



Cognitive Science 45 (2021) e12989

© 2021 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

ISSN: 1551-6709 online

DOI: 10.1111/cogs.12989

Easier Said Than Done? Task Difficulty's Influence on Temporal Alignment, Semantic Similarity, and Complexity Matching Between Gestures and Speech

Lisette De Jonge-Hoekstra,^{a,b}  Ralf F.A. Cox,^a Steffie Van der Steen,^b
James A. Dixon^c

^aDepartment of Developmental Psychology, Behavioural and Social Sciences, University of Groningen

^bDepartment of Orthopedagogy & Clinical Educational Science–Ortho, Education and Learning and Development, Behavioural and Social Sciences, University of Groningen

^cCenter for the Ecological Study of Perception & Action, Department of Psychological Sciences, University of Connecticut

Received 4 March 2021; received in revised form 8 April 2021; accepted 25 April 2021

Abstract

Gestures and speech are clearly synchronized in many ways. However, previous studies have shown that the semantic similarity between gestures and speech breaks down as people approach transitions in understanding. Explanations for these *gesture–speech mismatches*, which focus on gestures and speech expressing different cognitive strategies, have been criticized for disregarding gestures' and speech's integration and synchronization. In the current study, we applied three different perspectives to investigate gesture–speech synchronization in an easy and a difficult task: *temporal alignment*, *semantic similarity*, and *complexity matching*. Participants engaged in a simple cognitive task and were assigned to either an easy or a difficult condition. We automatically measured pointing gestures, and we coded participant's speech, to determine the temporal alignment and semantic similarity between gestures and speech. Multifractal detrended fluctuation analysis was used to determine the extent of complexity matching between gestures and speech. We found that task difficulty indeed influenced gesture–speech synchronization in all three domains. We thereby extended the phenomenon of gesture–speech mismatches to difficult tasks in general. Furthermore, we investigated how temporal alignment, semantic

Correspondence should be sent to Lisette De Jonge-Hoekstra, Department of Developmental Psychology, Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, Groningen, TS 9712, The Netherlands. E-mail: lisette.hoekstra@rug.nl

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

similarity, and complexity matching were related in each condition, and how they predicted participants' task performance. Our study illustrates how combining multiple perspectives, originating from different research areas (i.e., coordination dynamics, complexity science, cognitive psychology), provides novel understanding about cognitive concepts in general and about gesture–speech synchronization and task difficulty in particular.

Keywords: Gestures; Speech; Synchronization; Gesture–speech mismatches; Complexity matching; Multifractal detrended fluctuation analysis

1. Introduction

Gestures and speech are two salient aspects of multimodal communication in humans. When people tell a story, explain a difficult problem, or talk about daily affairs, they tend to move their hands in all kinds of ways. Many researchers have therefore proposed that gestures and speech are tightly coupled (e.g., Goldin-Meadow, 2003; McNeill, 1985). Moreover, this tight coupling has been conceptualized as gesture–speech synchronization (e.g., Iverson & Thelen, 1999; Pouw & Dixon, 2019b; Treffner & Peter, 2002). Gestures and speech synchronize in time, semantic content, emphasis, and emotional valence (for a comprehensive review, see Wagner, Malisz, & Kopp, 2014).

However, the semantic similarity between gesture and speech has been shown to break down as people approach transitions in understanding (e.g., an insight into a difficult problem; Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003). For instance, in a liquid conservation task a researcher pours equal amount of water into a wide glass and a narrow glass and asks a child which glass contains more water. When a child is about to learn the concept of conservation, they might say that there is more water in the narrow glass because the level of water is higher, while they gesture about the width of the glasses (Church & Goldin-Meadow, 1986). These instances of semantic dissimilarity are called *gesture–speech mismatches*.

Different explanations exist for the breakdown in the semantic similarity between gesture and speech when people approach transitions in understanding. Goldin-Meadow and colleagues' (Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003) explanations center around participants' conflicting cognitive strategies and hypotheses that are thought to exist just before participants achieve new insight into the problem they are working on (e.g., liquid conversation task). These conflicting strategies and hypotheses are then somehow differently expressed in gestures than in speech, during gesture–speech mismatches. However, Koschmann (2017) questions the existence of gesture–speech mismatches in the first place and suggests that they are an artifact of the disintegrated methodological coding systems that led to their discovery. Furthermore, Pouw, van Gog, Zwaan, and Paas (2017; also see Pouw, de Nooijer, van Gog, Zwaan, & Paas, 2014) highlight an explanatory gap in how an integrated gesture–speech system could produce disintegrated gesture–speech mismatches and suggest taking a dynamically embodied perspective to address this gap.

From a dynamically embodied, complex system's perspective, a change in understanding can be seen as a system of interrelated components that transitions from one stable state to a new, likely more advanced, stable state (Smith & Thelen, 2003; Stephen, Boncoddio, Magnuson, & Dixon, 2009; Stephen, Dixon, & Isenhower, 2009; Thelen & Smith, 1994, 2007; Van

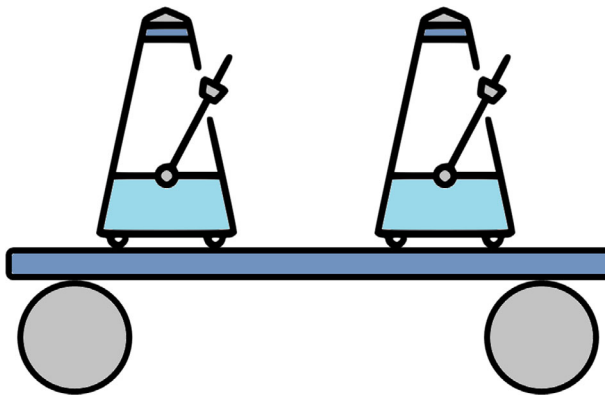


Fig. 1. Synchronization of coupled metronomes.

Geert, 2008, 2011). A transition from one stable state to another entails a reorganization of a system's components and their relations. This reorganization is elicited by a perturbation, that is, the learning situation. As put forward by De Jonge-Hoekstra et al. (2020), a metaphor for this reorganization is building a new LEGO-structure from an old structure, which is only possible when you break (perturb) the old structure and use the bricks to create a new structure. Taking such a dynamically embodied, complex system's perspective, De Jonge-Hoekstra, Van der Steen, Van Geert, and Cox (2016) suggest that difficult tasks perturb a system, thereby inducing a suboptimal coordination between gestures and speech, which could then lead to various forms of gesture–speech mismatches.

In this study, we empirically address whether task difficulty indeed affects gesture–speech synchronization. We will approach gesture–speech synchronization in three ways: (a) temporal alignment, (b) semantic similarity, and (c) complexity matching (explanation follows below). We will investigate how task difficulty affects temporal alignment, semantic similarity, and complexity matching between gestures and speech and how these different forms of gesture–speech synchronization are related. In addition, we will investigate how these three gesture–speech synchronization measures predict task performance.

1.1. Synchronization

Synchronization usually means that two (or more) systems start to behave in a similar way due to coupling (Pikovsky, Kurths, & Rosenblum, 2001). In cognitive science, synchronization comes in different forms, including temporal alignment, semantic similarity, and complexity matching. We will explain these three forms below, and describe how they have been linked to gesture–speech synchronization.

1.1.1. Temporal alignment

Temporal alignment is a well-known form of synchronization. A simple and widely used example of temporal alignment is two asynchronously ticking metronomes, which start to tick in synchrony when they are placed on a shelf on top of two cans that act like wheels (the movement of each metronome is transmitted through the wheels thus providing coupling; see

Fig. 1). Also within humans, body parts such as fingers (e.g., Haken, Kelso, & Bunz, 1985; Kelso, 1994) and legs (Clark, Whittall, & Phillips, 1988) have been shown to synchronize and temporally align in rhythmic patterns. Moreover, a recent study by Pouw, De Jonge-Hoekstra, and Dixon (2018) shows that speech is more rhythmic when it goes together with more gestures, suggesting a rhythmic synchronization between gestures and speech within humans. This paradigm of one-to-one temporal alignment of behavior has been applied to coordination between humans, where it has been found that humans tend to move in synchrony while rocking in rocking chairs (Richardson, Marsh, Isenhower, Goodman, & Schmidt, 2007), swinging pendulums (Richardson, Marsh, & Schmidt, 2005; Schmidt & O'Brien, 1997), or telling jokes (Schmidt, Nie, Franco, & Richardson, 2014), to name a few examples.

With regard to temporal alignment between gestures and speech, adult's gestures and speech are highly aligned in time (see Wagner et al., 2014, for an overview). In other words, most gestures beat in-phase and at the same rhythm as speech (Prieto & Roseano, 2018; also see Pouw et al., 2018). For gestures, this rhythm consists of changes in hand-movement velocity over time, while for speech, this rhythm refers to the organization and contrast of a sequence of repeated speech events and can be tracked acoustically through the amplitude envelope of speech (also see Fowler, 2010). To support the existence of temporal alignment between gestures and speech, several studies indicated that the moment of *maximum effort* in gestures goes together with changes in pitch (i.e., relative frequency, "highness" or "lowness") of speech (Kendon, 1972; Kita, Van Gijn, & Van der Hulst, 1998; Leonard & Cummins, 2011). Recent studies by Pouw and colleagues (Pouw & Dixon, 2019a, 2019b; Pouw et al., 2018) showed that this relation between maximum gestural effort and speech is actually a tight alignment of peak velocity in gestures and peak pitch in speech.

Some circumstances affect the temporal alignment between gestures and speech. Children's age is a robust correlate with the temporal alignment between gestures and speech. According to Iverson and Thelen (1999), the coupling between gestures and speech in infants emerges from natural oscillations of hand movements and vocal acts, which synchronize and become entrained over time (see also Esteve-Gibert & Prieto, 2014; Iverson & Fagan, 2004). As a consequence of this entrainment, the temporal alignment between gestures and speech becomes higher when infants and toddlers grow older (Butcher & Goldin-Meadow, 2000; also see Iverson & Thelen, 1999). Adults' gestures and speech are so tightly coupled in time that perturbing and delaying speech with a delayed auditory feedback also delays gestures (e.g., Rusiewicz, Shaiman, Iverson, & Szuminsky, 2013, 2014). Pouw and Dixon (2019) found that a delayed auditory feedback actually increases the temporal alignment between gestures and speech. Last, Bergmann, Aksu, and Kopp (2011) found that gestures and speech were more temporally aligned when their semantic content was more similar.

1.1.2. Semantic similarity

Semantic similarity refers to similarities in *meaning*. Humans can synchronize on a semantic level, whereby they align their "[...] understanding of the world with others [...]" (Dumas & Fairhurst, 2019, p. 10). Important to note is that semantic synchronization is not confined to (spoken) language but can take other action forms involving other body movements as well (Dumas & Fairhurst, 2019). Bodily forms of semantic, meaningful synchronization, such as playing give-and-take-games, or interpersonal movement coordination when a parent dresses

their child, seem to be essential for language development. Furthermore, differences in the semantic similarity of two people's words influence their bodily synchronization (see Shockley, Richardson, & Dale, 2009, for an overview).

Gestures and speech are considered to be semantically similar when a gesture is temporally aligned with a word or phrase, and both gesture and word/phrase convey the same meaning (cf. Wagner et al., 2014). Based on this definition, a distinction has been made between gestures that convey either *redundant*, *complementary*, *non-redundant*, or *mismatching*¹ semantic content to speech. Most of our gestures are either redundant (e.g., saying "The shelf is long" while gesturing that something is long) or complementary (e.g., saying "The shelf is [this] long" while gesturing the length of the shelf) to speech. Studies with participants from different languages show that the typical structure and semantic content of a language influence the semantic content of gestures (Allen et al., 2007; Kita & Özyürek, 2003), highlighting the usually strong semantic similarity between gesture and speech.

However, sometimes the semantic content of gestures and speech does not overlap and is thus non-redundant in general (Goldin-Meadow, Nusbaum, Garber, Church, 1993). Examples of non-redundant semantic content are a child who points to a cup while saying that they are thirsty, or a teacher who explains two strategies for a problem at the same time: one in speech and the other in gestures. In these examples, the semantic content of gestures and speech does not overlap, but their meaning is related and falls within an overarching theme ("drinking" and "problem solutions"). Mismatches between gestures and speech are a specific kind of non-redundant semantic content. As previously described, mismatches are known to occur when a child (or adult) learns a new strategy for a difficult cognitive problem (e.g., Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003; Goldin-Meadow, Alibali & Church, 1993). Similar to non-redundant semantic content, the meaning of gestures and speech during mismatches does not overlap but is related.

1.1.3. Complexity matching

Notwithstanding the impact and relevance of the synchronization examples above, involving temporal alignment and semantic similarity, complex systems in the real world often do not synchronize as one-to-one matching of behavior (Delignières, Almurad, & Roume, 2016). Complex systems, such as gestures and speech, can synchronize on many (time) scales of organization, which is called *complexity matching* (Stephen, Stepp, Dixon, & Turvey, 2008; West, Geneston, & Grigolini., 2008; see also Abney, 2016; Abney et al., 2014; Den Hartigh, Marmelat, & Cox, 2018). During complexity matching, the information exchange between complex systems is maximized (West et al., 2008). Complexity matching occurs when both systems are complex, and the degree of the two systems' complexity is similar.

1.1.3.1. Gestures and speech as complex systems. Gestures and speech are complex systems. They consist of many different and interacting components and scales and involve coordination of all these different components and scales of a system over time (e.g., Van Orden, Holden, & Turvey, 2003). Gestures' and speech's scales range from action potentials of neurons to overarching conversational goals and beyond (see also De Jonge-Hoekstra et al., 2016). For example, numerous muscles and bones in a person's arms, chest (and even legs),

lungs, and central nervous system are involved in each gesture. Importantly, speaking also involves a large number of components; it is estimated that we use more than 70 muscles for each syllable that we utter (e.g., Turvey, 2007).

Infants clearly show how complex gesturing and speaking actually is. Before the first pointing gestures emerge, infants have learned to control their eye movements to focus on an object (Adolph & Franchak, 2017), to use their hands to grasp things, and have learned about distances by crawling forward (Clearfield, 2004). All these actions and perceptions, which are great coordinative accomplishments in themselves, come together in their first pointing gestures. When infant's first words emerge, infants partly "build" on what they had accomplished for their first gestures (Esteve-Gibert & Prieto, 2014; Goldin-Meadow, 2007). However, uttering their first word involves another set of challenges too. Coordinating all different components to pronounce a specific syllable is a complex task as well, and it usually takes an infant many tries before they grasp the correct configuration. This process is nicely illustrated by Roy (Roy, Frank, DeCamp, Miller, & Roy, 2015; also see Roy, 2011), who showed how his son went from saying "gaaaa" to the word "water," over numerous trials, in about 6 months' time.

1.1.3.2. Complexity and fractal scaling. A complex system's coordination over time (e.g., Van Orden et al., 2003) can be more or less fluent. When the coordination of components and layers of a system over time is fluent, the changes of behavior at all different scales are related (e.g., Wijnants, 2014). In other words, variability across time scales is related and dependent, which means that changes on smaller time scales (e.g., neuronal level) influence changes on larger time scales (e.g., conversational goals) and vice versa. If one would plot that system's behavior over time (e.g., the time between word onsets during an affective conversation), one would see that small changes in the time series (visible as small *waves*) are nested within larger changes (larger *waves*; e.g., see Fig. 2a). Furthermore, if one would zoom in or out, the plotted time series would look similar at different levels of magnification. In other words, the variability at the level of milliseconds looks like the variability at the level of seconds, which looks like the variability at the level of minutes, and so forth. Objects that show such self-similarity, such as the Koch snowflake (Fig. 2c) or Romanesco broccoli (Fig. 2d) are also called *fractal* objects. Similarly, a nested and self-similar² structure of variability in the temporal domain is called (*mono*)*fractal* or *pink noise* (see Fig. 2a). Monofractal variability has been proposed as an index of optimal balance between rigid and random behavior and is often found in complex systems that change over time (Van Orden et al., 2003; Van Orden, Kloos, & Wallot, 2011; Wijnants, 2014). Indeed, many studies found that expert performance on repetitive motor tasks is more *pink* than non-expert behavior (e.g., Den Hartigh, Cox, & Gernigon, 2015; Kloos & Van Orden, 2010; Van Orden et al., 2011). Monofractal variability has thus been considered as an identifying feature of complex systems, corresponding to a systems' degree of complexity.

However, different from relatively repetitive motor tasks, more diverse human behavior shows sudden jumps and periods of relative stability mixed with intermittent bursts of variability (Dixon, Holden, Mirman, & Stephen, 2012; Ihlen & Vereijken, 2010; Kelty-Stephen, Palatinus, Saltzman, & Dixon, 2013; Stephen, Anastas, & Dixon, 2012). Moreover, these

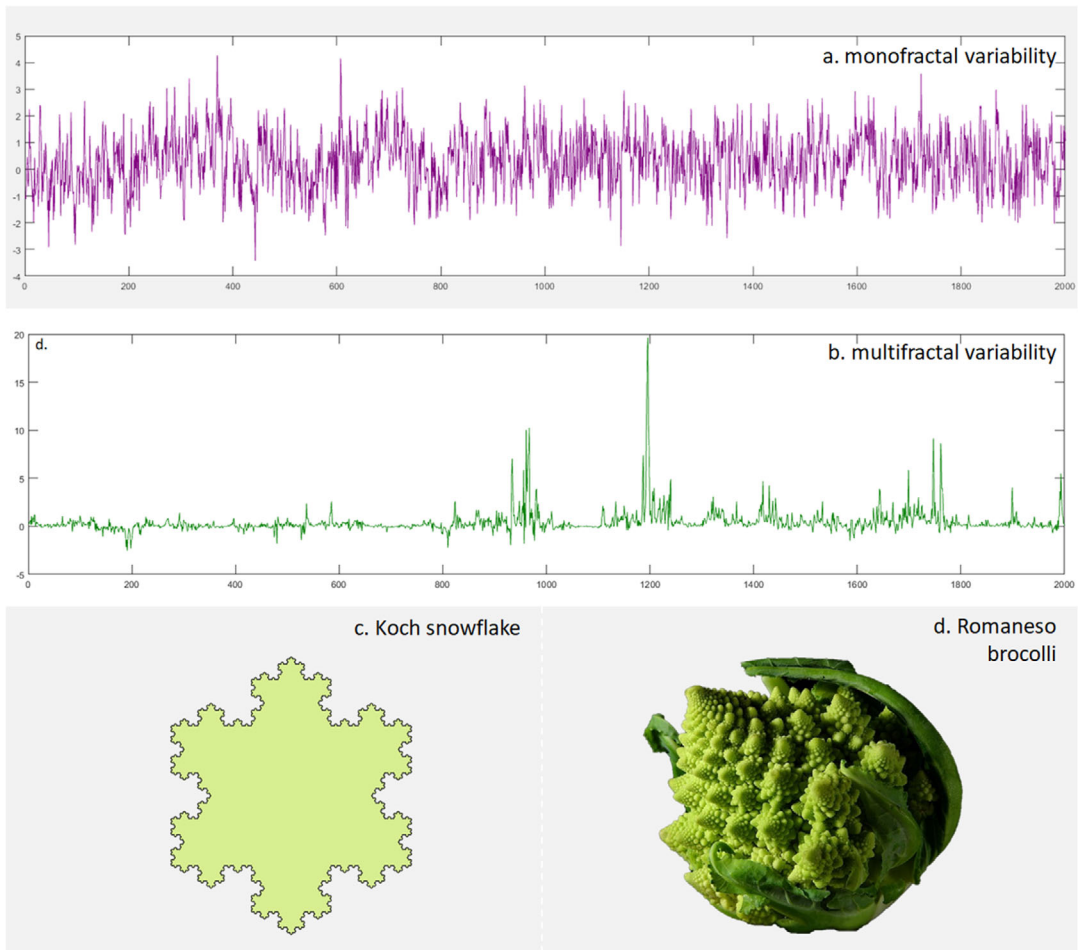


Fig. 2. Examples of fractal structures. Panel A shows a time series with a monofractal structure of variability (source: script in <https://doi.org/10.3389/fphys.2012.00141>). Panel B shows a time series with a multifractal structure of variability, whereby periods of monofractal variability are intermitted by periods of large fluctuations and periods of small fluctuations (source: script in <https://doi.org/10.3389/fphys.2012.00141>). Panel C displays the Koch snowflake (seventh iteration; source: bit.ly/2PGeRad). Panel D displays Romanesco broccoli (source: bit.ly/2wiEccN). The monofractal structures in panels A, C, and D are self-similar, which means that they look the same at different levels of magnification. The multifractal structure in panel D is less self-similar.

increases in variability have been related to transitions, which are a hallmark of human (and other complex systems') development. Examples of a sudden jump, which would go along with a burst in variability, are an abrupt change in conversation goals, or the "aha"-moment of acquiring new understanding (Dixon et al., 2012). Delignières et al. (2016), Dixon et al. (2012), Ihlen and Vereijken (2010), Kelty-Stephen et al. (2013), and Stephen et al. (2012) argue that timescales *themselves* also interact and that these interactions between timescales lead to these large changes in variability (for a clear and more in-depth explanation, please

see Kelty-Stephen et al., 2013). When variability with a monofractal (pink noise) structure is mixed with periods of changes in variability, these time series display a *multifractal* structure (see Fig. 2b). Therefore, identifying complex systems and establishing a system's degree of complexity should also incorporate multifractal variability (Delignières et al., 2016; Dixon et al., 2012; Ihlen & Vereijken, 2010; Kelty-Stephen et al., 2013; Stephen et al., 2012).

1.1.3.3. When does complexity matching occur? As previously described, complexity matching means that the degree of the system's complexity is similar due to coupling. In other words, when coupled systems match their complexity, the fractal structure of their temporal variability is alike.

Circumstances influence complexity matching. Abney et al. (2014) found that type of conversation influences whether complexity matching between two participants in speech occurs. Specifically, when participants discussed things that they had in common, the fractal scaling of participants' acoustic onset events was similar, and their speech thus showed complexity matching. However, no complexity matching was found when participants discussed issues on which they had different opinions. Furthermore, Almurad et al. (2017) investigated complexity matching between participants who were instructed to walk in synchrony. Participants walked either side-by-side or arm-in-arm (or independently), and the researchers measured the duration of the intervals between their strides. Participants in both (non-independent) conditions showed high levels of complexity matching, whereby arm-in-arm walking led to slightly higher levels of complexity matching than walking side-by-side. With regard to manual coordination between participants in terms of intervals between finger taps, fractal hand movements, and a larger magnitude of hand movement's variation, leads to stronger complexity matching between a leader and a follower than random hand movements (Coey et al., 2016). In addition, complexity matching is stronger when participants coordinate movements of both their hands than when they coordinate the movements of one of their hands to those of a partner (Coey et al., 2018). Most of these studies show that stronger coupling between systems goes together with higher levels of complexity matching (Cox, 2016).

Interestingly, research findings are mixed about whether complexity matching is functional in terms of task performance: While Fusaroli, Abney, Bahrami, Kello, and Tylén (2013) and Abney et al. (2014) found better task performance with higher levels of complexity matching, Schloesser, Kello, and Marmelat (2019) and Abney et al. (2015) found an inverse relation. With regard to gestures and speech, De Jonge-Hoekstra et al. (2016) suggest that difficult tasks may influence whether and how gestures and speech synchronize on multiple scales. This would imply that difficult tasks influence complexity matching between gestures and speech.

1.2. Current study

In this study, we investigated how a difference in task difficulty influences the synchronization between participant's gestures and speech, in terms of temporal alignment, semantic similarity, and complexity matching. We asked participants to repeatedly match targets of the same colors presented on a tablet with a touch screen, by means of pointing to these targets

and saying their location. Participants were assigned to either a predictable easy condition or to an unpredictable difficult condition.

Our first research question is: How does task difficulty influence temporal alignment, semantic similarity, and complexity matching between participant's gestures and speech? With regards to *temporal alignment*, Pouw and Dixon (2019b) found that gestures and speech became more synchronized in the more difficult delayed auditory feedback condition. We, therefore, expected that gestures and speech would be more synchronized in the difficult than in the easy condition (hypothesis 1A). Regarding *semantic similarity*, Goldin-Meadow and colleagues (e.g., Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003; Goldin-Meadow, Alibali et al., 1993) found that gestures and speech mismatch in semantic content when people are about to understand a task which they do not understand yet, and thus the task is difficult for them. We, therefore, expected less semantic similarity between gestures and speech in the difficult than in the easy condition (hypothesis 1B). With respect to *complexity matching*, there are no studies that directly investigated how task difficulty influences complexity matching. As described above, we do know that the level of complexity matching increases when the coupling between systems is stronger (Abney et al., 2014; Almurad, Roume, & Delignières, 2017; Coey et al., 2016, 2018). Our previously stated hypothesis 1A suggests that gestures and speech become more temporally aligned in the difficult condition, and thus a *stronger* coupling. However, our previously stated hypothesis 1B suggests that gestures and speech become less semantically similar in the difficult condition, and thus a *weaker* coupling. Because of this contradiction, we have no specific hypothesis for the influence of task difficulty on the level of complexity matching between gestures and speech.

Our second research question is: How are temporal alignment, semantic similarity, and complexity matching between gestures and speech related in the easy and difficult conditions? Bergmann et al. (2011) found that gestures and speech were more synchronized in time when their semantic content was more similar. This suggests that a higher temporal alignment between gestures and speech would go together with a higher semantic similarity. On the other hand, hypotheses 1A and 1B suggest a higher temporal alignment and a lower semantic similarity in the difficult condition. We, therefore, expected a positive relation between gestures' and speech's temporal alignment and semantic similarity in the easy condition (hypothesis 2A), and a negative relation between temporal alignment and semantic similarity in the difficult condition (hypothesis 2B). In line with a higher level of complexity matching when the coupling is stronger (Abney et al., 2014; Almurad et al., 2017; Coey et al., 2016, 2018) and in line with hypothesis 2A (positive relation between temporal alignment and semantic similarity in easy condition), for the easy condition, we expected a positive relation between gestures' and speech's temporal alignment, semantic similarity, and complexity matching as well (hypothesis 2C). Our expected negative relation between temporal alignment and semantic similarity (hypothesis 2B) in the difficult condition suggests an inverse relation in coupling strength. Therefore, we have no specific hypotheses about how complexity matching is related to either temporal alignment or semantic similarity in the difficult condition.

Our third research question is: How do temporal alignment, semantic similarity, and complexity matching between gestures and speech predict task performance? We assessed task performance in terms of the *time needed to finish the task*. Our experimental manipulation

of task difficulty will influence task performance, as difficult tasks typically take longer to perform. Therefore, we controlled for the influence of condition (task difficulty) when we investigated whether temporal alignment, semantic similarity, and complexity matching between gestures and speech predict task performance. According to Iverson and Thelen (1999; also see Butcher & Goldin-Meadow, 2000; Esteve-Gibert & Guellai, 2018) the temporal alignment between gestures and speech becomes higher when infants and toddlers grow older. As children's language skills change and become more advanced during that time too (e.g., Tamis-LeMonda et al., 1998, 2001), this could imply that more temporal alignment goes together with a better language performance. Mismatches are a form of semantic dissimilarity and predict better performance on *subsequent* tasks (e.g., Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003; Goldin-Meadow, Alibali et al., 1993). Findings of a link between complexity matching and task performance are mixed, whereby some studies found a positive relation (Abney, 2016; Abney et al., 2021; Fusaroli et al., 2013) while others found a negative relation (Schloesser et al., 2019). Taken together, these findings are not sufficiently conclusive to formulate hypotheses about how temporal alignment, semantic similarity, and complexity matching predict task performance.

2. Method

2.1. Participants

We included³ 30 participants (20 F, 10 M) between 18 and 27 years ($M = 20.70$, $SD = 2.39$) in our study. All participants were students with a Dutch nationality at a University in the Netherlands, who participated in the experiment in exchange for course credit or monetary compensation. The participants provided written consent. The ethical committee of the Psychology Department of the University of Groningen approved of the study.

2.2. Materials

Participants performed the task on a tablet (Lenovo MIIX 320-10ICR 1.44GHz x5-Z8350) with a 10.1 inch touchscreen (1280 × 800 pixels) and Windows 10 operating system. To facilitate pointing, the tablet was positioned in a 45° angle from the table using a tablet stand (see Fig. 3). The experiment was programmed using OpenSesame [version 3.0.0] (Mathôt, Schreij, & Theeuwes, 2012), which is an open-source program to build (social science) experiments. Using OpenSesame, we could run the task at the tablet (a detailed description follows below), and simultaneously record the time and x - and y -coordinates of participants' pointing (touching) at the screen as well as participants' speech signal.

Participants' speech was recorded at 44.1 kHz using a basic hands-free microphone that was plugged into the 3.5 mm audio jack of the tablet. We used Audacity [version 2.2.2] to normalize the volume of the speech signal and filter out background noise. Furthermore, we used Praat (Boersma & Weenink, 2018) [version 6.0.42] and RStudio [version 1.1.456] to calculate the *amplitude envelope* of the speech signal (He & Dellwo, 2016; Pouw & Trujillo, 2019, respectively; a detailed description follows below). The amplitude envelope that is cal-



Fig. 3. Set-up of the experiment.

culated by the R-script is identical to the amplitude envelope that is calculated by the Praat script (Pouw & Trujillo, 2019). We used a custom script in Matlab [version 2018a] to identify the start of syllables in the speech signal and to cut the audio recordings into smaller parts of one syllable each (a detailed description follows below). We used OpenSesame [version 3.0.0] (Mathôt et al., 2012) to manually code the semantic content of these syllables.

We used Matlab to carry out the analyses on the time series of pointing and the amplitude envelope of speech. We specifically used the multifractal detrended fluctuation analysis (MFDFA) package by Ihlen (2012) to perform MFDFA, to estimate the *temporal multifractality* of participant's gestures and speech. Furthermore, we used RStudio to carry out inferential statistics and the R-package ggplot2 (Wickham, 2016) to create plots of our data.

2.3. Procedure

Participants performed a tablet task (see Figs. 3 and 4), which can be found here: osf.io/dj5vr/ (Scripts & Materials > Tablet task). We instructed the participants to repeatedly (virtually) put a ring on a bar of the same color, by first pointing (touching) to the ring on the screen and thereafter to the top of the corresponding bar. Furthermore, each time that a participant pointed, we instructed them to say out loud the location of the ring and bar (left, middle, right) that they were pointing to in Dutch (“links,” “midden,” “rechts,” respectively). In addition, we instructed participants to perform the task as fast and accurately as possible (in accordance with Fitts, 1954). We randomly assigned the participants to either the easy ($n = 14$; see Fig., left panel) or the difficult condition ($n = 16$; see Fig. 4, right panel). In the easy condition, the color of the ring always corresponded to the color of the above bar (see Fig. 4, left panel). In the difficult condition, however, the color of the rings was random (see Fig. 4, right panel). Participants were not informed about the pattern being either random or non-random. Since it is impossible to understand a random pattern, participants in the difficult condition were constantly needed to reorganize to the new spatial arrangement. This state of reorganization shares similarities with the state of reorganization that precedes

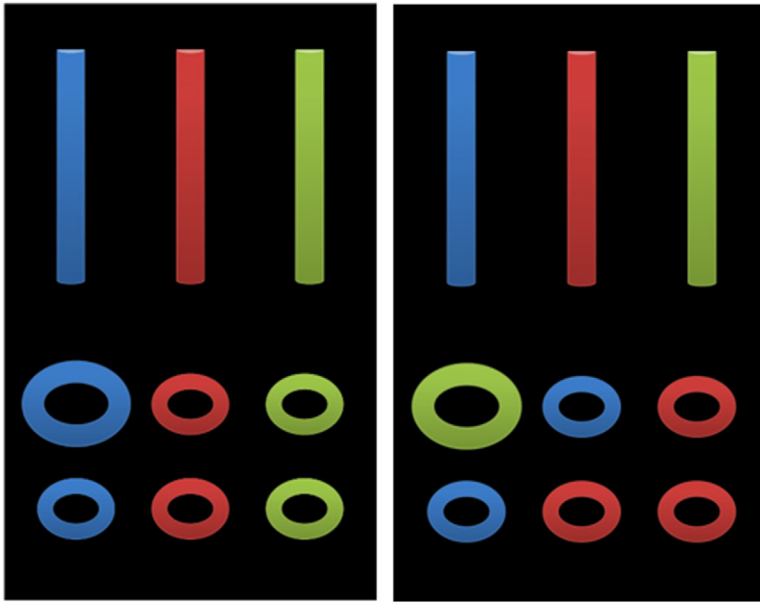


Fig. 4. Example of tablet task. The left panel displays the easy task, and the right panel displays the difficult task.

learning something new (see Kello, Beltz, Holden, & Van Orden, 2007; Stephen, Dixon, et al., 2009).

To register participants' pointing, we divided the screen into $3 \times 6 = 18$ (invisible) areas. Each top of the bar was positioned in an area at the top row of the screen, while each ring was positioned in an area at the second row from the bottom of the screen. The correct ring, that is, the ring that participants needed to point to during that trial, appeared larger on the screen as shown in Fig. 4 (upper left ring in both panels). Please note that the participants did not have to point to the correct ring or bar for the task to proceed. However, if participants failed to click on a *ring-area* or a *top of the bar-area*, the task did not proceed and the time and location of every first error were recorded.

During the task, the order in which the rings were presented alternated between left to right and right to left. For example, the correct order of the task in the left panel of Fig. 4 would be: [first row] left–left–middle–middle–right–right–[second row] right–right–middle–middle–left–left. The correct order of the task in the right panel of Fig. 4 would be: [first row] left–right–middle–left–right–middle–[second row] right–middle–middle–middle–left–left. Each time a participant finished with the last ring of a row, that row disappeared from the screen, the second row moved up, and a new row appeared at the bottom of the screen. The participants performed 540 repetitions of the task, which is identical to 180 rows of three rings and corresponding bars, or a total of 1080 times pointing and saying the location of either a ring or a bar. Before starting with the actual task, the participants completed a trial phase with 15 repetitions of the task, to get used to the set-up. The recordings of this trial phase were not included in the analysis.

2.4. Data preparation

To investigate the coupling between participants' gestures and speech, we recorded the time (ms), location (left/middle/right), and position (x - and y -coordinates) of their pointing, and their speech signal.

2.4.1. Gestures

For gestures, the above resulted in a time series⁴ of (a) the duration between pointing to rings and bars and vice versa, (b) a time series of the location of the pointing, and (c) a time series of distances between the exact locations that participants pointed to. With regards to distances between rings and bars, there are three possible distances⁵ that participants' fingers needed to travel while pointing: (a) a short distance of 608 pixels, when the ring and bar are vertically aligned; (b) a middle distance of 664 pixels, when the ring and bar are one location off (i.e., from the left ring to the middle bar) or (c) a long distance of 809 pixels, when the ring and bar are two locations off (i.e., from the left ring to the right bar). This third, long distance can only occur in the difficult condition, and therefore the frequency distribution of distances between targets differs between the two conditions.

From the work by Fitts (1954), we know that the distance (D) between targets, combined with the width (W) of targets (ring: 167 pixels; bar: 61 pixels), influences how difficult the movement between two targets (i.e., from ring to bar or vice versa) is to perform. Fitts referred to this as the index of difficulty (ID), which is given by the following formula: $ID = \log_2(\frac{2D}{W})$. Using this formula, from ring to bar, the ID for the short, middle, and long distance is 4.317, 4.444, and 4.729, respectively. From bar to ring, the ID for the short, middle, and long distance is 2.864, 2.991, and 3.276, respectively. In the current study, we aim to manipulate task difficulty by changing the overall task demand of matching targets of the same color when one of the targets' color was either random (difficult) or non-random (easy). However, any difference in movement time (MT) could potentially result from the difference in ID between targets. To remove this possible confound, and standardize this influence of the ID on each duration in our MT series, we divided each duration between pointing to two targets (MT) with the ID of that particular movement. These corrected durations between pointing to two targets corrected with the ID of each movement yielded a time series of MT/ID.

2.4.2. Speech

We recorded the participant's speech from the moment the first experimental trial was presented until the moment the participant finished with the last experimental trial. This yielded one long sound recording of what the participant said during the task. To increase the quality of the sound recording, we used Audacity to normalize the sound volume and filter out background noise. We subsequently used PRAAT (He & Dellwo, 2016) or R (Pouw & Trujillo, 2019) to calculate the *amplitude envelope* of the speech signal. The amplitude envelope basically is a smoothed outline of a speech signal's intensity (He & Dellwo, 2016), and its structure corresponds to the lower lip kinematics (He & Dellwo, 2017). In addition, we calculated the velocity of the speech signal's amplitude envelope, which captures how the amplitude envelope increases and decreases.

We identified the start of syllables by extracting the peaks in velocity of the amplitude envelope, using a custom MATLAB script (osf.io/dj5vr/; Scripts & Materials > Data preparation), and saved the audio between two velocity peaks as audio segments (i.e., one syllable per audio segment). The Dutch word “links” has one syllable, “midden” has two syllables, and “rechts” has one syllable. Due to individual differences in speaking, extracting one word or syllable per audio segment did not work perfectly for each participant, however.⁶ To ensure that MATLAB was not too sensitive (so as to cut one syllable into multiple audio segments) yet sensitive enough (so as to aggregate a maximum of five words into one audio segment), we manually tweaked a sensitivity parameter in the script (osf.io/dj5vr/; Scripts & Materials > Data preparation) for each audio recording. We subsequently coded the semantic content of the audio segments to identify the starting times of actual words.

We coded the semantic content of the audio segments using OpenSesame (osf.io/dj5vr/; Scripts & Materials > Data preparation). We loaded the audio segments into OpenSesame and coded whether a segment was (a) [the first half of] “links,” (b) [the first half of] “midden,” (c) [the first half of] “rechts,” (d) the second half of a word, (e) a sequence of multiple words, or (f) something else (i.e., other speech, a sigh). If a segment was (e) a sequence of multiple words, we coded the semantic content of the sequence of words in that segment. This coding of audio segments yielded a time series of words (segments) and their starting time. For (e) sequences of multiple words, we used the number of words in an audio segment to extract the same amount of velocity peaks of the amplitude envelope in that particular audio segment. We replaced the word sequences in the time series with the individual words and their velocity peaks. We removed the (f) other speech/sighs from the time series.

2.4.3. Combining gestures and speech

To investigate the temporal alignment and semantic similarity between gestures and speech, we aligned the time series of gestures and speech by linking the gestures to the word that was closest in time. To find the correct delay for each participant, we aligned the time series of gestures and speech for every delay between 10 and 1,000 ms, with steps of 10 ms, and calculated the amount of semantic content differences and the average asynchrony between gestures and speech (for an overview, osf.io/dj5vr/; Data). Since the amount of semantic content differences for each participant went down to a minimum and then went up again, we decided that the delay with the least amount of semantic content differences was the correct delay. If there were more delays with the least amount of semantic content differences, we picked the delay with the lowest average asynchrony between gestures and speech. The data files with the maximally aligned gestures and speech can be found here: osf.io/dj5vr/; Data > For analyses.

We calculated the difference between *amplitude* peaks (not *velocity* peaks) in the aligned time series to create a duration-time series for speech, and we used this time series to analyze the temporal alignment between gestures and speech. The amplitude peak of the amplitude envelope corresponds to the stressed syllable in a word (see Fig. 5). In each of the three words that the participants said, the first syllable of the word is stressed (links, midden, rechts). The amplitude peak, therefore, yielded a similar time point for each of the three words. Furthermore, to analyze semantic similarity, we used the semantic content time series of speech. We

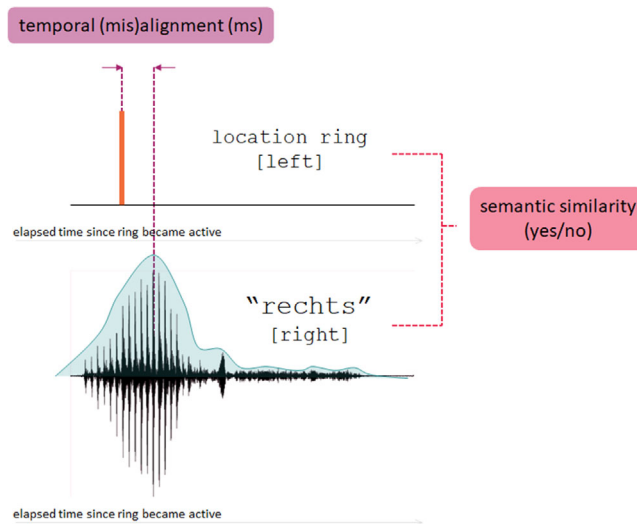


Fig. 5. Illustration of how we calculated temporal alignment and semantic similarity within a trial. The orange vertical line indicates the moment the participant's finger touched the screen when the participant pointed at the ring. The peak of the blue curve corresponds to the amplitude peak of the word that the participant said.

divided the duration time series of speech with the ID for that particular movement between ring and bar or vice versa to create an MT/ID time series for speech. We used this time series for speech to analyze complexity matching between gestures and speech.

2.5. Analysis

2.5.1. Calculating temporal alignment

For each trial, from ring to bar or bar to ring, we know the time between the moment the ring or bar became activated and (a) the moment that participants pointed to and touched a bar or ring, and (b) the amplitude peak of the word the participant said to indicate the ring's or bar's location. We compared these durations between the moment of pointing and the amplitude peak. For each participant, we calculated the average absolute difference between moments of pointing and amplitude peak and used this as our measure of temporal alignment. Please note that higher values correspond to lower temporal alignment. Fig. 5 displays how we estimated temporal alignment and semantic similarity within a trial. To check whether participant's temporal alignment was significantly higher than the chance level, for each participant we compared the empirical temporal alignment with the temporal alignment between their repeatedly shuffled durations of gestures and speech.

2.5.2. Calculating semantic similarity

For each trial, from ring to bar or bar to ring, we know whether participants' pointed to the left, middle, or right object, and which location they mentioned in the speech. We compared the location in gestures and speech location and identified whether they did or did not match.

We calculated the sum of mismatches in location and used this as our measure of semantic similarity. Please note that higher values correspond to lower semantic similarity. To check whether the participant's semantic similarity was significantly higher than the chance level, for each participant we compared the empirical semantic similarity with the semantic similarity between their repeatedly shuffled (mentioned) location of gestures and speech.

2.5.3. Calculating complexity matching

We applied MF DFA (Ihlen, 2012; Ihlen & Vereijken, 2010; Kantelhardt et al., 2002; Wallot, O'Brien, Haussmann, Kloos, & Lyby, 2014) to the time series of gestures and speech. MF DFA is a method to reliably approximate a time series *temporal multifractality*. MF DFA is an extension of detrended fluctuation analysis (DFA), which is a method to reliably approximate a time series' *temporal fractality*. An accessible explanation of MF DFA can be found in Appendix A.

In short, performing MF DFA on a time series yields a so-called multifractal spectrum (see Fig. 6; the details of going from time series to multifractal spectrum can be found in Appendix A). The width of this multifractal spectrum indicates the degree of temporal multifractality of the time series and is a measure of the multifractal structure of the time series' variability. In short, a higher degree of multifractal structure leads to a wider multifractal spectrum, while a lower degree of multifractal structure (or higher degree of monofractal structure) leads to a narrower multifractal spectrum. As previously described, complexity matching requires that the fractal structure of variability of the behavior of two complex systems matches. To investigate the degree of complexity matching between gestures and speech, we, therefore, calculated the difference in gestures' and speech's multifractal spectrum width. To check whether complexity matching between gestures and speech was significant, for each participant, we compared the actual difference in multifractal spectrum width with the difference in repeatedly sampled random pairs of gestures' and speech's multifractal spectrum width.

2.5.4. Monte Carlo permutation testing

We calculated all *p*-values using Monte Carlo (MC) permutation tests (Ninness et al., 2002; Todman & Dugard, 2001) because MC permutations tests do not require a specific underlying distribution of the data. By drawing 10,000 random samples from the original data, the probability that differences are caused by chance was measured. We used custom-made R scripts to calculate *p*-values using MC permutation tests (osf.io/dj5vr/; Scripts & Materials).

3. Results

3.1. Descriptives

Participants in the difficult condition performed the task on average within 987 s ($SD = 138$ s). While they always pointed to the correct location of the bar and ring, they said the incorrect location on average 119.8 out of 1080 trials ($SD = 29.6$), that is, 11%. A semantic dissimilarity was thus always a combination of a correct gesture and an incorrect utterance. In

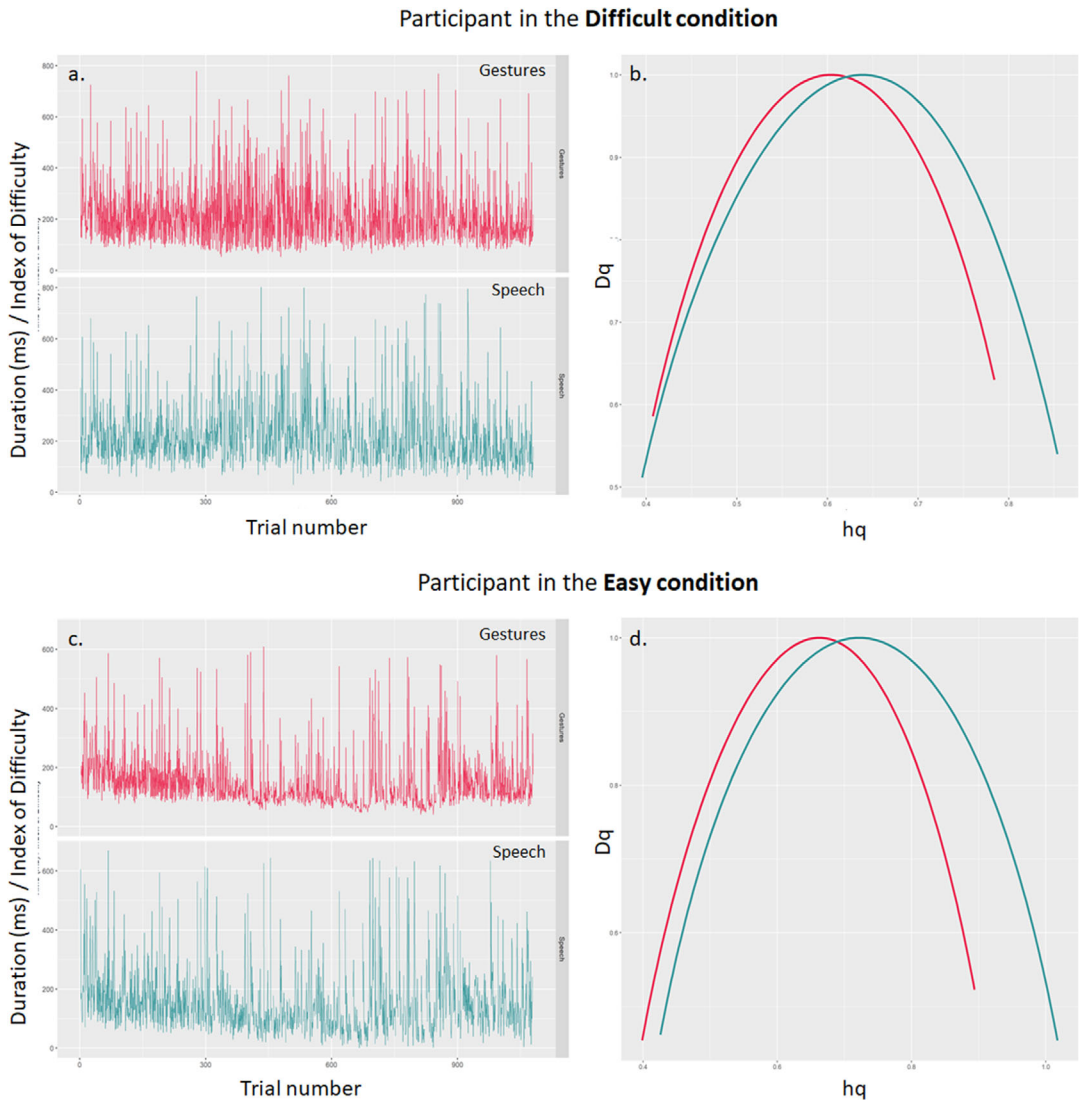


Fig. 6. Time series of duration (ms) divided by index of difficulty (ID) (panels A and C), and corresponding multifractal spectrums (panels B and D, respectively), for gestures (red) and speech (blue). Panels A and B illustrate the movement time/ID of time series gestures and speech and corresponding multifractal spectrums of a participant in the difficult condition and panels C and D of a participant in the easy condition. The difference in multifractal spectrum width is 0.081 for the participant in the difficult condition and 0.096 for the participant in the easy condition. We interpret this as more complexity matching between gestures and speech for the participant in the difficult condition, compared to the participant in the easy condition.

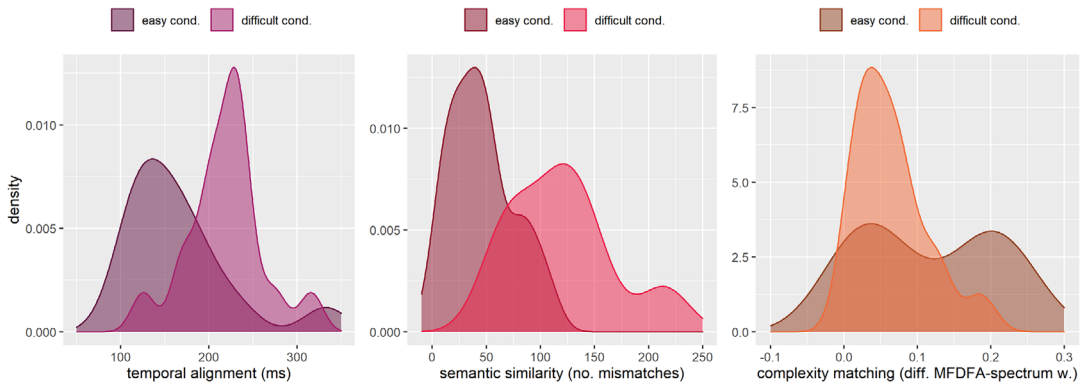


Fig. 7. Density plots of temporal alignment, semantic similarity, and complexity matching in the difficult and easy conditions.

the difficult condition, gestures' width of the MF DFA spectrum was on average 0.473 ($SD = 0.203$), and speech's width of the MF DFA-spectrum was on average 0.432 ($SD = 0.178$).

Participants in the easy condition performed the task on average within 749 s ($SD = 151$ s). Similar to the difficult condition, they always pointed to the correct location of the bar and ring, but they said the incorrect location on average 45.8 out of 1080 trials ($SD = 47.3$), that is, 4%. Gestures' width of the MF DFA spectrum was on average 0.618 ($SD = 0.169$), and speech's width of the MF DFA spectrum was on average 0.496 ($SD = 0.104$), in the easy condition.

3.2. RQ1: Task difficulty's influence on temporal alignment, semantic similarity, and complexity matching

With regard to temporal alignment, we found significantly less temporal alignment between participants' gestures and speech in the difficult condition ($M = 218.538$ ms, $SD = 43.652$) than in the easy condition ($M = 167.182$ ms, $SD = 62.322$), $p = .009$ ($\Delta_M = 51.356$, 95% $CI_{\Delta-MC} = -34.598, 35.322$), with a large effect size, $d = 0.955$ (see Fig. 7, left panel). This finding is opposite from our hypothesis 1A, as we expected that gestures and speech would be more temporally aligned in the difficult than in the easy condition. For all participants, the empirically observed temporal alignment between gestures and speech throughout the task was significantly higher than the temporal alignment between random pairs of their gestures' and speech's duration ($p < .001$).

For semantic similarity, we found significantly less semantic similarity between participants' gestures and speech in the difficult condition ($M_{mismatches} = 119.750$, $SD = 47.301$) than in the easy condition ($M_{mismatches} = 45.769$, $SD = 29.601$), $p < .001$ ($\Delta_M = 73.981$, 95% $CI_{\Delta-MC} = -32.661, 32.506$), with a very large effect size, $d = 1.875$ (see Fig. 7, center panel). This finding is in line with our hypothesis 1B, as we expected less semantic similarity between gestures and speech in the difficult than in the easy condition. For all participants, the empirically observed semantic similarity between gestures and speech throughout the task

was significantly higher than the semantic similarity between random pairs of their gestures' and speech's semantic content ($p < .001$).

With regard to complexity matching, we found more complexity matching between gestures and speech for participants in the difficult condition ($M_{diff. MFDFFA-spectrum\ width} = 0.065$, $SD = 0.049$) than in the easy condition ($M_{diff. MFDFFA-spectrum\ width} = 0.123$, $SD = 0.102$), $p = .026$ ($\Delta_M = -0.058$, 95% $CI_{\Delta-MC} = -0.049, 0.047$), with a medium to large effect size, $d = 0.726$ (see Fig. 7, right panel). When we visually inspected the density plot, participants in the difficult condition showed a striking peak around 0.0 and 0.1 in the difference of MFDFFA-spectrum width. However, participants in the easy condition showed no clear peak in the difference in MFDFFA-spectrum width but instead showed a wide range of values. In line with this, for 15 out of 16 participants in the difficult condition, we found the difference in MFDFFA-spectrum width to be significantly smaller ($p < .05$) than the difference in MFDFFA-spectrum width between random pairs of participants' gestures and speech, while we found this to be true for only eight out of 14 participants in the easy condition. Note that we did not make a prediction about the difference in complexity matching between the two conditions.

3.3. RQ2: Relations between temporal alignment, semantic similarity, and complexity matching

In the difficult condition, we found a significant, moderate, positive correlation between average temporal alignment (ms) and semantic similarity (number of gesture–speech mismatches), $r = .555$, $p = .014$ (95% $CI_{r-MC} = -0.422, 0.433$; see Fig. 8a). This finding is opposite from our hypothesis 2B, as we expected a negative relation between temporal alignment and semantic similarity in the difficult condition. We found a significant, moderate, negative correlation between average temporal alignment (ms) and complexity matching (difference in MFDFFA-spectrum width), $r = -.481$, $p = .031$ (95% $CI_{r-MC} = -0.430, 0.433$; see Fig. 8b). We did not state a hypothesis about the relation between temporal alignment and complexity matching. We found a non-significant, low, negative correlation between semantic similarity (number of gesture–speech mismatches) and complexity matching (difference in MFDFFA-spectrum width), $r = -.125$, $p = .336$ (95% $CI_{r-MC} = -0.414, 0.448$; see Fig. 8c). We did not state a hypothesis about the relation between semantic similarity and complexity matching. An overview of our findings with regards to research question 2 can be found in Fig. 9.

In the easy condition, we found a significant, moderate, positive correlation between average temporal alignment (ms) and semantic similarity (number of gesture–speech mismatches), $r = .653$, $p = .013$ (95% $CI_{r-MC} = -0.438, 0.511$; see Fig. 8d). This finding is in line with our hypothesis 2A, as we expected a positive relation between temporal alignment and semantic similarity in the easy condition. We found a non-significant, low, negative correlation between average temporal alignment (ms) and complexity matching (difference in MFDFFA-spectrum width), $r = -.205$, $p = .269$ (95% $CI_{r-MC} = -0.444, 0.489$; see Fig. 8e). This finding is not in line with our hypothesis 2C, as we expected a positive relation between temporal alignment and complexity matching. We found a non-significant, low, negative correlation between semantic similarity (no. of gesture–speech mismatches) and complexity matching (difference in MFDFFA-spectrum width), $r = -.211$, $p = .253$ (95%

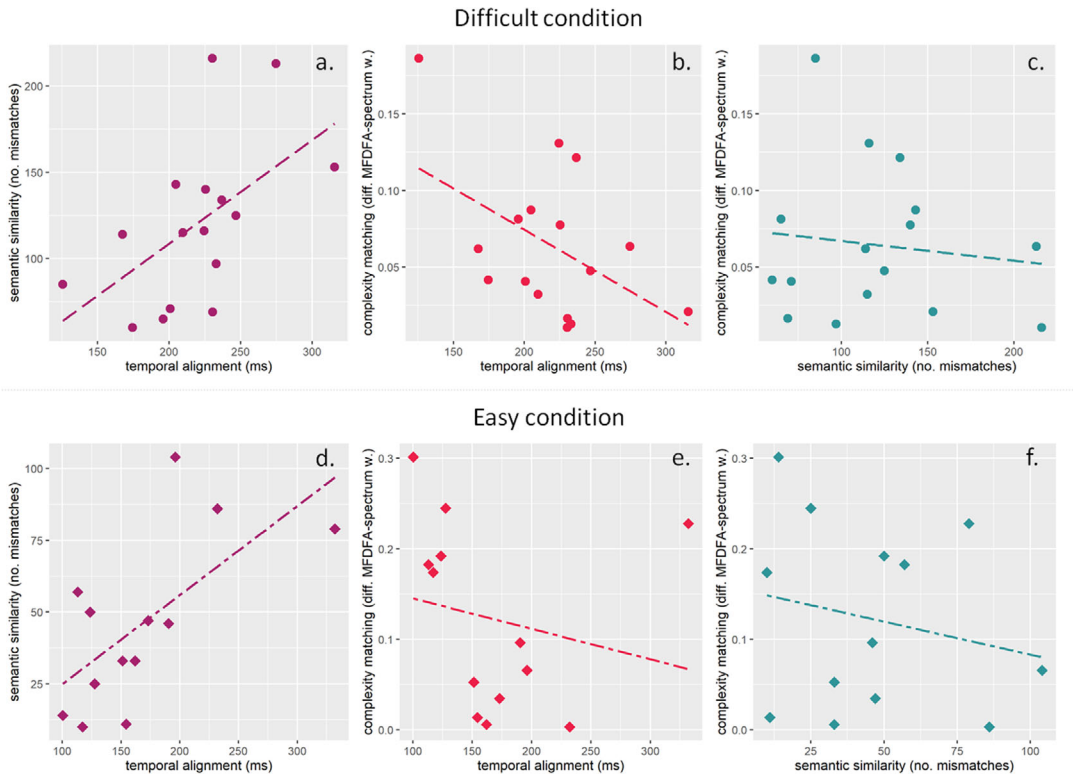


Fig. 8. Scatterplots of relations between the variables temporal alignment (ms), semantic similarity (no. of mismatches), and complexity matching (difference in MF DFA-spectrum width). Panels A–C display the relations in the difficult condition; panels D–F display the relations in the easy condition.

$CI_{r-MC} = -0.475, 0.477$; see Fig. 8f). This finding is not in line with our hypothesis 2D, as we expected a positive relation between semantic similarity and complexity matching.

3.4. RQ3: Predict task performance with temporal alignment, semantic similarity, and complexity matching

We performed a multiple linear regression to predict task performance (total time) based on temporal alignment, semantic similarity, and complexity matching.

With regard to the individual variables, greater temporal alignment significantly predicted better (i.e., a more speedy) task performance than condition alone, with R^2 increasing from 0.423 to 0.616, $p < .001$ ($\Delta_{R^2} = 0.192$, 95% $CI_{\Delta-MC} = 0.000, 0.082$). Less semantic similarity did not significantly predict better task performance than condition alone, with R^2 increasing from 0.423 to 0.425, $p = .764$ ($\Delta_{R^2} = 0.002$, 95% $CI_{\Delta-MC} = 0.000, 0.082$). Less complexity matching did not significantly predict better task performance than condition alone, with R^2 increasing from 0.423 to 0.456, $p = .214$ ($\Delta_{R^2} = 0.033$, 95% $CI_{\Delta-MC} = 0.000, 0.079$).

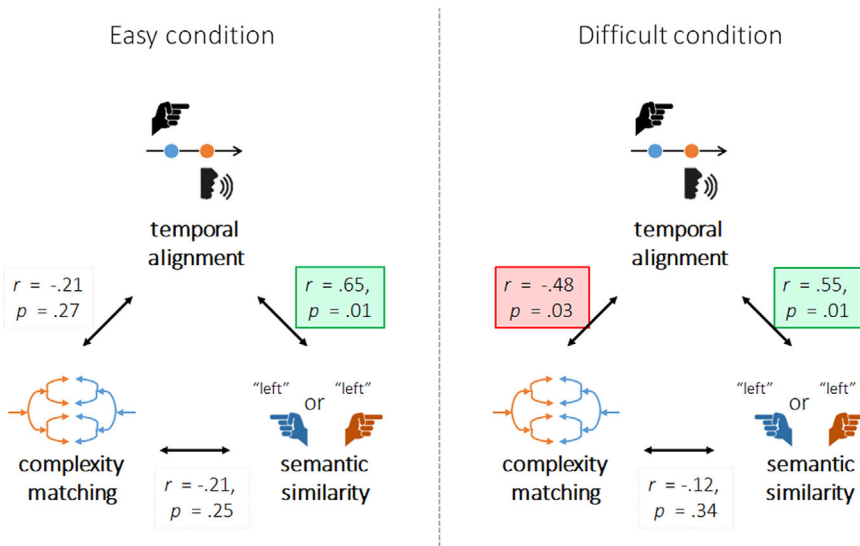


Fig. 9. Overview of the empirical relations between temporal alignment, semantic similarity, and complexity matching in the easy and difficult conditions.

Given that temporal alignment was a predictor of performance with the only condition in the model, we asked whether semantic similarity and complexity matching would contribute additional unique variance. When semantic similarity was included in the model with the condition and temporal alignment, we obtained a significant increase in R^2 from 0.616 to 0.734, $p = .003$ ($\Delta_{R^2} = 0.118$, 95% $CI_{\Delta-MC} = 0.000, 0.057$), whereby greater temporal alignment and less semantic similarity significantly predicted task performance. When we added complexity matching to the model containing condition and temporal alignment, we obtained a non-significant increase in R^2 from 0.616 to 0.619, $p = .628$ ($\Delta_{R^2} = 0.004$, 95% $CI_{\Delta-MC} = 0.000, 0.057$). Furthermore, when we added complexity matching to the model containing condition, temporal alignment, and semantic similarity, we obtained a non-significant increase in R^2 from 0.734 to 0.737, $p = .601$ ($\Delta_{R^2} = 0.003$, 95% $CI_{\Delta-MC} = 0.000, 0.040$).

4. Discussion

In this study, we investigated how a difference in task difficulty influences the synchronization between participant's gestures and speech, in terms of temporal alignment, semantic similarity, and complexity matching.

4.1. Summary of results

Our first research question was: How does task difficulty influence temporal alignment, semantic similarity, and complexity matching between participant's gestures and speech? We found significantly less temporal alignment and semantic similarity and more complexity

matching in the difficult than in the easy condition. With regard to complexity matching, we additionally observed a more peaked distribution of differences in MFDFA-spectrum widths in the difficult condition, while the distribution was clearly flattered in the easy condition. This suggests that for participants in the difficult condition, the fractal structure of variability of gestures' and speech' matches to a similar degree, which also points to complexity matching. Participants in the easy condition show a more variable degree of this matching, so no clear complexity matching.

Our second research question was: How are temporal alignment, semantic similarity, and complexity matching between gestures and speech related in the easy and difficult conditions? In the difficult condition, we found: (a) a moderate and significant positive relation between temporal alignment and semantic similarity; (b) a moderate and significant negative relation between temporal alignment and complexity matching; and (c) a low and non-significant negative relation between complexity matching and semantic similarity. In the easy condition, we found: (a) a moderate and significant positive relation between temporal alignment and semantic similarity, (b) a low and non-significant negative relation between temporal alignment and complexity matching, and (c) a low and non-significant negative relation between complexity matching and semantic similarity.

Our third research question was: How do temporal alignment, semantic similarity, and complexity matching between gestures and speech predict task performance? With regard to individual variables, we found that temporal alignment significantly predicted task performance, whereby more temporal alignment went together with better (i.e., a more speedy) task performance. Neither semantic similarity nor complexity matching significantly predicted task performance. With regard to combinations of variables, we found that temporal alignment and semantic similarity together predicted task performance better than temporal alignment alone, whereby more temporal alignment and less semantic similarity went together with better task performance. Adding complexity matching to the model did not significantly increase the model's exploratory power.

4.2. Phase synchronization

When two (weakly) coupled *oscillating* systems interact, their rhythm adjusts and their frequency entrains. This phenomenon is called *phase synchronization* (e.g., Pikovsky et al., 2001; Warren, 2006) and results in temporal alignment. We have viewed gestures and speech as two coupled systems throughout this paper. Akin to oscillating systems, we observed that participants in the easy condition rapidly got into a regular rhythm of gesturing and speaking. However, participants in the difficult condition struggled to get into and maintain a rhythm. In line with the higher temporal alignment that we found in the easy condition, we believe that the participant's gestures and speech also exhibited phase synchronization in the easy condition. Similarly, Pouw et al. (2019) found that rhythmical arm beating, but not wrist beating, entrained the amplitude envelope of speech. Although less pronounced than beating, participants in the easy condition of the current study also rhythmically moved their arms.

Pouw and Dixon (2019b) investigated temporal alignment between gestures and speech while participants told a story. As previously described, Pouw and Dixon (2019b) found an

increase in temporal alignment between participants' gestures and speech under delayed auditory feedback. Delayed auditory feedback is a delayed stimulus that entrains both gestures and speech, and gestures and speech become more synchronized to each other because they are entrained together. Pouw and Dixon (2019b) reasoned that delayed auditory feedback perturbs hand movements and speech and that the increase in gesture–speech synchrony is a way to stabilize rhythmic activity (such as gestures and speech) under disrupting circumstances (also see Pikovsky et al., 2001), that is, “stability through synergy” (Pouw & Dixon, 2019b, p. 28).

While the difficult task in our study did *disrupt* gestures' and speech's rhythm, task difficulty did not *entrain* gestures and speech. The nature of our perturbation was different from Pouw and Dixon (2019), and indeed we did not find more temporal alignment in the difficult than in the easy condition. However, we did find more complexity matching in the difficult than in the easy condition. Extending Pouw and Dixon's (2019) notion of “stability through synergy,” in the difficult condition, gestures and speech may have stabilized together by means of complexity matching, which entails coordination at multiple timescales, instead of entrainment; that is, coordination at a single timescale. Metaphorically speaking, the difficult condition might elicit a form of gesture–speech coordination that shares similarities with the coordination between a jazz-saxophonist and a -pianist while improvising together, which is characterized by “...a multitude of simple and complex rhythms, all interwoven extemporaneously into one cohesive sound” (i.e., complexity matching; Herby Hancock Institute of jazz, <https://bit.ly/2FIypCm>; also see Walton, Richardson, Langland-Hassan, & Chemero, 2015; Walton et al., 2011). The easy condition might elicit a form of gesture–speech coordination similar to clapping one's hands in a regular monotonous rhythm (i.e., entrainment). Furthermore, in the easy condition, entrainment may overrule complexity matching. This might suggest a trade-off between phase synchronization and complexity matching, which could be reflected in the negative relation between temporal alignment and complexity matching in the difficult condition that we found (also see Marmelat & Delignieres, 2012). In terms of our metaphor, if either the saxophonist or the pianist starts playing a regular monotonous rhythm, the other musician will be drawn to that regular and monotonous rhythm and will have a very hard time to maintain improvisation in all its complexity. We will discuss our findings' implications for the concept of complexity matching in the next paragraphs.

4.3. Complexity matching

While a body of research has shown that complexity matching exists between different human systems and under different circumstances (e.g., Abney, 2016; Abney et al., 2014; Almurad et al., 2017; Coey et al., 2016, 2018; Den Hartigh et al., 2018; Fusaroli et al., 2013; Marmelat & Delignieres, 2012; Ramirez-Aristizabal et al., 2018; Schloesser et al., 2019; Schneider, Ramirez-Aristizabal, Gavilan, & Kello, 2019), we are still grappling with what complexity matching *actually does* for people. In our study, we found more complexity matching between gestures and speech in the difficult than in the easy condition, and we interpreted this as a way for gestures and speech to stabilize together when entrainment is difficult to impossible. However, complexity matching did not predict participant's task performance

in terms of time to finish the task, and complexity matching was also not related to semantic content-alignment (i.e., number of speech errors). Apart from gestures and speech potentially being more stable, as we proposed, it is unclear whether and how participants benefited from more complexity matching.

Different studies about complexity matching during dyadic tasks do show that participants who demonstrated complexity matching benefited from this, in terms of reaching a collaborative goal (Abney et al., 2014; Fusaroli et al., 2013; see also Schloesser et al., 2019). Important to note is that the performance measures in the studies by Abney et al. (2014) and Fusaroli et al. (2013) are more sophisticated and captured higher-order goals than our simple performance measure of total time to perform the task did. In line with our findings, Schloesser et al. (2019) also found a weak and slightly negative relation between complexity matching—both within and between participants—and performance in terms of total time.

From a theoretical point of view, West et al. (2008) showed that complexity matching increases the information exchange between complex networks. However, as argued before by Abney (2016), we know little about what this *information* actually is and how to operationalize it. We could speculate that complexity matching only increases performance on tasks that involve the (re)organization of components to a higher-order structure. This higher-order structure could be the *common ground* that interacting people needed to establish during a conversation involving many different utterances (Abney et al., 2014), or the *joint decision* that people needed to converge to during a series of joint decision-making (Fusaroli et al., 2013). If it is true that complexity matching only increases performance on tasks that involve the (re)organization of components to a higher-order structure, this could hint that the information as proposed by West entails *interactions between components that form a synergy*.

An interesting study by Rigoli, Holman, Spivey, and Kello (2014; also see Schloesser et al., 2019) similarly suggests that information in complexity matching entails interactions between components that form a synergy. Rigoli et al. (2014) investigated participants who were asked to tap to a visual metronome by pressing a key. Rigoli et al. (2014) found complexity matching between the time series of participants' key press times and durations [*key press synergy*], and they found complexity matching between the time series of participants' pupil dilation and heart rate [*anatomic synergy*]. However, Rigoli et al. (2014) did not find complexity matching between key press and the anatomic time series. Rigoli et al. (2014), therefore, concluded that the key press and anatomic networks did not exchange information during the simple and relaxed task of tapping to the visual metronome. Similarly, in the easy (simple and relaxed) condition of the current study, we did not find complexity matching between gestures and speech, which suggests that these systems did not exchange information either. We did find complexity matching in the difficult condition, however, which suggests that the gestures and speech exchanged information and (re)organized as a synergy under these difficult task constraints. Future studies could investigate whether difficult tasks, involving higher-order goals, indeed elicit more complexity matching between systems than simple tasks. With regard to difficult tasks involving higher order-goals for children, one example is Piagetian conversation tasks, which have been used to study the interplay between gestures and speech before (e.g., Alibali et al., 2000; Church & Goldin-Meadow, 1986; De Jonge-Hoekstra et al., 2020; Pine, Lufkin, & Messer, 2004; Pine, Lufkin, Kirk, & Messer, 2007).

4.4. *Gesture–speech mismatches*

As previously described, Goldin-Meadow and colleagues (e.g., Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003; Goldin-Meadow, Alibali et al., 1993) found that children make gesture–speech mismatches (i.e., semantic *dissimilarities*) when they are on the verge of learning something new. Moreover, during these gesture–speech mismatches, children show a more advanced understanding of gestures than in speech. In the current study, we found more gesture–speech mismatches (i.e., less semantic similarity) in the difficult than in the easy condition, and these gesture–speech mismatches were always due to speech errors in semantic content. With our findings, we thus extend the phenomenon of gesture–speech mismatches from tasks in which people acquire understanding about cognitive problems to difficult cognitive tasks in general. Since a transition between “old” and “new” understanding was impossible in our experiment, participants’ gesture–speech mismatches were due to something different than competing cognitive understanding.

First, both in the current study and in previous studies, gestures had a clear spatial component that was directly linked to the physical properties of the task material (e.g., Bergmann & Kopp, 2010; De Jonge-Hoekstra et al., 2020; Hostetter & Alibali, 2008; Yeo & Alibali, 2018). This is not true for speech, however, and Smith and Gasser (2005) even propose that a too close resemblance between the physical structure of the environment and the structure of speech would limit speech’s functionality. Maybe difficult, cognitive tasks amplify this difference between gestures and speech in how they are coupled to the physical properties of the spatial environment, which could result in gesture–speech mismatches. Furthermore, we could question the extent to which speech actually needed to be functional in the current study. Participants performed the task individually and their speech did not have to be understandable for someone else (also see Fowler, 2010). Future studies could investigate how task constraints related to spatial structure and social context influence the occurrence of gesture–speech mismatches.

Second, participants had to verbally discriminate left from right in our experiment, which is known to be notoriously difficult for children and adults alike (e.g., Fisher & Camenzuli, 1987; McKinley, Dempster, & Gormley, 2015; Vingerhoets & Sarrechia, 2009). To our knowledge, no studies have investigated whether people find it difficult to discriminate between left and right using gestures as well. However, Abarbanell & Li (2020) recently found that instructing children to use gestures to discriminate between left and right benefits their performance on a rotation task more than instructing children to say the (Spanish) words (for) “left” and “right.” The authors explain this effect by gestures being directly linked to the spatial properties of a task, similar to our reasoning in the previous paragraph. This direct link between gestures and spatial properties of a task is particularly evident for deictic gestures, like the pointing of participants in our study. Therefore, discriminating between left and right using gestures was probably easier for the participants than using speech. Furthermore, while participants in the easy condition could just repeat the same sequence of words without much thought about their meaning, participants in the difficult condition needed to think about the words’ meaning constantly. Participants in the difficult condition were, therefore, more prone to confuse the words “left” and “right,” while they could correctly differentiate between left and right by means of pointing. This could explain why we found more gesture–speech

mismatches in the difficult than in the easy condition. Future studies need to investigate whether this phenomenon is more evident in tasks that require left-right discrimination as compared to spatial-temporal tasks in general as we argued in the previous paragraph.

Third, in line with Bergmann et al. (2011), we found a positive relation between temporal alignment and semantic similarity in both the difficult and easy conditions, which suggests that more temporal alignment goes together with less gesture–speech mismatches. However, it is yet unclear whether the temporal alignment is causally related to gesture–speech mismatches and what the direction of this potential relation would be. According to the information packaging hypothesis (Kita, 2000; also see Kita, Alibali, & Chu, 2017), gestures help to organize and “package” spatial information to both enable verbalization about this spatial information and ensure that the spatial information “fits” within the structure of speaking. When verbalization is challenging, speakers take more time to “package” information by means of gesturing. This would result in low temporal alignment between gestures and speech during gesture–speech mismatches as well as low temporal alignment in the difficult condition. This is in line with the positive relation between temporal alignment and semantic similarity and less temporal alignment, and also with less temporal alignment in the difficult condition, that we found. Follow-up studies could research the relation between gestures, speech, and gesture–speech mismatches in more detail, using methods to quantify the temporal direction of gesture–speech coupling, such as cross recurrence quantification analysis (see also De Jonge-Hoekstra et al., 2016). Moreover, in previous studies, temporal information usually has been disregarded when coding gesture–speech mismatches (e.g., Alibali et al., 2000).

4.5. Limitations

Our study has a number of limitations. We will address the limitations that we deem most important.

First, participant’s utterances during the experiment were very limited in scope and syntactic complexity (i.e., “left,” “middle,” “right”), which leaves open the question of how our findings will correspond to more typical, fluent, and syntactically complex speaking and gesturing. Previous studies have found complexity matching between participant’s fluent speech (Abney et al., 2014; Fusaroli et al., 2013). Furthermore, Abney et al. (2018) created spike trains of participant’s language and gesture events during fluent conversations and subsequently calculated the *burstiness* of both language and gesture events. Bursty processes are typical for complex dynamical systems (Barabási, 2005; Karsai, Kaski, Barabási, & Kertész, 2012), and in this sense, burstiness shares similarities with multifractality (albeit the scope of burstiness analysis is not multi-scaled). The methods used by Abney and colleagues (Abney et al., 2014, 2018; Fusaroli et al., 2013) provide viable directions for investigating complexity matching between gestures and speech in more typical and fluent speaking and gesturing.

Second, instead of changing the physical layout and order of the task, we could have increased task difficulty in a way that is closer to cognitive problem-solving. For instance, we could have asked participants to follow sets of rules about when to put which color ring on which color bar, and investigate how rules of varying difficulty influence gesture–speech coupling. However, such manipulation would not have perturbed participants continuously as

participants get used to rules, while the random order that we used in the current study did continuously perturb them.

Third, while we treated the trials from ring to bar and bar to ring equally, the instruction for these trials differed. For the trials from ring to bar, participants were instructed to select the bar that has the same color as the ring. For the trials from bar to ring, participants were instructed to select the enlarged ring. This difference in instruction could potentially lead to a different pattern of multifractal scaling for the trials from ring to bar and for the trials from bar to ring. In an interesting study, Kello et al. (2007) investigated a task whereby participants needed to press a key on a keyboard when they saw a stimulus on screen, thereby responding as fast as they could. Participants were allocated to either an easy, predictable condition, whereby the time between the stimuli was constant, or to a difficult, unpredictable condition, whereby the time between stimuli was random within a certain range. Kello et al. (2007) analyzed two time series: (a) A time series of the time between the appearance of the stimuli and pressing the key (reaction time), and (b) a time series of the time between pressing the key and releasing the key again (key-contact duration). The authors argue that participants only received instruction about reaction time (responding as fast as possible), while they received no instruction about key-contact duration. Kello et al. (2007) found the reaction times and key-contact durations in both conditions to be not or only weakly correlated. Furthermore, they found fractal scaling in both the reaction and the key-contact duration time series and in both conditions. The fractal scaling of the reaction time series of the difficult, predictable condition was lower than the fractal scaling of the other three time series. Although the study by Kello et al. (2007) shares some similarities with our study, there are notable differences as well. While pressing down a key as fast as possible and releasing a key correspond to a simple instruction versus no instruction, respectively (Kello et al., 2007), selecting a bar with the same color and selecting an enlarged ring correspond to a more complicated instruction versus a simple instruction, respectively (current study). Furthermore, while pressing down and releasing a key are two different motions, involving the contraction of different muscles (Kello et al., 2007), trials from bar to ring and ring to bar *both* involved pointing to a target and saying the location of that target (current study). A follow-up study could investigate whether the ring to bar trials differ from the bar to ring trials with regard to duration and multifractal scaling.

Fourth, our sample size is relatively small, which is largely due to failed audio recordings. However, we do have many datapoints per participant. Fifth, the number of measurements per participant (1,024) was on the small side for performing MF DFA (Ihlen & Vereijken, 2010), yet sufficient. Albeit challenging, we need to come up with ways to increase the number of measurements per participant while still keeping the task feasible for participants to do. Furthermore, Almurad and Delignières (2016) propose an alternative way of performing DFA (the monofractal variant of MF DFA), which allows for time series that are even shorter than 1,024 datapoints.

4.6. Conclusion

We aimed to investigate how task difficulty affects the synchronization between gestures and speech, thereby empirically addressing De Jonge-Hoekstra et al.'s (2016) proposal. By

doing so, we brought together different perspectives and ways of investigating gesture–speech synchronization. We found that task difficulty indeed influences gesture–speech synchronization in terms of temporal alignment, semantic similarity, and complexity matching. With our findings of less semantic similarity in the difficult condition, we extended the phenomenon of gesture–speech mismatches to difficult, cognitive tasks. Furthermore, we found more temporal alignment in the easy condition, which we related to phase synchronization between gestures and speech. We found more complexity matching between gestures and speech in the difficult condition, which we related to gestures and speech forming a more stable synergy under the influence of more difficult task constraints. Our findings add another piece to the puzzle of why complexity matching occurs in complex dynamical systems.

In sum, our study demonstrates how this perspective can be used to study the relation between gestures and speech, and gesture–speech mismatches—subjects that primarily have been studied from within cognitive psychology. While the body of research that tries to bridge between complex dynamical systems and coordination research, and cognitive psychology is steadily growing, we acknowledge that many gaps between the two perspectives still remain. We look forward to future work that continues to build connections between the two fields, and we hope that these future studies can build on our study.

Acknowledgments

We would like to thank Dr. Mark Span for constructing the tablet task in OpenSesame. Furthermore, we would like to thank all the students who helped with collecting data and coding words. In addition, we would like to thank Dr. Wim Pouw for his advice on how to automatically calculate temporal alignment between gestures and speech. Finally, we would like to thank Dr. Alexandra Paxton for the discussion we had about applying complexity matching to our gesture and speech data.

Open Research Badges



This article has earned Open Data and Open Materials badges. Data and materials are available at <https://osf.io/dj5vr/>.

Notes

1. Studies differ in whether they differentiate between non-redundant or mismatching content (Wagner, Malisz, & Kopp, 2014).
2. Strictly speaking, time series' variability usually is self-affine instead of self-similar because its dimensions are scaled by different amounts in the x - and y -directions. For purposes of brevity and clarity, we will use the term self-similar throughout the paper.

3. We recruited a total of 59 participants to participate in this experiment. However, due to technical issues with the tablets, for 29 participants, the audio was either not recorded or recorded with insufficient quality (e.g., loud ticks on the screen, background noise). After rigorous checks of the quality of all the audio recordings, we decided to include the 30 participants of which the audio-recordings were of high quality. For the analyses that we will conduct, with many datapoints, a sample of 30 participants is sufficiently large. We have the pointing data for all 59 participants, and we will use this data for other studies and research questions that do not involve speech.
4. A time series is a sequence of datapoints in chronological order.
5. The distances are calculated between the middle of the ring area and the middle of the top of the bar area.
6. Some participants pronounced a very loud “s” at the end of “links” and “rechts,” and therefore the MATLAB script identified two syllables within these words instead of one. Conversely, some participants mumbled the word “midden” (which is quite typical for people from the northern part of the Netherlands), and therefore the MATLAB script identified one syllable within this word, instead of two. In addition, participants differed in their range of speech amplitude during the task: Some spoke evenly loud during the whole task, while others intermitted softer and louder periods of speaking. Therefore, for some participants, a velocity peak in a softer part of the audio recording is not recognized as a velocity peak in a louder part of the audio recording. This resulted in MATLAB identifying multiple words as one syllable in the loud periods of speaking and multiple words per audio segment in the softer periods.

References

- Abarbanell, L., & Li, P. (2020). Unraveling the contribution of left-right language on spatial perspective taking. *Spatial Cognition & Computation*, 21(1), 1–38. <https://doi.org/10.1080/13875868.2020.1825442>
- Abney, D. H. (2016). The Complexity Matching hypothesis for human communication (Doctoral dissertation). <https://escholarship.org/uc/item/8kx4m274#main>
- Abney, D. H., Dale, R., Louwerse, M. M., & Kello, C. T. (2018). The bursts and lulls of multimodal interaction: Temporal distributions of behavior reveal differences between verbal and non-verbal communication. *Cognitive Science*, 42(4), 1297–1316. <https://doi.org/10.1111/cogs.12612>
- Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2014). Complexity matching in dyadic conversation. *Journal of Experimental Psychology: General*, 143(6), 2304–2315. <https://doi.org/10.1037/xge0000021>
- Abney, D. H., Paxton, A., Dale, R., & Kello, C. T. (2015). Movement dynamics reflect a functional role for weak coupling and role structure in dyadic problem solving. *Cognitive Processing*, 16(4), 325–332. <https://doi.org/10.1007/s10339-015-0648-2>
- Abney, D. H., Paxton, A., Dale, R., & Kello, C. (2021). Cooperation in sound and motion: Complexity matching in collaborative interaction. *Journal of Experimental Psychology: General*. Advance online publication. <https://doi.org/10.1037/xge0001018>
- Adolph, K. E., & Franchak, J. M. (2017). The development of motor behavior. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2), e1430. <https://doi.org/10.1002/wcs.1430>
- Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15(6), 593–613. <https://doi.org/10.1080/016909600750040571>

- Allen, S., Özyürek, A., Kita, S., Brown, A., Furman, R., Ishizuka, T., & Fujii, M. (2007). Language-specific and universal influences in children's syntactic packaging of Manner and Path: A comparison of English, Japanese, and Turkish. *Cognition*, *102*(1), 16–48. <https://doi.org/10.1016/j.cognition.2005.12.006>
- Almurad, Z. M. H., & Delignières, D. (2016). Evenly spacing in Detrended Fluctuation Analysis. *Physica A: Statistical Mechanics and Its Applications*, *451*, 63–69. <https://doi.org/10.1016/j.physa.2015.12.155>
- Almurad, Z. M. H., Roume, C., & Delignières, D. (2017). Complexity matching in side-by-side walking. *Human Movement Science*, *54*, 125–136. <https://doi.org/10.1016/j.humov.2017.04.008>
- Barabási, A. L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211. <https://doi.org/10.1038/nature03459>
- Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011), Bielefeld, Germany.
- Bergmann, K., & Kopp, S. (2010). Modeling the Production of Coverbal Iconic Gestures by Learning Bayesian Decision Networks. *Applied Artificial Intelligence*, *24*(6), 530–551. <https://doi.org/10.1080/08839514.2010.492162>
- Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer (6.0.42) [computer software]*. Retrieved from <http://www.praat.org/>
- Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech: When hand and mouth come together. In D. McNeill (Ed.), *Language and Gesture* (pp. 235–258). Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511620850.015>
- Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gestures and speech as an index of transitional knowledge. *Cognition*, *23*, 43–71.
- Clark, J. E., Whittall, J., & Phillips, S. J. (1988). Human interlimb coordination: The first 6 months of independent walking. *Developmental Psychobiology*, *21*(5), 445–456. <https://doi.org/10.1002/dev.420210504>
- Clearfield, M. W. (2004). The role of crawling and walking experience in infant spatial memory. *Journal of Experimental Child Psychology*, *89*(3), 214–241. <https://doi.org/10.1016/j.jecp.2004.07.003>
- Coey, C. A., Kallen, R. W., Chemero, A., & Richardson, M. J. (2018). Exploring complexity matching and asynchrony dynamics in synchronized and syncopated task performances. *Human Movement Science*, *62*, 81–104. <https://doi.org/10.1016/j.humov.2018.09.006>
- Coey, C. A., Washburn, A., Hassebrock, J., & Richardson, M. J. (2016). Complexity matching effects in bimanual and interpersonal syncopated finger tapping. *Neuroscience Letters*, *616*, 204–210. <https://doi.org/10.1016/j.neulet.2016.01.066>
- Cox, R. F. A. (2016, July). Complexity matching in a cooperative Wiimote game. 14th European Workshop on Ecological Psychology, Groningen, the Netherlands:
- Davis, T. J., Brooks, T. R., & Dixon, J. A. (2016). Multi-scale interactions in interpersonal coordination. *Journal of Sport and Health Science*, *5*, 10.
- De Jonge-Hoekstra, L., Van Der Steen, S., & Cox, R. F. A. (2020). Movers and shakers of cognition: Hand movements, speech, task properties, and variability. *Acta Psychologica*, *211*, 103187. <https://doi.org/10.1016/j.actpsy.2020.103187>
- De Jonge-Hoekstra, L., Van der Steen, S., Van Geert, P., & Cox, R. F. A. (2016). Asymmetric dynamic attunement of speech and gestures in the construction of children's understanding. *Frontiers in Psychology*, *7*(March), 1–19. <https://doi.org/10.3389/fpsyg.2016.00473>
- Delignières, D., Almurad, Z. M. H., Roume, C., & Marmelat, V. (2016). Multifractal signatures of complexity matching. *Experimental Brain Research*, *234*(10), 2773–2785. <https://doi.org/10.1007/s00221-016-4679-4>
- Den Hartigh, R. J. R., Cox, R. F. A., Gernigon, C., Van Yperen, N. W., & Van Geert, P. L. C. (2015). Pink noise in rowing ergometer performance and the role of skill level. *Motor Control*, *19*(4), 355–369. <https://doi.org/10.1123/mc.2014-0071>
- Den Hartigh, R. J. R., Marmelat, V., & Cox, R. F. A. (2018). Multiscale coordination between athletes: Complexity matching in ergometer rowing. *Human Movement Science*, *57*, 434–441. <https://doi.org/10.1016/j.humov.2017.10.006>

- Dixon, J. A., Holden, J. G., Mirman, D., & Stephen, D. G. (2012). Multifractal dynamics in the emergence of cognitive structure. *Topics in Cognitive Science*, 4(1), 51–62. <https://doi.org/10.1111/j.1756-8765.2011.01162.x>
- Dumas, G., & Fairhurst, M. T. (2019). Reciprocity and alignment: Quantifying coupling in dynamic interactions. 30. <https://doi.org/10.31234/osf.io/nmg4x>
- Esteve-Gibert, N., & Guellà, B. (2018). Prosody in the auditory and visual domains: A developmental perspective. *Frontiers in Psychology*, 9, 338. <https://doi.org/10.3389/fpsyg.2018.00338>
- Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, 57, 301–316. <https://doi.org/10.1016/j.specom.2013.06.006>
- Feldman, D. (2019). *Fractals and scaling (massive open online course)*. Retrieved from <https://www.complexityexplorer.org/courses/93-fractals-and-scaling-winter-2019/segments/7786?summary>
- Fisher, C. B., & Camenzuli, C. A. (1987). Influence of body rotation on children's left-right confusion: A challenge to bilateral symmetry theory. *Developmental Psychology*, 23(2), 187–189.
- Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391. <https://doi.org/10.1037/h0055392>
- Fowler, C. A. (2010). Embodied, embedded language use. *Ecological Psychology*, 22(4), 286–303. <https://doi.org/10.1080/10407413.2010.517115>
- Fusaroli, R., Abney, D. H., Bahrami, B., Kello, C. T., & Tylén, K. (2013, July 31). *Conversation, coupling and complexity: Matching scaling laws predict task performance*. Berlin: Cognitive Science.
- Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.
- Goldin-Meadow, S. (2007). Pointing sets the stage for learning language—And creating language. *Child Development*, 78(3), 741–745. <https://doi.org/10.1111/j.1467-8624.2007.01029.x>
- Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, 100(2), 279–297.
- Goldin-Meadow, S., Nusbaum, H., Garber, P., & Church, R. B. (1993). Transitions in learning: Evidence for simultaneously activated strategies. *Journal of Experimental Psychology: Human Perception and Performance*, 19(1), 92–107.
- Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51(5), 347–356. <https://doi.org/10.1007/BF00336922>
- He, L., & Dellwo, V. (2016). A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, San Francisco, CA, 8–12 September, pp. 530–534. <https://doi.org/10.21437/Interspeech.2016-1447>
- He, L., & Dellwo, V. (2017). Amplitude envelope kinematics of speech signal: Parameter extraction and applications. *Journal of the Acoustical Society of America*, 141, 3582.
- Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3), 495–514. <https://doi.org/10.3758/PBR.15.3.495>
- Ihlen, E. A. F. (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. *Frontiers in Physiology*, 3(June), 1–18. <https://doi.org/10.3389/fphys.2012.00141>
- Ihlen, E. A. F., & Vereijken, B. (2010). Interaction-dominant dynamics in human cognition: Beyond 1/f fluctuation. *Journal of Experimental Psychology: General*, 139(3), 28. <https://doi.org/10.1037/a0019098>
- Iverson, J. M., & Fagan, M. K. (2004). Infant vocal-motor coordination: precursor to the gesture-speech system? *Child Development*, 75(4), 1053–1066. <https://doi.org/10.1111/j.1467-8624.2004.00725.x>
- Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 6(11–12), 19–40.
- Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S., Bunde, A., & Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A*, 316(1–4), 87–114. [https://doi.org/10.1016/S0378-4371\(02\)01383-3](https://doi.org/10.1016/S0378-4371(02)01383-3)

- Karsai, M., Kaski, K., Barabási, A. -L., & Kertész, J. (2012). Universal features of correlated bursty behaviour. *Scientific Reports*, 2(1), 397. <https://doi.org/10.1038/srep00397>
- Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, 136(4), 551–568. <https://doi.org/10.1037/0096-3445.136.4.551>
- Kelso, J. A. S. (1994). The informational character of self-organized coordination dynamics. *Human Movement Science*, 13(3–4), 393–413. [https://doi.org/10.1016/0167-9457\(94\)90047-7](https://doi.org/10.1016/0167-9457(94)90047-7)
- Kelty-Stephen, D. G., Palatinus, K., Saltzman, E., & Dixon, J. A. (2013). A tutorial on multifractality, cascades, and interactivity for empirical time series in ecological science. *Ecological Psychology*, 25(1), 1–62. <https://doi.org/10.1080/10407413.2013.753804>
- Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. W. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177–210). Burlington: Elsevier. <https://doi.org/10.1016/b978-0-08-015867-9.50013-7>
- Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (1st ed., pp. 162–185). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511620850.011>
- 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., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1), 16–32. [https://doi.org/10.1016/S0749-596X\(02\)00505-3](https://doi.org/10.1016/S0749-596X(02)00505-3)
- Kita, S., Van Gijn, I., & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Lecture Notes in Artificial Intelligence* (pp. 23–35). Berlin: Springer-Verlag.
- Kloos, H., & Van Orden, G. (2010). Voluntary behavior in cognitive and motor tasks. *Mind and Matter*, 8(1), 19–43.
- Koschmann, T. (2017). Why would the discovery of gestures produced by signers jeopardize the experimental finding of gesture-speech mismatch? *Behavioral and Brain Sciences*, 40, E60. <https://doi.org/10.1017/s0140525x15002952>
- Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, 26(10), 1457–1471. <https://doi.org/10.1080/01690965.2010.500218>
- Marmelat, V., & Delignieres, D. (2012). Strong anticipation: Complexity matching in interpersonal coordination. *Experimental Brain Research*, 222, 137–148.
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324. <https://doi.org/10.3758/s13428-011-0168-7>
- McKinley, J., Dempster, M., & Gormley, G. J. (2015). ‘Sorry, I meant the patient’s left side’: Impact of distraction on left-right discrimination. *Medical Education*, 49(4), 427–435. <https://doi.org/10.1111/medu.12658>
- McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3), 350–371.
- Ninness, C., Newton, R., Saxon, J., Rumph, R., Bradfield, A., Harrison, C., ... Ninness, S. K. (2002). Small group statistics: A Monte Carlo comparison of parametric and randomization tests. *Behavior and Social Issues*, 12(1), 53–63. <https://doi.org/10.5210/bsi.v12i1.79>
- Peng, C. K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1), 82–87. <https://doi.org/10.1063/1.166141>
- Pikovsky, A., Kurths, J., & Rosenblum, M. (2001). *Synchronization: A universal concept in nonlinear sciences*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511755743>
- Pine, K. J., Lufkin, N., Kirk, E., & Messer, D. (2007). A microgenetic analysis of the relationship between speech and gesture in children: Evidence for semantic and temporal asynchrony. *Language and Cognitive Processes*, 22(2), 234–246. <https://doi.org/10.1080/01690960600630881>
- Pine, K. J., Lufkin, N., & Messer, D. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology*, 40(6), 1059–1067. <https://doi.org/10.1037/0012-1649.40.6.1059>

- Pouw, W. T. J. L., De Jonge-Hoekstra, L., & Dixon, J. A. (2018). Stabilizing speech production through gesture-speech coordination [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/arzne>
- Pouw, W. T. J. L., de Nooijer, J. A., van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology*, 5(April), 359. <https://doi.org/10.3389/fpsyg.2014.00359>
- Pouw, W. T. J. L., van Gog, T., Zwaan, R. A., & Paas, F. (2017). Are gesture and speech mismatches produced by an integrated gesture-speech system? A more dynamically embodied perspective is needed for understanding gesture-related learning. *Behavioral and Brain Sciences*, 40, e68. <https://doi.org/10.1017/S0140525X15003039>
- Pouw, W., & Dixon, J. A. (2019a). Quantifying gesture-speech synchrony. *Proceedings of the 6th Gesture and Speech in Interaction Conference*, Paderborn, Germany, pp. 75–80. <https://doi.org/10.17619/UNIPB/1-815>
- Pouw, W., & Dixon, J. A. (2019b). Entrainment and modulation of gesture-speech synchrony under delayed auditory feedback. *Cognitive Science*, 43(3), e12721. <https://doi.org/10.1111/cogs.12721>
- Pouw, W., Harrison, S. A., & Dixon, J. A. (2019). Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony. *Journal of Experimental Psychology: General*, 149(2), 391–404. <https://doi.org/10.1037/xge0000646>
- Pouw, W., & Trujillo, J. P. (2019). *Tutorial Gespın2019—Using video-based motion tracking to quantify speech-gesture synchrony*. <https://doi.org/10.17605/OSF.IO/RXB8J>
- Prieto, P., & Roseano, P. (2018). Prosody: Stress, rhythm, and intonation. In K. L. Geeslin (Ed.), *The cambridge handbook of Spanish linguistics*. (1st ed., pp. 211–236). Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781316779194.011>
- Ramirez-Aristizabal, A. G., Médé, B., & Kello, C. T. (2018). Complexity matching in speech: Effects of speaking rate and naturalness. *Chaos, Solitons & Fractals*, 111, 175–179. <https://doi.org/10.1016/j.chaos.2018.04.021>
- Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, 26(6), 867–891. <https://doi.org/10.1016/j.humov.2007.07.002>
- Richardson, M. J., Marsh, K. L., & Schmidt, R. C. (2005). Effects of visual and verbal interaction on unintentional interpersonal coordination. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1), 62–79. <https://doi.org/10.1037/0096-1523.31.1.62>
- Rigoli, L. M., Holman, D., Spivey, M. J., & Kello, C. T. (2014). Spectral convergence in tapping and physiological fluctuations: Coupling and independence of 1/f noise in the central and autonomic nervous systems. *Frontiers in Human Neuroscience*, 8, 713. <https://doi.org/10.3389/fnhum.2014.00713>
- Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41), 12663–12668. <https://doi.org/10.1073/pnas.1419773112>
- Roy, D. (2011). *Deb Roy: The birth of a word*. New York, NY: Films Media Group.
- Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of prosody and position on the timing of deictic gestures. *Journal of Speech, Language, and Hearing Research*, 56(2), 458–470.
- Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, 57, 283–300. <https://doi.org/10.1016/j.specom.2013.06.004>
- Schloesser, D. S., Kello, C. T., & Marmelat, V. (2019). Complexity matching and coordination in individual and dyadic performance. *Human Movement Science*, 66, 258–272. <https://doi.org/10.1016/j.humov.2019.04.010>
- Schmidt, R. C., Nie, L., Franco, A., & Richardson, M. J. (2014). Bodily synchronization underlying joke telling. *Frontiers in Human Neuroscience*, 8, 633. <https://doi.org/10.3389/fnhum.2014.00633>
- Schmidt, R. C., & O'Brien, B. (1997). Evaluating the dynamics of unintended interpersonal coordination. *Ecological Psychology*, 9(3), 189–206.
- Schneider, S., Ramirez-Aristizabal, A. G., Gavilan, C., & Kello, C. T. (2019). Complexity matching and lexical matching in monolingual and bilingual conversations. *Bilingualism: Language and Cognition*, 23(4), 1–13. <https://doi.org/10.1017/S1366728919000774>

- Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2), 305–319. <https://doi.org/10.1111/j.1756-8765.2009.01021.x>
- Smith, L. B., & Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences*, 7(8), 343–348. [https://doi.org/10.1016/S1364-6613\(03\)00156-6](https://doi.org/10.1016/S1364-6613(03)00156-6)
- Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1–2), 13–29. <https://doi.org/10.1162/1064546053278973>
- Stephen, D. G., Anastas, J. R., & Dixon, J. A. (2012). Scaling in cognitive performance reflects multiplicative multifractal cascade dynamics. *Frontiers in Physiology*, 3, 102. <https://doi.org/10.3389/fphys.2012.00102>
- Stephen, D. G., Boncoddò, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, 37(8), 1132–1149. <https://doi.org/10.3758/MC.37.8.1132>
- Stephen, D. G., & Dixon, J. A. (2011). Strong anticipation: Multifractal cascade dynamics modulate scaling in synchronization behaviors. *Chaos, Solitons & Fractals*, 44(1–3), 160–168. <https://doi.org/10.1016/j.chaos.2011.01.005>
- Stephen, D. G., Dixon, J. A., & Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6), 1811–1832. <https://doi.org/10.1037/a0014510>
- Stephen, D. G., Stepp, N., Dixon, J. A., & Turvey, M. T. (2008). Strong anticipation: Sensitivity to long-range correlations in synchronization behavior. *Physica A: Statistical Mechanics and Its Applications*, 387(21), 5271–5278. <https://doi.org/10.1016/j.physa.2008.05.015>
- Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child Development*, 72(3), 748–767. <https://doi.org/10.1111/1467-8624.00313>
- Tamis-Lemonda, C. S., Bornstein, M. H., Kahana-Kalman, R., Baumwell, L., & Cyphers, L. (1998). Predicting variation in the timing of language milestones in the second year: An events history approach. *Journal of Child Language*, 25(3), 675–700. <https://doi.org/10.1017/S0305000998003572>
- Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. London: MIT Press.
- Thelen, E., & Smith, L. B. (2007). Dynamic systems theories. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 1, pp. 258–312). New York: John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470147658.chpsy0106>
- Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Treffner, P., & Peter, M. (2002). Intentional and attentional dynamics of speech-hand coordination. *Human Movement Science*, 21, 641–697. [https://doi.org/10.1016/S0167-9457\(02\)00178-1](https://doi.org/10.1016/S0167-9457(02)00178-1)
- Turvey, M. T. (2007). Action and perception at the level of synergies. *Human Movement Science*, 26(4), 657–697. <https://doi.org/10.1016/j.humov.2007.04.002>
- Van Geert, P. (2008). Nonlinear complex dynamical systems in developmental psychology. In S. J. Guastello, M. Koopmans & D. Pincus (Eds.), *Chaos and complexity in psychology* (pp. 242–281). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139058544.009>
- Van Geert, P. (2011). The contribution of complex dynamic systems to development. *Child Development Perspectives*, 5(4), 273–278. <https://doi.org/10.1111/j.1750-8606.2011.00197.x>
- Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, 132(3), 331–350. <https://doi.org/10.1037/0096-3445.132.3.331>
- Van Orden, G., Kloos, H., & Wallot, S. (2011). Living in the pink: Intentionality, wellbeing, and complexity. In C. Hooker (Ed.), *Handbook of the Philosophy of Science* (Vol. 10, pp. 629–672). Amsterdam: Elsevier.
- Vingerhoets, G., & Sarrechia, I. (2009). Individual differences in degree of handedness and somesthetic asymmetry predict individual differences in left-right confusion. *Behavioural Brain Research*, 204(1), 212–216. <https://doi.org/10.1016/j.bbr.2009.06.004>
- 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>

- Wallot, S., O'Brien, B. A., Haussmann, A., Kloos, H., & Lyby, M. S. (2014). The role of reading time complexity and reading speed in text comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, 40(6), 1745–1765. <https://doi.org/10.1037/xlm0000030>
- Walton, A. E., Richardson, M. J., Langland-Hassan, P., & Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Frontiers in Psychology*, 06, 313. <https://doi.org/10.3389/fpsyg.2015.00313>
- Walton, A. E., Washburn, A., Langland-Hassan, P., Chemero, A., Kloos, H., & Richardson, M. J. (2018). Creating time: Social collaboration in music improvisation. *Topics in Cognitive Science*, 10(1), 95–119. <https://doi.org/10.1111/tops.12306>
- Warren, W. H. (2006). The dynamics of perception and action. *Psychological Review*, 113(2), 358–389. <https://doi.org/10.1037/0033-295X.113.2.358>
- West, B. J., Geneston, E. L., & Grigolini, P. (2008). Maximizing information exchange between complex networks. *Physics Reports*, 468(1–3), 1–99. <https://doi.org/10.1016/j.physrep.2008.06.003>
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
- Wijnants, M. L. (2014). A review of theoretical perspectives in cognitive science on the presence of $1/f$ scaling in coordinated physiological and cognitive processes. *Journal of Nonlinear Dynamics*, 2014, 962043. <https://doi.org/10.1155/2014/962043>
- Wijnants, M. L., Cox, R. F. A., Hasselman, F., Bosman, A. M. T., & Van Orden, G. (2012). A trade-off study revealing nested timescales of constraint. *Frontiers in Physiology*, 3(May), 116. <https://doi.org/10.3389/fphys.2012.00116>
- Yeo, A., & Alibali, M. W. (2018). Does visual salience of action affect gesture production? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(5), 826–832. <https://doi.org/10.1037/xlm0000458>

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A Detailed and accessible description of Multifractal Detrended Fluctuation Analysis (MFDFA)