Pre-registered Report

Quantifying gesture-speech synchrony: Exploratory data report and pre-registration

Wim Pouw [1, 2] & James A. Dixon[1]

THIS IS A PREPRINT (VERSION 1) – MANUSCRIPT UNDER REVIEW

Center for the Ecological Study of Perception and Action, University of Connecticut[1]

Department of Psychology, Education, & Child Studies, Erasmus University Rotterdam[2]

Author note: Correspondence should be addressed to Wim Pouw (wimpouw@uconn.edu).

Open data & Pre-registration: Experiment code, (raw) data, and analyses scripts supporting this pre-registered report are available on the Open Science Framework (https://osf.io/5ja6y).

Funding: The research has been funded by The Netherlands Organisation of Scientific Research (NWO; Rubicon grant "Acting on Enacted Kinematics", Grant Nr. 446-16-012; PI Wim Pouw).

Abstract

Spontaneously occurring speech is often seamlessly accompanied by *hand gestures*. Detailed observations of video data suggest that speech and gesture are tightly synchronized in time, consistent with a dynamic interplay between body and mind. However, spontaneous gesture-speech synchrony has rarely been objectively quantified beyond analyses of video data, which do not allow for identification of kinematic properties of gestures. Consequently, the point in gesture which is held to couple with speech, the so-called moment of "maximum effort", has been variably equated with the peak velocity, peak acceleration, peak deceleration, or the onset of the gesture. In the current pre-registered report, we provide novel evidence from motion-tracking and acoustic data that peak velocity is closely aligned, and shortly leads, the peak pitch (F0) of speech. We propose to replicate this in a more comprehensive sample, so as to provide a rigorous quantification of gesture-speech synchrony.

Keywords: co-speech gesture, speech production, motion tracking

Background

Humans across all known cultures tend to move their hands during speaking (Kendon, 2004; McNeill, 2005), suggesting a fundamental connection between communicative vocalizations and hand movements (Iverson & Thelen, 1999; Gentilucci & Corballis, 2006). Because humans perceive and produce such hand gestures seamlessly as part of vocal communication (Cooperrider, 2018; Willems, Özyürek, & Hagoort, 2006), it is easy to overlook that gesturing is a highly complex motor skill that appears to serve a variety of functions. Perhaps most notably, *iconic* gestures may provide embodied depictions of what is referenced or contextualized in speech (Kendon, 2004; McNeill, 2005; Streeck, 2008). So-called *beat* gestures are rhythmic movements that stress the part of an utterance that has semantic or emotional salience for the speaker (Leonard & Cummins, 2010). *Pointing* gestures direct attention to objects (imaginary or real) in the environment (Kita, 2003). All these types of gestures not only promote interpersonal understanding, they have also been found to support cognitive processes of the gesturer, such as spatial problem solving and fluent speech production (Kita, Alibali, & Chu, 2017; Pouw, de Nooijer, van Gog, Zwaan, & Paas, 2014). Although it is far from clear how gestures perform such functions, there is one fundamental aspect of gesture that is undisputedly central to its functioning: gestures are performed in synchrony with speech (Kendon, 2004; McNeill, 2005). Without synchrony with speech, gestures would fail to unambiguously point to objects or portray them through depiction, and be meaningless as markers of semantic or emotional salience (Quine, 1968).

Although it is widely accepted that synchrony is fundamental to gesture's functioning, fine-grained quantification of gesture-speech synchrony as it occurs

spontaneously during speaking is currently lacking. There is promising evidence that the moment of "maximum effort" within a gesture is closely timed with the prosodic contrasts made in speech, but such evidence has varying degrees of objectivity and generalizability. Specifically, the evidence is based either on: a) artificial data (i.e., gestures produced by the experimenter (e.g., Leonard & Cummins, 2010), b) pointing gestures that are produced in a repetitive way outside the context of fluid speech (e.g., Rochet-Capellan, Laboissiere, Galvan, & Schwartz, 2008), or c) analyses of video recordings that do not allow for quantification of kinematic properties of gesture production (Loehr, 2004). To be clear, such research has been crucial in the study of gesture-speech synchrony, but also solicits an important next research objective: A fine grained quantification of the synchrony of spontaneous gesture kinematics relative to speech.

For example, the most promising evidence for gesture-speech synchrony relies on methodology involving experimenter judgments of the intensity of gestural hand movements, the "maximum effort" of a gesture (Loehr, 2012; Wagner, Malisz, & Kopp, 2014). The maximum effort is the supposed to be the moment at which there is an energetic peak in the gesture stroke. However, as Wagner and colleagues (2014) conclude, the concept of maximum effort is an ambiguous spatiotemporal marker of a gesture:

> "[the maximum effort is studied] with varying degrees of measurement objectivity and with varying definitions of what counts as an observation of maximum effort. Most definitions evoke a kinesthetic quality of effort or *peak effort* (Kendon, 2004) correlated with abrupt changes in visible movement either as periods of movement acceleration or strokes (Kita, van Gijn, & van

> der Hulst, 1998), as sudden halts or *hits* (Shattuck-Hufnagel, Veilleux, & Renwick, 2007), or as maximal movement extensions in space called *apexes* (Leonard & Cummins, 2008)." p. 221 (original emphasis)

As such, there is a need for a more fine-grained quantification of spatio-temporal properties of gesture in the form of specific *measurable* energetic peaks (e.g., peak acceleration, peak velocity). Such energetic peaks may provide the much sought after "anchor point" in gesture, the property of gesture that supposedly couples to a property of speech, thus creating synchrony. In the current pre-registered and exploratory data report, we provide preliminary evidence for key objective anchor points to study gesture and speech synchrony, and propose a larger scale replication of our findings. This should provide a novel quantification of temporal coordination of spontaneous gesture and speech. In addition to fundamental insights about how speech and gesture arise, the applied importance of quantifying synchrony of gesture and speech is immediately evident for the field of psychopathology and speech pathology. Such fields have already attempted to relate measures of gesture-speech synchrony to the diagnosis of certain pathologies (e.g., De Marchena & Eigsti, 2010). Other immediate applications of reliable quantifications of synchrony could one day be found in education (Ianì, Cutica, & Bucciarelli, 2017) and deception research (Pérez-Rosas, Abouelenien, Mihalcea, & Burzo, 2015).

Previous Research

A short review of the most relevant research on gesture-speech synchrony is provided below (for a comprehensive review see Wagner et al., 2014). The review is

focused on objective quantifications of synchrony for three major families of gesture, namely, beat, iconic, and pointing gestures.

Beat gestures

Beat gestures are rhythmic baton-like co-speech hand movements that typically consist of a short-burst of an up-down movement (Leonard & Cummins, 2008; McNeil, 2005). These gestures convey no obvious iconic semantic content; they do not depict. Rather, these gestures can provide a meaningful prosodic context for speech, by kinesthetically stressing certain utterances during discourse. Beat gestures are an ideal test bed for assessing synchrony, as this is typically understood to be their main feature, and are relatively stereotypical (although not invariant) in their movement (as compared to iconic gestures for example).

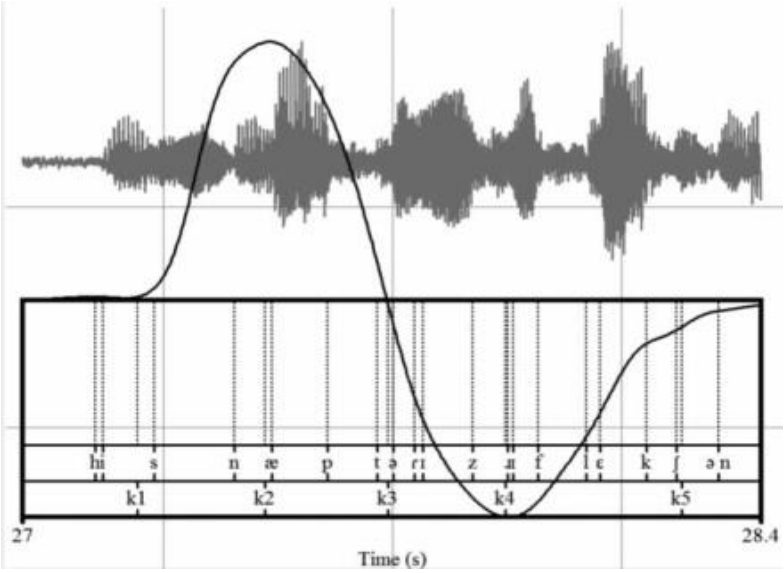
Leonard & Cummins (2010) found that people tend to perceive other's beat gestures as asynchronous or "out of beat" with speech, if a beat gesture is produced after the occurrence of peak pitch of the concomitant speech segment. Subjects were less sensitive to gesture-speech asynchronies when beats led peak pitch in time (see also Treffner, Peter, & Kleidon, 2008). Peak pitch refers to the maximum that is reached in the fundamental frequency of speech (*F0*; i.e., pitch)[1] during a (gesture-speech) event, and such rises in pitch are reliable markers of prosodic stress (Hewlett & Beck, 2013). Leonard & Cummins (2010) further hypothesized that this asymmetry with regards to perceptions of synchrony would

[1] Pitch refers to the frequency (per second) with which a soundwave produced by voiced speech event completes it cycle (Hewlett & Beck, 2013). Rises in pitch occur when the vocal chords are tensed. Next to prosodic prominence, rises in pitch are highly correlated with speech intensity.

be present in production of beat gestures as well. In a ground breaking motion-tracking study on gesture production, the experimenters repeatedly produced a "beat gesture [as] to coincide, in as natural a manner as possible" with speech (Leonard & Cummins, 2010; p. 1465) while tracking the motion of their own hand. From the motion-tracking data, several key events were extracted (see Figure 2). They found that movement extremum and peak velocity in gesture, preceded the peak in pitch. Beat gestures typically lasted 1 second from onset to termination, with onset of gesture leading the onset of stressed syllable by about 300ms. Peak velocity of the upward movement was most closely aligned with peak pitch (with a lead of peak velocity of about 50-60 ms)[2].

[2] Exact means and variances for timing of gesture with respect to peak pitch were not numerically reported, which we therefore estimate from the graphs.

Figure 1. Beat gesture velocity trace and pitch



Note. Reproduced figure from Leonard & Cummins showing an example of a single velocity trace of a beat gesture (2010; permission for reproduction granted from publisher) showing that the peak velocity (*k2*) synchronized with peak pitch when pronouncing the word "snapped". *k1* = onset, *k3* = point of maximum extension, *k4* = peak velocity, *k5* = point of termination.

As Leonard & Cummins (2010) acknowledge, their research setting was artificial in nature, as gestures were produced in succession by the experimenter as invariantly as possible, and with awareness of the relationship under study (i.e., not a situation of spontaneous gesticulation). However, similar observations have been made qualitatively in more natural contexts as (Loehr, 2004) and have recently been observed in a comprehensive quantitative video-analysis study (Loehr, 2012) where conversations between two interlocutors were examined. Loehr (2012) used an annotation procedure to uncover different phases in speech intonations, most notably pitch accent. Gesture phases were also annotated by identifying the time at which the maximum effort occurred;

multiple raters visually judged maximum effort based on video recordings. Maximum effort was defined in line with Kendon's (2004) terminology as the "kinetic 'goal' of the gesture stroke" (p. 77, Loehr, 2012). Based on these annotations, it was observed that both beat and iconic gestures were indeed closely coordinated with intonation phases. More specifically, it was estimated based on the timing data derived from the annotations that the gesture apex occurred 17ms before the pitch accents, with a relatively large standard deviation of about 270ms. These relatively high standard deviations further confirm initial qualitative findings from McClave (1994) who suggested that gesture and speech are not perfectly synchronized – gesture movement peaks occur roughly around prosodic peaks in speech.

Iconic gestures

Iconic gestures are those co-speech hand movements that present an idea, event, or object through iconic resemblance between movement and some referent (Streeck, 2008). Such resemblance relations can be more or less abstract. For example, one can use a rolling motion of the hand with the utterance, "the ball *rolled* down the hill," (concrete relation) or sweep the hand from left to right with the utterance, "I am not liking the idea" (metaphorical relation).

Given that iconic gestures have a semantic dimension to them, the temporal coordination of iconic gestures and speech has primarily been analyzed on this level. That is, gesture researchers have related the timing of a gesture event with the segment of speech that is about the same thing as the gesture – this speech segment is the so-called "lexical affiliate" of a gesture (e.g., Church, Kelly, & Holcombe, 2014; Morrel-Samuels &

Krauss, 1992). A lexical affiliate need not always be identifiable and often gestures may present something novel that is not explicitly referred to in speech.

As an example of lexical affiliate research, consider a study by Church and colleagues (2014) in which participants either: a) gestured while verbally explaining a specific task, or b) physically demonstrated the task while verbally explaining it (e.g., how to throw a dart). The onset of each gesture (throwing gesture) relative to its lexical affiliate (the spoken word "throw") was determined from analysis of video. Likewise, the onset of each demonstrated action (i.e., throwing the dart) relative to the lexical affiliate ("throw") was also determined. The onsets of a gesture were interpreted as indicating the inception of an idea reflected in gesture, and should be carefully distinguished from either the point where the gesture reaches its highest velocity, or its final rest point after the key stroke. On average, they found that gesture and speech onset differed by about 593 ms, where gesture onset precedes the onset of the lexical affiliate 89% of the time. By contrast, in the demonstration condition, action onset and the lexical affiliate had much more varied timing, suggesting that speech and gesture are indeed synchronized to an exceptional degree.

Despite its obvious semantic component, iconic gestures still seem to synchronize on the prosodic level of speech as well (Kendon, 2004; McNeil, 2005). Although no motion-tracking studies have been done, Loehr (2012) does provide evidence that maximum effort in gestures in general synchronizes closely with accented speech (see also Mendoza-Denton & Jannedy, 2011; Shattuck-Hufnagel et al., 2007). However, possible temporal differences with respect to speech between beat gestures and iconic gestures were not

directly assessed, nor statistically tested. Thus, the extent to which beat and iconic gestures synchronize in similar ways to speech remains an open question.

Pointing gestures

Pointing gestures (or "deictic" gestures) have been the most extensively studied with motion-tracking methods. Pointing gestures may direct attention to things in the environment, and may further be used in discourse to organize imagined referents in space (Kita, 2003). Pointing gestures, in most cultures, take the form of an extended index finger which is intentionally aligned toward a relevant object or space.

Using motion tracking, Rochet-Capellan and colleagues (2008) assessed how pointing gestures are coordinated with orofacial articulation in a naming task. Participants pointed to a target while verbally producing its artificial CVCV name (i.e., consonant-vowel-consonant-vowel; e.g., baba) with a stress on the first syllable or the second syllable (e.g., **bá**ba vs. ba**bá**). It was found that the gesture's movement extremum (i.e., referred to by the authors as the pointing apex) was synchronized with the maximum jaw aperture (jaw apex) for the stressed syllable (i.e., first or second depending on condition). Gesture onset led jaw apex by about 149-223 ms, but the gesture apex actually followed the jaw apex for the stressed syllable by about 11-129 ms (see Esteve-Gibert & Prieto, 2014) for related findings from video analysis, showing close alignment of peak pitch with pointing apex).

Further, in a similar paradigm, Rusiewicz, Shaiman, Iverson, & Szuminsky (2014; see also Krivokapić, Tiede, Tyrone, 2017) found that the total time for a pointing gesture to reach a stop is dependent on contrasting pitch accents of the target word. When there was stressed intonation versus no intonation in the spoken sentences, total pointing gesture time (gesture onset to gesture apex) was increased. Additionally, in half of the trials

participants heard their own speech with a 200 millisecond delay, which typically elongates spoken responses. It was found that the time between gesture onset, and the vowel-to-vowel midpoint of the gesturally referenced target word was increased, as compared to when speech was affected by delayed auditory feedback (although gestures were lengthened overall when speech was lengthened). Thus gesture adjusted to the elongation of speech in the perturbed conditions, almost to extent of full compensation (see also Chu & Hagoort, 2014).

It should be noted that the pointing gestures that have been studied above are a rather exotic character in the family of gesture. This is because they are referring to something in the immediate environment, and the pointing movement is continuously controlled by visual perception of the target (Chu & Hagoort, 2014). However, pointing gestures that occur during spontaneous speech are recruited to organize objects in imagined space, and therefore such narrative pointing gestures are not visually controlled by the environment. It is therefore an open question to what extent current research on pointing gestures reveals the dynamics of spontaneous narrative pointing gestures.

Summary and Current Approach

The short review above (and our assessment of the broader literature) yields several calls for research. Firstly, research on the synchrony of *spontaneous* gestures with fluid speech has still to fully utilize motion-tracking methodology. As Wagner and colleagues (2014) conclude: "…there are no studies on fluid, continuous gestural movements, and the way in which they are aligned with continuous pitch contours" (p. 223). Indeed, the predominant method has been video analysis which yields relatively low resolution, and suboptimal identification of movement properties, as they need to be

identified by eye. This has also led to what is now an ambiguous anchor point in gesture - the maximum effort. The few studies that do provide high kinematic detail (e.g., Leonard & Cummins, 2010; Rochet-Capellan et al., 2008) have been done in artificial environments and primarily cover exogenously controlled pointing gestures, excluding spontaneous gestures. These initial studies were, of course, a sensible starting point for research and have provided important insights about potential markers of coupling for gesture (peak effort) and speech (peak pitch), but it is currently unclear how the gesture-speech system operates in a noisy environment. That is, we have remarkably little evidence about how gesticulation actually behaves in the wild. These first assessments thus call for: a) higher kinematic resolution of studying gesture together with speech, b) in more ecologically valid contexts, and c) with a more objective identification of possibly relevant energetic peaks as anchor points for gesture and speech.

Subjects in the current exploratory study (N = 4) retold the narrative of a cartoon they had just watched, a common gesture-elicitation method (McNeill, 2005), which yielded about 230 gesture events. We employed high resolution motion-tracking of the dominant hand (240 Hz) during narration (non-dominant hand was not used for gesturing). From the movement time series, we identified energetic peaks during each gesture event (peak velocity, peak acceleration, peak deceleration), providing an objective measurement of gesture kinematics. Gesture identification was performed using ELAN (Lausberg & Sloetjes, 2009) so as to categorize different gestures, and to define the onset of a gesture based on assistance of hand-movement time series (see method and Crasborn, SLoetjes, Auer, & Wittenburg, 2006). Similar to previous studies (e.g., Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2008), we further extracted pitch (*F0*) from acoustic

data (using PRAAT, Boersma, 2001) so as to identify peaks of pitch within relevant gesture-speech events, which we show is a reliable anchor point for gesture-speech synchrony. Gesture-speech synchrony was quantified by the difference (*D*) in milliseconds between peak pitch and the relevant gesture anchor points (e.g., gesture onset, peak velocity). In the current study, we focused on three major gesture types, namely beat, iconic, and narrative pointing gestures. This exploratory study will allow us to answer a host of classic questions that have not been quantitatively studied to the current extent, including: What reliable kinematic anchor point in a gesture event is most closely synchronized with peak pitch? How strong is the synchrony between gesture and speech? Do beat, iconic, and pointing gestures differ in gesture-speech synchrony? To what extent are there individual differences in gesture-speech synchrony? Our initial results here provide strong hints about the answers to these issues.

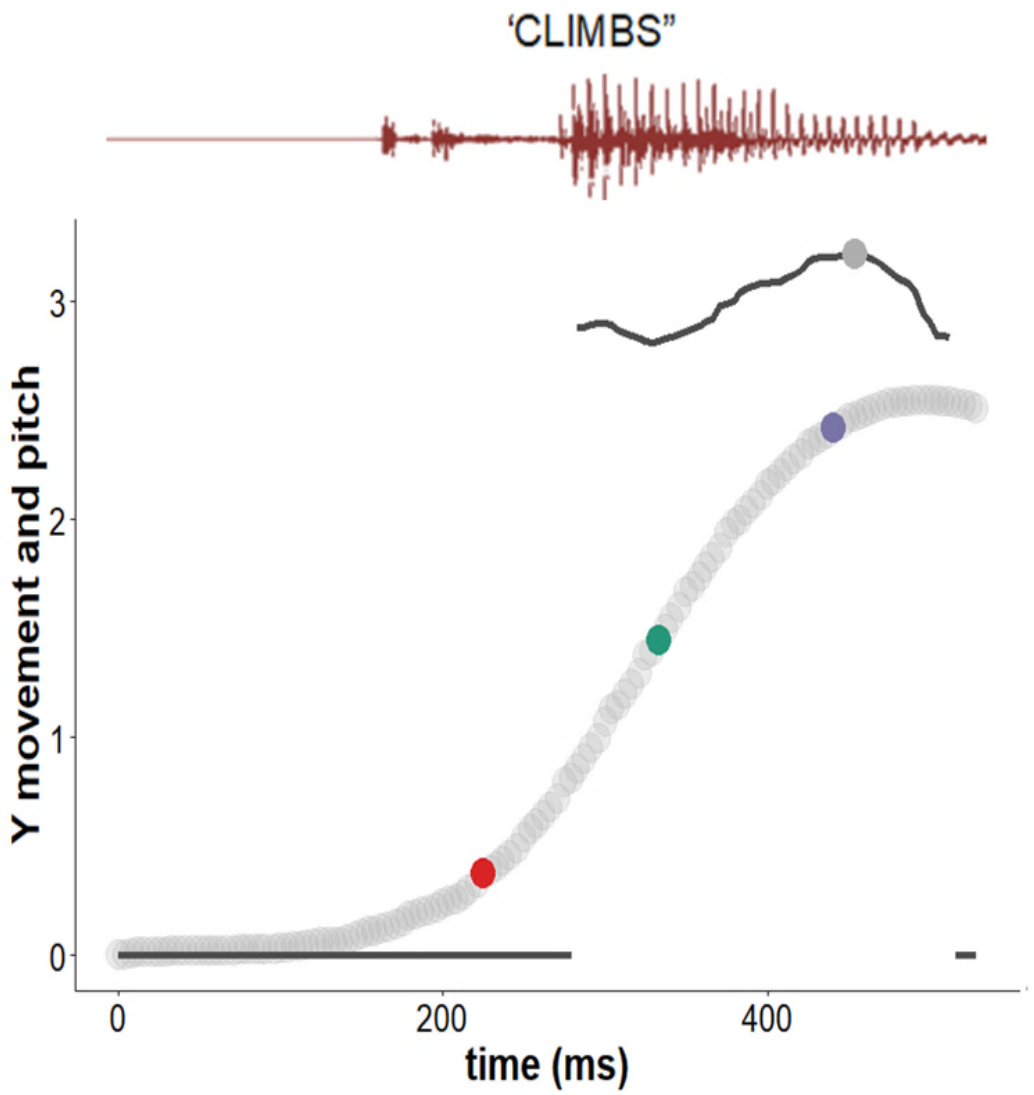
Method

Due to word limit constraints, a detailed method section is included as supplemental materials. The extended method section includes experimental procedure, apparatus information, and data preparation and aggregation procedures.

Four male right-handed graduate students at the University of Connecticut participated in this study (ages = 30, 38, 23, 34). Two participants were native speakers of American English and two were native speakers of Spanish with high proficiency in spoken and written English. In total, we collected movement and speech data from about 15 minutes of narration.

For each gesture-speech event, the peak velocity, peak acceleration, peak deceleration, and peak pitch of speech were extracted by a custom-written function in R. Figure 2 shows an example of the peak-finding results for the "CLIMBS the wall" iconic gesture also referenced in Figure A of the supplemental materials.

Figure 2. Visual example peak extraction method



Note. Example of change y-axis position (grey) and pitch track (black) over time (ms; centered and scaled) for the "CLIMBS the wall" gesture (see supplemental materials). Red dot = peak acceleration, green dot = peak velocity, purple dot = peak deceleration, solid grey dot = peak pitch. We have super imposed the raw sound waveform in red above. The pitch (F0) reflects the vocal fold opening at pronouncing the "I" in "climbs".

Results

Descriptive

A total of 231 gesture events were observed (beat = 152, iconic = 44, pointing = 31, undefined/abandon = 4). Average time for gesture events was 829 ms (*SD* = 602 ms); beat gesture *M* = 739 (*SD* = 398), iconic gesture *M* = 947 (SD = 789), pointing gesture *M* = 667 (*SD* = 443). Table 1 provides an overview of the production rates of the different gestures, as well as speech rate (spoken words per minute narration). It is important to note that these gesture ratios are very comparable to other studies that have used the same retelling of cartoon procedure (see e.g., McNeill, 2005, p. 42, where a comparable 41% of iconic gestures was found). This serves as evidence that in the current sample the glove and measuring apparatus did not seem to greatly alter spontaneous gesture tendencies.

Table 1. Gesture and speech rates

PPN	**Beat p/m**	**Iconic p/m**	**Pointing p/m**	**Other p/m**	**Pitch (F0) Mean**	**Speech rate p/m**	**Time of Narration**
1	18.6	8.5	2.7	0.6	103.77	395	5.17
2	3.6	12.7	0.3	0.3	106.84	143	3.08
3	10.9	9.8	3.6	2.5	107.46	125	2.76
4	3.8	3.3	1.5	2.0	107.09	151	3.99
M (*SD*) %	9.22 *(7.11)* 43%	8.57 *(3.93)* 40%	2.03 *(1.44)* 10%	1.43 *(1.07)* 7%	105.9 *(11.3)*	204 *(128)*	3.74 *(1.09)*

Note. The gesture and speech rates (words spoken) are given for frequency *per minute* (p/m). "PPN" refers to participant number.

Gestures and Peak Pitch

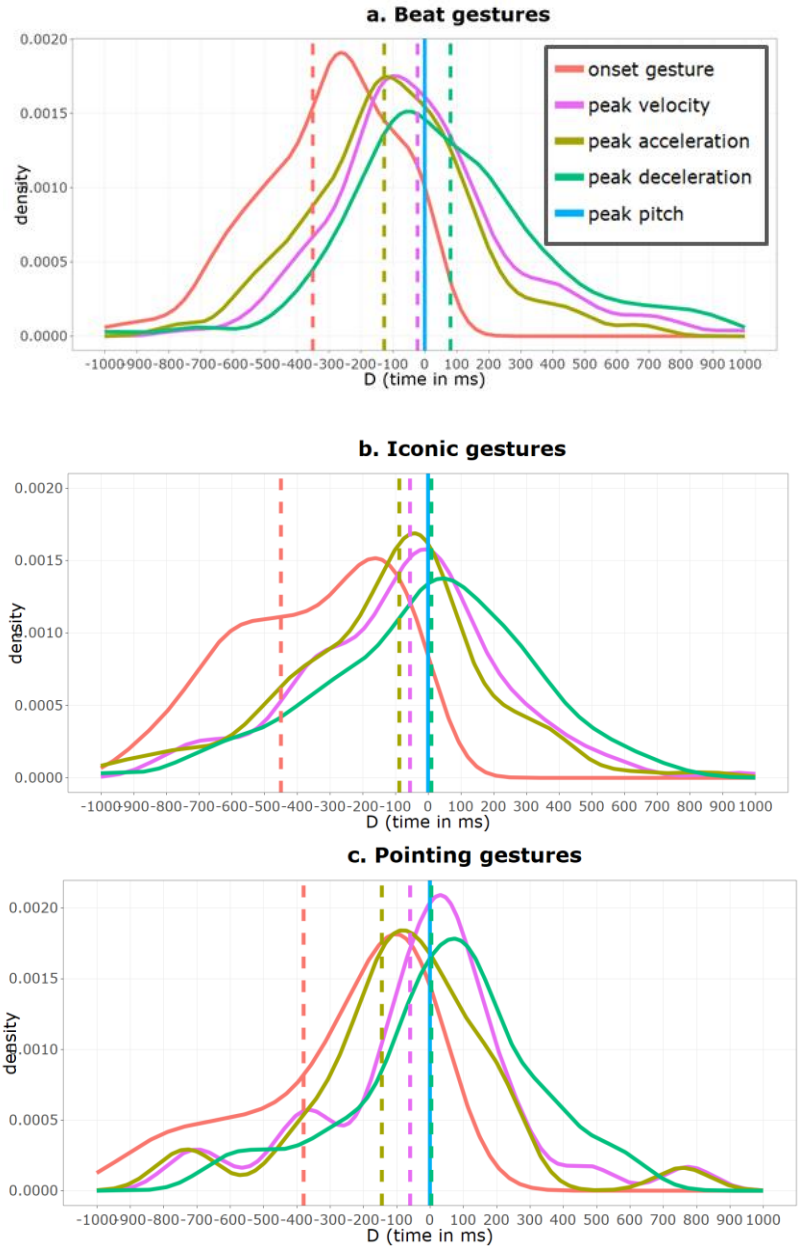
We first assessed the temporal relation between speech (peak pitch) and properties of gesture. Table 2 shows the mean difference in milliseconds, *D*, between the peak pitch and the different kinematic properties of gesture - gesture onset, peak velocity, peak acceleration and peak deceleration, for each gesture type separately. Figure 3 shows the relative frequency distributions of *D* for these gesture properties relative to peak pitch (which defines the zero point).

Table 2. Mean Difference, *D*, in milliseconds (peak pitch – gesture property)

Kinematic property	BEAT	ICONIC	Pointing	F-test Diff *p* (corrected) Bayes Factor
Onset				
M (SD)	-349 *(314)*	-449*(426)*	-379 (427)	*F* (2, 6) = 1.28
95% CI [lower, upper]	[-400, -299]	[-526, -374]	[-536, 223]	*p* > .99 BF_{01} = 3.04
Peak velocity				
M (SD)	-22 *(334)*	-55 *(589)*	-59 (349)	*F* (2, 6) = 1.39
95% CI [lower, upper]	[-75, 31]	[-159, 49]	[-185, 67]	*p* > .99 BF_{01} = 3.08
Peak acceleration				
M (SD)	-126 *(377)*	-87 *(630)*	-144 (380)	*F* (2, 6) = 0.612
95% CI [lower, upper]	[-187, -66]	[-199, 23]	[-282, -7]	*p* > .99 BF_{01} = 7.69
Peak deceleration				
M (SD)	81 *(367)*	10 *(499)*	4 (344)	*F* (2, 6) = .718
95% CI [lower, upper]	[22, 138]	[-78, 98]	[-120, 129]	*p* = .104 $BF_{01 =}$ 0.22

Note. P-values are Bonferroni adjusted for four comparisons. We have also computed Bayes Factors (BF) with non-informative default prior widths p(M) = 0.5 using R package "BayesFactor" (Rouder, Morey, Speckman, & Province, 2002) as to provide a measure of evidence for the null-hypothesis over the alternative hypothesis. The BF's are reported for the likelihood of the observed data provided the null-hypothesis true (no differences per gesture type) versus the alternative hypothesis (BF_{01}). If BF_{01} > 3 this can be treated as moderate/substantial evidence (3-10 as strong evidence; see Rouder, Morey, Verhagen, Swagman, Wagenmakers, 2016), and can be read as the observed data being 3 (or more) times more likely under the null-model versus the alternative model.

Figure 3. Distribution of D's: Gesture properties relative to peak pitch



Note. Frequency distributions of *D* for each gesture property. D is the difference in the timing of that gesture property relative to timing of peak pitch (blue line at zero). The peak of the distributions are the *mode* of *D*. The dotted lines are *mean D*. Negative values of *D* indicate that the gesture property occurred before peak pitch. As can be seen, gesture properties generally seem to lead peak pitch in time.

A flat distribution curve of *D* would be an indication of a random occurrence of a kinematic property of gesture with regards to peak pitch. We obtain a clearly non-uniform distribution of *D* for beat, iconic, and pointing gestures, showing an impressive temporal coupling between gesture and speech prosody. Furthermore, the data show that gesture's peak velocity, peak acceleration and gesture onset, all lead peak pitch in time (and is followed by peak deceleration). Gesture onset and peak acceleration are clearly not the point at which gestures synchronize with peak pitch. For each gesture property separately (i.e., onset, peak velocity, peak acceleration, peak deceleration), we performed a within-subjects ANOVA to assess differences in *D* between each gesture type (3 levels: beat vs. iconic vs. pointing gesture events; see Table 2). In the current sample, we did not find statistically significant differences between gesture types for D^3. This suggests that all the gestures types addressed here in this exploratory sample are roughly comparable in the degree to which they synchronize with peak pitch. The Bayesian Analyses further show that the observed data were 3 times or more likely under the null-hypothesis (absence of effect of gesture type) for gesture onset, peak velocity, peak acceleration. However, for peak deceleration we did not find substantial evidence for the null-model, suggesting that peak deceleration may differ in *D* between gesture types (when tested with larger samples).

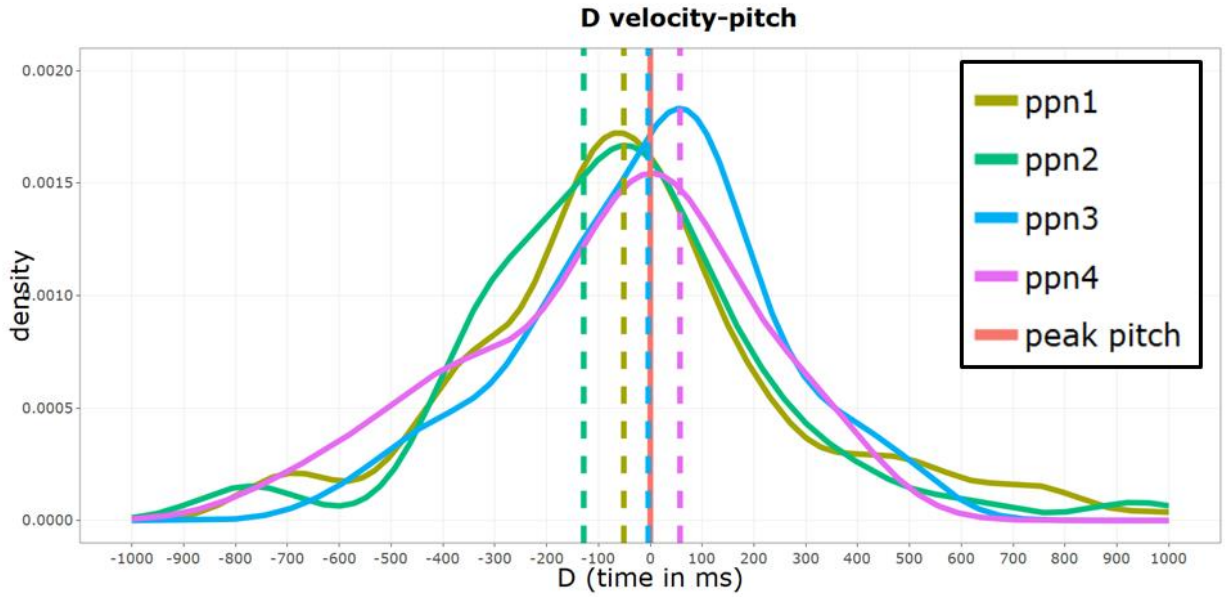
[3] Throughout the manuscript all reported p-values are corrected for multiple comparisons when appropriate and are based on two-tailed tests.

A further question that arises is whether there is one particular gesture property that is most closely coordinated with peak pitch in speech. Since we did not find reliable statistical differences in *D* between gesture types, we collapsed all beat, iconic, and pointing gesture events for the following analyses. With this combined data, we performed a within-subjects ANOVA with gesture property (peak velocity vs. peak acceleration vs. peak deceleration) as a within-subjects variable, and *D* as the dependent variable. We found that these gesture properties differed reliably in their *D*'s, F (2, 6) = 17.54, $p < .001$. Paired post-hoc comparisons (p-values Bonferroni corrected) revealed that peak velocity shortly led peak pitch (M_D = -39, SD_D = 454, 95%CI[-90: 11]), as compared to peak deceleration which followed peak pitch ($p < .001$; M_D = 44 , SD_D = 424, 95%CI[-3 : 92]). Peak acceleration was furthest from peak pitch (M_D = -113, SD_D = 494, 95%CI[-168 : -58]), and was statistically different from peak velocity and peak deceleration (*ps* < .001). As can be seen, both peak velocity and peak deceleration have 0 in their confidence intervals, suggesting that both closely synchronize with peak pitch, with peak velocity shortly leading (39 ms), and peak deceleration shortly following (44 ms) peak pitch.

It is further informative to provide a descriptive overview of individual differences in gesture-speech synchronization. Because the *D* for peak velocity seems to be a good general marker of gesture-speech synchronization for all gesture types combined, we have plotted in Figure 4 the *D* for peak velocity separately for each participant for all gesture events. As can be seen from Figure 4, each participant showed synchronization (judging from non-uniformity of the distribution curves of *D*), with some individual variability in synchronization (SD_D = 87 ms). However, there were no statistical differences between participants' *D* for peak velocity (p = .197; BF_{01} = 40.21). Note, that the current sample

involves only four subjects and both native and non-native speakers. A more comprehensive understanding of individual differences (and possible effects of language proficiency) would require a larger sample size.

Figure 4. Distribution of *D*'s for peak velocity relative to peak pitch



Note. Individual differences for frequency distributions of *D* for peak velocity relative to peak pitch for all gestures combined.

Discussion

This exploratory study has provided the following preliminary implications with regards to classic questions in gesture research. These implications should be regarded as tentative; our goal is to replicate and extend these findings in a larger sample (see method and pre-registration).

Firstly, gesture-speech synchrony is obviously occurring, as indicated by clear peaks in the distributions of difference in timing (*D*) between peak pitch and kinematic gesture properties. This synchrony with speech is remarkable given that beat, iconic, and pointing gestures each serve different functions. The current results suggest that regardless of their role in discourse, all gestures tend to emerge in synchrony with speech. However, it is clear from the relatively large standard deviations of *D* that gesture-speech synchrony is not a 1-1 coupling, suggesting a more loose temporal relation between gesture and speech (Loehr, 2004, 2012; McClave, 1994].

Secondly, we have disambiguated gesture's anchor point with speech, by objectively assessing which energetic peak in manual movement most closely aligns with energetic peak pitch. Most clearly, gesture onset, and peak acceleration are not most closely synchronized with peak pitch. For all gestures, peak velocity is closely synchronized with peak pitch (gestures lead speech with 39 milliseconds), but most notably for beat gestures. For iconic and pointing gestures peak deceleration could also be a good anchor point for studying gesture-speech synchronization.

Finally, evidence of synchronization is obtained for all participants, which is interesting as the current sample contained both native and non-native speakers of

American English. However, more data is needed to assess the variability of gesture-speech synchrony for native and non-native speakers.

In addition to these preliminary conclusions, this exploratory study serves as a proof of concept for a follow-up replication study. First, the current motion-tracking apparatus did not dramatically impede gesture tendencies. We observed rates of gesturing that are comparable to previously reported video-data research. Informally, we observed that participants seemed to gesture just as they would normally. A second proof of concept is that we can process a relatively large amount of data by automated means. For example, peak pitch seems to be a good anchor point of speech for gesture. Notwithstanding that other more advanced linguistic annotation techniques of prosody may provide more depth of analysis (Loehr, 2012), these techniques are very time consuming and require a subjective experimenter identification, which is one of the reasons such time-intensive micro-analysis have been undertaken with relatively small segments of data (e.g., Loehr, 2012 analyzed video data for 146 seconds). Thus with the current method we can process a lot of gesture-speech events, thereby allowing for a larger-scale quantification of gesture-speech synchrony. Using the current methodology, we propose to collect more data to provide a comprehensive quantification of the temporal coupling between gesture and speech. We further aim to explore possible relations of gesture-speech synchrony as function of language proficiency.

Pre-registration Proposed Study

Ethics Statement

This proposed study will be performed in accordance with the ethical guidelines established by the Institutional Review Board of the University of Connecticut.

Research Objective, Sample size & Power

The main objective is to see whether the observed synchrony between speech and gesture can be replicated in a larger sample. Specifically, we aim to provide a reliable point estimate of the degree of synchrony (*D*) between gesture and speech. Given that peak velocity is a reliable anchor point for all gesture types combined (smallest *D* to peak pitch in exploratory data), we will provide a population mean estimate for this *D*. Based on a required 95% confidence level (α = .05) and an estimate of the population standard deviation derived from our exploratory data (observed *SD* of synchrony for peak velocity and peak pitch was 334 ms), we can compute the required sample size, assuming we are sampling from a normal distribution, and allowing for a two-sided error boundary of 100 ms:

$$\text{Minimal Required } n = \left(\frac{Z\alpha/_2 \, \sigma}{E}\right) = \left(\frac{1.95 \times 334}{200}\right) = 10.713$$

Participants

We will recruit 12 undergraduates (6 female/6 male) of the University of Connecticut under the condition that they are monolingual native speakers of American English. Additionally, for exploratory purposes we will recruit an additional of 12 undergraduates (6 female/6 male) of the University of Connecticut under the condition that they are native speakers of Spanish and learned American English as their second

language. This dataset should provide us with at least 1000 gesture events from about 80 minutes of narrative based on the gesture rates of our exploratory data, which would produce the most comprehensive dataset on gesture-speech synchrony currently available.

Procedure and Analyses Plan

Participants will receive course credit or a small monetary reward for their participation. We propose to directly replicate the current methods and analyses as reported in the exploratory data report (with a slight amendment to the gesture analysis procedure; see below). We do not foresee any reasons for exclusion of participants, but will report any exclusions should they be made. We will also perform additional analyses that provide insight into the effect of such exclusions on our interpretations.

The exact data manipulation- and statistical analyses procedures performed for the exploratory data will be re-performed with the current proposed study. Note, that the analyses are documented in the R data preparation and analysis scripts that are provided with this pre-registered report. All exploratory analyses and exploratory covariates, such as possible differences of gender or language proficiency will be reported as such in the final research report. All data should be collected within 4 weeks (annotation procedures and analyses will take an additional 4-8 weeks).

Gesture Analysis

For 5 randomly chosen participants (ca. 20% of the video data), beat, iconic, and pointing gestures will be annotated by a second rater. We will compute and report a modified Cohen's Kappa (Holle & Rein, 2013) as provided by ELAN Annotation software (Lausberg & Sloetjes, 2009) to assess reliability of gesture annotation between multiple raters. Following common standards, if reliability fails to reach .75 or higher, annotation

procedure will be reevaluated and adjusted and the data will be recoded to reach a higher convergence between raters.

Data Availability

All the quantitative data of the proposed study (excluding the video data) will be made publically available at the Open Science Framework (https://osf.io/5ja6y). Due to privacy restrictions (that are not applicable to the exploratory data) we cannot share the raw sound and video data from the participants, but we will share all quantitative data.

References

Boersma, P. (2001). PRAAT, a system for doing phonetics by computer. *Glot International* 5 (9/10), 341-345.

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

Church, R. B., Kelly, S., & Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Language, Cognition, & Neuroscience, 29*(3), 345-354. doi: 10.1080/01690965.2013.857783.

Cooperrider, K. (2018). Foreground gesture, background gesture. *Gesture*, *16*(2), 176-202. doi: 10.1075/gest.16.2.02coo.

Crasborn, O., Sloetjes, H., Auer, E., & Wittenburg, P. (2006). Combining video and numeric data in the analysis of sign languages with the ELAN annotation software. In C. Vetoori (Ed.), *Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios* (pp. 82-87). Paris: ELRA.

de Marchena, A., & Eigsti, I. M. (2010). Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. *Autism Research, 3*(6), 311-322. doi: 10.1002/aur.159.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, *56*(3), 850-864. doi: 10.1044/1092-4388.

Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience & Biobehavioral Reviews, 30*(7), 949-960. doi: 10.1016/j.neubiorev.2006.02.004.

Hewlett, N., & Beck, J. M. (2013) *An Introduction to the Science of Phonetics*. New York: Routledge.

Holle, H., & Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. In H. Lausberg (Ed.), *Understanding Body Movement. A Guide to Empirical Research on Nonverbal Behaviour. With an Introduction to the NEUROGES Coding System*. (New York: Peter Lang).

Ianì, F., Cutica, I., & Bucciarelli, M. (2017). Timing of gestures: Gestures anticipating or simultaneous with speech as indexes of text comprehension in children and adults. *Cognitive Science, 41*(6), 1549-1566. doi: 10.1111/cogs.12381.

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.

Kendon, A. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press, 2004.

Kendon, A. Some relationships between body motion and speech. In A. Siegman, B. Pope, (Eds.), *Studies in Dyadic Communication*, pp. 177–210 (New York: Pergamon, 1972).

Kita, S. (2003). *Pointing: Where Language, Culture, and Cognition Meet*. New Jersey: Lawrence Erlbaum

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. doi: 10.1037/rev0000059.

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.), *Gesture and Sign Language in Human-Computer Interaction*, Proceedings International Gesture Workshop Bielefeld, Germany, September 17-19, 1997. doi: 10.1007/BFb0052986.

Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology*, *8*(1), 1-36. doi: 10.5334/labphon.75.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, *41*(3), 841-849. doi: 10.3758/BRM.41.3.841.

Leonard, T., Cummins, F. (2010). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. doi: 10.1080/01690965.2010.500218.

Loehr, D. P. (2004). Gesture and intonation (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, *3*(1), 71-89. doi: 10.1515/lp-2012-0006.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research*, *23*(1), 45-66. doi: 10.1007/BF02143175.

McNeill, D. *Gesture and Thought*. Chicago: University of Chicago press, 2005.

Mendoza-Denton, N., & Jannedy, S. (2011). Semiotic layering through gesture and intonation: A case study of complementary and supplementary multimodality in

political speech. *Journal of English Linguistics*, *39*(3), 265-299. doi: 10.1177/0075424211405941.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615. doi: 10.1037/0278-7393.18.3.615.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., & Burzo, M. (2015).Deception detection using real-life trial data. In Z. Zhang & P. Cohen, *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (pp. 59-66). doi: 10.1145/2818346.2820758.

Plotly Technologies Inc. (2015). Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC. URL: https://plot.ly

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*, 1-14. doi: 10.3389/fpsyg.2014.00359.

Quine, W. V. O. (1968). Ontological relativity. *Journal of Philosophy*, *65*, 185–212.

Rochet-Capellan, A., Laboissiere, R., Galvan, A., Schwartz, J. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research*, *51* (6), 1507–1521. doi: 10.1044/1092-4388.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2002). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356-374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, & E. J. Bayesian (2016). analysis of factorial designs. *Psychological Methods*, *22*, 304-321. doi: 10.1037/met0000057.

Richardson, M. (n.d.). Retrieved from http://xkiwilabs.com/software-toolboxes/)

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. doi: 10.1016/j.specom.2013.06.004.

Shattuck-Hufnagel, S., Y., Y., Veilleux, N., & Renwick, M. , (2007). A method for studying the time alignment of gestures and prosody in American English: 'hits' and pitch accents in academic-lecture-style speech. In A. Esposito, M. Bratanic, E. Keller, M. Marinaro, (Eds.), *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*. NATO Security Through Science series E: Human and Societal Dynamics, vol. 18. Amsterdam: IOS Press.

Streeck, J. (2008). Depicting by gesture. *Gesture 8* (3), 285-301. doi: 10.1075/gest.8.3.02str.

Treffner, P., Peter, M., & Kleidon, M. (2008). Gestures and phases: The dynamics of speech-hand communication. *Ecological Psychology, 20* (1), 32-64. doi: 10.1080/10407410701766643.

Wagner, P., Malisz, Z., & Kopp, S (2014). Gesture and speech in interaction: An overview. *Speech Communication, 57*, 209-232. doi: 10.1016/j.specom.2013.09.008.

Willems, R. M., Özyürek, A., & Hagoort, P. (2006). When language meets action: The neural integration of gesture and speech. *Cerebral Cortex*, *17*(10), 2322-2333. doi: 10.1093/cercor/bhl141.

Competing interests

The authors have no competing interests to report.

Author Contributions: WP and JD have designed the study. WP has conducted the exploratory study and analyses with supervision of JD. WP has written the current pre-registered report with critical revisions by JD.

Supplemental Materials

Quantifying gesture-speech synchrony: Exploratory data report and pre-registration

Extended Method Exploratory Study

Procedure

Participants were first equipped with a glove for the dominant hand that allowed us to attach the motion sensor of the Polhemus Liberty via Velcro to the index finger. Then a full clip of Tweety and Sylvester "Canary road" was watched. This cartoon clip is often used in gesture research, which lasts about 350 seconds. Participants were informed beforehand that they would later retell the narrative to the experimenter. The glove was attached prior to watching the video so that the subject could get used to wearing it. After watching the clip, participants were asked to retell the narrative of the cartoon while holding their non-dominant hand in their pocket as the recording equipment was running. No instructions were provided about hand gesturing.

Apparatus

Motion tracking. We used a Polhemus Liberty (Polhemus Corporation, Colchester, VT, USA) with a single motion-sensor collecting 3D position data at 240Hz (~0.13 mm spatial resolution). The motion sensor was attached to the top of the participant's index finger (at the height of the fingernail). This allowed us to capture arm movements together with movements of the wrists and fingers. We recorded the motion of only one hand to simplify data collection and analysis. Since our peak-finding function could be sensitive to small but significant jumps in position data due to noise, we applied a low-pass

Butterworth filter to the position velocity and accaleration traces with a cut-off of 10Hz (e.g., Leonard & Cummins, 2010).

Audio. Instead of using the noisier sound stream of the video camera, we obtained speech data by using a RT20 Audio Technica Cardioid microphone (44.1kHz) which suppresses surrounding noises including any unintended experimenter noise (e.g., coughs).

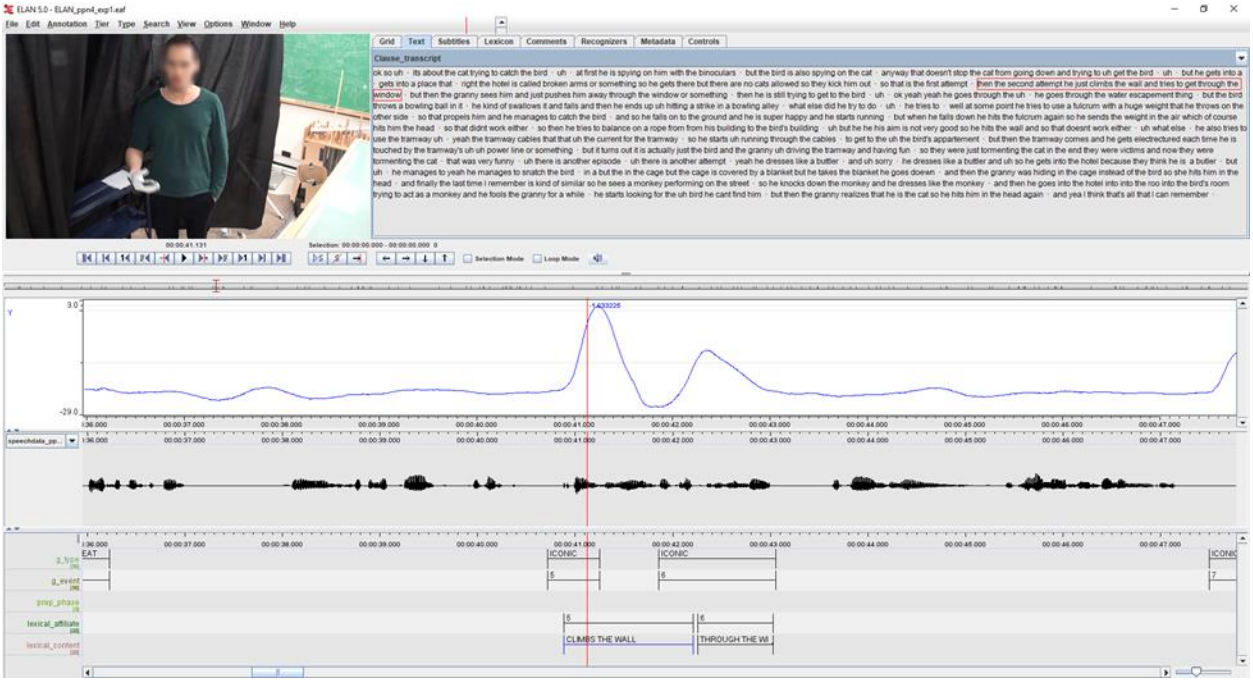
Motion & audio recording. We used C++ to simultaneously call and write audio and movement data. We modified a C++ script made publicly available by Michael Richardson (Richardson, n.d.) in which we included scripts to enable recording of sound from a microphone (using toolbox SFML for C++ https://www.sfml-dev.org/).

Camera. We videotaped participants using Sony Digital HD Camera HDR-XR5504 Recorder, sampling at 29.97 frames per second.

Data Preparation

Gesture annotation phase. In the annotation phase, the first author transcribed speech and identified gesture events (see figure A). For the annotation phase, we loaded in the video data, audio data, as well as the time series of the motion tracking into ELAN (Crasborn et al., 2006). ELAN allows the user to visually present the movement time series along with the video data. As such, the emergence of gesture could be identified based on the actual movement data rather than the lower resolution method of identifying movement on the basis of changes in movement per video frame, which can be difficult (see Figure A for an example of the annotation interface with movement data). As introduced by Crasborn and colleagues (2006), this provides clear advantages over traditional gesture video analysis.

Figure A. Software interface of the linguistic annotation program ELAN



Note. Example shown of a participant in mid gesture. The event where Sylvester "CLIMBS the wall" [lexical affiliate] whereby the participant performed an upward motion reaching the top point, subsequently the gesture is retracted by letting it fall (retraction phase). This is followed by another gesture that depicts a crawling motion with downward trajectory. The blue on white panel shows the movement data of the participant (here movement on the y axis) as collected from the Polhemus Liberty together with speech data and gesture annotation panels.

The procedure of marking a gesture in the current dataset was as follows. Gesture onset was identified by spotting a gesture in the video, categorizing it as either a beat, iconic, or pointing gesture (based on gesture categorization guidelines by McNeill, 2005). In cases where the gesture was not of a clear nature, it was categorized as "undefined"; we

also categorized "abandoned" gestures that were not completed (see e.g., Kita, Alibali, & Chu, 2017). After having spotted a gesture, the experimenter would go back to beginning of the gesture event and seek out the onset of the gesture (first fluent change from static position), on the basis of the time series of the kinematic data (with the use of x and y axis, and velocity trace). The gesture event was marked as ending at the place where the gesture completed its main stroke, thus not including a possible post-stroke hold, and not including a retraction phase. Excluding these optional end-phases of gesture allowed us to ensure that our peak-finding functions do not pick out possible energetic peaks in the retraction phase (which is generally known not to coordinate meaningfully with speech).

Speech Pitch. We extracted pitch time series of the audio recording using PRAAT with default range suitable for males 75-500 Hz[4] (Boersma, 2001). We matched the sampling rate of pitch with that of the motion tracker (1 sample per 4.16 milliseconds).

Speech content. For exploratory purposes, also using ELAN, speech was transcribed and lexical affiliates of iconic gesture were identified if possible, but not when gestures did not clearly refer to what was mentioned in text.

Data aggregation and analysis. We wrote a custom code in R (R core Team 2013) to aggregate the ELAN, PRAAT, and motion tracking data. We interpolated the movement data to match the pitch data with an interpolation function in *R*. Using a custom-made function, we automatically read in ELAN gesture and speech annotation files so that these events were marked in the movement and pitch time series. Plots in this data report were produced using R package Plotly (Plotly Technologies Inc, 2015) and ggplot2 (Wickham,

[4] For the proposed study which will include female participants, we will extract pitch series within a range of 100-500 Hz.

2009). Density plots (Figure 3 and Figure 4) were produced with the ggplot2 "geom_density" plotting function, which draws on the "1d Kernel Density Estimate" function called "stat_density" (R code available at https://osf.io/5ja6y).

Data Availability

All (raw) data, pitch data (PRAAT), annotation data (ELAN), experiment code (C++), data preparation code (R), & analyses code(R) generated for the exploratory study will be publicly available on the Open Science Framework (https://osf.io/5ja6y).