Stabilizing Speech Production through Gesture-Speech Coordination

Wim Pouw[1,2], Lisette de Jonge-Hoekstra[1, 3] & James Dixon[1]

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

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

[3]The University of Groningen, Developmental Psychology

Word count (in text): 7.342

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

Open data: The analyses and data preparation code, as well as (raw) anonymized quantitative data are available on the open science framework (https://osf.io/9pj4d/).

Funding: This 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) and The Prins Bernhard Cultuurfonds (reference nr.: 40021263/PHR/ILE; PI Lisette de Jonge-Hoekstra).

Acknowledgements: We would like to thank Andy Lücking and Stefan Kopp for their helpful correspondence concerning questions we had about the SaGA dataset.

Abstract

Hand-gestures are seamlessly coordinated with speech. Yet, there is only anecdotal support for gestures' functional role in speech production. Here we explore temporal aspects of speech production when people use hand gesture. We performed exploratory analyses with a naturalistic German-speaking sample from The Bielefeld Speech and Gesture Alignment Corpus (SaGA), which consisted of 67 minutes of narration data and over 500 gesture events (N = 6). We found that the rhythmic timing of speech (defined as the mean and standard deviations of speech onset intervals) is highly correlated with the likelihood of gesturing. Furthermore, we utilized deep learning methods to track gesture motion, and extracted the amplitude envelope of speech, so as to gauge the degree of (continuous) gesture-speech synchrony. We then performed a continuous time-series analysis (recurrence quantification analysis; RQA) to index how temporal properties of speech change when gesture and speech are more or less synchronized. Our analyses revealed that when gesture and speech were more synchronized, the temporal structure of speech was more ordered and less complex, as indexed by classic measures of dynamic temporal stability (e.g., Entropy, Ratio of %Determinism/Recurrence). We suggest that a fundamental gesture-speech relation is rooted in entrainment, which yields stability in the temporal structure of speech.

Keywords: synchrony, hand-gesture, deep learning, motion tracking, acoustic analyses, recurrence quantification analysis

Introduction

One of the major ongoing questions in research on multimodal language is: Why do humans use hand gestures? A productive stream of research in this regard has focused on individual differences in cognitive predispositions, where it is found that lower working memory capacities consistently relate to a higher likelihood of gesturing (Chu, Meyer, Foulkes, & Kita, 2014; Gillespie, James, Federmeier, & Watson, 2014). Such results dovetail with findings which show that gestures support spatial problem solving by providing productive sensorimotor loops. Activating these loops may help stabilize otherwise unstable passive problem-solving processes (e.g., Chu & Kita, 2011; Pouw, Mavilidi, van Gog, Paas, 2016; Pouw, de Nooijer, van Gog, Zwaan, Paas, 2014). In addition, participants may be able to pick up invariant relations from such sensorimotor loops, and use these relations in problem solving. For example, simulating the alternating movements of interlocking gears increases the likelihood of discovering the concept of alternation (Stephen, Dixon, & Isenhower, 2009; see also Chu & Kita, 2008).

Yet this relatively productive line of research on the function of gestures in spatial problem-solving contexts does not readily translate into what is *the* paradigmatic context of gesture, communication. Gestures mostly occur in communicative contexts and primarily during speech production, as well as when speech is abstract and non-spatial in nature (McNeill, 2005). Thus evidence is still lacking for theories that propose a pivotal role for gesture in speech-production processes (e.g., Krauss, 1998; McNeill, 1992, 2005). Indeed, Hoetjes, Krahmer, & Swerts (2014) conclude that there is little convincing evidence for direct (beneficial) effects of gesture on speech properties (e.g., fluency, acoustic parameters), especially in the case of spontaneous gestures in natural contexts.

Moreover, in Hoetjes and colleagues' (2014) own comprehensive study, effects of gesture (vs. no gesture) were absent for a panoply of speech properties related to semantic fluency and acoustic correlates of prosody. In their study, participants explained tying a knot with and without gesture. Hoetjes and colleagues found no indication that gesturing affected speech duration, number of words spoken, speech rate, number of filled pauses, and acoustic properties, such as pitch (minimum and maximum F0, pitch range, or the mean). Corroborating such results, in a follow-up experiment, listeners that were provided speech samples of the participants of the previous study, could not hear the difference between gesture and no gesture conditions.

Although there is a lack of evidence for gesture's support of speech production, it can be argued that gesture researchers have been primarily interested in the semantic content of gesture and speech, focusing on representational gestures that depict an absent state of affairs (e.g., Goldin-Meadow & Brentari, 2017). By extension, the cognitive benefit of gesture has primarily been attributed to "priming", "informing", or "activating" mental semantic representations in some way or another (Krauss, 1998; Goldin-Meadow & Brentari, 2017; Kita, Alibali, & Chu, 2017; Hostetter, Alibali , & Kita, 2007; cf. Pouw, van Gog, Zwaan, Paas, 2017), so as to successfully summon concepts that the speech production system fails to achieve unimodally (i.e., semantic fluency). As such, it has been theorized that primarily semantics are in gesture's functional domain (e.g., Hoetjes et al., 2014; Hosetter & Alibali, 2008). Of course, there has been research on non-semantic aspects of the gesture-speech relation. For example, several studies have tapped into basic acoustic properties of speech that correlate with gesturing (e.g., Cravotta, Busà, & Prieto, 2018; Hoetjes et al., 2014; Krahmer & Swerts, 2007). However, this research has not yet been

translated into a mechanism of how gestures might support speech production through these acoustic effects, if any are found at all (Hoetjes et al., 2014).

We think the focus on semantics might have overshadowed one salient aspect of gesture-speech coordination, that is clearly reflected in the phenomenon of so-called beat gestures. Beat gestures do not have depictive qualities, but rather "beat" with the rhythm of speech (prosody). However, "beat gesture" is not a real exclusive gesture category that is isolable from gesticulation as such (Prieto, Cravotta, Kushch, Rohrer, & Vilá-Giminéz, 2018; Shattuck-Hufnagel & Ren, 2018), but rather a pervasive property of most types of gestures (e.g., Danner, Barbosa, & Goldstein, 2018; Krivokapić, Tiede, Tyrone, Goldenberg, 2016). That is, many, if not most, gestures beat with the rhythm of speech, even when they reserve degrees of freedom for deictic, iconic or metaphorical expression (Loehr, 2004; McNeill, 2005; Wagner, Malisz, & Kopp, 2014). Despite the pervasiveness of beat-like aspects of gestures, it is safe to say that gesture researchers are currently in the dark about the possible function of beat-like aspects for the gesturer. Beat-like aspects of gestures have, therefore, been primarily understood as a communicative tool, as they highlight information by movement intensity peaks (e.g., peak velocity, peak deceleration), in a similar way as acoustic peaks that constitute prosodic contrasts in speech (Krahmer & Swerts, 2007; Leonard & Cummins, 2010).

There is preliminary evidence for supporting functions of beat-like gestures for the gesturer. For example, in a recent motion-tracking and acoustic analysis study (Pouw & Dixon, under review) we have found that when speech is disrupted due to a delayed auditory feedback, gesture (beat and iconic) and speech become more synchronized (cf. McNeill, 1992). That is, gestures' peak velocity was less variably aligned with peak in F0 (or

pitch) when speech was perturbed by an interfering signal in the form of delayed auditory feedback (as opposed to no interference). We speculated that gesture might aid in keeping a *stable rhythm*, such that speech and gesture entrain one another to attain a more stable rhythmic regime that is less affected by interfering signals. This may be compared to findings in speech pathology research, which showed that people who stutter improve drastically in their speech fluency when trained to speak to the external stable rhythm of a metronome (e.g., Brady, 1971; Davidow, 2014). By analogy, beat gesture may serve as a physical metronome that is (bidirectionally) coupled with speech.

This general idea is also supported by studies on second-language learning, which show that learners improve their pronunciation if they make beat gestures while learning to speak the second language (as rated by independent raters; Gluhareva & Prieto, 2017; see also Kushch, 2018; Prieto, Llanes-Coromina, & Rohrer 2018). Therefore, it is possible that one of the primary functions of gesture is to entrain to speech, with a stable *gesture-speech system* as a result (see also Iverson & Thelen, 1999; Rusiewicz, 2011; Rusiewicz & Esteve-Gibert, 2018; Treffner, Peter, Kleidon, 2008). Entrainment, then, is the outcome of the *emergent interplay of two systems that are simply more stable together.*

In the current paper, we explored a two-part question with a variety of analyses.

Part A) Is the rhythmic timing of speech related to whether speech co-occurs with gesture?

Part B) If so, then does the degree of gesture-speech coupling predict changes in temporal structure of speech?

For part A) we assessed whether the presence or absence of gesture was related to consistent changes in *speech onset intervals*. For part B) we *first* assessed gesture-speech synchrony through motion-acoustic analyses (e.g., Pouw & Dixon, under review; Danner et al., 2018); specifically, we assessed the consistency of timing between the peak velocity of a gesture and the peak amplitude envelope of speech. For the gesture motion analyses, we applied novel deep learning methods to quantify hand motions from video (Mathis et al., 2018; see also Pouw, Trujillo, Dixon, 2018). We *then* used this measure of gesture-speech synchrony to predict temporal changes in speech as indexed by non-linear time-series analyses called *Recurrence Quantification Analysis* (e.g., Jackson, Tiede, Riley, Whalen, 2016; Jackson, Tiede, Beal, & Whalen, 2016; Webber & Marwan, 2015).

These analyses were performed on an open data set called the Bielefeld *Speech and Gesture Alignment* Corpus (SaGA) which contains rich annotations of speech and gesture observed from direction-giving dialogs (Lücking , Bergmann, Hahn, Kopp, Rieser, 2010; Lücking, Bergman, Hahn, Kopp, & Rieser, 2013; for implementations see Bergmann & Kopp, 2010). As this data has annotations of word onset as well as hand gesticulation, it was ideally suited to test our exploratory hypothesis that properties of gesturing and of speech rhythm are related.

Method

Original Study

We requested access to the SaGA dataset from the BAS CLARIN Repository (Reichel, Schiel, Kisler, Draxler, Pörner, 2016) which contains a sample of 6 German-speaking participants (67.76 minutes of video material; 5 males), consisting of 280 minutes of video material of direction-giving dialogs (which has not yet been made freely available; see Lücking, et al., 2013). The SaGA dataset has been extensively tested for inter-rater reliability (for full details see Lücking, et al., 2013), and has so far been primarily used for machine learning purposes (e.g., Bergmann & Kopp, 2010) and recently by Hassemer & Winter (2018) for semantic analyses of gesture.

The SaGA dataset comes with numerous of annotations made available in ELAN (Lausberg & Sloetjes, 2009; for full detail of the SaGA dataset see the manual; Bergmann et al., 2014). We extracted *speech transcripts* that were based on PRAAT-assisted timing of word onset and duration (Boersma, 2001). We also extracted the *gesture annotations*, which indicated when participants gestured with either the right or the left hand, or both.

Procedure Original Study

The participants were "router-givers" who were giving directions in the context of a virtual reality (VR) town. The route-giver was provided a virtual tour along several landmarks and was subsequently asked to describe the route and the landmarks to the "follower" who was sitting in front of the speaker. For full details of the procedure see (Lücking et al., 2013).

Analyses Part A: Likelihood of gesture and the rhythm of speech

The following analyses were used to assess whether the temporal structure of speech was related to whether speech co-occurred with gesture (see Figure 1 for a graphical overview of our analyses procedures).

Annotations (speech transcript, gesture events for left and/or right hands) were sampled into a time series using a custom written script in R, at 25Hz (40ms per data point)[1]. As a measure of the rhythm of speech, we computed from the speech transcript annotations the mean and the standard deviation of the inter-speech-onset-interval, for each sentence within separated blocks of 100 *fluidly* spoken words (*see below*). Inter-speech-onset-interval is simply the time (in ms) between the onset of each spoken word and the next word of that sentence. Inter-speech-onset-interval was calculated for all successive words in a sentence.

The *mean* of the inter-speech-onset-interval (i.e., the average of inter-speech-onset-intervals across the sentences in a block) reflects the relative speed/tempo of speech (higher inter-speech-onset-intervals indicate slower speech rate). The *standard deviation* of inter-speech-onset-intervals reflects the variability around the mean of inter-speech-onset-interval. Lower standard deviations of inter-speech-onset-interval reflect that speech word onsets are *more consistently paced*, i.e., speech follows a more isochronous (i.e., equally paced) rhythm when the standard deviation of inter-speech-onset-interval is low.

As a measure of gesture likelihood, we computed the relative occurrence of gesture during each of the 100 spoken words. For example, if for a 100 word block, 80 of the words spoken were occurring while gesturing, the gesture likelihood measure would give a .8

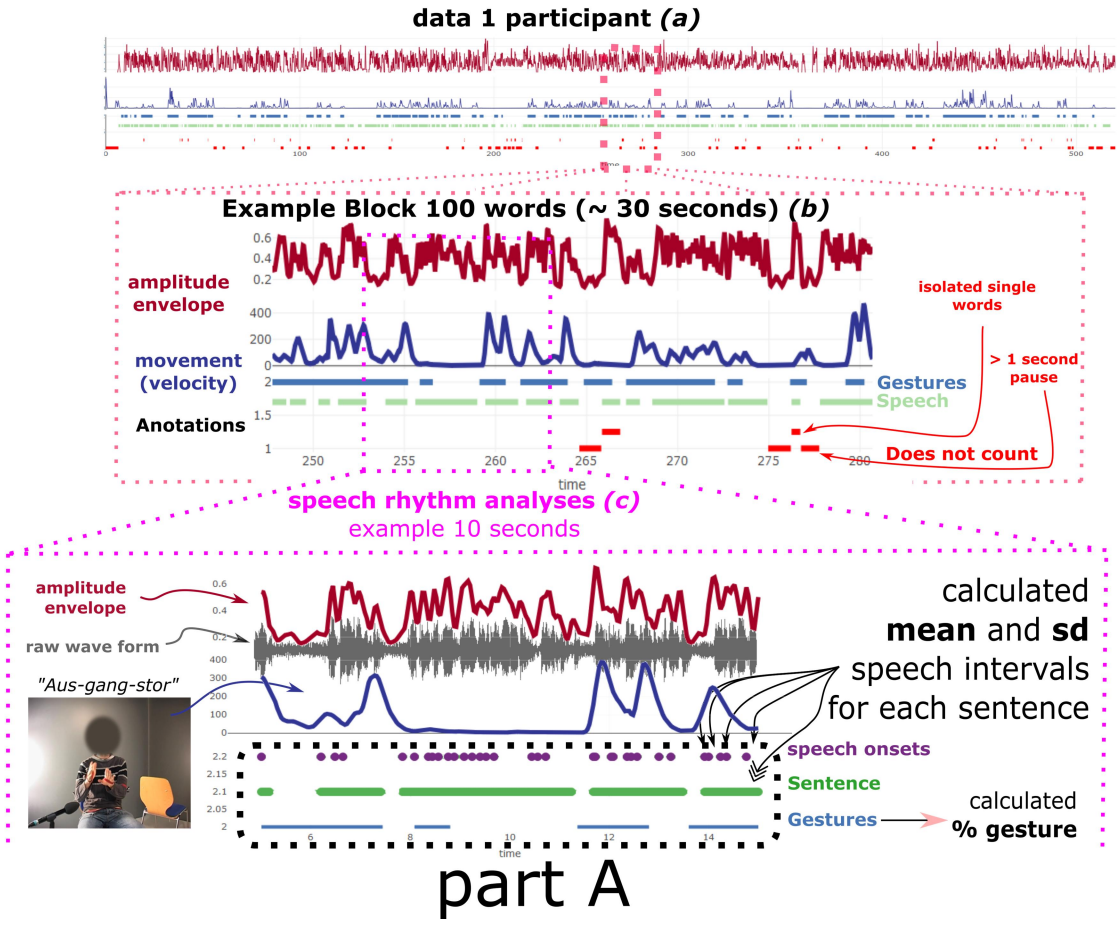
[1] Data analyses scripts and data can be retrieved at: https://osf.io/9pj4d/.

gesture likelihood (i.e., range = [0-1]; 0 = no gesture, 1 = always gesture). When reporting the analyses, we will also introduce a scaled version of this measure which controls for the amount of time over which speech occurred. Note that we chose a fixed number of words (i.e., 100) to ensure that the amount of speech was controlled for, and therefore cannot serve as a simple explanation of differences in gesture likelihood.

Our rhythmicity and gesture-likelihood measures were only computed on segments with continuously occurring speech. All spoken words that were not part of a multi-word sentence (which could be filler words like "uhm") were excluded for data analyses. Furthermore, there were segments for each participant where there was no speech, such as pauses, breaths, or when the interlocutor asked for a clarification. To ignore these non-speech segments, anytime a pause took longer than 1 second we interpreted this as sentence boundaries (see e.g., McClave, 1994 for a similar cut-off time). Obviously, time between sentence boundaries was not used in calculations for inter speech intervals. This ensures that our measures predominately track the rhythm of *fluid continuous speech*.

Descriptives. The current procedure resulted in a parsed dataset of a total of 79 fluid-speech blocks of 100 words each (i.e., total of 7900 words spoken). These blocks had a mean gesture likelihood of 61.7%. Averaged across all blocks, the mean inter-speech-onset-interval was 236ms (*SD* = 34), and the average standard deviation of the ISI was 160 ms (*SD* = 92 ms).

Figure 1. Graphical overview pre-processing and analyses

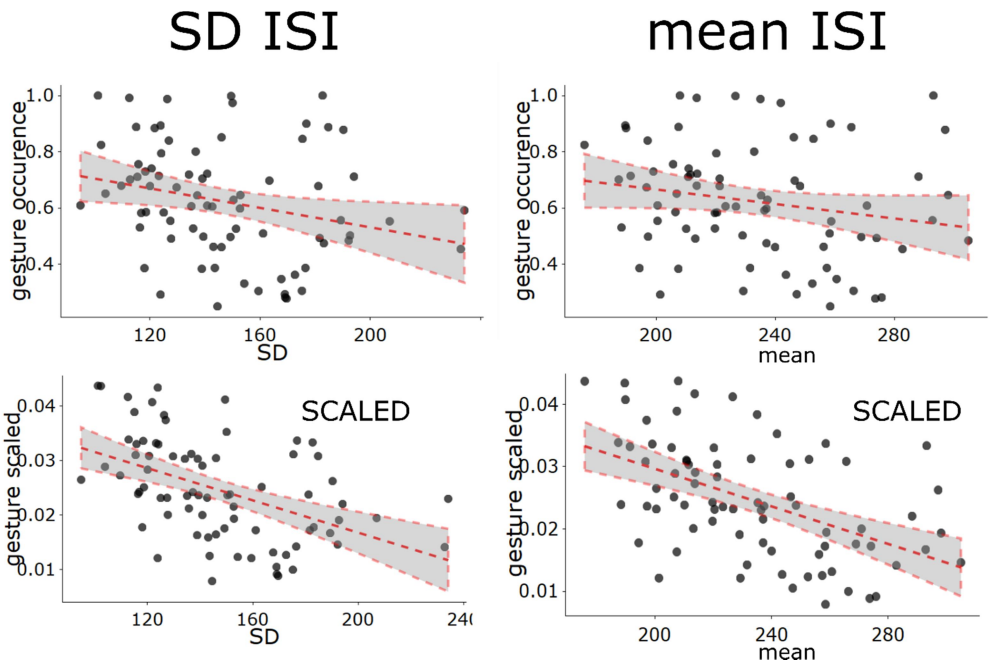


Note Figure 1. Schematic showing the processing steps for part A of the analysis. Panel (a) shows the full annotation and time-series data for one participant. Panel (b) shows a 100-word block segment within the longer time series. In dark red the amplitude envelope time series is shown which we extracted from the raw audio waveform; see panel (c) for example of how amplitude envelope relates to raw audio waveform. The blue time series shows an example the velocity time series of gesture, which indicates movement intensity (of the right hand). The time series data will be submitted for the analyses of Part B. For the current analyses for Part A the speech onsets and the gesture likelihood are important; Panel (c) shows a subsection of a 100-word block in which we illustrate how speech intervals were calculated within sentences (mean and SD) and how they are related to gesture occurrence.

Results Part A

As described, two summary statistics are important for inter-speech-onset interval (ISI). Firstly, lower standard deviations of ISI indicate a more consistent rhythmic timing in speech. Secondly, the mean of ISI provides the mean tempo of the speech rhythm. Figure 2 shows the correlations between gesture likelihood (y-axis) and the SD of word-onset-interval (left top panel) as well as the mean of the word-onset-interval (right top panel). Note that for the forthcoming analyses two prominent outlier data points were excluded (see https://osf.io/n2erk/ for supplemental figures with outliers), and thus final dataset consists of 77 blocks.

Figure 2. Relationship with speech rhythmic timing and gesture likelihood



Note Figure 3. Every point (n = 77) reflects measures calculated for a window of 100 words spoken (77 windows in total). Likelihood of gesture is based on either a left or right hand gesture (or both) occurring during speech. The graphs show that if words are spoken in a more consistent rhythm (lower SD ISI), and if words are spoken with a higher tempo (lower mean), we find a higher percentage of the time spent gesturing. This is regardless of whether we scale gesture likelihood by the time duration of that block (see scaled below). In fact, the effect seems to be more pronounced for the scaled measure.

The graphical results show that when speech is more consistently rhythmic (lower SD ISI), there is a higher likelihood of gesture, r = -.26, 95%CI[-.45, -.035], t(75) = -4.712, p = .024. When rhythmic of speech has a higher tempo, a higher gesture occurrence was

observed, but not to a statistically reliable degree, r = -.19, 95%CI[-.40, -.027], t(75) = -5.05, p = .083.

Note, that our method excludes the possibility that our measures are affected by significant differences speech duration (as we omitted all silent pauses longer than 1 second), as well as differences in the amount of words spoken (i.e., the amount of speech content), as all blocks are of the same size of 100 words. Yet, it is possible that the percentage of gesturing during a given block of 100 words is dependent on the *time duration* of the block. A simple calculation shows that shorter ISI's give a shorter time of the block of 100 words (e.g., 100 words x *250ms mean interval* = 25 seconds; 100 words x *200 ms mean* interval = 20 seconds). To assess whether amount of time is affecting our gesture likelihood, we scaled the gesture likelihood by the amount of time in milliseconds (of that block). Importantly, the relationship for gesture likelihood (now scaled for time) reveals a reliable correlation for standard deviation in ISI, r = -.48, 95%CI[-.63, -.28], t(75) = -4.712, p < .001. The same is true for the correlation between mean ISI and the scaled gesture occurrence, r = -.50, 95%CI[-.65, -.32], t(75) = -5.05, p < .001).

Next, we ask whether the speed (mean) or the consistency (SD) of ISI rhythm is a better predictor of gesture occurrence (scaled for time). When only mean-ISI was entered as a predictor for gesture occurrence, this did not reliably improve predictions as

compared to a model that included the overall mean, change in $\chi 2$ [4] = 3.795, p = .051. However, adding SD-ISI as a second predictor to the previous model, did reliably improve predictions in gesture occurrence, change in $\chi 2$ [5] = 6.58, p = .011. This final model showed that SD-ISI had a significant effect (*b* =-0.0025 [95%CI:-0.0044, -0.0006], *t* (69) = -2.59, *p* = .012), but mean-ISI did not (*b* =-.0005 [95%CI: -0.0015, 0.0025], *t* (69) = 0.497, *p* = .620)[2]. We note that despite the correlation between mean-ISI and SD-ISI, multi-collinearity (e.g., Field, Miles, Field, 2012) was low (VIF =2.305).

In sum, the current analyses reveal that the consistency of the rhythmic timing of speech on the word-level is significantly correlated with gesture likelihood. Specifically, we provide evidence that the lower variability of inter-speech-onset intervals (ISI's) is related to more gesture occurrences.

[2] Note further that if we enter SD-ISI as a first predictor and mean-ISI as a second predictor to the model, similar results obtained. Such that SD-ISI was the only statistically reliable predictor.

Intermediate Discussion Part A

The previous rhythmic-timing analyses show that speech which co-occurs with more gesture unfolds in a temporally different way as compared to speech with less co-occurring gestures. Specifically, the intervals of word-onset are more uniformly rhythmic (more consistently timed) when speech is accompanied by more gesturing. This analysis provides a promising indication that gesture and temporal structure of speech are intimately related. However, more evidence is needed to claim that gesture-speech coordination affects the temporal structure of speech.

One limitation of the current analyses is that speech rhythm or prosody is a multidimensional property that depends on a range of acoustic properties (e.g., Fundamental Frequency, Amplitude Envelope) that play out dynamically over time (e.g., Cummins, 2009). In the previous analyses we have however reduced speech to the statistical regularity of word-onsets as a proxy for rhythm. This is a rather strong simplification of rhythmic timing, if only because the *perception* of speech rhythm is very more likely to be determined by perception of syllable-centers rather than word onsets. Note further that words will have different number of syllables, and the syllable lengths themselves can vary widely, especially in stress-timed languages like German (Roach, 1982). Thus our conclusions would be much stronger if we also find changes in temporal

structure changes in speech on an acoustic dimension that is relevant to listeners. To resolve this, we will look at another more sensitive property of the rhythm of speech, called the amplitude envelope, which has been understood as an important determinant for the rhythm of speech (Chandrasekeran, Trubanova, Stillitttano, Caplier, & Ghazanfar, 2009; Tilsen, & Arvaniti, 2013).

In the next analyses, we will apply a more continuous approach to measuring the movement properties of gesture as well as the gross amplitude fluctuations in the speech signal. This allows us to compute a measure of gesture-speech synchrony by relating peak velocity of a gesture with peaks in the amplitude envelope, thereby creating a continuous measure of the strength of gesture-speech coupling. If gesture and speech are mutually entraining such that they jointly support language production, the strength of gesture-speech coupling should predict properties of the temporal structure of the speech signal. To test this idea, we performed Recurrence Quantification Analysis (RQA; (e.g., Jackson et al., 1996; Riley, Balasubramaniam, & Turvey, 1999) on the amplitude envelope. RQA provides measures of the stability and complexity of continuous speech. We then relate these RQA measures of the structure of speech to the degree of gesture-speech coupling (synchrony of peaks in gesture and speech).

Analyses Part B: Continuous time-series analyses of gesture-speech synchrony and dynamic properties of speech

For a graphical overview of the analyses procedures for part B see Figure 3. Firstly we will introduce the continuous measures for gesture motion and speech. Subsequently, we will introduce the motion-acoustic analyses with which we measured gesture-speech synchrony. Then we introduce the time-series analyses (RQA) to assess temporal structure of continuous speech.

Continuous Speech: amplitude envelope

For continuous tracking of speech, we obtained the amplitude envelope from the audio stream of the video data[3]. The amplitude envelope is a measure that tracks gross changes (the envelope) in the amplitude of the raw waveform, while ignoring fast time-scale changes. The amplitude envelope has been suggested as an adequate measure to track the rhythmic structure of speech (Tilsen, & Arvaniti, 2013), and has further been found to correlate with labial (lip) movements during speech production (Chandrasekeran, Trubanova, Stillitttano, Caplier, & Ghazanfar, 2009). The amplitude envelope was computed in PRAAT, using the PRAAT script made available by He & Dellwo (2016, 2017). This script applies a Hilbert transform to compute the envelope, and provides a scaled measure in Hilbert Units ranging from (0 to 1). A score 0 is the lowest amplitude level obtained for that

[3] Note, that in the current data there are segments wherein the interlocutor spoke to the participant, and in such cases audio was cut out by the original researchers that produced the SAGA corpus. Since those cuttings were not occurring often when the participant was producing fluid speech and gesture, such cuts are not likely to affect our analyses.

time series, and 1 is the highest amplitude obtained. The amplitude envelope was resampled at 25Hz to match sampling rate of the annotations and the motion tracking time series.

Continuous Gesture movement: Deep Learning Motion Tracking

Although the SaGA study originally involved motion-tracking of both hands (using the VR system), this data was not deemed reliable by the original researchers and was thus not part of the open dataset. Therefore, to track motion of the hands we utilized a novel deep learning method called 'Deep Lab Cut' developed by Mathis and colleagues (2018). This deep learning method has already been successfully applied in animal motor control research, such as the study of drosophila and mouse behavior, and has a good performance for temporal estimation of movement intensity peaks (Pouw, Trujillo, Dixon, 2018). We used this method to train a deep neural network to recognize 2D palm positions (x, y coordinates for each hand) from the original SaGA video data. The deep neural network was already pre-trained for animal pose estimation (ResNet with 50 layers; He, Zhang, Ren, & Sun, 2016). We re-trained the network on the basis of 150 randomly chosen frames that were hand-annotated (using ImageJ; Rueden, Schindelin, & Hilner, 2017) for left and right hand-palm positions. We re-trained the network for 250,000 iterations, yielding a mean deviance from computer estimated hand-palm positions versus experimenter annotated positions of about 4 pixels, which is highly accurate performance. The reader can go here (https://osf.io/bxdt6/) to see a sample of the high-accuracy motion tracking that 'Deep Lab Cut' affords (or download the complete motion-tracking dataset at OSF). To remove artifacts of jitters and sudden incorrect estimations in the movement signal, we smoothed movement time series (x, y) and time-derivatives (velocity), with a first-order low-pass

Butterworth filter (10 Hz). To reduce the complexity of the analyses, we only took into account movement of the right hand (2D velocity computed from x, y coordinates); all participants gestured with their right hand predominately. For further details of the deep learning motion tracking 'DeepLabCut', see Mathis & colleagues (2018) with respect to accuracy, and the website (http://www.mousemotorlab.org/deeplabcut/) of the Adaptive Motor Control Lab for tutorials and implementations of the method.

Motion-Acoustic Analyses: Gesture-speech synchrony

In previous research (Pouw & Dixon, under review), we quantified gesture-speech synchrony by temporally comparing movement intensity peaks in gesture (e.g., peak acceleration, peak velocity) with prosodic markers in speech (e.g., fundamental frequency [F0; perceived as pitch]; see Danner et al., 2018, for comparable approach). We used the amplitude envelope[4] as a prosodic marker to be related to gesture motion by monitoring abrupt changes in this signal (positive peaks in the time-derivative of the amplitude envelope, i.e, amplitude-envelope *change*). Peaks in amplitude-envelope change indicate sudden fluctuations in the amplitude of the speech signal, which we observed to often coincide with syllable onsets. To compute asynchronies for each gesture-speech event, we obtained: a) when the highest peak in velocity of the gesture occurred, and b) when the highest positive peak in the amplitude envelope-change was found in the speech signal. The difference between these two times (gesture peak velocity – speech amplitude-envelope change) gives a measure of the temporal synchrony between gesture and speech (Δt, in

[4] We found that F0 estimation with the current data was not as reliable as the amplitude envelope estimation, and therefore chose intensity changes as a prosodic marker.

ms). We computed Δt for all gesture-speech events within a 100-word block, and then computed the mean and standard deviation of the Δt within that block (Figure 1). The mean Δt gives a measure of the lag between gesture and speech. The standard deviation of Δt gives a measure of the strength of the coupling between gesture and speech for that block (i.e., stronger coupling relationships should be more stable).

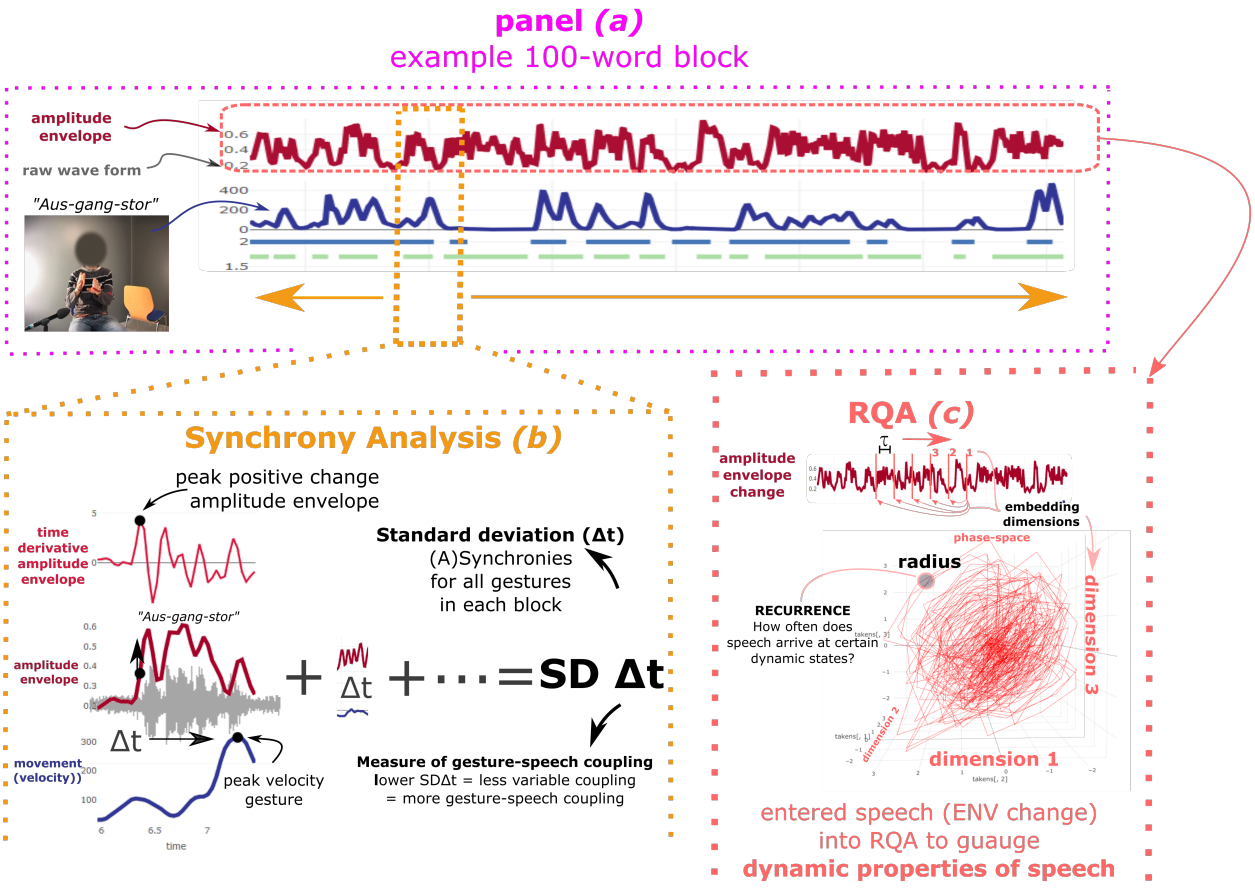
Averaged across blocks, the mean synchrony between speech and gesture (for blocks that contained more than 50% gesture) was -62 ms (*SD* = 292), indicating that peak velocity gesture *preceded* peak in amplitude envelope change by 62 ms. The average standard deviation of gesture-speech synchrony for all blocks was 828 ms (*SD* = 324).

Dynamic analysis of speech: Recurrence Quantification Analyses

Because the standard deviation of Δt provides a measure coupling between gesture and speech, we hypothesized that it would predict aspects of the temporal structure of the speech signal itself. To test this overarching hypothesis, we first analyzed the amplitude envelope of the speech signal using Recurrence Quantification Analysis. This well-established non-linear method, originating from physics, and often used in perception and action research (Webber & Zbilut, 2005; Riley et al., 1999) has also been applied in speech research (Jackson et al., 2016; Lieshout & Namasivayam, 2010). For example, based on RQA indices of temporal structure of lip movement, Jackson and colleagues were able to predict whether a spoken sentence was embedded in a more or less complex linguistic structure. Lieshout & Namsivayam showed that predictability of lip-tongue movements during speech production (as indexed by RQA) was related to whether speech was produced in a higher or lower tempo. Our analyses tie into these previous RQA analyses with movements of

speech-articulators, given that the amplitude envelope that we have extracted is known to be correlated closely to lip movements during speech (Chandrasekaran, et al., 2009).

Figure 3. Overview analyses Part B (time series analyses)



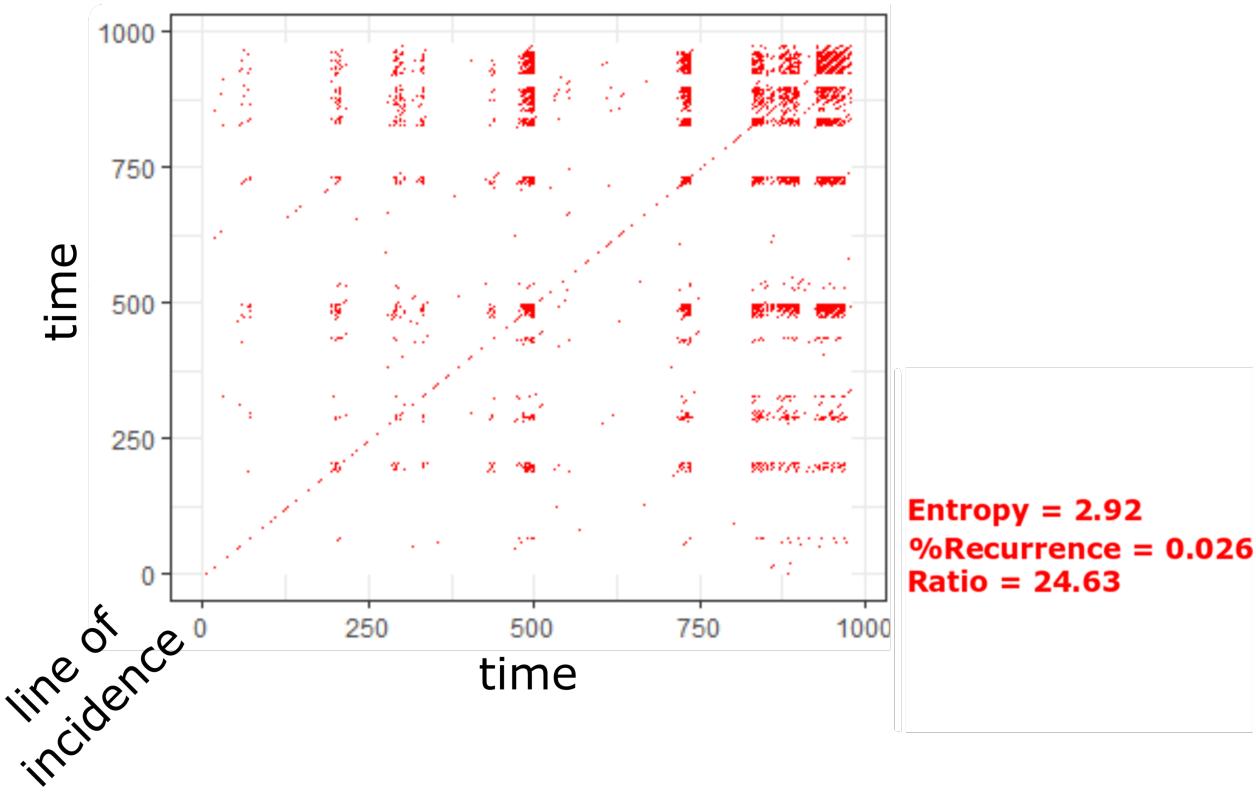
Note. Panel (a) shows a 100-word block (as shown in Figure 1 as well). Recall that amplitude envelope is the dark red line. The blue line shows the velocity of gesture movement. Panel (b) shows that for each gesture-speech event in the 100-word block the peak in velocity is determined, as well as the peak of positive amplitude envelope change, which is then compared for synchrony between the peaks in milliseconds. For all gesture-speech events occurring in a 10- word block, the standard deviation for G-S synchronies are calculated (but only for blocks with more than 50% gesture; see results below). Panel

(c) shows a schematic explanation of Recurrence Quantification Analysis (RQA) which we performed on the amplitude envelope change time series. In this example, it is shown how the signal is assessed through time, by plotting the signal with 2 delays (i.e. 2 times 40ms) of itself over time twice, yielding a 3-dimensional reconstructed "phase-space" (see section RQA-section in method for further information and the supplemental materials https://osf.io/4qt7w/). For this block the speech signal visited a region of phase-space within the size of the radius (i.e. the greater the radius, the larger this region) for three times, showing a recurrent state. RQA calculates these recurrences for all regions of state-space, and can do this for state spaces with more than 3 dimensions. For example, for our current analyses we assessed recurrences for a 12-dimensional state-space. For our analyses, the results of panel (d) and (e) are related to assess the relation between gesture-speech synchrony and the temporal structure of speech.

General Procedure RQA. Note that a more detailed overview of RQA method can be found in the supplemental materials: https://osf.io/4qt7w/. RQA tracks the degree to which a system revisits regions in its phase space (i.e., set of all possible states) over time. RQA is often combined with a method called phase-space reconstruction (Abarbanel, 1996). Phase-space reconstruction involves embedding *one* signal against delayed versions of itself, thereby reconstructing a topological version of the *whole* systems' (n-dimensional) phase-space (see Figure 3 panel e). RQA builds a record of each time the system returns to a region of phase space (i.e., a "recurrence), indicated by points in the corresponding recurrence plot (see Figure 4).RQA quantifies aspects, such as line

structures, of this recurrence plot to assess dynamic properties of system behavior, such as stability/variability and predictability/randomness.

Figure 4. Example of a recurrence plot of the speech signal



Note. This is a recurrence plot of actual data of the same speech time series that is shown in 100-word block panel of Figure 1. Each red-colored dot indicates that the signal revisited a region in phase space that was visited earlier. The patterns of recurrences in this plot are very complex, resulting in a varied range of line structures. This complexity is captured by the Entropy measure. %Recurrence, indicates that the degree to which the system returns to previous states. Lastly, Ratio is the %Determinism scaled by %Recurrence rate (which measures density of recurring points), and is an indication for a system's predictability (Zbilut & Webber, 2006).

To perform RQA we used functions from the R package 'nonlinearTseries' (Garcia, 2015), which also allowed us to estimate important parameter settings that define phase-space reconstruction methods, namely the time lag, number of embedding dimensions, and radius via the standard methods (see Webber & Marwan, 2015, for details; also see supplemental materials: https://osf.io/4qt7w/). We used a time lag of $\tau = 2$ (80 milliseconds), and an embedding dimension of $m = 12$. We determined that a radius of 1.2 yielded desirable spread of data while recurrence rates were kept relatively low with a mean percentage %recurrence = 4.14% (SD = 3.38)[5].

Key measures RQA. To gauge the temporal structure of continuous speech, we focused on the following three RQA indices (also see Figure 4). Percent recurrence (%Recurrence) indicates the degree to which the system repeatedly visits states. For example, in Figure 3 for the RQA example, the system re-visits a region in phase-space three times, and the state corresponding to this region is therefore a point in the recurrence plot (see Figure 3). Higher %Recurrence indicates that the system is attracted to a certain stable behavioral regime, as it keeps visiting previously occupied regions in state-phase.

Recurrence plots can further differ in the degree that the points form diagonal lines. A diagonal line indicates that the system repeatedly visits the same *sequence* of states. The RQA measure 'Entropy' is the Shannon entropy of the frequency distribution of these

[5] Note that the pattern of results we observed were robust to changes in these parameters. It has been argued that RQA results are relatively robust for different parameter changes of dimensionality and time lag (Webber & Zbilut, 2005).

diagonal line lengths in the recurrence plot. For simple oscillatory systems that behave in a well-ordered manner (e.g., periodic sine waves), the entropy tends to be very low, as the diversity in line lengths is limited. For more complex signals (e.g., Lorenz system; speech signal), sequences of previously visited regions in phase space occur in a more disordered manner, and entropy will be generally higher. Thus, Entropy allows us to quantify the variety and complexity of the system's continuous behavior.

Another measure we use is Ratio, which is a scaled version of %Determinism. %Determinism is the percentage of points in the recurrence plot that are part of diagonal lines. Thus, %Determinism tracks whether the signal consistently returns to previous trajectories. As such, high %Determinism suggests a more predictable system, while low determinism indicates a less predictable system. Ratio is simply the %Determinism scaled by the amount or density of observed recurrences, and has been proposed (Webber & Zbilut, 1994) as a measure of non-linear stability (e.g., Vohs, Mohr, Kryzwanek, 2002). Lower Ratio indicates a lower stability.

In sum, we will focus on three RQA measures, %Recurrence, Entropy, and Ratio, which track the tendency of the system to revisit states, the complexity of the signal, and its stability, respectively. With these measures of the temporal structure of speech (through RQA) now in place, we can now assess whether such measures are related to the gesture-speech synchrony measure that we obtained via the motion-acoustic analyses. That is, does the degree of gesture-speech coupling predict changes in temporal structure of speech?

Results Part B

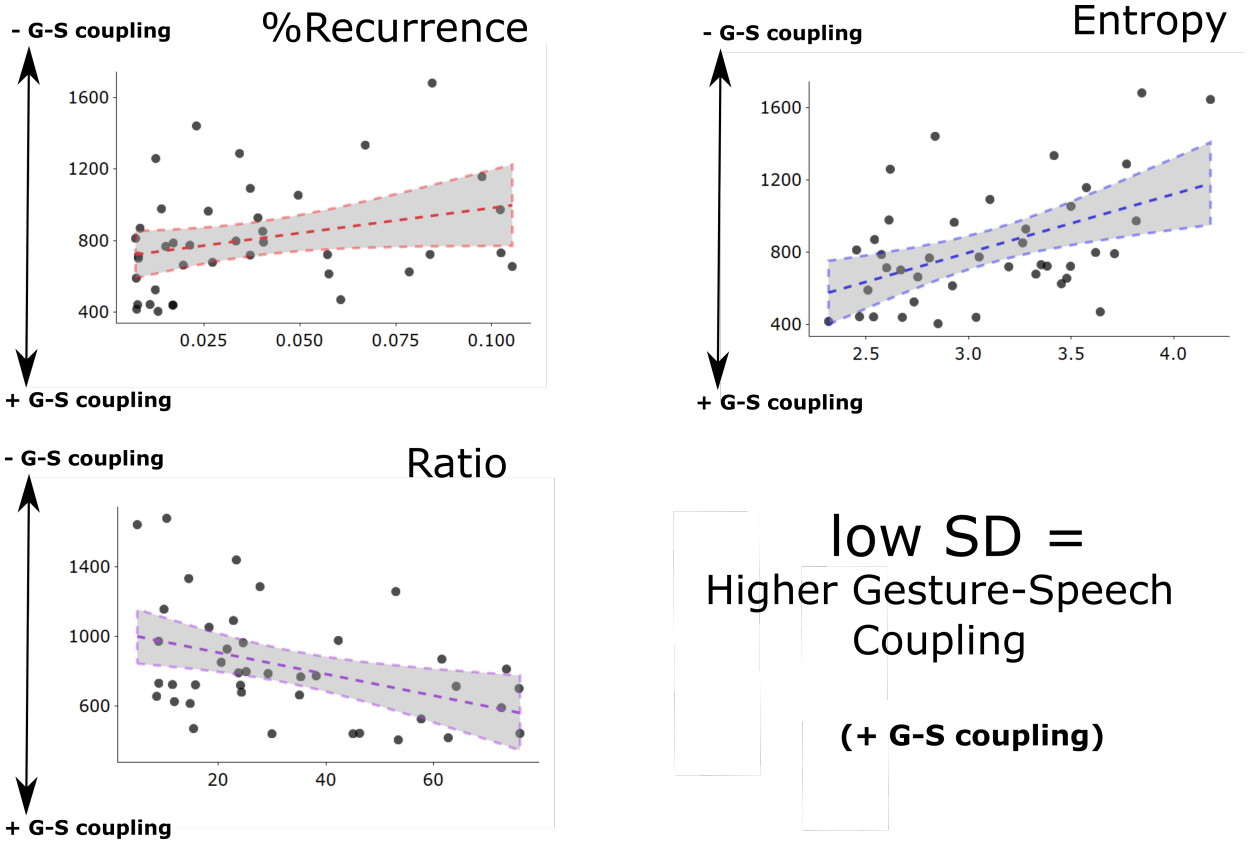
As described in the previous sections, we computed the coupling strength as the *standard deviation of the gesture-speech asynchronies* (i.e., between (a) the peak velocity of the gesture and (b) the peak positive change in the amplitude envelope), for all gestures in each block of 100 words. Lower standard deviations indicate that gestures are *more strongly* (i.e., less variably) coupled with speech, and the gesture-speech synchrony is thus more consistent (an indication of higher gesture-speech coupling). Note that we can only perform these analyses for blocks in which speech that co-occurred with gesture to some substantial degree. We, therefore, performed these analyses on blocks in which gesture and speech co-occurred at least 50% of time; 41 of the 77 blocks were retained for analysis under this criteria.[6].

The results show that the degree of gesture-speech coupling strength was reliably correlated to the temporal structure of speech as indicated by our RQA analyses. Namely

[6] Note that we also looked at blocks where 70 percent of the time gestures occurred (n = 19), and this yielded similar results as the ones that we are about to present.

there were reliable relationships between gesture-speech coupling strength and Entropy, Recurrence and Ratio[7] (see Figure 5 as well as explanation below).

Figure 5. RQA and gesture-speech synchrony



Note Figure 5. Bivariate relationships between Entropy, Recurrence, and Ratio, and the SD of Gesture-Speech (G-S) Synchrony. Lower values of SD G-S Synchrony indicate stronger gesture-speech coupling.

[7] For these analyses we excluded one impossible outlier on the gesture-speech synchrony dimension, e.g., see graph with outlier (Entropy and Synchrony): https://osf.io/zeth4/.

Recurrence was not significantly related to gesture-speech coupling strength (r = .289, 95%CI[.-.025, .550] , *t*[38] = 0.29, *p* = .071), thus gesture-speech coupling does not appear to significantly increase the degree to which the system returns to previous states (although we excluded one outlier that did increase the current effect; see supplementary graph https://osf.io/s6bd2/). We did find a reliable positive correlation between gesture-speech coupling strength and Entropy (*r* = .477, 95%CI[.198,.684] , *t*[39] = 3.386, *p* = .002), indicating that when gesture-speech coupling is stronger (i.e. lower SD gesture-speech synchrony), the temporal complexity of speech is lower. Ratio was reliably and negatively related to gesture-speech coupling strength, (r = -.409, 95%CI[-.116, -.637] , *t*[39] = -2.80, *p* = .008). This suggests that the stability of the speech signal was greater when speech was more strongly coupled with gestures[8].

[8] For the reader familiar to RQA (or see e.g. Marwan et al., 2007 for more information), note that other RQA measures also showed reliable correlations with gesture-speech coupling. For example %Determinism was also related to gesture-speech coupling, (*r* = .328, 95%CI[-.023, .557] , *t*[39] = 2.17, *p* = .036), wherein higher predictability of speech was found for higher-gesture speech coupling. Laminarity (*r* = .360, 95%CI[.059, .60] , *t*[39] = 2.412, *p* = .021) and Average Line Length (*r* = .53, 95%CI[-.276, .720] , *t*[39] = 3.912, *p* < .001) also

General Discussion

With an open dataset (SaGa dataset) containing more than 500 gestures and 30 minutes of non-stop speech, we assessed whether the temporal structure of speech is related to gesturing. Specifically, we investigated a two-part research question:

Part A) Is the rhythmic timing of speech related to whether speech co-occurs with gesture?

Part B) If so, then does the degree of gesture-speech coupling predict changes in temporal structure of speech?

Pertaining to part A, we found that when gestures occur more during blocks of fluid speech, that speech is more rhythmically timed (i.e., lower standard deviations of inter-speech onset intervals or ISI's), and has a higher tempo (i.e., lower mean ISI). Furthermore, when we scaled these metrics by the time duration of the 100-word block, the relation between gesture likelihood and speech rhythm was maintained, suggesting that time available for possible gesturing does not change our conclusions. Further analyses revealed

---

showed reliable correlations. For simplicity, we have however focused on the three variables (%Recurrence, Ratio, Entropy).

that the relation between speech rhythmicity was a better predictor of gesture likelihood as compared to the tempo of speech intervals.

Given that Part A showed a relationship between gesture likelihood and speech rhythmicity, in Part B we asked whether the degree of gesture-speech coupling might relate to the temporal structure of the speech signal. Through motion-tracking, we obtained the peak velocity of each gesture, and measured its temporal synchrony (i.e. coupling) with the peak change in the amplitude envelope of speech (co-occurring with that gesture). We found that the strength of the coupling between gesture and speech was related to the aspects of the temporal structure of speech as indexed by Recurrence Quantification Analyses (RQA). Namely, when gestures and speech were more strongly coupled, speech was a) more stable (lower Ratio between %Determinism), and b) less complex (lower Entropy). These continuous time-series analyses confirm our previous analyses that the rhythmic structure of speech is altered when speech is more strongly coupled with gesture.

Note that, before we performed RQA, we correlated gesture-speech coupling strength with the SD of ISI (the discrete measure used in Part A), but these did not show reliable relations (see supplemental figure https://osf.io/64edx/). As such, it seems that the non-linear RQA analyses can uncover changes in temporal structure that are not readily picked up by rhythmic timing measures such as inter-speech onset intervals.

Both sets of analyses (i.e., Parts A and B) demonstrate that when gesture and speech are coupled, speech becomes more stably uniform and ordered in its temporal structure. Speech on the word-level (i.e., lower SD inter-speech intervals) is more rhythmically timed when it is co-occurring with gesture. At the prosodic-acoustic level (i.e., amplitude envelope), speech is more predictable (lower Ratio) and is less complex (lower Entropy, i.e. walk less diverse paths) when gesture-speech coupling is stronger (i.e., lower standard deviations gesture-speech [a]synchronies). In other words, when speech is more stable and ordered in structure, gesture and speech are more likely to be strongly coupled.

However, clearly more research is needed to gain a better understanding of the temporal dimension of speech in relation to gesture. For example, from the current study it is, strictly speaking, not clear whether *gestures* are affecting speech's temporal structure, or vice versa. We suggest that the current results might best be understood as evidence of coupled systems, both of which serve the higher-order goal of stable language production. Gesture and speech belong, in this sense, to a family of synchronization phenomena (Pikovsky, Rosenblum, & Kurths, 2001) which show that when sub-system oscillators (e.g., walking & breathing) are coupled, make more efficient use of energy (e.g., oxygen uptake) when performing their task (see e.g., Amazeen, Amazeen, & Beek, 2001).

More research is also needed to address the function of gesture and speech in a way that binds semantic analyses (e.g., De Jonge-Hoekstra, Van der Steen, Van Geert & Cox, 2016) and rhythmic analyses of speech (e.g. Tilsen, 2009; Shattuck-Hufnagel & Ren, 2018). Namely, although we find that gesture-speech coupling predicts complexity (Entropy) of speech, we do not know whether complexity of the *content* of speech changed as well. With regard to semantic complexity, it has been found that when sentences are embedded within more complex linguistic structures (versus less complex linguistic embedding), that the kinematics of lip movements becomes less predictable (lower %Determinism) (Jackson et al., 2016). Relatedly, van Lieshout and Namasivayam (2010) have shown that when speech productions are faster, that synchronization (relative phases) between the tongue and lip closures become less predictable (less %Determinism). More research and theoretical considerations are, however, needed to tie these findings on semantic complexity, speech kinematics, and gesture-speech rhythm into a coherent whole. In any case, it is clear that the dynamic temporal structure of speech as indexed by RQA is related to a myriad of factors (production, rhythm, semantic), one of which seems to be gesticulation. Thus, a long-term goal of work in this area should be uniting different approaches to gesture-speech coupling. Future research should consider multiple, intertwined functions of gestures, such as rhythmicity, recruiting and simulating (physical)

properties of the environment, social coordination and expressing emotions (e.g., Gunes & Pantic, 2010).

The rhythms of gesture and speech manifested *within* an individual may further provide opportunities for coupling *between* individuals. Kotz, Ravignani and Fitch (2018) for instance emphasize that rhythm plays an important role in social coordination among many different species. When considering studies on humans in specific, rhythm is evident in social coordination from timescales ranging from less than a second (e.g. Hale, Ward, Buccheri, Oliver & Hamilton, 2018) to 20-30 seconds (Jaffe et al., 2001). Furthermore, Raja (2017) suggests that the central nervous system resonates with these rhythms of inter-personal coordination. Thus, future research could focus on how gesture-speech rhythms do not only arise through self-entrainment, but also through *interpersonal* entrainment that allow for new or improved linguistic stabilities on a dyadic level.

To provide a mechanism for self-entrainment, Pouw, Harrison, & Dixon (2018) found that when gestures have higher physical impetus on the body, peaks in the amplitude envelope and the fundamental frequency of phonation occur that increase over time. Such peaks are directly related to prosodic dimensions of speech and have been used as such in the study of the gesture and speech prosody correlation (for a review see Wagner et al., 2014). These results suggest that bodily resonances can travel (Turvey & Fonseca, 2014)

from the upper limbs (and associated anticipatory postural adjustments) to alveolar muscles that determine airflow of the larynx (which affects phonation), thereby directly, physically entraining the speech system on the prosodic dimension. These related findings sketch a plausible explanation for why humans gesture. The central idea is that stabilities can spontaneously emerge from physical gesture-speech resonances, with the strength of this coupling being related to fluency (rhythmicity) of speech, as laid out in this paper. Note that if gestures have a role to play in the emergence of speech rhythm, such a role is far from trivial. For example, the rhythm of speech is a central defining feature that is shared between (but also differentiates among) particular languages (Brookshire, Lu, Nusbaum, Goldin-Meadow, & Casasanto, 2017; Tilsen & Arvaniti, 2013), and speaking more rhythmically has been found to add to listeners' understanding when the speech signal is degraded with noise (Wang, Kong, Zhang, Wu, & Li, 2018).

Conclusion

We started this paper with the question: Why do humans use hand gestures? The current results show that gesture and speech are intimately related on the temporal dimension. We provided evidence consistent with the idea that gesture-speech coupling affords a stable and more uniform rhythm to emerge in language production.

References

Amon, M. J., Pavlov, O. C., & Holden, J. G. (2018). Synchronization and fractal scaling as foundations for cognitive control. *Cognitive Systems Research*, *50*, 155-179. doi: 10.1016/j.cogsys.2018.04.010

Amazeen, P. G., Amazeen, E. L., & Beek, P. J. (2001). Coupling of breathing and movement during manual wheelchair propulsion. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1243.

Bergmann, K., & Kopp, S. (2010). Modeling the production of coverbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence*, *24*(6), 530-551. doi: 10.1080/08839514.2010.492162

Bergmann, K., Damm, O., Freigang, F., Fröhlich, C., Hahn, F., …, Wittwer, N. (2014). *Documentation – SaGAland*. Retrieved from URL: https://www.phonetik.uni-muenchen.de/Bas/BasSaGADoku.pdf

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

Brady, J. P. (1971). Metronome-conditioned speech retraining for stuttering. *Behavior Therapy*, *2*(2), 129-150. doi: 10.1016/S0005-7894(71)80001-1

Brookshire, G., Lu, J., Nusbaum, H. C., Goldin-Meadow, S., & Casasanto, D. (2017). Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences*, *114*(24), 6352-6357. doi: 10.1073/pnas.1620350114

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

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, *137*(4), 706. doi: 10.1037/a0013157

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*(1), 102. doi: 10.1037/a0021790

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694. doi: 10.1037/a0033861

Cravotta, A., Busà, M. G., & Prieto, P. (2018). Restraining and encouraging the use of hand gestures: Effects on speech. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 206-210).

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, 37(1), 16-28. doi: 10.1016/j.wocn.2008.08.003

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

Davidow, J. H. (2014). Systematic studies of modified vocalization: the effect of speech rate on speech production measures during metronome-paced speech in persons who

stutter. *International journal of language & communication disorders*, *49*(1), 100-112. doi: 10.1111/1460-6984.12050

De Jonge-Hoekstra, L., Van der Steen, S., Van Geert, P., & Cox, R. F. (2016). Asymmetric dynamic attunement of speech and gestures in the construction of children's understanding. *Frontiers in psychology*, *7*, 473.doi: 10.3389/fpsyg.2016.00473

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications.

Garcia, C. A. (2015). *nonlinearTseries: Nonlinear Time Series Analysis*. R Package Version 0.2, 3.

Gillespie, M., James, A. N., Federmeier, K. D., & Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: Evidence from individual differences. *Cognition*, *132*(2), 174-180. doi: 10.1016/j.cognition.2014.03.012

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, *21*(5), 609-631. doi: 1362168816651463

Goldin-Meadow, S., & Brentari, D. (2017). Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and Brain Sciences*, *40,* 1-60. doi: 10.1017/S0140525X15001247

Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, *1*(1), 68-99. doi: 10.4018/jse.2010101605

Hale, J., Ward, J. A., Buccheri, F., Oliver, D., & Hamilton, A. (2018, June 7). Are you on my wavelength? Interpersonal coordination in naturalistic conversations. https://doi.org/10.31234/osf.io/5r4mj

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

He, L., & Dellwo V. (2017). Amplitude envelope kinematics of speech signal: parameter extraction and applications. In: Trouvain, Jürgen; Steiner, Ingmar; Möbius, Bernd. Elektronische Sprachsignalverarbeitung 2017. Dresden: TUDpress, 1-8.

He, L., & Dellwo, V. (2016). A Praat-Based Algorithm to Extract the Amplitude Envelope and Temporal Fine Structure Using the Hilbert Transform. In *Proceedings Interspeech* 2016 (pp. 530-534), San Francisco. doi: 10.21437/Interspeech.2016-1447

Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture?. *Speech Communication*, 57, 257-267. doi: 10.1016/j.specom.2013.06.007

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes*, *22*(3), 313-336. Doi: 10.1080/01690960600632812

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495-514.

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.

Jackson, E. S., Tiede, M., Beal, D., & Whalen, D. H. (2016). The Impact of Social–Cognitive Stress on Speech Variability, Determinism, and Stability in Adults Who Do and Do Not Stutter. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1295-1314. doi: 10.1044/2016_JSLHR-S-16-0145

Jackson, E. S., Tiede, M., Riley, M. A., & Whalen, D. H. (2016). Recurrence quantification analysis of sentence-level speech kinematics. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1315-1326. doi: 10.1044/2016_JSLHR-S-16-0008

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat, P., & Stern, D. N. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development*, i-149.

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

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in Cognitive Sciences*, *22*(10), 896-910. doi: 10.1016/j.tics.2018.08.002

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396-414. doi: 10.1016/j.jml.2007.06.005

Krauss, R.M., 1998. Why do we gesture when we speak? *Current Directions in Psychological Science*, 7, 54–60. doi: 10.1111/1467-8721.ep13175642

Krivokapić, J., Tiede, M. K., Tyrone, M. E., & Goldenberg, D. (2016). Speech and manual gesture coordination in a pointing task. In *Proceedings Speech Prosody 2016*, 1240-1244.

Kushch, O. (2018). *Beat gestures and prosodic prominence: impact on learning*. Dissertation, Universitat Pompeu Fabra.

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.

Lücking, A., Bergman, K., Hahn, F., Kopp, S., & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, *7*(1-2), 5-18. doi: 10.1007/s12193-012-0106-8

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2010). The Bielefeld speech and gesture alignment corpus (SaGA). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Mathis, A., Mamidanna, P., Abe, T., Cury, K. M., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). Markerless tracking of user-defined features with deep learning. *arXiv preprint arXiv:1804.03142*.

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

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago press.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*(5-6), 237-329.

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

Pouw, W. & Dixon, J. A. (under review). Entrainment and modulation of gesture-speech synchrony under delayed auditory feedback. doi: 10.31234/osf.io/avj7m

Pouw, W. Harrison, S., & Dixon, J. A. (2018). Gesture-Speech Physics: The Biomechanical Basis of Gesture-Speech Synchrony. Preprint retrievable from doi: 10.31234/osf.io/tgua4

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*, 359. doi: 10.3389/fpsyg.2014.00359

Pouw, W. T. J. L., Mavilidi, M. F., van Gog, T., & Paas, F. (2016). Gesturing during mental problem solving reduces eye movements, especially for individuals with lower visual working memory capacity. *Cognitive Processing*, *17*(3), 269-277. doi: 10.1007/s10339-016-0757-6

Pouw, W., Trujillo, J., & Dixon, J. A. (2018). The Quantification of Gesture-speech Synchrony: A Tutorial and Validation of Multi-modal Data Acquisition Using Device-based and Video-based Motion Tracking. Preprint retrievable from: https://doi.org/10.31234/osf.io/jm3hk

Pouw, W. T. J. L., Van Gog, T., Zwaan, R. A., & Paas, F (2017). Are gesture and speech mismatches produced by an integrated gesture-speech system? A more dynamically embodied perspective is needed for understanding gesture-related learning. *Behavioral and Brain Sciences*, *40*. doi: 10.1017/S0140525X15003039.

Prieto, P., Llanes-Coromina, J., & Rohrer, P. L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In *Klessa K, Bachan J, Wagner A, Karpiński M, Śledziński D. Proceedings of the 9th International Conference on Speech Prosody; 2018 June 13-16; Poznań, Poland.[Lous Tourils]: ISCA; 2018. p.*

498-502. DOI: 10.21437/SpeechProsody. 2018-101. International Speech Communication Association (ISCA).

Prieto, P., Cravotta, A., Kushch, O., Rohrer, P., & Vilà-Giménez, I. (2018). Deconstructing beat gestures: a labelling proposal. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 201-205).

Raja, V. (2018). A theory of resonance: towards an ecological cognitive architecture. *Minds and Machines*, 28(1), 29-51. doi: 10.1007/s11023-017-9431-8

Reichel, U. D., Schiel, F., Kisler, T., Draxler, C., & Pörner, N. (2016). The BAS speech data repository. Retrieved from http://real.mtak.hu/45970/1/RSKDP_LREC2016.pdf

Riley, M. A., Balasubramaniam, R., & Turvey, M. T. (1999). Recurrence quantification analysis of postural fluctuations. *Gait & Posture*, *9*(1), 65-78. doi: 10.1016/S0966-6362(98)00044-7

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. *Linguistic Controversies*, 73-79.

Rusiewicz, H. L. (2011). Synchronization of speech and gesture: A dynamic systems perspective. *In proceedings 2nd Gesture and Speech in Interaction* (GESPIN), Bielefeld, Germany.

Rusiewicz, H., L., & Esteve-Gibert, N. (2018). Temporal coordination of prosody and gesture in the development of spoken language production. In P. Prieto & N. Esteve-Gibert (Eds.), *The Development of Prosody in First Language Acquisition*. Amsterdam: John Benjamins.

Rueden, C. T., Schindelin, J. & Hiner, M. C. et al. (2017). ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*, *18*(1), 529. doi:10.1186/s12859-017-1934-z.

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

Stephen, D. G., Dixon, J. A., & Isenhower, R. W. (2009). Dynamics of representational change: Entropy, action, and cognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1811. doi: 10.1037/a0014510

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science*, *33*(5), 839–879. https://doi.org/10.1111/j.1551-6709.2009.01037.x

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628-639. doi: 10.1121/1.4807565

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

Turvey, M. T., & Fonseca, S. T. (2014). The medium of haptic perception: A tensegrity hypothesis. *Journal of Motor Behavior*, *46*(3), 143-187. doi: 10.1080/00222895.2013.798252

van Lieshout, P., & Namasivayam, A. (2010). Speech motor variability in people who stutter. In B. Maassen and P. van Lieshout (Eds.), *Speech motor control: New developments in basic and applied research* (pp. 191 – 214). Oxford, UK: Oxford University Press.

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.

Wang, M., Kong, L., Zhang, C., Wu, X., & Li, L. (2018). Speaking rhythmically improves speech recognition under "cocktail-party" conditions*. The Journal of the Acoustical Society of America*, *143*(4), EL255-EL259. doi: 10.1121/1.5030518

Webber J. R., C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*, 26-94. retrieved from: http://www.saistmp.com/publications/spiegorqa.pdf

Webber Jr, C. L., & Zbilut, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. *Journal of Applied Physiology*, 76(2), 965-973.

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

Hassemer, J., & Winter, B. (2018). Decoding Gestural Iconicity. *Cognitive Science.* doi: 10.1111/cogs.12680

Zbilut, J. P., & Webber, C. L. (2006). Recurrence Quantification Analysis. *Wiley Encyclopedia of Biomedical Engineering*. doi:10.1002/9780471740360.ebs1355