more prominently expressed.

How we utilize action and gesture kinematics likely also depends on other social factors, such as the type of addressee with whom we are interacting or the natural constraints of the context. Previous studies, which have provided part of the theoretical foundation for the work described in this thesis, have shown that actions and gestures are produced differently depending on, amongst other factors, the shared knowledge between two individuals (Schubotz, Özyürek, & Holler 2019), the expertise-level and age of the addressee (Brand et al., 2002; Campisi & Özyürek, 2013; Fukuyama et al., 2015) and the social role of the person performing the act (McEllin et al., 2018). Understanding how these various factors fit together would help us to understand how communicative behavior is flexibly adapted to the situation, and whether there are commonalities amongst the strategies employed.

7.4.2 Beyond Communication in Movement

How different situations lead to different multimodal strategies is important for understanding the complexity of human communication, but cannot be fully answered with the paradigms utilized in this thesis. However, I believe this work provides a useful foundation for better understanding multimodal language. The production experiment in **Chapter 2** provides evidence for a similar modulation of kinematics, regardless of the specific action or gesture being performed. As discussed in **Chapter 6,** quantifying kinematic features based on markerless motion tracking data provides an ecologically valid test of how different contexts influence the way that we produce actions and gestures. This motion capture approach has the additional benefit of providing stimuli for future studies, allowing one to collect data in a relatively unconstrained manner, such as during conversation, for use in comprehension experiments. The fine-grained kinematic information available from motion capture nicely compliments the fine-grained acoustic information that has long been used for studying (psycho-) linguistics. For example, future research could look at how gesture kinematics and speech acoustics change depending on the context in which they are produced, how the two dynamics influence one another, and how this dynamic unfolds at different levels, such as sentence-, interaction-, and discourse-level.

7.4.3 How the Brain Extracts Meaning from Movement

The study presented in **Chapter 3** provided some first evidence that our expectations about how an action or gesture is normally performed can help us to infer underlying

communicative intentions. However, this is a necessary simplification of how such intention inference likely occurs. A somewhat more complete model of how intention recognition is cognitively achieved is provided in Figure 26. Beyond simply assuming that everyone has a relatively similar idea about how actions typically occur, we should also consider that these internal representations not only differ between individuals but are also influenced by recent experiences (Jacquet et al., 2016). Furthermore, our expectations not only help us make inferences, but they also bias our perception at the level of kinematics (Hudson et al., 2018). Returning to the idea of our natural inclination to learn from novel information in the environment, kinematic exaggeration may be required to push the perceptual system of our addressee away from their biases so that they see the *way* that we are performing the act. The extent of exaggeration may therefore be directly related to the extent of visual perceptual bias that likely occurs in an observer. Future research taking into account the state of prior expectations, and their influence on our perception of an action or gesture, would help to build a more complete model of how the brain processes the communicative information embedded in movement.

In most studies of intention recognition, including those presented here, there are only two possible choices. In the case of my experiments, observers knew that they only had to judge whether an action or gesture was performed in a more-communicative or less-communicative manner. In real-life scenarios, we are more likely to see behavior that has a larger number of potential underlying causes. Some of these may be social intentions, such as communication, deception, or competition. However, the action or gesture may also be influenced by other factors that may be less intentional, for example atypical movement patterns seen in Parkinson's disease (Alberts, Saling, Adler, & Stelmach, 2000), or even cultural factors, such as the taboo of left-handed gestures in Ghana (Kita & Essegbey, 2007). It is therefore crucial to understand how our knowledge about our addressee shapes our expectations about them, and further how we are able to make sense of the open-endedness of real-life intention inference. Recent experiments are showing that the brain can flexibly switch between prior expectations and incoming visual information in order to understand what someone is doing (Chambon et al., 2017). Expanding this model to include multiple intentions could be a useful avenue to understand how different neural systems allow us to focus on the most useful information available, allowing us to accurately understand another person's intentions.

7.4.4 Clarifying Meaning in Interaction

An intention to communicate typically implies that you are intending to engage in an interaction with another person. This interaction might be very short, in the case of an instruction that is meant to produce a response behavior in the addressee. The interaction may also be longer, involving dialogue and mutual exchange of information. In the current thesis, I have on the one hand investigated communicative actions and gestures produced without any addressee feedback, and on the other hand recognition of intention or identification of gestures without any adaptation of the actor to the observer's needs. This split between producer and addressee was a useful experimental control, but limits our interpretation of results in the context of natural interactions.

Chapter 5 provided a more interactive, multimodal setting, showing that speakers do indeed adapt their communicative strategy not only to the intention to communicate, but also to specifically compensate for difficulty in communicating. I believe this shows the importance for future research to investigate communication as an interactive process. Indeed, other researchers have recently pushed for more "second person", interactional studies, as opposed to the "third person", purely observer-based experiments typically used in social neuroscience (Risko, Richardson, & Kingstone, 2016; Schilbach et al., 2013). I believe that the current thesis has used experimentally controlled, yet ecologically valid methods to provide important insights into how communicative behavior is produced and understood. These findings should serve as the basis for further research utilizing more interactive settings.

One challenging, yet highly relevant avenue of research would be to capture multimodal communicative behavior from two interacting individuals. Such an approach would be highly valuable for understanding social interaction in a more complete way, modeling both the inter- and intra-individual processes. This would allow us to understand how communicative behavior is dynamically adapted to the partner's behavior, the context, common ground, or other features of the interaction. Such research would be useful for advancing social robotics or used to create more socially attuned virtual avatars. While **Chapter 6** utilizes a two-person interactive setting, a dynamical systems approach could allow us to look at not only what one person is doing, but how their behavior is adapted to the behavior of their partner, and how the interaction evolves over time.

7.4.5 Lending a Hand to Communication

To take the idea of studying communication in interaction a step closer to real-world situations, we should consider how communicative behavior is affected by situations that affect the sending or receiving of communicative signals. In **Chapter 6** I used interfering noise to see how gesture production was affected and whether noise changed the way speech and gesture went together. In this case, the gesture signal is the prominent signal, as it is not affected by the background noise and is prominent enough to be seen from across a room. How gesture and lip movements go together may be further affected by physical proximity between speaker and addressee. In order to manipulate the prominence of gestures, we could also consider situations where the view is partially obscured, for example due to objects or people between the speaker and addressee, or due to decreased light. Combining such studies with noise would help to disentangle how physical context can shape the way speech and gesture are modulated for communication and thus how much the speaker takes his or her addressee's viewpoint into account when planning a communicative utterance.

Beyond looking at what speakers do in adverse communicative situations, it is important to understand which behaviors are actually useful to an addressee. I showed in **Chapter 4** that kinematic modulation may support gesture identification. An interesting question left open is whether the strategies employed by a speaker in noise or reductions of visibility also support better comprehension. For example, while increases in the fundamental frequency of speech are commonly reported in response to noise, this increase does not seem to contribute to an increased intelligibility of the speech (Lu & Cooke, 2009). Instead, a general increase in higher frequencies relative to lower frequencies (i.e. spectral tilt) or other factors may better explain the higher intelligibility of "Lombard" compared to normal speech. Similarly, the extensive coupling of the communicative articulators likely leads to modulations that are actually not useful to the addressee. Investigating which modulations are useful for improving comprehension is a crucial step as it would have important implications for other areas are research, such as social robotics design and a better understanding of multimodal language comprehension.

7.4.6 Towards More Quantitative Studies of Social Behavior

The kinematic features that we calculated and utilized in the current thesis are based on the gross movements of the arms and hands (see the discussion in Chapter 4, Experiment II, for more on *gross kinematics*). The movements and shaping of the fingers and the overall configuration of the hands in relation to one another, however, are also highly important in conveying semantic and intentional information. An interesting question is how much information we utilize from each of these sources. For example, I showed in **Chapter 4** that the availability of information in the visual scene impacted how well participants were able to identify gestures. In **Chapters 2** and **3**, the amount of visual information also changed the set of kinematic features that were used for intention recognition. This suggests that we likely selectively utilize the information that we believe is most useful. I believe that our understanding of communicative behavior could therefore greatly benefit from determining which sources of information are more useful to observers.

Overall, I believe the results presented in **Chapters 2-4** show that even these gross kinematics, which only coarsely capture everything that is happening in a complex action or gesture, are still meaningful to observers. However, there is also evidence that much more fine grained kinematic features, such as the configuration and kinematics of the fingers, provide enough information to inform concrete intention recognition (Becchio et al., 2018; Cavallo et al., 2016), at least in some cases (Naish et al., 2013). I believe an interesting direction for future research is to quantify both the finger and hand kinematics as well as the more high-level, but coarse-grained, kinematic features that I have discussed in this feature. It is likely that such fine-grained kinematics would inform intention recognition at the level of predicting future actions in a given sequence, while gross kinematics inform our overall perception of the action.

Beyond simply collecting and analyzing more sources of information (e.g. hands, fingers, arms), I believe it is also important to look in more detail at what information the kinematics are actually carrying. In **Chapter 4** I showed that increasing segmentation led to more accurate identification of the gesture. An interesting next step would be to determine whether this segmentation occurs more strongly at certain points in the gesture, and whether is paired with changes in the trajectories of the movements as well. I believe it is likely that communicatively intended gestures

are kinematically modulated in a goal-oriented way. In this view, communicative gestures should emphasize information that is relevant. In some cases, such as demonstrating an entirely novel action (e.g. a new toy to a child), each component of the complete action would be emphasized. In other cases, it may be more specific. For example, it may be that we are showing a friend how to use our stove, which is similar to theirs but requires you to push in the knob before rotating it. In this case, the initial grasp and the final turn are not relevant for the demonstration. The pushing of the knob, however *is*, and would more likely be kinematically modulated. The modulation would emphasize that this component of the action is relevant for our friend to see, and would simultaneously show exactly how we did it. An interesting extension to the literature would there be to test whether the relevant aspects of a communicative action or gesture can be identified purely based on the kinematics. In other words, future research should address the temporal specificity of a communicative intention in its influence on action and gesture kinematics.

7.5 Conclusions

The aim of this thesis was to investigate the kinematic profile of communicatively intended actions and gestures, and how an addressee can utilize the information embedded in communicative kinematics. In Chapter 2, I showed that the intention to communicate systematically modulates the kinematics of both actions and gestures, and that both eye-gaze behavior and kinematics can signal the intention to communicate. In Chapter 3, I show that observers can use their expectations about how a movement is typically performed in order to infer the underlying communicative intention. Beyond simply reading intentions, Chapter 4 shows that communicative kinematics also make these movements easier to comprehend. However, I suggest that kinematic modulation is more than simply signaling one's intention, or clarifying meaning when one wishes to be more communicative. Chapter 5 shows that kinematic modulation is part of the larger, dynamically coupled speech-gesture production system, where degradation of one modality (e.g. speech) leads to a compensatory and potentially dynamically coupled response in gesture kinematics. This means that kinematic modulation is not just a reflection of an intention to communicate, but is a response to communicative need. It signals our intentions, clarifies meaning, and pushes that meaning through noise. These results show the importance of looking at meaningful movements, such as actions and gestures, both at a kinematic level and at the level of their interaction with other

communicative signals, such as speech, lip movements, and eye-gaze.

Given that movement qualities play such an important role in conveying meaning and making ourselves understandable to others, this work has several implications beyond the study of communicative action and gesture. Clinical populations, such as those with Autism or Parkinson's disease may experience communicative difficulties that can be explained by differences in how movement is produced and/or perceived. As social robotics becomes an increasing prominent aspect of society, we must understand how movement kinematics shape our perceptions of such social robots, and how our own movement kinematics should be taken into account by robots in order to understand our unspoken intentions.

All in all, communication is not just what we say or do. It is also the *way* we move, providing a glimpse into our intentions and giving shape to the ideas we wish to communicate.

References

Alberts, J. L., Saling, M., Adler, C. H., & Stelmach, G. E. (2000). Disruptions in the reach-to-grasp actions of Parkinson's patients. *Experimental Brain Research*, *134*(3), 353–362. https://doi.org/10.1007/s002210000468

Ansuini, C., Cavallo, A., Bertone, C., & Becchio, C. (2014). The visible face of intention: why kinematics matters. *Frontiers in Psychology*, *5*, 815. https://doi.org/10.3389/fpsyg.2014.00815

Ansuini, C., Cavallo, A., Bertone, C., & Becchio, C. (2015). Intentions in the brain: The unveiling of Mister Hyde. *Neuroscientist*, Vol. 21. https://doi.org/10.1177/1073858414533827

Ansuini, C., Cavallo, A., Koul, A., D'Ausilio, A., Taverna, L., & Becchio, C. (2016). Grasping others' movements: Rapid discrimination of object size from observed hand movements. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 918–929. https://doi.org/10.1037/xhp0000169

Ansuini, C., Santello, M., Massaccesi, S., & Castiello, U. (2006). Effects of end-goal on hand shaping. *Journal of Neurophysiology*, *95*(4), 2456–2465. https://doi.org/10.1152/jn.01107.2005

Anzulewicz, A., Sobota, K., & Delafield-Butt, J. T. (2016). Toward the Autism Motor Signature: Gesture patterns during smart tablet gameplay identify children with autism. *Scientific Reports*, *6*(1), 31107. https://doi.org/10.1038/srep31107

Auksztulewicz, R., & Friston, K. (2015). Attentional Enhancement of Auditory Mismatch Responses: a DCM/MEG Study. *Cerebral Cortex*, *25*(11), 4273–4283. https://doi.org/10.1093/cercor/bhu323

Bangerter, A. (2004). *Using Pointing and Describing to Achieve Joint Focus of Attention in Dialogue*. Retrieved from https://journals.sagepub.com/doi/pdf/10.1111/j.0956-7976.2004.00694.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bavelas, J., Coates, L., & Johnson, T. (2002). Listener Responses as a Collaborative Process : The Role of Gaze. *Journal of Communication*. Retrieved from www.mirc.com

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone:

Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495–520. https://doi.org/10.1016/j.jml.2007.02.004

Beattie, G., & Shovelton, H. (2002). An experimental investigation of the role of different types of iconic gesture in communication: A semantic feature approach. *Gesture*, *1*(2), 129–149. https://doi.org/10.1075/gest.1.2.03bea

Becchio, C., Cavallo, A., Begliomini, C., Sartori, L., Feltrin, G., & Castiello, U. (2012). Social grasping: From mirroring to mentalizing. *NeuroImage*, *61*(1), 240–248. https://doi.org/10.1016/j.neuroimage.2012.03.013

Becchio, C., Koul, A., Ansuini, C., Bertone, C., & Cavallo, A. (2018). Seeing mental states: An experimental strategy for measuring the observability of other minds. *Physics of Life Reviews*, Vol. 24, pp. 67–80. https://doi.org/10.1016/j.plrev.2017.10.002

Becchio, C., Manera, V., Sartori, L., Cavallo, A., & Castiello, U. (2012). Grasping intentions: from thought experiments to empirical evidence. *Frontiers in Human Neuroscience*, *6*(May), 1–6. https://doi.org/10.3389/fnhum.2012.00117

Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. *Cognition*, *106*(2), 894–912. https://doi.org/10.1016/j.cognition.2007.05.004

Becchio, C., Sartori, L., & Castiello, U. (2010). Toward You: The Social Side of Actions. *Current Directions in Psychological Science*, *19*(3), 183–188. https://doi.org/10.1177/0963721410370131

Beugher, S. De, Brône, G., & Goedemé, T. (2018). A semi-automatic annotation tool for unobtrusive gesture analysis. *Language Resources and Evaluation*, *52*(2), 433–460. https://doi.org/10.1007/s10579-017-9404-9

Biswas, K. K., & Basu, S. K. (2011). Gesture recognition using Microsoft Kinect®. *The 5th International Conference on Automation, Robotics and Applications*, 100–103. https://doi.org/10.1109/ICARA.2011.6144864

Blakemore, S., & Decety, J. (2001). From the Perception of Action to the Understanding of Intention. *Nature Reviews Neuroscience*, *2*, 561–567. https://doi.org/10.1038/35086023

Blokpoel, M., van Kesteren, M., Stolk, A., Haselager, P., Toni, I., & van Rooij, I. (2012). Recipient design in human communication: simple heuristics or perspective taking? *Frontiers in Human Neuroscience*, *6*, 253. https://doi.org/10.3389/fnhum.2012.00253

Boer, M. de, Toni, I., & Willems, R. M. (2013). What drives successful verbal communication? *Frontiers in Human Neuroscience*, *7*, 622. https://doi.

org/10.3389/fnhum.2013.00622

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer*. Retrieved from http://www.praat.org/

Bostanci, E., Kanwal, N., & Clark, A. F. (2015). Augmented reality applications for cultural heritage using Kinect. *Human-Centric Computing and Information Sciences*, *5*(1), 20. https://doi.org/10.1186/s13673-015-0040-3

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for 'motionese': modifications in mothers' infant-directed action. *Developmental Science*, *5*(1), 72–83. https://doi.org/10.1111/1467-7687.00211

Brand, R. J., & Shallcross, W. L. (2008). Infants prefer motionese to adult-directed action. *Developmental Science*, *11*(6), 853–861. https://doi.org/10.1111/j.1467-7687.2008.00734.x

Brand, R. J., Shallcross, W. L., Sabatos, M. G., & Massie, K. P. (2007). Fine-Grained Analysis of Motionese: Eye Gaze, Object Exchanges, and Action Units in Infant-Versus Adult-Directed Action. *Wiley Blackwell*.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating Action Understanding: Inferential Processes versus Action Simulation. *Current Biology*, *17*(24), 2117–2121. https://doi.org/10.1016/J.CUB.2007.11.057

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, *40*(8), 1129–1138. https://doi.org/10.1016/S0028-3932(02)00008-8

Caligiore, D., Pezzulo, G., Miall, R. C., & Baldassarre, G. (2013). The contribution of brain sub-cortical loops in the expression and acquisition of action understanding abilities. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2504–2515. https://doi.org/10.1016/J.NEUBIOREV.2013.07.016

Campisi, E., & Özyürek, A. (2013). Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children. *Journal of Pragmatics*, *47*(1), 14–27. https://doi.org/10.1016/j.pragma.2012.12.007

Candidi, M., Curioni, A., Donnarumma, F., Sacheli, L. M., & Pezzulo, G. (2015). Interactional leader–follower sensorimotor communication strategies during repetitive joint actions. *Journal of The Royal Society Interface*, *12*(110), 20150644. https://doi.org/10.1098/rsif.2015.0644

Cañigueral, R., & Hamilton, A. F. de C. (2019). The Role of Eye Gaze During Natural Social Interactions in Typical and Autistic People. *Frontiers in Psychology*, *10*, 560. https://doi.org/10.3389/fpsyg.2019.00560

Cary, M. S. (1978). The Role of Gaze in the Initiation of Conversation. *Social Psychology*, *41*(3), 269. https://doi.org/10.2307/3033565

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., & Becchio, C. (2016). Decoding intentions from movement kinematics. *Scientific Reports*, *6*(November). https://doi.org/10.1038/srep37036

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection. *NIPS 2007*. https://doi.org/10.1016/j.visres.2015.04.007

Chambon, V., Domenech, P., Jacquet, P. O., Barbalat, G., Bouton, S., Pacherie, E., … Farrer, C. (2017). Neural coding of prior expectations in hierarchical intention inference. *Scientific Reports*, *7*(1), 1–16. https://doi.org/10.1038/s41598-017-01414-y

Chang, C.-Y., Lange, B., Zhang, M., Koenig, S., Requejo, P., Somboon, N., … Rizzo, A. (2012). Towards Pervasive Physical Rehabilitation Using Microsoft Kinect. *6th International Conference on Pervasive Computing Technologies for Healthcare*, 159–162. https://doi.org/10.4108/icst.pervasivehealth.2012.248714

Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A., & Henson, R. (2016). Silent Expectations: Dynamic Causal Modeling of Cortical Prediction and Attention to Sounds That Weren't. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *36*(32), 8305–8316. https://doi.org/10.1523/JNEUROSCI.1125-16.2016

Chu, M., & Kita, S. (2011). The Nature of Gestures' Beneficial Role in Spatial Problem Solving. *Journal of Experimental Psychology: General*, *140*(1), 102–116. https://doi.org/10.1037/a0021790

Chu, M., & Kita, S. (2015). Co-thought and Co-speech Gestures Are Generated by the Same Action Generation Process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 257–270. https://doi.org/http://dx.doi.org/10.1037/xlm0000168

Church, R. B., Alibali, M. W., & Kelly, S. D. (Eds.). (2017). *Why Gesture?* https://doi.org/10.1075/gs.7

Ciaramidaro, A, Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B. G., & Walter, H. (2007). The intentional network: how the brain reads varieties of intentions. *Neuropsychologia*, *45*(13), 3105–3113. https://doi.org/10.1016/j.neuropsychologia.2007.05.011

Ciaramidaro, Angela, Becchio, C., Colle, L., Bara, B. G., & Walter, H. (2013). Do you mean me? Communicative intentions recruit the mirror and the mentalizing

system. *Social Cognitive and Affective Neuroscience*, *65*(3), 461–468. https://doi.org/10.1017/CBO9781107415324.004

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, *7*(4–5), 507–525. https://doi.org/10.1177/1461445605054404

Clark, R. A., Vernon, S., Mentiplay, B. F., Miller, K. J., McGinley, J. L., Pua, Y., … Bower, K. J. (2015). Instrumenting gait assessment using the Kinect in people living with stroke: reliability and association with balance tests. *Journal of NeuroEngineering and Rehabilitation*, *12*(1), 15. https://doi.org/10.1186/s12984-015-0006-8

Clayman, S. E. (2013). Turn-constructional units and the transition-relevance place. In *The Handbook of Conversation Analysis* (pp. 150–166). Retrieved from http://claire-bull.artistwebsites.com.

Cooke, M., Lecumberri, M. L. G., & Barker, J. (2008). *The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception*. https://doi.org/10.1121/1.2804952

Crasborn, O., Van Der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages*. *Sign Language & Linguistics*, *11*(1), 45–67. https://doi.org/10.1075/sl&l.11.1.04cra

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In *Processes of Change in Brain and Cognitive Development: Attention and Performance* (Vol. 21, pp. 249–274). https://doi.org/10.1.1.103.4994

Csibra, G., & Gergely, G. (2007). "Obsessed with goals": Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*(1), 60–78. https://doi.org/10.1016/j.actpsy.2006.09.007

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Cuijpers, R. H., Schie, H. T. Van, Koppen, M., Erlhagen, W., & Bekkering, H. (2006). Goals and means in action observation: A computational approach. *Neural Networks*, *19*(3), 311–322. https://doi.org/10.1016/j.neunet.2006.02.004

Da Gama, A., Fallavollita, P., Teichrieb, V., & Navab, N. (2015). Motor Rehabilitation Using Kinect: A Systematic Review. *Games for Health Journal*, *4*(2), 123–135. https://doi.org/10.1089/g4h.2014.0047

Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006). Lombard speech: Auditory (A), Visual (V) and AV effects. In *Proceedings of Speech prosody* (Vol. 2). Retrieved from http://www.isca-speech.org/archive

de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary Systems for Understanding Action Intentions. In *Current Biology* (Vol. 18). https://doi.org/10.1016/j.cub.2008.02.057

de Ruiter, J. P. (2006). Can gesticulation help aphasic people speak, or rather, communicate? *Advances in Speech Language Pathology*, *8*(2), 124–127. https://doi.org/10.1080/14417040600667285

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

de Ruiter, J. P., Noordzij, M. L., Newman-Norlund, S., Newman-Norlund, R., Hagoort, P., Levinson, S. C., & Toni, I. (2010). Exploring the cognitive infrastructure of communication. *Interaction Studies*, *11*, 51. https://doi.org/10.1075/is.11.1.05rui

DeBeer, C., Carragher, M., Nispen, K. van, Ruiter, J. de, Hogrefe, K., & Rose, M. (2015). Which gesture types make a difference? Interpretation of semantic content communicated by PWA via different gesture types. *GESPIN 4*, pp. 89–93.

Diersch, N., Mueller, K., Cross, E. S., Stadler, W., Rieger, M., & Schütz-Bosbach, S. (2013). Action Prediction in Younger versus Older Adults: Neural Correlates of Motor Familiarity. *PLoS ONE*, *8*(5), e64195. https://doi.org/10.1371/journal.pone.0064195

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*, *89*, 45–60. https://doi.org/10.1016/j.cortex.2017.01.016

Donnarumma, F., Dindo, H., & Pezzulo, G. (2017). Sensorimotor Coarticulation in the Execution and Recognition of Intentional Actions. *Frontiers in Psychology*, *8*, 237. https://doi.org/10.3389/fpsyg.2017.00237

Dragan, A. D., Lee, K. C. T., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308. https://doi.org/10.1109/HRI.2013.6483603

Dragan, A., & Srinivasa, S. (2014). Integrating human observer inferences into robot motion planning. *Autonomous Robots*, *37*(4), 351–368. https://doi.org/10.1007/s10514-014-9408-x

Drijvers, L., & Özyürek, A. (2017). Visual Context Enhanced: The Joint Contribution of Iconic Gestures and Visible Speech to Degraded Speech Comprehension. *Journal of Speech Language and Hearing Research*, *60*(1), 212. https://doi.

org/10.1044/2016_JSLHR-H-16-0101

Ekman, P., & Rosenberg, E. L. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.

Farroni, T., Csibra, G., Simion, F., & Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9602–9605. https://doi.org/10.1073/pnas.152159999

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9

Fernández-Baena, A., Susín, A., & Lligadas, X. (2012). Biomechanical Validation of Upper-Body and Lower-Body Joint Movements of Kinect Motion Capture Data for Rehabilitation Treatments. *2012 Fourth International Conference on Intelligent Networking and Collaborative Systems*, 656–661. https://doi.org/10.1109/iNCoS.2012.66

Fitzpatrick, M., Kim, J., & Davis, C. (2011a). The effect of seeing the interlocutor on speech production in noise. In *Data Processing* (Vol. 8). Retrieved from http://www.isca-speech.org/archive

Fitzpatrick, M., Kim, J., & Davis, C. (2011b). *The Intelligibility of Lombard Speech : Communicative setting matters* (Vol. 128). Retrieved from http://www.isca-speech.org/archive

Friston, K. (2011). What Is Optimal about Motor Control? *Neuron*, *72*(3), 488–498. https://doi.org/10.1016/j.neuron.2011.10.018

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7

Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron*, *50*(4), 531–534. https://doi.org/10.1016/J.NEURON.2006.05.001

Fry, D. B. (1975). Simple Reaction-Times to Speech and Non-Speech Stimuli. *Cortex*, *11*(4), 355–360. https://doi.org/10.1016/S0010-9452(75)80027-X

Fukuyama, H., Qin, S., Kanakogi, Y., Nagai, Y., Asada, M., & Myowa-Yamakoshi, M. (2015). Infant's action skill dynamically modulates parental action demonstration in the dyadic interaction. *Developmental Science*, *18*(6), 1006–1013. https://doi.org/10.1111/desc.12270

Galati, A., & Galati, A. (2015). Speakers adapt gestures to addressees' knowledge:

Implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, *29*(4), 435–451. https://doi.org/10.1080/01690965.2013.796397

Galna, B., Barry, G., Jackson, D., Mhiripiri, D., Olivier, P., & Rochester, L. (2014). Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait and Posture*, *39*(4), 1062–1068. https://doi.org/10.1016/j.gaitpost.2014.01.008

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech & Language*, *28*(2), 580–597. https://doi.org/10.1016/j.csl.2013.07.005

Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. In *Journal of Speech Language and Hearing Research* (Vol. 53). https://doi.org/10.1044/1092-4388(2009/08-0138)

Garnier, M., Ménard, L., & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? *The Journal of the Acoustical Society of America*, *144*(2), 1059–1074. https://doi.org/10.1121/1.5051321

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292. https://doi.org/10.1016/S1364-6613(03)00128-1

Gerwing, J., & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture*, *4(2)*, 157–195. https://doi.org/10.1075/gest.4.2.04ger

Gielniak, M. J., & Thomaz, A. L. (2012). Enhancing interaction through exaggerated motion synthesis. *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '12*, 375. https://doi.org/10.1145/2157689.2157813

Goldenberg, G., Hartmann, K., & Schlott, I. (2003). Defective pantomime of object use in left brain damage : apraxia or asymbolia ? *Neuropsychologia*, *41*, 1565–1573. https://doi.org/10.1016/S0028-3932(03)00120-9

Goldenberg, G., Hermsdörfer, J., Glindemann, R., Rorden, C., & Karnath, H. O. (2007). Pantomime of tool use depends on integrity of left inferior frontal cortex. *Cerebral Cortex*, *17*(12), 2769–2776. https://doi.org/10.1093/cercor/bhm004

Goldin-Meadow, S. (2017). Using our hands to change our minds. *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 8. https://doi.org/10.1002/wcs.1368

Gonzalez Rothi, L. J., Heilman, K. M., & Watson, R. T. (1985). Pantomime comprehension and ideomotor apraxia. *Journal of Neurology Neurosurgery, and Psychiatry*, *48*, 207–210. Retrieved from http://jnnp.bmj.com/content/jnnp/48/3/207.full.pdf

Grèzes, J., & Decety, J. (2002). Does visual perception of object afford action? Evidence from a neuroimaging study. *Neuropsychologia*, *40*(2), 212–222. https://doi.org/10.1016/S0028-3932(01)00089-6

Grice, H., Cole, P., & Morgan, J. (1975). *Logic and conversation*.

Gullberg, M., & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behavior*, *33*(4), 251–277. https://doi.org/10.1007/s10919-009-0073-2

Hamilton, A. F. de C. (2016). Gazing at me: the importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1686), 20150080. https://doi.org/10.1098/rstb.2015.0080

Hamilton, A. F. de C., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *26*(4), 1133–1137. https://doi.org/10.1523/JNEUROSCI.4551-05.2006

Hermsdörfer, J., Li, Y., Randerath, J., Goldenberg, G., & Johannsen, L. (2012). Tool use without a tool: Kinematic characteristics of pantomiming as compared to actual use and the effect of brain damage. *Experimental Brain Research*, *218*(2), 201–214. https://doi.org/10.1007/s00221-012-3021-z

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*, *45*(13), 1707–1724. https://doi.org/10.1016/J.VISRES.2004.12.021

Hickok, G. (2012). Computational neuroanatomy of speech production. In *Nature Reviews Neuroscience* (Vol. 13). https://doi.org/10.1038/nrn3158

Hickok, G. (2013). Do mirror neurons subserve action understanding? *Neuroscience Letters*, *540*, 56–58. https://doi.org/10.1016/j.neulet.2012.11.001

Hillebrandt, H., Blakemore, J., & Roiser, J. P. (2013). Dynamic Causal Modelling of effective connectivity during perspective taking in a communicative task. *NeuroImage*, *76*, 116–124. Retrieved from http://eprints.bbk.ac.uk/6562/1/6562.pdf

Hilliard, C., & Cook, S. W. (2016). Bridging gaps in common ground: Speakers design their gestures for their listeners. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, *42*(1), 91–103. https://doi.org/10.1037/xlm0000154

Hoetjes, M., & Carro, I. M. (2017). Under load : The effect of verbal and motoric cognitive load on gesture production. *Journal of Multimodal Communication.*, *4*(1–2). https://doi.org/10.14746/jmcs

Hoetjes, M., Krahmer, E., & Swerts, M. (2015). On what happens in gesture when communication is unsuccessful. *Speech Communication*, *72*, 160–175. https://doi.org/10.1016/j.specom.2015.06.004

Holladay, R. M., Dragan, A. D., & Srinivasa, S. S. (2014). Legible Robot Pointing. *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium On*.

Holler, J., & Beattie, G. (2005). Gesture use in social interaction: how speakers' gestures can reflect listeners' thinking. *2nd Conference of the International Society for Gesture Studies (ISGS): Interacting Bodies*, 1–12. Retrieved from http://gesture-lyon2005.ens-lyon.fr/article.php3?id_article=259

Holler, J., Kokal, I., Toni, I., Hagoort, P., Kelly, S. D., & Özyürek, A. (2015). Eye'm talking to you: Speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Social Cognitive and Affective Neuroscience*, *10*(2), 255–261. https://doi.org/10.1093/scan/nsu047

Holler, J., Shovelton, H., & Beattie, G. (2009). Do Iconic Hand Gestures Really Contribute to the Communication of Semantic Information in a Face-to-Face Context? *Journal of Nonverbal Behavior*, *33*(2), 73–88. https://doi.org/10.1007/s10919-008-0063-9

Holler, J., & Wilkin, K. (2011). An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses. *Journal of Pragmatics*, *43*(14), 3522–3536. https://doi.org/10.1016/j.pragma.2011.08.002

Hostetter, A. B., & Alibali, M. W. (2004). On the tip of the mind: Gesture as a key to conceptualization. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt0bq3923m/qt0bq3923m.pdf

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Huber, J. E., & Chandrasekaran, B. (2006). Effects of Increasing Sound Pressure Level on Lip and Jaw Movement Parameters and Consistency in Young Adults. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1368–1379. https://doi.org/10.1044/1092-4388(2006/098)

Hudson, M., McDonough, K. L., Edwards, R., & Bach, P. (2018). Perceptual teleology: Expectations of action efficiency bias social perception. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1884). https://doi.org/10.1098/rspb.2018.0638

Humphries, S., Holler, J., Crawford, T. J., Herrera, E., & Poliakoff, E. (2016). A third-person perspective on co-speech action gestures in Parkinson's disease. *Cortex*, *78*, 44–54. https://doi.org/10.1016/J.CORTEX.2016.02.009

Hussein, M. A., Ali, A. S., Elmisery, F., & Mostafa, R. (2014). Motion control of robot by using Kinect sensor. *Research of Applied Sciences, Engineering and Technology*, 1384–1388.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the Intentions of Others with One's Own Mirror Neuron System. *PLoS Biology*, *3*(3), e79. https://doi.org/10.1371/journal.pbio.0030079

Innocenti, A., de Stefani, E., Bernardi, N. F., Campione, G. C., & Gentilucci, M. (2012). Gaze direction and request gesture in social interactions. *PLoS ONE*, *7*(5), e36390. https://doi.org/10.1371/journal.pone.0036390

Jacquet, P. O., Roy, A. C., Chambon, V., Borghi, A. M., Salemme, R., Farnè, A., & Reilly, K. T. (2016). Changing ideas about others' intentions: updating prior expectations tunes activity in the human motor system. *Scientific Reports*, *6*(1), 26995. https://doi.org/10.1038/srep26995

Junqua, J.-C., Fincke, S., & Field, K. (1999). The Lombard effect: a reflex to better communicate with others in noise. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 2083–2086 vol.4. https://doi.org/10.1109/ICASSP.1999.758343

Kampe, K. K. W., Frith, C. D., & Frith, U. (2003). "Hey John": Signals conveying communicative intention toward the self activate brain regions associated with "mentalizing," regardless of modality. *The Journal of Neuroscience*, *23*(12), 5258–5263. https://doi.org/10.1002/hbm.21164

Kelly, S. D., Byrne, K., & Holler, J. (2011). Raising the Ante of Communication: Evidence for Enhanced Gesture Use in High Stakes Situations. *Information*, *2*(4), 579–593. https://doi.org/10.3390/info2040579

Kelly, S. D., Healey, M., Özyürek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review*, *22*(2), 517–523. https://doi.org/10.3758/s13423-014-0681-7

Kelly, S. D., Ozyurek, A., & Maris, E. (2010). Two Sides of the Same Coin: Speech and

Gesture Mutually Interact to Enhance Comprehension. *Psychological Science*, *21*(2), 260–267. https://doi.org/10.1177/0956797609357327

Kelly, S. W., Burton, A. M., Riedel, B., & Lynch, E. (2003). Sequence learning by action and observation: Evidence for separate mechanisms. *British Journal of Psychology*, *94*(3), 355–372. https://doi.org/10.1348/000712603767876271

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*(1), 55–68. https://doi.org/10.1017/S030500090001343X

Kendon, A. (1986). Current issues in the study of gesture. In J.-L. Nespoulous, P. Perron, A. R. Lecours, & T. S. Circle (Eds.), *The biological foundations of gestures: Motor and semiotic aspects* (1st ed., pp. 23–47). Psychology Press.

Kendon, A. (2004). *Gesture: visible actions as utterance*. Cambridge: Cambridge University Press.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, *8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2

Kim, J., Davis, C., Vignali, G., & Hill, H. (2005). *Auditory-Visual Speech Processing A visual concomitant of the Lombard reflex*. Retrieved from http://www.isca-speech.org/archive

Kim, J., Sironic, A., & Davis, C. (2011). Hearing Speech in Noise: Seeing a Loud Talker is Better. *Perception*, *40*(7), 853–862. https://doi.org/10.1068/p6941

Kipp, M. (2001). ANVIL A Generic Annotation Tool for Multimodal Dialogue. *Eurospeech*. Retrieved from http://www.mirlab.org/conference_papers/International_Conference/Eurospeech 2001/papers/page1367.pdf

Kita, S. (2000). How representational gestures help speaking. In *Language and Gesture* (1st ed., pp. 162–185).

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245–266. https://doi.org/10.1037/rev0000059

Kita, S., & Essegbey, J. (2007). Pointing left in Ghana. *Gesture*, *1*(1), 73–95. https://doi.org/10.1075/gest.1.1.06kit

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? *Journal of Memory and Language*, *48:1*, 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3¨A

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, *22*(8), 1212–1236. https://doi.org/10.1080/01690960701461426

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1371*, 23–35. https://doi.org/10.1007/BFb0052986

Koelewijn, T., van Schie, H. T., Bekkering, H., Oostenveld, R., & Jensen, O. (2008). Motor-cortical beta oscillations are modulated by correctness of observed action. *NeuroImage*, *40*(2), 767–775. https://doi.org/10.1016/j.neuroimage.2007.12.018

Koul, A., Cavallo, A., Ansuini, C., & Becchio, C. (2016). Doing it your way: How individual movement styles affect action prediction. *PLoS ONE*, *11*(10), e0165297. https://doi.org/10.1371/journal.pone.0165297

Kuznetsova, A. (2016). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1. https://doi.org/10.18637/jss.v082.i13

Lacadie, C. M., Fulbright, R. K., Arora, J., Constable, R. T., & Papademetris, X. (2007). Brodmann Areas defined in MNI space using new Tracing Tool in BioImage Suite. *Proceedings of the 14th Annual Meetng of the Organization for Human Brain Mapping*, *36*(1), 6494.

Lane, H., & Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. In *Journal of Speech and Hearing Research* (Vol. 14). https://doi.org/10.1044/jshr.1404.677

Lewkowicz, D., Quesque, F., Coello, Y., & Delevoye-Turrell, Y. N. (2015). Individual differences in reading social intentions from motor deviants. *Frontiers in Psychology*, *6*, 1175. https://doi.org/10.3389/fpsyg.2015.01175

Lisman, J. E., & Grace, A. A. (2005). The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron*, *46*(5), 703–713. https://doi.org/10.1016/J.NEURON.2005.05.002

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42. https://doi.org/10.1016/J.COGNITION.2016.12.007

Lombard, E. (1911). Le signe de l'élévation de la voix. *Annales Des Maladies de L'Oreille et Du Larynx*, *37*, 101–119.

Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, *51*(12), 1253–1262. https://doi.org/10.1016/J.SPECOM.2009.07.002

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-Reading Aids Word Recognition Most in Moderate Noise: A Bayesian Explanation Using High-Dimensional Feature Space. *PLoS ONE*, *4*(3), e4638. https://doi.org/10.1371/journal.pone.0004638

Macleod, A., & Summerfield, Q. (2009). *British Journal of Audiology Quantifying the contribution of vision to speech perception in noise*. https://doi.org/10.3109/03005368709077786

Manera, V., Becchio, C., Cavallo, A., Sartori, L., & Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research*, *211*(3–4), 547–556. https://doi.org/10.1007/s00221-011-2649-4

Manthey, S., Schubotz, R. I., & Von Cramon, D. Y. (2003). Premotor cortex in observing erroneous action: An fMRI study. *Cognitive Brain Research*, *15*(3), 296–307. https://doi.org/10.1016/S0926-6410(02)00201-X

Marsh, L. E., De, A. F., & Hamilton, C. (2011). Dissociation of mirroring and mentalising systems in autism. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2011.02.003

Marsh, L. E., Mullett, T. L., Ropar, D., & Hamilton, A. F. d. C. (2014). Responses to irrational actions in action observation and mentalising networks of the human brain. *NeuroImage*, *103*, 81–90. https://doi.org/10.1016/j.neuroimage.2014.09.020

McEllin, L., Knoblich, G., & Sebanz, N. (2018, November 20). Distinct kinematic markers of demonstration and joint action coordination? Evidence from virtual xylophone playing. *Journal of Experimental Psychology: Human Perception and Performance*, pp. 885–897. https://doi.org/10.1037/xhp0000505

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McNeill, D. (1994). Hand and Mind: What Gestures Reveal about Thought. In *Leonardo* (Vol. 27). https://doi.org/10.2307/1576015

Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes*, *22*(4), 473–500. https://doi.org/10.1080/01690960600696916

Meyer, D. E., Abrams, R. a, Kornblum, S., Wright, C. E., & Smith, J. E. (1988). Optimality

in human motor performance: ideal control of rapid aimed movements. *Psychological Review*, *95*(3), 340–370. https://doi.org/10.1037/0033-295X.95.3.340

Molenberghs, P., Mattingley, J., Cunnington, R., & Mattingley, J. B. (2011). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neuroscience and Biobehavioral Reviews*, *36*, 341–349. https://doi.org/10.1016/j.neubiorev.2011.07.004

Nagels, A., Kircher, T., Steines, M., & Straube, B. (2015). Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Human Brain Mapping*, *36*(5), 1925–1936. https://doi.org/10.1002/hbm.22746

Naish, K. R., Reader, A. T., Houston-Price, C., Bremner, A. J., & Holmes, N. P. (2013). To eat or not to eat? Kinematics and muscle activity of reach-to-grasp movements are influenced by the action goal, but observers do not detect these differences. *Experimental Brain Research*, *225*(2), 261–275. https://doi.org/10.1007/s00221-012-3367-2

Newman-Norlund, R., Van Schie, H. T., Van Hoek, M. E. C., Cuijpers, R. H., & Bekkering, H. (2009). *The role of inferior frontal and parietal areas in differentiating meaningful and meaningless object-directed actions*. https://doi.org/10.1016/j.brainres.2009.11.065

Nichols, S., & Stich, S. P. (2003). *Mindreading*. https://doi.org/10.1093/0198236107.001.0001

Novack, M. A., & Goldin-Meadow, S. (2017). Gesture as representational action: A paper about function. *Psychonomic Bulletin & Review*, *24*(3), 652–665. https://doi.org/10.3758/s13423-016-1145-z

Novack, M. A., Wakefield, E. M., & Goldin-Meadow, S. (2016). What makes a movement a gesture? *Cognition*, *146*, 339–348. https://doi.org/10.1016/j.cognition.2015.10.014

Ondobaka, S., & Bekkering, H. (2012). Hierarchy of idea-guided action and perception-guided movement. *Frontiers in Psychology*, *3*(December), 579. https://doi.org/10.3389/fpsyg.2012.00579

Ondobaka, S., De Lange, F. P., Wittmann, M., Frith, C. D., & Bekkering, H. (2015). Interplay between conceptual expectations and movement predictions underlies action understanding. *Cerebral Cortex*, *25*(9), 2566–2573. https://doi.org/10.1093/cercor/bhu056

Ortega, G., & Ozyürek, A. (2016). Generalisable patterns of gesture distinguish semantic categories in communication without language. *Proceedings of the*

38th Annual Meeting of the Cognitive Science Society, 1182–1187. Retrieved from http://hdl.handle.net/2066/159292

Osiurak, F., Jarry, C., Baltenneck, N., Boudin, B., & Le Gall, D. (2012). Make a gesture and I will tell you what you are miming. Pantomime recognition in healthy subjects. *Cortex*, *48*(5), 584–592. https://doi.org/10.1016/j.cortex.2011.01.007

Oztop, E., Wolpert, D., & Kawato, M. (2005). Mental state inference using visual control parameters. Cognitive Brain Research, 22(2), 129–151. https://doi.org/10.1016/J.COGBRAINRES.2004.08.004

Ozyürek, A. (2010). The role of iconic gestures in production and comprehension of language: evidence from brain and behavior. In *Gesture in Embodied Communication and Human-Computer Interaction* (1st ed.).

Özyürek, A. (2002). Do Speakers Design Their Cospeech Gestures for Their Addressees? The Effects of Addressee Location on Representational Gestures. *Journal of Memory and Language*, 46(4), 688–704. https://doi.org/10.1006/JMLA.2001.2826

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society B*, 369, 20130296. https://doi.org/10.1098/rstb.2013.0296

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, 107(1), 179–217. https://doi.org/10.1016/J.COGNITION.2007.09.003

Pan, X., & Hamilton, A. F. d. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. British Journal of Psychology, 109(3), 395–417. https://doi.org/10.1111/bjop.12290

Paraskevopoulos, G., Spyrou, E., & Sgouropoulos, D. (2016). A Real-Time Approach for Gesture Recognition using the Kinect Sensor. *Proceedings of the 9th Hellenic Conference on Artificial Intelligence - SETN '16*, 1–4. https://doi.org/10.1145/2903220.2903241

Pedersoli, F., Benini, S., Adami, N., & Leonardi, R. (2014). XKin: an open source framework for hand pose and gesture recognition using kinect. *The Visual Computer*, *30*(10), 1107–1122. https://doi.org/10.1007/s00371-014-0921-x

Peeters, D., Chu, M., Holler, J., Hagoort, P., & Özyürek, A. (2015). Electrophysiological and Kinematic Correlates of Communicative Intent in the Planning and Production of Pointing Gestures and Speech. *Journal of Cognitive Neuroscience*, *27*(12), 2352–2368. https://doi.org/10.1162/jocn_a_00865

Peeters, D., Holler, J., & Hagoort, P. (2013). Getting to the Point : The Influence of

Communicative Intent on the Kinematics of Pointing Gestures. *The 35th Annual Meeting of the Cognitive Science Society*, 1127–1132.

Peirce, J., Gray, J. R., Simpson, S., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*. Retrieved from http://www.forskningsdatabasen.dk/en/catalog/2443061653

Pezzulo, G., & Dindo, H. (2013). Intentional strategies that make co-actors more predictable: The case of signaling. *Behavioral and Brain Sciences*, *36*(04), 371–372. https://doi.org/10.1017/S0140525X12002816

Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS ONE*, *8*(11), e79876. https://doi.org/10.1371/journal.pone.0079876

Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., & Castelfranchi, C. (2018). The body talks: Sensorimotor communication and its brain and kinematic signatures. *Physics of Life Reviews*. https://doi.org/10.1016/J.PLREV.2018.06.014

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/J.TICS.2018.01.009

Pick, H. L., Siegel, G. M., Fox, P. W., Garber, S. R., & Kearney, J. K. (1989). Inhibiting the Lombard effect Effects of noise on speech production: Acoustic and perceptual analyses. *Citation: The Journal of the Acoustical Society of America*, *85*, 894. https://doi.org/10.1121/1.397561

Pittman, A. L., & Wiley, T. L. (2001). Recognition of Speech Produced in Noise. *Journal of Speech, Language, and Hearing Research*, *44*(3), 487–496. https://doi.org/10.1044/1092-4388(2001/038)

Pouw, W., Harrison, S. J., & Dixon, J. A. (2019). Gesture-Speech Physics: The Biomechanical Basis for the Emergence of Gesture-Speech Synchrony. *Journal of Experimental Psychology: General*. https://doi.org/10.31234/OSF.IO/TGUA4

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Quesque, F., Lewkowicz, D., Delevoye-Turrell, Y. N., & Coello, Y. (2013). Effects of social intention on movement kinematics in cooperative actions. *Frontiers in Neurorobotics*, *7*(OCT). https://doi.org/10.3389/fnbot.2013.00014

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., & Alku, P. (2013). *Analysis*

and Synthesis of Shouted Speech. Retrieved from https://pdfs.semanticscholar.org/e114/9cdf4a7990c5afd2bbf0158326f0056df8e3.pdf

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the Fourth Wall of Cognitive Science. *Current Directions in Psychological Science*, *25*(1), 70–74. https://doi.org/10.1177/0963721415617806

Rizzolatti, G., Cattaneo, L., Fabbri-Destro, M., & Rozzi, S. (2014). Cortical mechanisms underlying the organization of goal-directed actions and mirror neuron-based action understanding. *Physiological Reviews*, *94*(2), 655–706. https://doi.org/10.1152/physrev.00009.2013

Rose, M. L., Mok, Z., & Sekine, K. (2017). Communicative effectiveness of pantomime gesture in people with aphasia. *International Journal of Language and Communication Disorders*, *52*(2), 227–237. https://doi.org/10.1111/1460-6984.12268

Rostolland, D. (1982). Acoustic Features of Shouted Voice. *Acta Acustica United with Acustica*, *50*(2). Retrieved from https://www.ingentaconnect.com/content/dav/aaua/1982/00000050/00000002/art00006

Runeson, S., & Frykholm, G. (1983). Kinematic Specification of Dynamics as an Informational Basis for Person-and-Action Perception: Expectation, Gender Recognition, and Deceptive Intention. *Journal of Experimental Psychology: General Copyright*, *112*(4), 585–615. Retrieved from http://psycnet.apa.org/fulltext/1984-22257-001.pdf

Sacheli, L. M., Tidoni, E., Pavone, E. F., Aglioti, S. M., & Candidi, M. (2013). Kinematics fingerprints of leader and follower role-taking during cooperative joint actions. *Experimental Brain Research*, *226*(4), 473–486. https://doi.org/10.1007/s00221-013-3459-7

Sartori, L., Becchio, C., Bara, B. G., & Castiello, U. (2009). Does the intention to communicate affect action kinematics? *Consciousness and Cognition*, *18*(3), 766–772. https://doi.org/10.1016/j.concog.2009.06.004

Sartori, L., Becchio, C., & Castiello, U. (2011). Cues to intention: The role of movement information. *Cognition*, *119*(2), 242–252. https://doi.org/10.1016/j.cognition.2011.01.014

Schiffer, A.-M., Krause, K. H., & Schubotz, R. I. (2014). Surprisingly correct: Unexpectedness of observed actions activates the medial prefrontal cortex. *Human Brain Mapping*, *35*(4), 1615–1629. https://doi.org/10.1002/hbm.22277

Schiffer, A.-M., & Schubotz, R. I. (2011). Caudate Nucleus Signals for Breaches of Expectation in a Movement Observation Paradigm. *Frontiers in Human*

Neuroscience, *5*, 38. https://doi.org/10.3389/fnhum.2011.00038

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, *36*(04), 393–414. https://doi.org/10.1017/S0140525X12000660

Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, *44*(5), 718–730. https://doi.org/10.1016/j.neuropsychologia.2005.07.017

Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *The Journal of the Acoustical Society of America*, *85*(1), 295–312. https://doi.org/10.1121/1.397737

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology : CB*, *18*(9), 668–671. https://doi.org/10.1016/j.cub.2008.03.059

Senju, A., & Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127–134. https://doi.org/10.1016/j.tics.2008.11.009

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M.A.J.(2012). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24*(2), 426-434. https://doi.org/10.1093/cercor/bhs324

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, *33*(1), 115–125. https://doi.org/10.1111/j.1551-6709.2008.01006.x

Southgate, V., Chevallier, C., & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental Science*, *12*(6), 1013–1019. https://doi.org/10.1111/j.1467-7687.2009.00861.x

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Spunt, R. P., & Lieberman, M. D. (2013). The Busy Social Brain: Evidence for Automaticity and Control in the Neural Systems Supporting Social Cognition and Action Understanding. *Psychological Science*, *24*(1), 80–86. https://doi.org/10.1177/0956797612450884

Spunt, R. P., Satpute, A. B., & Lieberman, M. D. (2011). Identifying the What, Why, and How of an Observed Action: An fMRI Study of Mentalizing and Mechanizing during Action Observation. *Journal of Cognitive Neuroscience*, *23*(1), 63–74. https://doi.org/10.1162/jocn.2010.21446

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799

Stapel, J. C., Hunnius, S., & Bekkering, H. (2012). Online prediction of others' actions: The contribution of the target object, action context and movement kinematics. *Psychological Research*, *76*(4), 434–445. https://doi.org/10.1007/s00426-012-0423-2

Stapel, J. C., Hunnius, S., & Bekkering, H. (2015). Fifteen-month-old infants use velocity information to predict others' action targets. *Frontiers in Psychology*, *6*, 1092. https://doi.org/10.3389/fpsyg.2015.01092

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise The use of visible speech cues for improving auditory detection of spoken sentences. *Citation: The Journal of the Acoustical Society of America*, *26*, 1197. https://doi.org/10.1121/1.1907309

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition*, *13*(6), 657–665. https://doi.org/10.1080/13506280500410949

Titze, I. R., & Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946. https://doi.org/10.1121/1.402929

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907–915. https://doi.org/10.1038/nn1309

Todorov, E., & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, *5*(11), 1226–1235. https://doi.org/10.1038/nn963

Tomasello, M. (2010). *Origins of human communication*. Cambridge: MIT Press.

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. *Cognition*, *180*, 38–51. https://doi.org/10.1016/j.cognition.2018.04.003

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2019). The communicative advantage: how kinematic modulation supports semantic comprehension of pantomimes. *Psychological Research*.

Trujillo, J. P., Vaitonyte, J., Simanova, I., & Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, *51*(2), 769–777. https://doi.org/10.3758/s13428-018-1086-8

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, *8*(6), 769–800. https://doi.org/10.1080/13506280042000144

van Elk, M., van Schie, H., & Bekkering, H. (2014). Action semantics: A unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge. *Physics of Life Reviews*, *11*(2), 220–250. https://doi.org/10.1016/j.plrev.2013.11.005

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*(3), 829–858. https://doi.org/10.1002/hbm.20547

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*, 564–584. https://doi.org/10.1016/j.neuroimage.2009.06.009

van Rooij, I., Kwisthout, J., Blokpoel, M., Szymanik, J., Wareham, T., & Toni, I. (2011). Intentional communication: computationally easy or difficult? *Frontiers in Human Neuroscience, 5*, 52. https://doi.org/10.3389/fnhum.2011.00052

Vannuscorps, G., & Caramazza, A. (2016). Typical action perception and interpretation without motor simulation. *Proceedings of the National Academy of Sciences*, *113*(1), 86–91. https://doi.org/10.1073/PNAS.1516978112

Vesper, C., & Richardson, M. J. (2014). Strategic communication and behavioral coupling in asymmetric joint action. *Experimental Brain Research*, *232*(9), 2945–2956. https://doi.org/10.1007/s00221-014-3982-1

Vesper, C., Schmitz, L., & Knoblich, G. (2017). Modulating Action Duration to Establish Nonconventional Communication. *Journal of Experimental Psychology: General*, *146*(12), 1722–1737. Retrieved from http://dx.doi.org/10.1037/xge0000379.supp

Vesper, C., van der Wel, R. P. R. D., Knoblich, G., & Sebanz, N. (2011). Making oneself predictable: reduced temporal variability facilitates joint action coordination. *Experimental Brain Research*, *211*(3–4), 517–530. https://doi.org/10.1007/s00221-011-2706-z

Vollmer, A.-L., Lohan, K. S., Fischer, K., Nagai, Y., Pitsch, K., Fritsch, J., … Wredek, B. (2009). People modify their tutoring behavior in robot-directed interaction for action learning. *2009 IEEE 8th International Conference on Development and Learning*, 1–6. https://doi.org/10.1109/DEVLRN.2009.5175516

Volman, I., Noordzij, M. L., & Toni, I. (2012). Sources of variability in human communicative skills. *Frontiers in Human Neuroscience*, *6*, 310. https://doi.org/10.3389/fnhum.2012.00310

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wang, Y., & Hamilton, A. F. de C. (2012). Social top-down response modulation (STORM): a model of the control of mimicry in social interaction. *Frontiers in Human Neuroscience*, *6*, 153. https://doi.org/10.3389/fnhum.2012.00153

Wasenmüller, O., & Stricker, D. (2017). *Comparison of Kinect V1 and V2 Depth Images in Terms of Accuracy and Precision*. https://doi.org/10.1007/978-3-319-54427-4_3

Willems, R. M., de Boer, M., de Ruiter, J. P., Noordzij, M. L., Hagoort, P., & Toni, I. (2010). A Dissociation Between Linguistic and Communicative Abilities in the Human Brain. *Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0956797609355563

Williamson, R. A., & Brand, R. J. (2014). Child-directed action promotes 2-year-olds' imitation. In *Journal of Experimental Child Psychology* (Vol. 118). https://doi.org/10.1016/j.jecp.2013.08.005

Winner, T., Selen, L., Murillo Oosterwijk, A., Verhagen, L., Pieter Medendorp, W., van Rooij, I., & Toni, I. (2019). Recipient Design in Communicative Pointing. *Cognitive Science*, 12733. https://doi.org/10.1111/cogs.12733

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). *ELAN: a Professional Framework for Multimodality Research*. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:60436:2/component/escidoc:60437/LREC 2006_Elan_Wi

Woll, B. (2014). Moving from hand to mouth: echo phonology and the origins of language. *Frontiers in Psychology*, *5*, 662. https://doi.org/10.3389/fpsyg.2014.00662

Woll, B., & Sieratzki, J. S. (1998). Echo phonology: Signs of a link between gesture and speech. In *Behavioral and Brain Sciences* (Vol. 21). https://doi.org/10.1017/s0140525x98481263

Wurm, X. M. F., & Lingnau, A. (2015). Decoding Actions at Different Levels of Abstraction. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(20), 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15.

Zhang, C., & Hansen, J. (2007). Analysis and classification of speech mode: whispered through shouted. *INTERSPEECH*, 1–4. Retrieved from http://crss.utdallas.edu

Zollinger, S. A., & Brumm, H. (2011). The Lombard effect. *Current Biology*, *21*(16),

R614–R615. https://doi.org/10.1016/j.cub.2011.06.003

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

Nederlandse Samenvatting

Mensen zijn sociale dieren. In het dagelijks leven hebben we veel contact met andere mensen waarin we gebruik maken van verschillende manieren van communiceren. Bij het woord communicatie denken we vaak aan taal (gesproken, geschreven, of gebarentaal), maar communicatie is veel meer. Mensen communiceren ook zonder woorden. Stel je voor dat je in een restaurant zit met een aantal vrienden. Als een van je vrienden zijn glas oppakt en in de lucht brengt herken je waarschijnlijk snel of hij dit doet om een toost uit te brengen of om een slok te nemen. Dit komt omdat je zijn intentie herkent voordat de actie compleet is, waardoor je snel en passend kan reageren. Zelfs als deze vriend zijn hand zou opheffen *alsof* hij een toost uitbrengt zou jij dat waarschijnlijk ook begrijpen. Dat mensen dit kunnen is deels wat sociale interactie zo effectief maakt. Hierdoor kunnen we efficiënt communiceren, onze acties met anderen coördineren (bijvoorbeeld je eigen glas opheffen in reactie op een toost), elkaar beïnvloeden en van elkaar leren.

Het menselijk vermogen om intenties te herkennen noemen we 'sociaal signaleren'. Deze term verwijst naar de diverse manieren waarop mensen sociale signalen afgeven die het anderen in staat stelt onze gemoedstoestand (bijvoorbeeld of we geïrriteerd of juist blij zijn) en intenties te herkennen en daar gepast op te reageren. Sommige vormen van sociaal signaleren zijn al wetenschappelijk onderzocht, zoals studies naar lichaamstaal, die bijvoorbeeld verklaren hoe mensen hun lichaam naar iemand toe draaien tijdens een gesprek om een grote mate van betrokkenheid te communiceren. Sommige signalen zijn subtieler, zoals minimale verschillen in de uitvoering van onze handelingen. Zo kunnen mensen een glas oppakken zowel met de intentie om te drinken of om een toost uit te brengen. In het tweede geval wordt het glas gepakt met een duidelijk sociale intentie die wordt gecommuniceerd door de manier waarop ons lichaam en onze ogen bewegen. Iemand anders herkent daardoor onze intentie, voordat de handeling is afgerond, en kan daar gepast op reageren.

In dit proefschrift onderzoek ik hoe mensen deze complexe signalen uitvoeren en begrijpen. Ik bestudeer voornamelijk acties met voorwerpen in relatie tot handbewegingen zonder voorwerpen om intenties beter te begrijpen. Hierdoor heb ik meer inzicht verkregen in hoe onze concrete en sociale doelen onze bewegingen formeren, en hoe beweging samengaat met andere signalen zoals oogbewegingen

om communicatie te ondersteunen.

De invloed van een communicatieve intentie op onze kinematica, oftewel de manier van bewegen, en ook op de herkenbaarheid van deze intentie is tot nu toe vooral onderzocht door te kijken naar vrij simpele bewegingen, bijvoorbeeld naar iets wijzen of een voorwerp oppakken. In **Hoofdstuk 2** breid ik dit onderzoek verder uit naar meer complexe handelingen zoals acties met voorwerpen en iconische handgebaren. Door gebruik van bewegingsopnames, oftewel *motion tracking*, kijk ik naar de rol van communicatieve intentie in het moduleren van de kinematica van acties en handgebaren. Verder test ik of deze bewegingsmodulatie voldoende is om de communicatieve intentie van de actie of het handgebaar herkenbaar te maken voor iemand anders.

In **Hoofdstuk 3** onderzoek ik hoe de dynamiek van het brein zorgt voor de herkenning van een communicatieve intentie op basis van alleen de kinematica van een handbeweging. Functionele *magnetic resonance imaging,* een soort hersenscan, wordt gebruikt om naar de activatie van, en verbindingen tussen, verschillende hersengebieden. Deze hersenscans werden gemaakt terwijl participanten een taak uitvoerden waarin ze moesten beslissen of verschillende handbewegingen uitgevoerd waren met of zonder een communicatieve intentie.

In **Hoofdstuk 4** richt ik me op de semantische kant van bewegingen, oftewel de betekenis daarvan. Daarin kijk ik of het communicatieve moduleren van de kinematica van een beweging ook invloed heeft op het begrijpen van de betekenis van de handbeweging. Eerdere bevindingen geven aan dat de kinematica van het reiken naar een voorwerp ervoor zorgt dat een kijker de volgende actie al kan voorspellen. Kenmerken zoals hoe punctueel de handeling is, dat wil zeggen hoe duidelijk de grenzen zijn tussen individuele bewegingen, en hoe groot een handbeweging is zorgen ervoor dat de betekenis duidelijk te herkennen is. In twee experimenten laat ik specifieke stukken van handbewegingen zien waardoor de hoeveelheid visuele informatie steeds kleiner wordt. Hiermee test ik de specifieke rol en timing van de kinematica in hoe deze het begrip van handbewegingen ondersteunt.

In **Hoofdstuk 5** kijk ik naar hoe mensen handbewegingen, mondbewegingen, en spraak moduleren en coördineren tijdens interacties in een lawaaierige omgeving. Als mensen moeten praten in lawaai overdrijven ze de acoustische (intensiteit en toonhoogte) en visuele (mondbewegingen) delen van hun spraak, wat het

Lombard Effect heet. Luisteraars hebben niet alleen profijt van deze acoustische en visuele modulaties, maar ook van de handbewegingen van de spreker. Het was nog niet bekend of de spreker zijn handbewegingen op een vergelijkbare manier als de spraak moduleert. Bovendien was het niet duidelijk of deze modulatie van handbewegingen gebruikt zou worden als onderdeel van een algemene verhoging van de communicatieve inspanning van de spreker of als onderdeel van een strategische adaptatie van de meest nuttige signalen. In dit hoofdstuk beschrijf ik een experiment waarin wij gebruik hebben gemaakt van een interactieve communicatie taak samen met bewegings-, audio-, en filmopnames om erachter te komen hoe spraak en handbewegingen samenkomen om communicatie te ondersteunen tijdens lawaai.

In **Hoofdstuk 6** richt ik me op de mogelijkheden en implicaties van het gebruik van *motion tracking* (i.e. bewegingsopnames) om de kinematica van betekenisvolle bewegingen zoals acties en handgebaren te onderzoeken. *Motion tracking* is al eerder gebruikt voor het onderzoeken van *motor control*, oftewel het besturen van het lichaam, en ook van een aantal niet kinematische kenmerken van handbewegingen. Echter, het hoge aantal vrijheidsgraden in de analyse maakt het moeilijk om dit soort methoden voor meer complexe, naturalistische bewegingen toe te passen. Om de kinematische kenmerken te kwantificeren die nuttig zijn voor het onderzoeken van betekenisvolle bewegingen heb ik een analytisch kader ontwikkeld en beschrijf hier de implicaties en mogelijkheden voor toekomstige onderzoek.

De experimenten in mijn proefschrift laten zien dat bewegingsmodulatie een belangrijk deel is van communicatie. Het is een manier om onze intenties te signaleren, het maakt de betekenis van een handbeweging duidelijker, en het zorgt ervoor dat deze betekenis ook duidelijk is in een lawaaierige omgeving. Uit deze resultaten zien we dat het belangrijk is om betekenisvolle handbewegingen te onderzoeken op het niveau van kinematica en de interactie met andere communicatieve signalen, zoals spraak, mond- en oogbewegingen. Omdat beweging een grote rol speelt in communicatie, heeft dit onderzoek ook implicaties op klinische groepen, zoals mensen met autisme of de ziekte van Parkinson. Deze groepen kunnen problemen hebben met effectieve communicatie, wat misschien verbonden is met verschillen in hoe ze bewegingen uitvoeren en/of geïnterpreteerd. Als sociale robots een steeds groter aspect van onze samenleving vormen moeten we ook begrijpen hoe de manier van bewegen onze perceptie van robots kan beïnvloeden en hoe robots rekening moeten houden met de manier waarop wij bewegen.

In conclusie, communicatie is niet alleen wat wij zeggen of doen. Het is ook *hoe* wij bewegen, wat onze intenties herkenbaar kan maken en vorm kan geven aan de ideeën die we willen overbrengen.

ACKNOWLEDGMENTS

Over the last 4.5 years my life has been consumed by my PhD. Looking back, I am incredibly proud of and excited about the work I have done and everything I have learned. But none of this would have been possible without the tons of support, whether direct or indirect, from so many people around me. This chapter is dedicated to them, as I try to express how important all of you were to this endeavor.

First of all, I have to thank my direct supervising team. You all took a chance on me, with someone who was completely new to the field. I will always be grateful you gave me this opportunity. **Irina**, from the beginning you were always ready to answer any questions I had, and help in any way you could. Your insights always helped to keep me grounded and focused. From technical discussions about statistics and neuroimaging designs to backing me up any time I needed, I couldn't have asked for a better daily supervisor. **Asli**, thank you for all the support, guidance, and patience. We didn't always see eye-to-eye on everything, but you always made time for in-depth discussions and feedback sessions. I probably wasn't the easiest PhD for you, running back and forth between DCC and MPI, and having a very difficult-to-spell last name ;) but I'm glad you continued to put in the effort. I always appreciate your enthusiasm and ambition, and how much you have supported me in building my network and connecting me with different projects. Thank you for challenging me and helping me to grow as a researcher over these past years. It has been an honour to work with you, and I am happy to not be going too far in the near future! **Harold**, you've always pushed me to see the bigger picture, both in my research and my career. More than anything, I want to thank you for your support and for helping me to develop myself and continue to think ahead. From long discussions on what 'intentions' are, to late-night card games and Donders after-parties, I will really miss working with you!

Thank you to my amazing reading committee, **Ivan Toni, Cristina Becchio, and Sotaro Kita**, for taking the time to read and evaluate my thesis. Whether published work or direct discussions, you have all been very influential on the development of this work, and I am very honored to have you three appraising the finished product. I look forward to discussing all of this with you at the defense!

Odile, Ysbrand, Niels: The first real part of my academic career started with you! Bedankt voor all jullie begeleiding, steun, en voor alles wat ik van jullie heb geleerd. **Chris, Froukje, Stella, Anita, Premika, Femke, Ires, Sonja, Tim, Mardien,** it was a pleasure working with all of you! A special thanks to **Sarah,** for all the fun times, constant encouragement, and lovely dinner parties. **Jos,** my first full experiments and Matlab coding all started with you. Thank you for all your patience and support in this, and for starting me down the (usually very fun) rabbit hole of coding.

All the members, past and present, of SENSe (formerly known as ANC): **Stan, Sybrine, Johan, Egbert, Sari, Birgit, Sylvia, Lukas & Lukas, Peta, Olympia, Irina, Lara, Josh,** as well as everyone from MLC: **Renske, Gerardo, Louise, Marlou, Beyza, Zeynep, Ezgi, Dilay, Vicky, Francie, Tom, Erce,** thank you all for all the feedback, discussions, and general fun times around DCC and MPI. **Gerardo**, muchas gracias por todo! I really enjoyed our Spanish lunches, but most of all I have to thank you for that email you

sent at the beginning of my PhD, when you asked me and Asli if we had thought of using the Kinect for my first experiment. It was a challenging start, but it defined the rest of my PhD work in a way I could not have imagined. To the Research Assistants, **Birgit** and **Renske,** thank you for all of your help and fun conversations. Birgit, thanks also for all the fun you contributed to lab meetings and end-of-year celebrations, it's been great working with you. I'm sure I will continue to see you around for the coming time!

Samantha, you were a fantastic confederate, from all the times being greeted by me (for the first time, every time) in the waiting room, and for the many hours sitting in the experiment room. **Ksenija, Lydia,** thank you for the assistance with data collection and work on my first projects! To my bachelor students, **Selina**, **Eva, Thomas,** it was really a pleasure working with you on the memory project. **Julija,** thanks for all your hard work on our validation study. **Emma, Clarissa, Charlotte**, thank you for all your work on the Lowlands data. **Emma,** it was a pleasure to supervise you on your masters thesis on this project! **Yingdi, Lauren,** thanks for your work on the hand-shape project.

Of course nothing would ever get done without the super helpful admin and technical groups. In particular, **Jolanda, Vanessa, Miriam, and Gerard** van DCC**,** heel erg bedankt voor alles! **Alex,** thank you for always being so helpful and patient! **Johan,** your one-way screen completed the Lowlands experiment set-up, so thank you for this, and everything else you have helped with. **Paul**, thanks for all your help and training at the MR lab! **Ayse**, thanks for your support with participant forms, payments, and all the administration at DCCN. **Julia, Sander**, thank you for answering all my questions relating to Language in Interaction. **Kevin,** for all the hard work you put into the IMPRS. You made it a really great experience to be part of this graduate school. **Dennis,** for the mentoring along the way. We didn't talk often, but it was always good to know that if I needed anything, you were always available. **Dr.Kan,** bedankt voor de prettige samenwerking! Aan **alle participanten** van mijn onderzoeksprojecten, bedankt voor het meedoen en voor alle feedback die jullie hebben gegeven. Jullie hebben dit mogelijk gemaakt!

Antonia, thank you for providing a space for me to work on this thesis and on my autism project. It made a very inspiring final 2 months of my PhD life! **Victoria**, it was great sharing an office with you. Thanks for contributing to a great time in London, and for your feedback while I created my last two figures!

Monja, Tim, Mo, Suzanne, thanks for sharing the Works Council experience with me, and for the continued dinner parties afterwards!

Daniel, I'm really glad you've stuck around after our Masters time in Amsterdam. Thank you for joining me and Hedwig on so many of our adventures to Belgium, for the great discussions about life and science, and for always helping out with the animals! You've been a great friend.

Lukas, from discussions on statistics to the nature of the universe, I really love your

enthusiasm towards knowledge. I don't think many people would be so excited about discussing solutions to overly complex statistical problems, but I'm glad I can come to you for inspiration! Thanks for being my paranymph. I look forward to continuing our discussion/beer tasting sessions!

Linda, these 4.5 years definitely would not have been as fun without you around. Thanks for pulling me into this crazy speech-gesture-noise project that has grown so much bigger than I would have expected when we first started it for Kletskoppen. Crazy weekend science festivals, wine & webinar evenings, and everything else: I'm looking forward to continuing this as a postdoc! (and yes, that : was very intentional) We still need to write that Nature paper after all..

To my **parents,** thank you for always believing in me and supporting me in my dreams. You always said I could do anything I put my mind to, and that I should do whatever makes me happy. Thank you for believing in me, so I could also believe in myself and ultimately end up here. To the rest of my family, thank you for bearing with me while I've been so very out of touch for so long. On that note, I have to give a very special to you, **Pops,** for making it possible to visit back home as often as I have. To my Dutch family (Trujillos, van der Meers, Leijhs), thank you for everything, but in particular for making me feel so at home here in the Netherlands. **Arnout & Joke**, thank you for your very helpful comments on the Nederlandse samenvatting!

Finally, **Hedwig/Heddy.** This whole experience has been so crazy and demanding, I am so grateful to have gone through this together with you, to share the excitement and frustration of every step along the way. Thank you for supporting me on the late nights, the weeks away at conferences, and the not-always-convenient working hours. There's no one I would have rather had by my side for this. And now that I'm done.... You're up next!

About the Author

James Trujillo was born on September 28, 1988, in Anaheim, California, in the U.S. In 2011 he completed a Bachelor of Arts degree in Psychology with a minor in Biology at Northeastern State University, in Oklahoma. In the same year he came to the Netherlands to start the Master of Science program for Neurosciences at Vrije Universiteit in Amsterdam. The program was completed in 2013 after an internship at the Erasmus University in Rotterdam with prof. Jos van der Geest studying the effect of eye-position on saccadic adaptation dynamics, and a second internship at the VU Medical Centre with prof. Ysbrand van der Werf and prof. Odile van den Heuvel studying working memory in Parkinson's disease using fMRI. After this, James stayed at the VU medical centre until 2014 as a research assistant in the same lab, working on data acquisition and analysis for a follow-up fMRI study on executive functioning in Parkinson's disease as well as data analysis on a project investigating emotion regulation in patients with OCD. In 2015, he started a joint PhD position at the Donders Institute for Brain Cognition and Behavior and the Centre for Language Studies, working under prof. Harold Bekkering, prof. Asli Özyürek and dr. Irina Simanova. During this time, he also joined the University Works Council and was involved in various science outreach projects such as language festivals, science festivals, and instructional webinars. He also has been involved in internal collaborations (e.g. Drijvers & Trujillo, 2018) and external collaborations (e.g. Pouw, Trujillo, Dixon, 2019). In the fall of 2019 he joined dr. Judith Holler's CoSI lab to expand his work to the neurocognitive mechanisms and behavioral dynamics of face-to-face social interactions.

Publications

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

Trujillo, J.P., Simanova, I., Özyürek, A., & Bekkering, H. (2019). Seeing the unexpected: How brains read communicative intent through kinematics. *Cerebral Cortex*, in press.

Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A. (2019). The communicative advantage: how kinematic signaling supports semantic comprehension. *Psychological Research*, 1-15.

Trujillo, J.P., Vaitonyte, J., Simanova, I., Özyürek, A. (2019). Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research. *Behavior Research Methods*, 51(2): 769-777.

Pouw, P., **Trujillo., J.P.**, Dixon, J.A. (2019). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*. 1-18.

Drijvers, L., **Trujillo, J.P.** (2018). Commentary: Transcranial magnetic stimulation over left inferior frontal and posterior temporal cortex disrupts gesture-speech integration. *Frontiers in Human Neuroscience*, 12: 256

Trujillo, J.P., Gerrits, N.J.H.M., Veltman, D.J., Berendse, H.W., van der Werf, Y.D., van den Heuvel, O.A. (2015). Reduced neural connectivity but increased task-related activity during working memory in de novo Parkinson patients. Human Brain Mapping, 36(4): 1554-1566.

Trujillo, J.P., Gerrits, N.J.H.M., Vriend, C., Berendse, H.W., van den Heuvel, O.A., van der Werf, Y.D.(2015). Impaired planning in Parkinson's disease is reflected by reduced brain activation and connectivity. Human Brain Mapping, 36(9): 3703-3715.

de Wit, S.J., van der Werf, Y.D., Mataix-Cols, D., **Trujillo, J.P.**, van Oppen, P., van den Heuvel, O.A. (2015). Emotion regulation before and after transcranial magnetic stimulation in obsessive compulsive disorder. *Psychological Medicine*, 45(11): 3059-3073.

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit:

http://www.ru.nl/donders/graduate-school/phd/