Routledge
Taylor & Francis Group

Discourse Processes

ISSN: 0163-853X (Print) 1532-6950 (Online) Journal homepage: https://www.tandfonline.com/loi/hdsp20

Gesture Networks: Introducing Dynamic Time Warping and Network Analysis for the Kinematic Study of Gesture Ensembles

Wim Pouw & James A. Dixon

DISCOURSE PROCESSES
2020, VOL. 57, NO. 4, 301–319
https://doi.org/10.1080/0163853X.2019.1678967

Routledge
Taylor & Francis Group

Gesture Networks: Introducing Dynamic Time Warping and Network Analysis for the Kinematic Study of Gesture Ensembles

Wim Pouw [a,b] and James A. Dixon[a]

[a]Center for the Ecological Study of Perception and Action University of Connecticut; [b]Department of Psychology, Educational, and Child Studies, Erasmus University Rotterdam

ABSTRACT

We introduce applications of established methods in time-series and network analysis that we jointly apply here for the kinematic study *of gesture ensembles*. We define a gesture ensemble as the set of gestures produced during discourse by a single person or a group of persons. Here we are interested in how gestures kinematically relate to one another. We use a bivariate time-series analysis called dynamic time warping to assess how similar each gesture is to other gestures in the ensemble in terms of their velocity profiles (as well as studying multivariate cases with gesture velocity and speech amplitude envelope profiles). By relating each gesture event to all other gesture events produced in the ensemble, we obtain a weighted matrix that essentially represents a network of similarity relationships. We can therefore apply network analysis that can gauge, for example, how diverse or coherent certain gestures are with respect to the gesture ensemble. We believe these analyses promise to be of great value for gesture studies, as we can come to understand how low-level gesture features (kinematics of gesture) relate to the higher-order organizational structures present at the level of discourse.

Introduction

Hand gestures come in a variety of forms that support their various functions. Gestures can serve to point out objects in the environment, they can signal an object's presence through iconic reference, or they may beat with the rhythm of speech. Despite these different apparent functions, we know that gesture typologies are to some extent artificial. For example, even iconic and pointing gestures, both of which reserve degrees of freedom for referential expression, often still exhibit beat-like functions as they couple with prosodic contrasts in speech (Esteve-Gibert & Prieto, 2013; Shattuck-Hufnagel & Prieto, 2019; Shattuck-Hufnagel & Ren, 2018). Complicating matters further, gesture's role in discourse operates on multiple time scales. Beat gestures may beat with the rhythm of speech (Leonard & Cummins, 2011; McClave, 1994; Wagner, Malisz, & Kopp, 2014); such rhythmic gestures couple with speech on a time scale that transcends single gesture events. In a similar vein, some gestures have recurring forms that are repeatedly used to maintain discourse cohesion (McNeill et al., 2001); such cohesion is established by producing recurrent kinematic features in gestures over the time span of a narrative or discourse. In sum, gesticulation is a complex, multiscale, spatiotemporal phenomenon, which means that multilevel approaches are need to understand its dynamics.

Despite the general acknowledgment by theoreticians that gestures are not isolated events and have a story to tell in terms of their higher-order organization during discourse (Kendon, 2004;

CONTACT Wim Pouw ✉ wimpouw@gmail.com Center for the Ecological Study of Perception and Action, University of Connecticut, 406 Babbidge Road, Unit 1020, Bousfield Building, room 367

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hdsp.

McNeill, 2005), there is currently a lack of quantitative approaches in gesture studies that relate lower-order gestural features with the higher-order organizations of what we might call *gesture ensembles*. We define gesture ensembles here as a researcher-defined set of gestures that can be produced by a person or a group of persons to be investigated for possible structural properties that these gestures possibly have in common. For example, one such prototypical structural property are gestures that are similar in their spatial trajectories, which McNeill (2005) called "catchments." In the current article we aim to innovate the quantitative study of gesture ensembles. We hope therefore to fortify ongoing innovations in the quantitative study of gesture kinematics together with speech dynamics (Alviar, Dale, & Galati, 2019; Danner, Barbosa, & Goldstein, 2018; Pouw, Trujillo, & Dixon, in press) by enriching the investigation into how gestural low-level events feed into higher-level linguistic structures (Krivokapić, 2014; Ravignani et al., 2019; Shattuck-Hufnagel & Ren, 2018).

Study of gesture as a multiscale phenomenon

Pioneering the quantitative investigation of the idea of coherent subsets in gestural ensembles, gesture researcher and theorist David McNeill and computer scientist Francis Queck introduced new applications of methods of how to relate low-level gesture features with higher-order phenomena in gesture (McNeill et al., 2001; Quek, 2004; Quek, Bryll, McNeill, & Harper, 2001; Quek et al., 2002; Xiong, Quek, & McNeill, 2003). Central to this research program was the concept of "catchment" gestures, which are gestures that are recognizably similar in form and repeat during discourse (McNeill, 2000). Such catchments were of interest given the assumption "that the recurrence of imagery in a speaker's thinking will generate recurrent gesture features. Recurrent images suggest a common discourse theme" (Quek et al., 2002, p. 178). The goal was to uncover the organization of a gestural discourse, by matching recurrent low-level features such as hand displacement and location, as well as symmetry relations between left- and right-hand gesticulation, with recurring topics in the discourse (Quek, 2004).

This search for relevant commonalities in low-level features between different gestures has been aided by recent developments in human–computer interaction research, where machine learning has been successfully used to differentiate between, for example, beat versus iconic gestures or between gestural phases, such as the main stroke versus the retraction phase (Madeo, Lima, & Peres, 2017). However, so far the idea of catchments as a quantifiable construct that can be of interest for understanding gesture dynamics under psychological experimentation has not gained ground. Rather the concept of catchments in psychological research is still primarily quantified through multirater judgments of gesture similarity (Kimbara, 2008; Mol, Krahmer, Maes, & Swerts, 2012). Multirater judgments might in some cases be optimal to judge gesture similarities. For example, when the meaning of a gesture needs to be recognized for recurrence in a discourse, we would need a human interpreter. However, low-level feature recurrences of gesture (e.g., kinematic profiles) cannot be reliably based on experimenter judgments. Therefore, the potential way in which higher-order organization emerges from low-level dynamic gestural features is not studied at the moment, whereas detecting such higher-order organization might be crucial for understanding discourse processes.

This situation highlights the need for innovation in the *multiscale* study of gesture ensembles, where multiple levels that comprise gesture activity are connected. Such innovation could be part of a toolkit including already adopted multiscale approaches to gesture–speech dynamics. For example, (bivariate) spectral analyses methods (cf. Xiong et al., 2003) have reappeared in the gesture literature. These methods allow for the decomposition of gesture and speech time series into several dominant time scales to further gauge on which temporal scales gesture and speech couple their activity (Danner et al., 2018; Pouw & Dixon, 2019). Such spectral decomposition approaches help quantify and correlate gesture and speech activity that happens on the syllable, clause, and sentence time scales. Others have used approaches in dynamical systems research whereby changes in temporal structures of categorical events or continuous time series (recurrence quantification analysis; Wallot, 2017; Webber & Marwan, 2015) are the object of study as opposed to averaged metrics of speech or gesture activity (De Jonge-Hoekstra, Van der Steen, Van Geert, & Cox, 2016; Fusaroli & Tylén, 2016; Pouw, De Jonge-Hoekstra, & Dixon, 2018). These recurrence quantification analysis approaches are especially potent for understanding global gesture patterns that reflect stability of recurrent behavior through a time-dependent analysis of local fluctuations in gesture activity. As can be seen, all these approaches are heavily influenced by dynamical systems thinking of "emergence" (Pikovsky, Kurths, & Rosenblum, 2001) and therefore have as their common denominator the quantification of meaningful relationships between a system's local fast-scale activity with that of global slower forming patterns, without reducing one scale into another. The new gesture analysis we are about to propose falls within this endeavor.

In the current article we introduce established analyses in time-series and network analyses as a novel approach to the study of the emergent structure of gesture kinematics. We call this approach *gesture network analysis*. We use two previously obtained datasets that include motion-tracking data of spontaneous gestures produced during narration (Lücking, Bergmann, Hahn, Kopp, & Rieser, 2010; Pouw, De Jonge-Hoekstra, & Dixon, 2018). We then assessed similarity in structure of the velocity profiles of gestures using dynamic time warping (DTW; Giorgino, 2009; Mueen & Keogh, 2016). By performing these analyses for each gesture relative to each other gesture in the set of all gestures produced in an ensemble, we are left with a weighted matrix where each pair of gestures is given a similarity score. Such a weighted matrix can thus be treated as a network of similarity relationships. This opens up the study of gestures via a network approach whereby gestures are understood relative to its position in the network of the gesture ensemble (i.e., the set of all gestures). We apply these analyses to assess a number of research questions serving to demonstrate the versatility of the current methodology.

In the next section the basic procedure of gesture network analysis is introduced. This procedure first requires a quantification of similarities between two time series, for which we introduce DTW. The second procedure is the construction and statistical evaluation of networks from the produced DTW similarity metrics.

Methods

Datasets

We reanalyze two previously collected motion-tracking gesture datasets to assess gesture ensemble dynamics under varying contexts. For each dataset, "gesture events" were recorded and identified by human annotators whereby an event consisted of the meaningful dominant hand stroke that coordinated and co-occurred with speech.

DAF dataset

The first dataset is fully reported in Pouw and Dixon (2019) and concerns a within-subject kinematic study of gesture's coordination with speech when retelling a cartoon to the experimenter (who was not speaking back) during a delayed auditory feedback (DAF) manipulation of 140 ms or retelling without such a manipulation (NO DAF). In this study participants' hand movements were recorded with a Polhemus Liberty wired motion tracker. Only the dominant hand was allowed to gesture to simplify analysis. We originally found that gesture–speech synchrony was more pronounced under DAF but that gestures were also attracted to synchronize with the auditory delay. For the current data we excluded 3 participants of the original experiment as these had fewer than 10 gestures per condition given that they produced shorter narratives (i.e., gestures per minute were not necessarily lower for these participants). The dataset therefore consisted of seven participants (M age = 18.8 years, 2 women) producing 532 gestures in total (mean gesture count, 76; i.e., the mean size of the gesture ensemble).

Speech and Gesture Alignment Corpus dataset

The second dataset concerns the Bielefeld Speech and Gesture Alignment Corpus (SAGA) as reported in Lücking et al. (2010), which was further enriched with videography motion tracking by Pouw et al. (2018). The dataset involves German-speaking participants retelling a route navigation they had learned in a virtual reality simulation. Participants were route-givers, explaining the route to a conversational partner who was tasked to listen to be able to navigate the route afterward. The SAGA dataset is richly annotated, and here we are interested in annotations relating gesture type of the dominant hand, as well the semantic reference of speech co-occurring with gesture (see below). The dataset consisted of six participants (5 males; no age reported) consisting of 864 gestures in total (mean gesture count, 144; i.e., the mean size of the gesture ensemble).

Gesture network analysis

Step 1. Choosing a kinematic variable

We use 3D speed of the dominant hand (hereinafter *velocity*) as the key kinematic parameter under investigation. We chose velocity as it is a simple one-dimensional vector that captures gross intensity in gesture. Note, however, that what kinematic variable or variables should be chosen depends on the research question. When interested in the absolute spatial trajectory or geometrical form of a gesture (Shattuck-Hufnagel & Ren, 2018) one can choose 3D (x, y, z) or 6D (adding pitch, yaw, roll) motion variables. Such analysis does require the researcher to further specify frames of reference that allow for meaningful comparison (e.g., body-centric frame of reference or joint angles). If interested in scale-free geometrical forms (e.g., Cook & Tanenhaus, 2009), such that for example a *smaller*-sized circling gesture trajectory is deemed equal in kind as a *larger*-sized circling gesture, the researcher should normalize the time series. Alternatively, in some cases the researcher wants to normalize *time* for each gesture if interested in the absolute trajectory of a gesture while ignoring how slow or fast such a trajectory is completed. As mentioned, for simplicity, however, we use the untransformed velocity of the dominant hand as the main kinematic variable.

Step 2. Dynamic time warping

DTW enables a quantification of similarity between two univariate or multivariate time series (Giorgino, 2009; Mueen & Keogh, 2016; Müller, 2007; Silva, Batista, & Keogh, 2016). It is applied to “time series” (i.e., an ordered sequence of values of n length along some ordering dimension t), and “time warping” refers to the procedure of computing the “distance” after nonlinear alignment (i.e., warping) of observations between the time series. For example, for two gesture time series it can be assessed how the first time series (e.g., velocity of a gesture event 1) must be stretched or truncated so that values of time series 1 are maximally aligned with values of time series 2 (e.g., velocity of a gesture event 2), where maximal alignment means which observation in time series 1 is a closest match (in time and absolute value) with an observation in time series 2. Then, the *distance* for each matched observation is computed and finally summed yielding a measure of the dissimilarity between both time series. The stretching and truncating (i.e., matching of the observations) follows several rules; for example the order of observations after warping must be preserved (for formal definitions see Müller, 2007 and Silva et al., 2016).

Figure 1 shows an example of two time series and the relative distances between them as computed by `R` package `dtw` (Giorgino, 2009); we use this `R` package for DTW analyses throughout the article. For computing the DTW distance measure, the distances for each matched observation are summed and optionally then normalized for the length of the time series. Normalization is needed given that distance is the result of adding up distances for each matched observation of the time series. Therefore, longer time series will also have higher distance scores if not normalized for time. Therefore, we normalize the distance measure (by the cumulative length of the two time series compared; Giorgino, 2009) since we compare a great many time series with variable lengths.

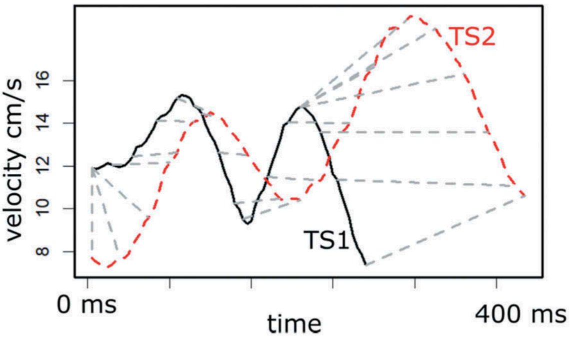

Figure 1. Example DTW distance graph of two velocity profiles. This graph is produced by R package dtw and concerns two gesture velocity profiles (TS1 = gesture 1 velocity; TS2 = gesture 2 velocity) that were extracted from the DAF dataset (Pouw & Dixon, 2018). At regular time intervals a non-normalized DTW distance is indicated with the gray dashed lines, but note that DTW computes these distances for each observed value.

We have briefly mentioned that the warping procedure is defined by rules that constrain how the cross-matching of values between time series is performed. Although for the current demonstration article we do not modify the default rules as set by R package dtw (Giorgino, 2009), it should be noted that there are a host of customizations to DTW depending on unique features of the researcher's data. We highlight one common departure from generic DTW. Recently the so-called end-points constraint has come under scrutiny (Silva et al., 2016). The end-point constraint dictates that the matching procedure should be performed from beginning to end of the time series. However, often the beginning and end of an event (e.g., a start and ending of a gesture event) are not the most important points of interest, and often it is difficult to define when an event (e.g., gesture) really starts and ends. As a consequence, distances that are computed for the beginning and end points may inadvertently inject noise into the distance estimate. Indeed, we prioritize the comparison of most important part of an event that defines the most variance in the signal (i.e., the central chunk of information in the gesture event). Therefore, Silva et al. (2016) reviewed several techniques and introduced their own solution (called ψ-DTW) that relaxes the end-point constraint; this solution has been found to increase accuracy and computation speed in several classification tasks.[1]

Multivariate DTW. DTW was originally developed for speech recognition purposes, whereby differences in, for example, syllable length could be ignored while commonalities in acoustic structure recognized. However, such analyses were performed not with a set of one-dimensional time series but with multidimensional time series containing spectral coefficients. Indeed, we can imagine that DTW analysis is equally applicable when the time series has an extra dimension. Figure 2 shows the same time series as above, but this time a speech component (amplitude envelope[2]) is added to the "state-space" for the time series. State-space here merely means a space onto which states (e.g., gesture and speech states) of a system can be mapped. In the current multivariate case a state at time *t* is not only defined by the velocity of a gesture at *t* but also the amplitude envelope of speech at *t*. Again, we can imagine that DTW can be applied here in a similar fashion by computing the relative distances between the two multivariate time series that need to be applied in warping one onto the other but this time in a multidimensional state-space. Note that this multivariate DTW estimate we use assumes some kind of dependence. We assume a dependence between speech and gesture, as they are a strongly coupled systems, one affecting the other and vice versa. However, we could also opt for an independent comparison, whereby we assume that speech and gesture are two states of two different systems that define an outcome of interest. For such an independent approach we simply calculate the distance between time series for each dimension separately as to then sum the independent distance estimates (Mueen & Keogh, 2016). Note that for

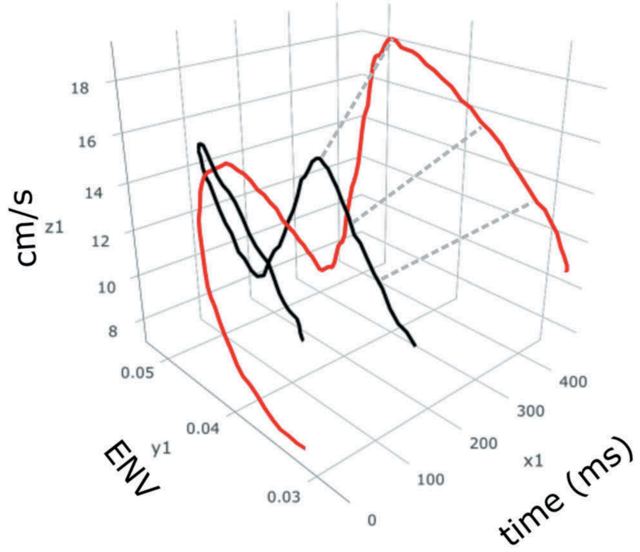

Figure 2. Example multivariate distance graph. This graph (produced by the authors) provides an example of multivariate state space where DTW can be applied. The gestures velocity profiles (cm/s) of Figure 1 are replicated here, but this time, differences in concomitant speech (ENV, amplitude envelope) are added as another dimension. The gray lines now reflect how DTW would compute distances in such a multivariate state-space.

all the multivariate DTW analyses performed, we z-normalized the amplitude envelope and the gesture velocity for each participant's range such that distances are computed on comparable dimensions.

Step 3. Network analyses

Weighted networks. The third step is extracting DTW's output for each possible comparison of gesture pairs in the gesture ensemble and construct matrices that can be used as input for network analyses. For network analyses and graphing we use the `R` package `igraph` (Csardi & Nepusz, 2006). Figure 3A shows an example of how a weighted matrix is constructed containing for each cell DTW results for a gesture pair comparison. The resulting matrix of values is often referred to as a weighted matrix (or distance matrix), as each cell of the matrix expresses a continuous "weight" of the relationship between two nodes (i.e., gesture nodes). The relationships are also called "edges" and are graphically expressed as connections between nodes, and the length of those edges reflects (in the case of weighted networks) the degree of dissimilarity (i.e., distance) between two (gesture) nodes. Figure 3A shows such a network graphic expression of the weighted matrix, where it is important to understand that each edge connecting each node pair has a certain length that reflects the DTW distance between these two gesture events' velocity profiles[3] (or whatever gestural dimension that might be compared).

Measures for weighted networks. For simplicity we only use two network measures that reflect two properties of how a gesture node is situated relative to the gesture ensemble. First, how central is the position of gesture in the gesture ensemble network? Second, how diverse are the relationships of the gesture node relative to the gesture ensemble? Both measures are standard measures computed by `igraph` (Csardi & Nepusz, 2006).

Mean distance. The *mean distance* is a simple measurement of the average DTW distance of a gesture to all other produced gestures in the ensemble. It is computed by averaging all distances for each gesture, and in network graphic terms lower mean distance expresses the centrality of the gesture in the gesture ensemble. Namely, when gestures are least dissimilar to other gestures (i.e.,

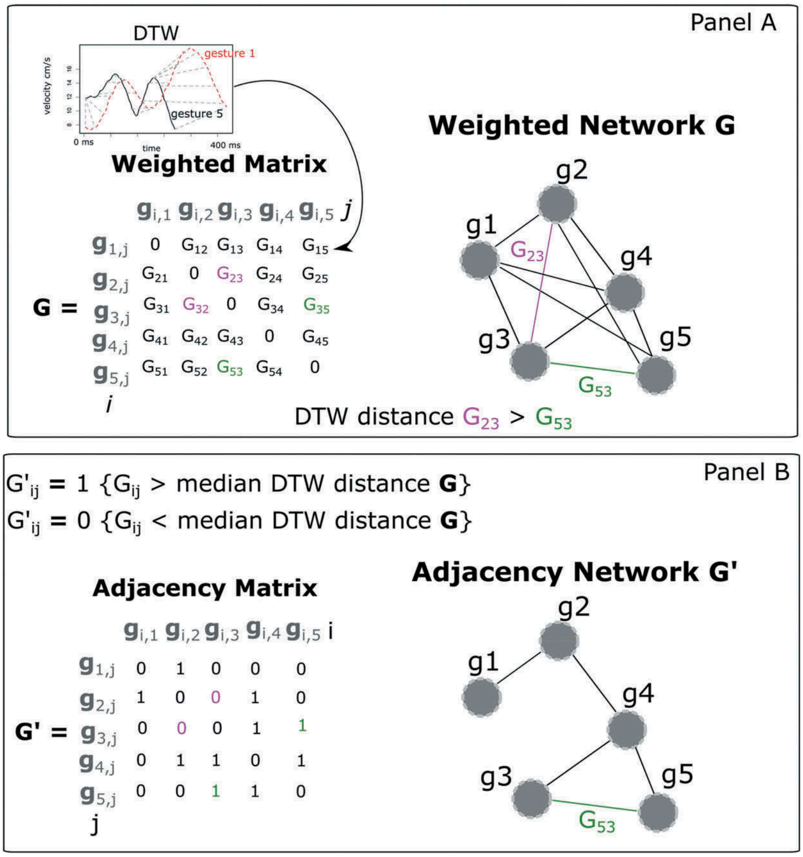

Figure 3. Example data processing for network analysis of gesture ensembles. (A) Weighted matrix **G** reflecting a gesture ensemble of five gestures. Each matrix cell G_{ij} consists of DTW distance values. For example, cell G_{15} contains the DTW distance of gesture event g1 and g5. The diagonal of the matrix always contains 0 as the DTW distance between an identical gesture event is 0. On the right-hand side the weighted matrix is expressed in a network graph, whereby each DTW distance is expressed as an edge connection between gesture events. Gesture events are expressed as the nodes of the graph. It can be seen that G_{23} (in purple) has a higher-length edge than G_{53} (in green), which means that gesture event g2 and g3 are more dissimilar (higher DTW distance) than gesture event g5 and g3. (B) Result of dichotomizing the weighted matrix **G** into what is called an adjacency matrix **G'**. We can for example dichotomize each cell of **G** to a 0 or a 1 provided the value is under or above the median DTW distance of **G**. This results in an adjacency matrix **G'**, which now leads to empty edge connections between nodes. This is clearly reflected in the network graph of **G'** where now only short edge lengths survive. This adjacency network further provides detection of subnetworks called cliques. It can be seen that gesture events g3-g5-g4 now form a subgroup, which could mean that these gestures are meaningfully related (as they are comparable in their kinematics).

have a lower mean distance to all gestures in the ensemble) they will occupy a position in a gesture network such that distances between all nodes are minimized (i.e., more central position).

Diversity. *Diversity* is a more complex measure tracking variability of a pattern and is used in network analysis to quantify the relative diversity or complexity of a node's relationships (Eagle, Macy, & Claxton, 2010). The measure is the scaled Shannon entropy of the weights of the node's edges, which in our case means the entropy of the gesture's DTW distances to all the other gestures in the ensemble. Higher entropy indicates that a node (gesture) is more diversely connected (more diverse DTW distances) to all other nodes (gestures) in the network; in other words, the gesture has more variable similarity relations and there is no typical way to describe its relationships. A lower entropy indicates that a node has more uniform relationship with all other nodes in the network; for example, all gestures have the same similarity (then entropy = 0), or if there are gestures that are very alike and gestures that are not alike but nothing in between.

Adjacency networks. We mostly work in this article with weighted networks. Weighted networks are a special kind of network as each node is connected *to some weighted degree* with all other nodes. However, we might be interested in studying how networks may be parsed into subnetworks. One way to accomplish this is transforming a weighted matrix into an adjacency matrix, whereby each value of a cell is given a 0 or a 1 based on some distance threshold, for example, any DTW distance values (i.e., edge weights/distance) higher than the median DTW distance for a gesture ensemble should be treated as unconnected (0) and all other DTW distance values under the median are similar enough to be treated as connected (1). To see how a weighted matrix is transformed into an adjacency matrix and what this means for the network graph please see Figure 3B. Basically, an adjacency matrix is a dichotomized version of weighted matrix, where each value in weighted matrix is transformed to a 0 or 1 for the adjacency matrix depending on some threshold (e.g., median split). A 0 in the adjacency matrix indicates that a gesture pair (two nodes) will not have an edge connection, and a 1 indicates that the two gesture nodes do get a connection. If the network graph in Figure 3A and B are compared, it can be seen that the adjacency network has now fewer connections between its nodes as large distances between gesture velocity profiles are treated as unrelated events.

The reason for transforming a weighted network with continuous similarity scores into an adjacency matrix with 0 (no edge connection) and 1 (edge connection) primarily lies in making apparent hidden subnetworks within the larger ensemble. We could for example in our case wonder whether gesture velocity profiles that are so dissimilar from each other should be linked with an edge as this might obfuscate possible higher order structures (e.g., subnetworks) that are difficult to detect otherwise in a weighted network graph where nodes are all connected (but spatially dispersed based on similarity distance). Thus, although to some extent arbitrary, the result of dichotomizing a weighted matrix into an adjacency matrix will allow for further study of sub-networks.

Results

Now that the general procedure for gesture network analysis has been laid out, we provide four research-specific implementations to demonstrate the versatility of gesture network analysis. First, we use a simple approach to DTW with univariate time series (gesture velocity) where we assess how different categories of gestures (e.g., beat vs. iconic) are situated in a gesture network. Second, we use a multivariate DTW approach, wherein we also include a speech time series so as to see how a behavioral perturbation can affect a network constructed out of gesture + *speech* events (as shown in Figure 2). For the third analysis, we combine annotations about what is said in speech with gesture network analysis, whereby we can show that referents in speech can predict dynamic aspects of gesture (i.e., positions of co-occurring gestures in the gesture network). In the final proof-of-concept "mock" analysis, we show that gesture network analysis is especially promising for assessing large-scale gesture dynamics at levels that transcend individual people, whereby we construct a large gesture network of an ensemble of gestures produced by multiple persons who speak the same or a different language. This final analysis promises to provide a way to assess whether gestures are affected in their kinematics when produced under different spoken language.

Simple approach with univariate time series: gesture categorization

Figure 4 shows an example of a network of a gesture ensemble produced by single participant, with colored nodes for three types of gestures that were coded in the DAF dataset (Pouw & Dixon, 2019). Three types of gestures were coded. First, beat gestures are those gestures that do not have depictive or symbolic content and oscillate with the prosody of speech. Second, iconic gestures are those gestures that have depictive or symbolic content. Third, undefined gestures are those gestures that did not fit the above categories. For this analysis we include all gestures that are produced under DAF and NO DAF conditions.

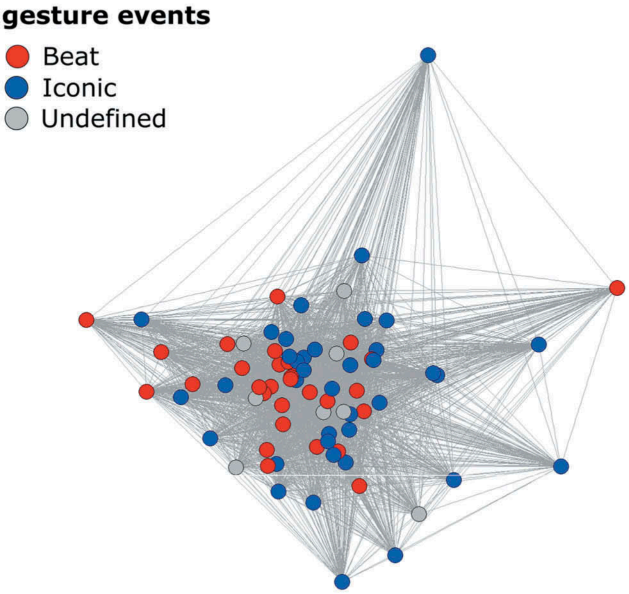

Figure 4. Example gesture ensemble for single participant. This gesture network reflects beat (red), iconic (blue), and undefined gestures (gray) which are compared with each other. It can be seen that the iconic gestures are more spatially dispersed as compared with beat gestures, which seem to occupy a more cluttered region. Note that spatial dispersion means a higher dynamic time warping distance (i.e., higher dissimilarity in velocity profiles). For all the different networks for each participant see here: https://osf.io/teuyn/.

An assumption in the literature is that iconic gestures are more complex in their trajectories as compared with beat gestures. In the current network analyses this translates into clear-cut predictions about how gesture nodes should be situated to other nodes in a network. Namely, iconic gestures should be connected to other gestures with farther distances (they are more likely to be dissimilar to the other gestures), and they are more likely to hold more diverse similarity relationships with the other nodes. To assess this we constructed a weighted matrix with DTW distances between velocity profiles of gestures (only iconic and beat gestures) for each participant individually and extracted the distance and diversity measures. This procedure enables the comparison of differences in gesture categorization within participants.

To test the differences for gesture categorization as shown in Figure 5, we performed mixed regression analyses with `R` package `nlme` (Pinheiro, Douglas, Debroy, Yes, & Yes, 2011). All models have participant as random intercept. A model containing gesture type (iconic vs. beat) as a predictor for distance improved model fit as compared with a base model predicting the overall mean, change in χ^2 [1] = 14.95, $p < .001$. The model showed that iconic gestures were farther spatially removed from all other nodes as compared with beat gestures, b Iconic = 0.025, 95% CI [0.013, 0.038], $t(475) = 20.16$, $p < .001$. Thus, we can conclude that iconic gestures are more dissimilar in their velocity profile to all other gestures in the ensemble, whereas beat gestures have a higher likelihood to have a similar velocity profile as other gestures.

We would equally predict that iconic gestures will have more variable similarity with other gestures, because they are often more complex in their trajectories (as compared with beat gestures). We tested this, and a model containing gesture type (iconic vs. beat) as a predictor for diversity improved model fit as compared with a base model predicting the overall mean, change in χ^2 [1] = 59.50, $p < .001$. Indeed, iconic gestures had a higher diversity score as compared with beat gestures, b Iconic = 0.013, 95% CI [0.010, 0.016], $t(475) = 7.96$, $p < .001$.

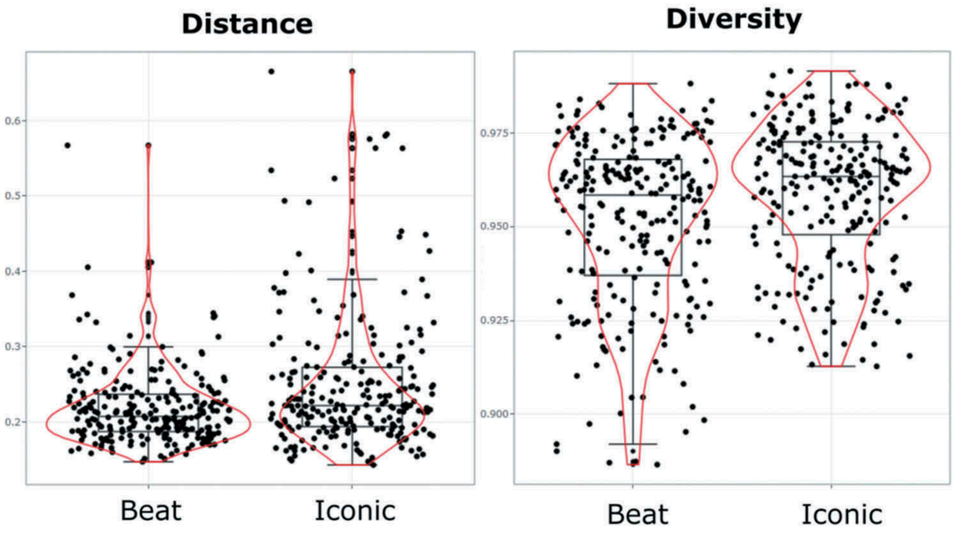

Figure 5. Distance and diversity scores for beat and iconic gestures. The boxplots contain summary data (average and quartiles) for distance and diversity for the gestures that were produced. The violin plots indicate the smoothed density distribution. The jitter plot indicates the individual observations for distance and diversity estimates, whereby they are randomly jittered on the horizontal axis as to increase visibility. As can be seen beat gestures have shorter distances to other gestures (i.e., are more similar in their velocity profiles) and show a lower diversity (have less diversity relative to other gestures in the ensemble).

Multivariate approach: gesture–speech trajectories when speech is perturbed

Using the DAF dataset again, we constructed a similarity network not only on the basis of comparing a velocity profile of a gesture but also on comparing a concomitant speech event during the gestures using the amplitude envelope time series. This multivariate approach of DTW provides a comparison of the similarity of *gesture–speech* events (see Figure 2 again for reference). Figure 6 shows an example of such a gesture–speech ensemble network. In the current application, we want to see if a gesture–speech ensemble is affected by DAF of speech (which affects speech fluency) as compared with a control condition with no DAF. In the original study (Pouw & Dixon, 2019), we obtained that DAF perturbation led to more gesture–speech synchrony and a slight gesture–speech offset. The dis-synchronization of gesture was moved toward the auditory delay. We have no specific hypothesis of what will happen to the gesture–speech ensemble when under DAF. On the one hand, it can be reasoned that gesture-speech events become more variable as compared with other gesture–speech events when under a perturbation, as gesture and speech is pushed out of its usual routine. On the other hand, we might expect that gesture–speech events will become more uniform, given that we have found earlier that there is stability under perturbance in terms gesture–speech synchrony (although note that we are assessing uniformity on different levels of analyses—the level of gesture ensembles rather than gesture–speech synchrony within single events).

We performed network analyses for each subject's gesture ensemble as exemplified in Figure 6. As shown in Figure 7, we find that a DAF speech perturbation lead to more dissimilarity (higher distances between nodes) and more diversity as compared with gesture–speech events that were not produced under a speech perturbation. Namely, a model predicting distances based on condition (NO DAF vs. DAF) reliably accounted for more variance as compared with a base model predicting the overall mean, χ^2 [1] = 29.55, $p < .001$. We also added gesture type (Iconic vs. Beat) and its interaction with condition to the model, but this did not lead to improved model fit. The best fitting model with condition as predictor shows that DAF perturbation led to higher distances (more overall dissimilarity) for gesture–speech events as compared with the NO DAF condition, *b* NO DAF vs. DAF = −0.002, 95% CI [−0.0028, −0.0018], $t(475) = -3.86$, $p < .001$.

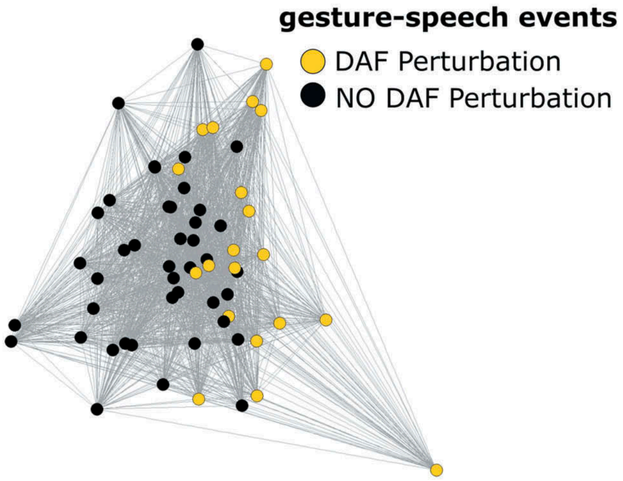

Figure 6. Gesture–speech network with DAF manipulation. The current gesture–speech network comprises gesture–speech events produced by single participant. Each node reflects a gesture (velocity) and speech (amplitude envelope) event that is compared with the other gesture–speech events. In gold, gesture–speech events were produced under DAF manipulation; in black are the gesture–speech events produced under control NO DAF condition. It can be seen that gesture–speech events produced within DAF or within NO DAF conditions are closer to each other. For all the different networks for each participant see here: https://osf.io/3csfu/.

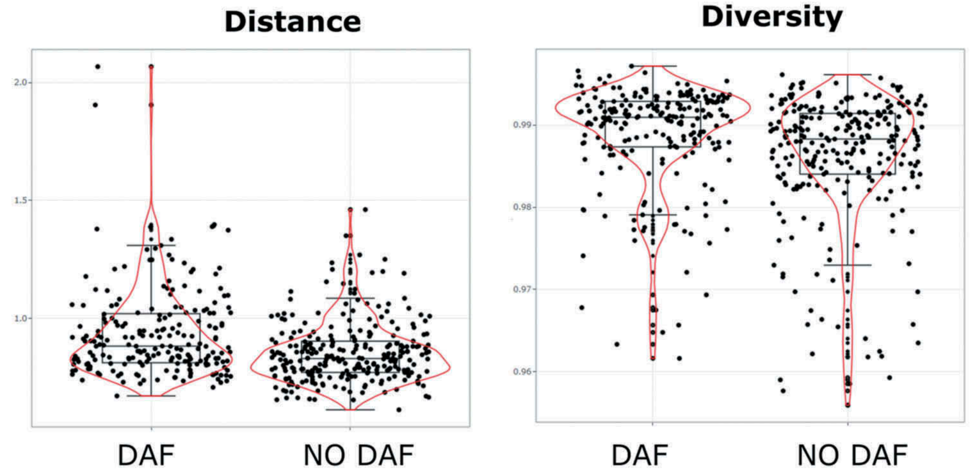

Figure 7. Effects of speech perturbation on distance and diversity measures.

We also found that condition was reliably more predictive for diversity of gesture–speech event nodes as compared with a base model, change in χ^2 [1] = 14.70, p < .001. Although an interaction of condition and gesture type was not reliable as an added predictor, adding gesture type as a main effect to the model containing condition did improve model fit, change χ^2 [1] = 20.79, p < .001. The final model shows that DAF speech perturbation lead to more diverse distance relations of gesture–speech events, b DAF vs. NO DAF = −0.002, 95% CI [−0.002, −0.001], t(475) = −4.22, p < .001. Further, Iconic gesture–speech events had more diverse distance relations, b Beat vs. Iconic = 0.0012, 95% CI [0.001, 0.003], t(475) = −4.56, p < .001.

Thus, interestingly, we can obtain new insights in how gesture and speech are affected in their dynamics relative to the ensemble of gesture–speech events. We find that perturbing the speech system leads to more diverse and dissimilar gesture–speech trajectories as compared with the gesture ensemble.

Speech annotation-enriched gesture networks: relating semantic processes to gesture kinematics

The German-speaking SAGA dataset contains codings of the references that were made in speech (see section 3.4.3, of documentation Bergmann et al., 2014). That is, all possible landmarks that were referred to during the route-navigation speech task were coded. We quantified how often references were made during iconic[4] gesturing by a participant. Namely, some references to objects that were likely more important in the route navigation are talked about more often than other referents during gesturing. We asked whether gestures that were produced with semantic references occurring repeatedly during a discourse are more likely to be different from gestures that co-occur during novel references (i.e., references that are made less often). Indeed, recall from the introduction this is precisely what "catchments" were about, where gestures with recurring semantics have recurring kinematics. If gestures co-occur with a recurring reference in speech, we would then predict that such gestures have lower distance scores, as they are more like most other gestures and occupy a more central position in the gesture network. To test this we first assessed for each gesture event in the participant's gesture ensemble which speech reference was concurrently made (if two references were made, then the reference which was talked about longer during the gesture was used). Having the number of "occurrences" of references, we could then relate the distance measure with the recurrences.

We find that that gestures which co-occur with speech references that are more prevalent during the discourse seem to be more similar in their kinematics to most other gestures. Figure 8 seems to further show that this relationship follows a power-law relationship. This means that the relationship between number of speech references and gesture network position is nonlinear; particularly for the first levels of increasing speech references, the network position is heavily affected. After a particular

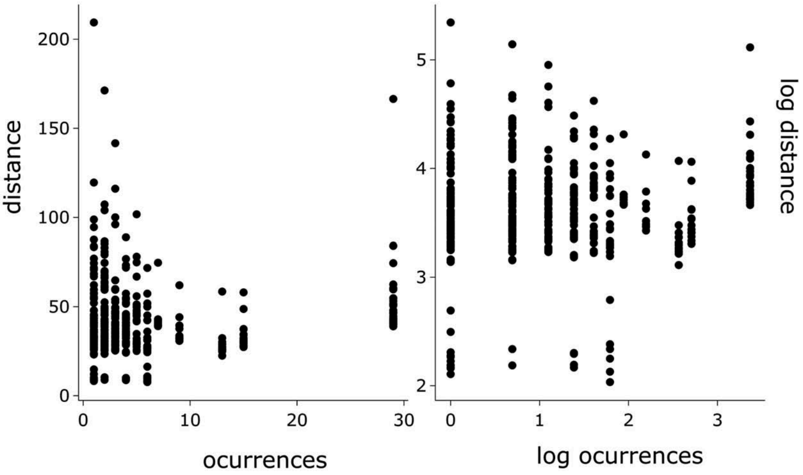

Figure 8. Relationship with distance measures as compared to the number of occurrences of reference in speech. This graph shows whether gesture distance to the gesture ensemble was affected by whether that gesture was produced with a speech referent that occurred more or less (occurrences). The graphs seem to show a relationship such that especially within about 15 references it seems that gestures become more similar relative to the gesture ensemble when the speech referral is less novel (as it occurred more during the discourse). The right-hand plot shows the log-log plot for occurrences and distance.

higher level of speech references has been reached, further increases in references does not further affect a gesture's network position (or at least less so as compared with increases from lower levels of speech references). If there is a power-law relationship, there should be a linear relationship between the logarithm of distance and the logarithm of occurrences. Indeed, when regressing (random intercept participant) log(occurrences) on the log(mean distance), we obtain that this model fits better than a model predicting the overall mean of log(distance), change in χ^2 [1] = 8.64, p = .003. The final model indicated that when occurrences of a speech referent increases, there is a lower distance (higher similarity) with the other gestures in the ensemble, b = −0.044, 95% CI [−0.072, −0.015], t(420) = −2.987, p = .003. The power law indicates that at a certain extreme of occurrences, the negative relationship with distance becomes saturated; in other words, at some point more recurrences do not predict lower distance. Note that a linear model yielded a marginally reliable result indicating a similar conclusion (p = .076).

Proof-of-concept large-scale analysis: language-specific gesture kinematics in groups of speakers

As juxtaposed in Methods, so far the network graphs we have used are organized such that all gesture events are connected to all other gesture events (as we have performed DTW for each gesture comparison). This means that all gesture events are linked and only spatially separated by their similarity distances. We can, however, force the gesture-similarity network to break ties when gestures are too dissimilar to each other with some arbitrary threshold. In this way we can force subnetworks to form that may be more revealing than simply spatially separating them. We use this analysis as a proof-of-concept "mock" analysis to show how our approach can study kinematic similarities in gestures in large groups, groups of speakers producing gestures while speaking different languages. Note though, due to large differences in the datasets used, we cannot and do not draw any conclusions or apply any inferential statistics on the results.

The "mock" research question that we want to assess is whether gestures produced within a particular language can be differentiated purely by gesture kinematics. In other words, can we detect regularities in kinematics of gesture ensembles that reflect whether they belong to a particular language? This question can now be addressed quantitatively via gesture network analysis. Given that we have two datasets where participants spoke a different language, namely the American-English DAF dataset (Pouw & Dixon, 2018) and the German SAGA dataset (Lücking et al., 2010), we could approach this question via network analysis, and specifically we might want to use dichotomized adjacency networks for this. Before we proceed, however, it is important to note that these results are not conclusive in any way as the current datasets are different in a number of ways next to the difference in spoken language. Namely, in the DAF dataset, participants could only move their dominant hand whereas in the SAGA dataset they could move both hands. Furthermore, different motion-tracking technology was used. In the DAF dataset a high-sampling 3D motion tracker was used, whereas in the SAGA dataset we obtained 2D motion tracking using videography motion-tracking (Mathis et al., 2018), which has a lower sampling rate. Finally, retelling a cartoon narrative as in the DAF dataset is different and will most likely solicit different types of gestures, as compared with explaining a route to a particular landmark as was the case for the SAGA dataset. Thus, the method and nature of the tasks was different on a number of dimensions.

To assess whether gesture kinematics of Germans could be differentiated from North Americans, we constructed a large network comparing all velocity profiles of iconic gestures produced in the two datasets (a total of 802 iconic gestures, of which 542 German gestures). We chose iconic gestures as these gestures are likely to be most differentiable and have a higher variability as compared with for example beat gestures. The velocity profiles were z-normalized for each dataset, such that absolute differences in motion-tracking output were not taken into account for DTW analyses. After constructing a weighted matrix (802*802 = 643,204 cells) we transformed this matrix into an adjacency matrix where each value under the median was given a value of 1 and all values reflecting high DTW distances (above median) were given a 0. The adjacency matrix was plotted using `igraph` shown in Figure 9.

As can be seen in Figure 9, iconic gestures that were produced in the North American speaking dataset were much more interconnected with each other (as now represented by their relative positions to each other) as compared with their connections with gestures produced in the German-speaking dataset. Although we want to showcase the promise of this method to study language-modulated gesture kinematics, we should not make anything out of the current results, given the fact that these gestures are produced under completely different circumstances. Rather we aim to show with this demonstration that the current network approach is able to provide a graphical description of how gesture ensembles are produced by a different group of people. Thus, whether the current pattern in the data reflect true language differences (or methodological differences) we must leave for more controlled experimentation. Note that if we wanted to test the current reliability of the cluttering, we could assess whether for example North American gestures are more likely to have edges with other North American gestures (as compared with German gestures). Other more advanced network analysis measures concern the identification of subcliques of networks (subnetworks) and see whether they correlate with predefined labels (Csardi & Nepusz, 2006). In a weighted network approach we could compare the mean distance of German–German connections as compared with German–American connections.

Discussion

The goal of the current article was to show the potential of novel applications of bivariate time series and network analyses for the kinematic study of gesture ensembles. This novel approach in gesture

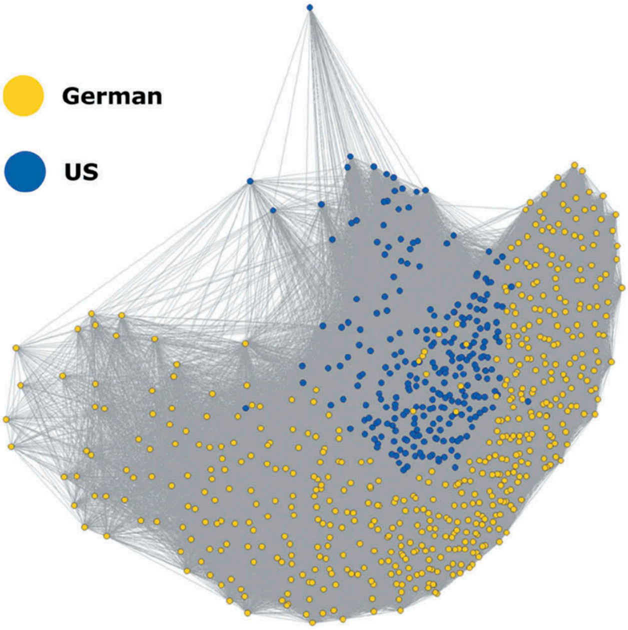

Figure 9. A gesture ensemble network of groups of speakers (Germans and North Americans). The current adjacency network is built from dichotomizing (based on a median threshold) a weighted network containing the DTW analyses of the velocity profiles of gesture pairs. The geographical position of each node is now determined by the number of connections it has with the other nodes. It can be seen that for example the most highly placed blue (North America – US) node only has connections to other gestures that were also produced by North Americans but not to the gold (German) nodes (in a weighted network it would still connect to all other German nodes). Nodes that have no connection with other nodes are not plotted here. It can be seen that there is a clear case of cluttering where iconic gestures from the North American dataset are more likely to be connected with other North American gestures.

research we may call *gesture network analysis*. These analyses allow for the study of how low-level features of multimodal language (e.g., gesture velocity profiles, concomitant speech) are interconnected in a way that reflects higher-order structure on a discourse level. Namely, all analyses have shown structural properties in gesturing that *emerge over time* that are apparent when looking at the kinematic inter-relations between gestures. Although the current results are exploratory and need to be replicated, we have obtained novel insights into how gesture categorization, speech perturbations, and semantics of co-current speech affect gesture ensemble dynamics. Finally, we have shown in a proof-of-concept fashion, how one could study with gesture network analysis whether the language spoken is reflected in the higher-order dynamics at the level of gesture ensemble kinematics. However, again note that these final analyses do not provide any evidence for this and was purely demonstrative for the potential of gesture network analysis. Below we summarize the results and further emphasize the potential of each of these analyses for future research.

Gesture categorization

The current analyses showed that iconic-labeled gestures showed much more dissimilarity from all other gestures in the ensemble and showed higher diversity as compared with beat-labeled gestures. This is a straightforward result that dovetails with general assumptions about these gesture types. A further potential of this type of analysis is to seek further differentiations of more fine-grained gesture typologies (e.g., character-viewpoint gestures, metaphoric gestures) in gesture networks. Thus, the current analyses provide a way to quantitatively ground the existence of the myriad of gesture-typologies that theorists have proposed (Kendon, 2004; McNeill, 2005).

Gesture–speech trajectories when speech is perturbed

The speech perturbation analyses were aimed to show that the effect of experimental manipulations can be studied on the ensemble level and that networks can be constructed out of *gesture–speech* ensembles as well by applying multivariate DTW analyses. For this analysis we defined the time series of each event by gesture velocity as well as amplitude envelope of concomitant speech. We found that perturbing speech with a DAF (which yields speech disfluency) affected how gesture–speech events were situated relative to the gesture–speech ensemble. Namely, under DAF gesture–speech events were more different from the other gesture–speech events in the ensemble and also showed a higher diversity of relationships with the events in the ensemble. This is an interesting result that shines new light on our earlier findings, which showed that gesture and speech were more synchronized under the DAF. When combining these results, it seems that when under perturbation gesture–speech synchrony is stable, whereas the gesture and speech trajectories become more diverse as compared with the ensemble. Again, such results need to be replicated, but the potential of probing experimental manipulations in the current fashion might have great implications for understanding gesture and speech as a multiscale phenomenon.

Relating semantic processes to gesture kinematics

Possibly one the most daunting explanations in gesture research is how low-level features of gestures are reflecting meaning at the level of discourse. Our current analysis provides one way into this research question by assessing how gesture kinematics on the ensemble level might change in structure as a function of the concomitant informative value of referents in speech. Using the meticulous semantic coding in the SAGA dataset (Lücking et al., 2010), we find that gestures that are produced during more-often-referred-to objects or landmarks are also more similar to other gestures in terms of their velocity profiles. In other words, we find larger distances for gestures that co-occur with novel references. This result can be explained in that gestures must be maximally informative when new information in speech is conveyed, which dovetails with findings that have shown that gestures are reduced in their production rate when they are referring to something that is becoming common ground in the discourse (Hoetjes, Koolen, Goudbeek, Krahmer, & Swerts, 2015) as well as with research showing that gestures are more prevalent when speech referents are less accessible (Debreslioska & Gullberg, 2019). The informativeness of a gesture might then be in part defined by its relative novel trajectory relative to the gesture ensemble. Gestures that are more similar to the rest of the gesture ensemble are less informative, and thus they are more likely to co-occur with speech that covers common ground.

Language-specific gesture kinematics in groups of speakers

With the demonstration of the group-level gesture network analysis, we hope to have sparked the interest for the quantitative assessment of large group-level discourse processes. For our mock question we were interested in whether gesture kinematics of U.S. or German speakers were different at the group level by relating them in on large-scale gesture network. This mock research hypothesis that gesture kinematics is directly constrained by language spoken is theoretically plausible, as we know that gestures might be changed in terms of their syntactical combination as a function of the language spoken (Kita, Alibali, & Chu, 2017; Kita & Özyürek, 2003). The current analyses, unfortunately, cannot provide any evidence for this given the many differences of the datasets that we used here next to the differences in spoken language. But the point is that these analyses as introduced here provide a methodological route to such evidence.

Other implementations of the current approach

The emphasis in the current article lies in providing one particular implementation of gesture network analysis. However, it is easy to develop other implementations for this approach. For example, where we have used DTW results as input for network analysis, attractor dynamics analysis (Borjon, Abney, Smith, & Yu, 2018) could also be performed and used as network's input. Such an analysis is interesting for gauging the *variability* of movement. For instance, we can compare two gestures on multiple kinematic dimensions of a gesture event (e.g., as state space containing acceleration + position) to gauge whether they occupy similar regions of state-space, regardless of their temporal ordering. Such analyses would not so much gauge whether gestures are similar in their absolute trajectory through time but rather whether gestures use *similar degrees of freedom* to perform different types of gestures. Another way to make comparisons between gestures is based on their spectral coherence (Pouw & Dixon, 2019). Then we could assess the question of whether some gestures oscillate at a shared frequency, whereas other gestures oscillate at different or more complex combinations of frequencies. Thus, there are endless ways you can construct a weighted matrix containing comparisons between gestures on some relevant dimension.

Note further that with respect to network analysis, we have not introduced analyses that map properties of the network as a whole (e.g., subnetwork analyses, overall connectivity measures). Instead, we have focused on quantifying gestures' position relative to other gestures in the network. The now-ignored subnetwork analyses provide means to test whether the gesture ensemble contains unique clustered subnetworks, which would indicate that underlying dynamics force gestures to bifurcate in one form or another. Thus, discovery of subnetworks in gesture ensembles could be helpful in motivating categorical differences between different gestures. Another potent application of the current network approach is to apply it to dyadic conversation research (Garrod & Pickering, 2009) or research on musical gestures (Hospelhorn & Radinsky, 2017; Pearson, 2013). For example, the current analysis can be used to assess gestural alignment between dyads, whereby two persons become more aligned in the way they gesture as time develops, gauged by the degree to which which a dyadic gesture network contracts as (DTW-calculated) distances between gestures become smaller. Finally, we note that network graphing can be valuable for qualitative analysts of gesture and conversation as well, as network graphing allows for an intuitive visual representation of gesture ensemble that might help in describing conversation dynamics.

In conclusion, we believe the versatility of gesture networks is endless, and we have argued that they can provide novel insights about multimodal language within and between individuals. We hope that the current approach becomes part of the general toolkit of gesture researchers, leading to potential novel discoveries about how higher-order discourse structure can emerge from gesture kinematics.

Notes

1. Unfortunately, ψ-DTW has not been implemented in `R` package `dtw` yet, but MATLAB code is made available by Silva et al. (2016).
2. The amplitude envelope referred to here is retrieved by taking the modulus of the Hilbert transform of the speech signal and then smoothing it with a low-pass filter (He & Dellwo, 2017). Please see https://osf.io/uvkj6/ for a helpful R script for extracting the amplitude envelope from audio files (Pouw & Trujillo, 2019). Consequently, this metric tracks the rough (i.e., envelope) intensity (i.e., amplitude) changes in speech. The amplitude envelope is a key dynamic property of the rhythm of speech and highly correlates with articulatory kinematics (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009).
3. Of note, the 2D graphical expression of the weighted matrix must always be an approximation of the weighted edges given that basic laws of triangles are violated when having to graph several weighted relationships of nodes in 2D. Igraph enables the approximation of the topology of a 2D representation of the weighted matrix, through a technique called multidimensional scaling. That network statistics are, however, performed with nonapproximated weights.
4. We only assessed iconic gestures as these are the gestures that are meaningfully related to what is said in speech (rather then how it is said as in the case of beat gestures).

Funding

This research has been funded by Netherlands Organisation for Scientific Research (NWO; Rubicon grant “Acting on Enacted Kinematics,” grant no. 446-16-012; PI Wim Pouw) https://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/91/28191.html.

Open data

The data and analysis script supporting this study are publicly available on OSF page https://osf.io/u46qd/.

ORCID

Wim Pouw http://orcid.org/0000-0003-2729-6502

References

Alviar, C., Dale, R., & Galati, A. (2019). Complex communication dynamics: Exploring the structure of an academic talk. *Cognitive Science*, *43*(3). doi:10.1155/S1110865704405101

Bergmann, K., Damm, O., & Freigang, F., Wittwer, *N*. (2014, January 31st). Documentation - Sagaland Version 2. Retrieved from https://www.phonetik.uni-muenchen.de/Bas/BasSaGADoku.pdf

Borjon, J. I., Abney, D. H., Smith, L. B., & Yu, C. (2018). Developmentally changing attractor dynamics of manual actions with objects in late infancy. *Complexity*, *2018*, 1–13. doi:10.1155/2018/4714612

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. doi:10.1371/journal.pcbi.1000436

Cook, S. W., & Tanenhaus, M. K. (2009). Embodied communication: Speakers' gestures affect listeners' actions. *Cognition*, *113*(1), 98–104. doi:10.1016/j.cognition.2009.06.006

Csardi, G., & Nepusz, T. (2006). *The igraph software package for complex network research. InterJournal, Complex Systems*, *1695*(5), 9.

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, *71*, 268–283. doi:10.1016/j.wocn.2018.09.007

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*. doi:10.3389/fpsyg.2016.00473

Debreslioska, S., & Gullberg, M. (2019). Discourse reference is bimodal: How information status in speech interacts with presence and viewpoint of gestures. *Discourse Processes*, *56*(1), 41–60. doi:10.1080/0163853X.2017.1351909

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, *328*(5981), 1029–1031. doi:10.1126/science.1186605

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(2012/12-0049)

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–171. doi:10.1111/cogs.12251

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292–304. doi:10.1111/j.1756-8765.2009.01020.x

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, *031*(i07), Retrieved from https://ideas.repec.org/a/jss/jstsof/v031i07.html

He, L., & Dellwo, V. (2017). Amplitude envelope kinematics of speech: Parameter extraction and applications. *The Journal of the Acoustical Society of America*, *141*(5), 3582. doi:10.1121/1.4987638

Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., & Swerts, M. (2015). Reduction in gesture during the production of repeated references. *Journal of Memory and Language*, *79–80*, 1–17. doi:10.1016/j.jml.2014.10.004

Hospelhorn, E., & Radinsky, J. (2017). Method for analyzing gestural communication in musical groups. *Discourse Processes*, *54*(7), 504–523. doi:10.1080/0163853X.2015.1137183

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior*, *32*(2), 123–131. doi:10.1007/s10919-007-0044-4

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. doi: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. doi:10.1016/S0749-596X(02)00505-3

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130397. doi:10.1098/rstb.2013.0397

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, *26*(10), 1457–1471. doi:10.1080/01690965.2010.500218

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The bielefeld Speech and Gesture Alignment Corpus (SaGA). *LREC 2010 Workshop: Multimodal Corpora–Advances in Capturing, Coding and Analyzing Multimodality*. Retrieved from https://pub.uni-bielefeld.de/publication/2001935

Madeo, R. C. B., Lima, C. A. M., & Peres, S. M. (2017). Studies in automated hand gesture analysis: An overview of functional types and gesture phases. *Language Resources and Evaluation*, *51*(2), 547–579. doi:10.1007/s10579-016-9373-4

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Weygandt Mathis, M., & Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*, 1281–1289. doi: 10.1038/s41593-018-0209-y

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

McNeill, D. (2000). Catchments and contexts: Non-modular factors in speech and gesture production. In Ed., D. McNeill *Language and gesture* (pp. 312–328).Cambridge: Cambridge University Press doi:10.1017/CBO9780511620850.019

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D., Quek, F., McCullough, K.-E., Duncan, S. D., Furuyama, N., Bryll, R., … Ansari, R. (2001). Catchments, prosody and discourse. *Gesture*, *1*(1), 9–33. doi:10.1075/gest.1.1.03mcn

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language*, *66*(1), 249–264. doi:10.1016/j.jml.2011.07.004

Mueen, A., & Keogh, E. (2016). Extracting optimal performance from dynamic time warping. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2129–2130). doi:10.1145/2939672.2945383

Müller, M. (2007). *Information retrieval for music and motion*. Heidelberg, Germany: Springer.

Pearson, L. (2013). Gesture and the sonic event in Karnatak music. *Empirical Musicology Review*, *8*(1), 2–14. doi:10.18061/emr.v8i1.3918

Pikovsky, A., Kurths, J., & Rosenblum, M. (2001). *Synchronization: A universal concept in nonlinear sciences*. Cambridge: Cambridge university press.

Pinheiro, J., Douglas, S. V., Debroy, U. T. S., Yes, L., & Yes, L. (2011). *Package "nlme". Linear and nonlinear mixed effects models, version, 3-1*.

Pouw, W., De Jonge-Hoekstra, L., & Dixon, J. (2018). *Stabilizing Speech Production through Gesture-speech Coordination*. doi:10.31234/osf.io/arzne

Pouw, W., & Dixon, J. A. (2019). Entrainment and modulation of gesture–Speech synchrony under delayed auditory feedback. *Cognitive Science*, *43*(3), e12721. doi:10.1111/cogs.12721

Pouw, W., Trujillo, J., & Dixon, J. A. (in press). *The quantification of gesture-speech synchrony: A tutorial and validation of multi-modal data acquisition using device-based and video-based motion tracking. Behavior Research Methods*.

Pouw, W, & Trujillo, J. P. (2019). *Materials Tutorial Gespin2019 - Using Video-based Motion Tracking to Quantify Speech-gesture Synchrony*. doi: 10.17605/OSF.IO/RXB8J

Quek, F. (2004). The catchment feature model: A device for multimodal fusion and a bridge between signal and sense. *EURASIP Journal on Advances in Signal Processing*, *2004*(11), 769219. doi:10.1155/S1110865704405101

Quek, F., Bryll, R., McNeill, D., & Harper, M. (2001). Gestural origo and loci-transitions in natural discourse segmentation. *IEEE Workshop on Cues in Communication* (p. 8). Hawaii.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., … Ansari, R. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, *9*(3), 171–193. doi:10.1145/568513.568514

Ravignani, A., Dalla Bella, S., Falk, S., Kello, C. T., Noriega, F., & Kotz, S. A. (2019). Rhythm in speech and animal vocalizations: A cross-species perspective. *Annals of the New York Academy of Sciences*. doi:10.1111/nyas.14166

Shattuck-Hufnagel, S., & Prieto, P. (2019). Dimensionalizing co-speech gestures. *Proceedings of the International Congress of Phonetic Sciences 2019* (p. 5). Melbourne, Australia.

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, *9*. doi:10.3389/fpsyg.2018.01514

Silva, D. F., Batista, G. A. E. P. A., & Keogh, E. (2016). On the effect of endpoints on dynamic time warping. *SIGKDD Workshop on Mining and Learning from Time Series, II*. San Francisco, USA.

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

Wallot, S. (2017). Recurrence quantification analysis of processes and products of discourse: A tutorial in R. *Discourse Processes*, *54*(5–6), 382–405. doi:10.1080/0163853X.2017.1297921

Webber, C. L., & Marwan, N. (2015). *Recurrence quantification analysis: Theory and best practices*. Heidelberg, Germany: Springer.

Xiong, Y., Quek, F., & McNeill, D. (2003). Hand motion gestural oscillations and multimodal discourse. *Proceedings of the 5th International Conference on Multimodal Interfaces* (pp. 132–139). Vancouver, Canada