This paper was presented at the 6th Gesture and Speech in Interaction Conference that was held at Paderborn University, Germany from September 11-13, 2019.

To cite this paper:

Pouw, W. & Dixon, J. A. (2019). Quantifying gesture-speech synchrony. In: Grimminger, A. (Ed.): *Proceedings of the 6th Gesture and Speech in Interaction – GESPIN 6* (pp. 75-80). Paderborn: Universitaetsbibliothek Paderborn. doi:10.17619/UNIPB/1-815

Quantifying gesture-speech synchrony

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

[1]Center for the Ecological Study of Perception and Action, University of Connecticut
[2]Department of Psychology, Education, & Child Studies, Erasmus University Rotterdam
wimpouw@uconn.edu, james.dixon@uconn.edu

Abstract

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

1. Introduction

Humans across all known cultures tend to move their hands during speaking (Kendon, 2004), suggesting a fundamental connection between communicative vocalizations and hand movements (Iverson & Thelen, 1999). There is one fundamental aspect of gesture that is central to its functioning: gestures are performed in synchrony with speech. Without synchrony with speech, gestures would fail to unambiguously point to objects or portray them through depiction, and be meaningless as markers of semantic or emotional salience (Quine, 1968).

Although it is widely accepted that synchrony is fundamental to gesture's functioning, fine-grained quantification of gesture-speech synchrony as it occurs spontaneously during speaking is currently lacking (see however Danner, Barbosa, Goldstein, 2018; Pouw & Dixon, 2018). There is abundant research showing that the moment of "maximum effort" within a gesture is closely timed with the prosodic contrasts made in speech, but such evidence has varying degrees of objectivity and generalizability. Specifically, the primary evidence is based either on: a) artificial data (e.g., gestures produced by the experimenter, e.g., Leonard & Cummins, 2010), b) pointing gestures that are produced in a repetitive way outside the context of fluid speech (e.g., Rochet-Capellan, Laboissiere, Galvan, & Schwartz, 2008), or c) analyses of video recordings that do not allow for quantification of kinematic properties of gesture production (Loehr, 2012). To be clear, such research has been crucial in the study of gesture-speech synchrony, but also solicits an important next research objective given the technological advancements in current day and age (e.g., Danner et al., 2018; Pouw, Trujillo, Dixon, in press): A fine grained quantification of the synchrony of spontaneous gesture kinematics relative to speech.

For example, the most promising evidence for gesture-speech synchrony relies on methodology involving experimenter judgments of the intensity of gestural hand movements, the "maximum effort" of a gesture (Loehr, 2012; Wagner, Malisz, & Kopp, 2014). The maximum effort is the supposed to be the moment at which there is an energetic peak in the gesture stroke. However, as Wagner and colleagues (2014) conclude, the concept of maximum effort is an ambiguous spatiotemporal marker of a gesture "[the maximum effort is studied] with varying degrees of measurement objectivity and with varying definitions of what counts as an observation of maximum effort. Most definitions evoke a kinesthetic quality of effort or *peak effort* (Kendon, 2004) correlated with abrupt changes in visible movement either as periods of movement acceleration or strokes (Kita, van Gijn, & van der Hulst, 1998), as sudden halts or *hits* (Shattuck-Hufnagel, Veilleux, & Renwick, 2007), or as maximal movement extensions in space called *apexes* (Leonard & Cummins, 2008)" (p. 221, original emphasis).

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

2. Current approach

Subjects in the current exploratory study (N = 4) retold the narrative of a cartoon they had just watched, a common gesture-elicitation method (McNeill, 2005), which yielded about 230 gesture events. We employed high resolution motion-tracking of the dominant hand (240 Hz) during narration (non-dominant hand was not used for gesturing). From the movement time series, we identified energetic peaks during each gesture event (peak velocity, peak acceleration, peak deceleration), providing an objective measurement of gesture kinematics. Gesture identification was performed using ELAN (Lausberg & Sloetjes, 2009) so as to categorize different gestures, and to define the onset of a gesture based on assistance of hand-movement time series (see method and Crasborn, SLoetjes, Auer, & Wittenburg, 2006). Similar to previous studies (e.g., Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2010), we further extracted pitch (*F0*) from acoustic data so as to identify peaks of pitch within relevant gesture-speech events, which we show is a reliable anchor point for gesture-speech synchrony. Gesture-speech synchrony was quantified by the difference (*D*) in milliseconds between peak pitch and the relevant gesture anchor points (e.g., gesture onset, peak velocity). In the current study, we focused on three major gesture types, namely beat, iconic, and narrative pointing gestures. This exploratory study will allow us to answer a host of classic questions that have not been quantitatively studied to the current extent, including: What reliable kinematic anchor point in a gesture event is most closely synchronized with peak pitch? How strong is the synchrony between gesture and speech? Do beat, iconic, and pointing gestures differ in gesture-speech synchrony? To what extent are there individual differences in gesture-speech synchrony? For this conference contribution, we will report results of a larger scale study with 50 participants.

3. Method & Results

Four male right-handed graduate students at the University of Connecticut participated in this study (ages = 30, 38, 23, 34). Two participants were native speakers of American English and two were native speakers of Spanish with high proficiency in spoken and written English. In total, we collected movement and speech data from about 15 minutes of narration. Note that this much narration is considerable relative to other comprehensive studies of temporal coordination of gesture and speech, which have been naturally time-constrained because of the time-intensiveness of video analytic annotation (e.g., Loehr, 2012)

3.1. Apparatus

Motion tracking. We used a Polhemus Liberty (Polhemus Corporation, Colchester, VT, USA) with a single motion-sensor collecting 3D position data at 240Hz (~0.13 mm spatial resolution). The motion sensor was attached to the top of the participant's index finger (at the height of the fingernail). This allowed us to capture arm movements together with movements of the wrists and fingers. We recorded the motion of only one hand to simplify data collection and analysis.

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

Motion & audio recording. We used C++ to simultaneously call and write audio and movement data. We modified a C++ script made publicly available by Michael Richardson (Richardson, n.d.) in which we included scripts to enable recording of sound from a microphone (using toolbox SFML for C++ https://www.sfml-dev.org/).
Camera. We videotaped participants using Sony Digital HD Camera HDR-XR5504 Recorder, sampling at 29.97 frames per second.

3.2. Procedure

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

3.3. Data Preparation

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

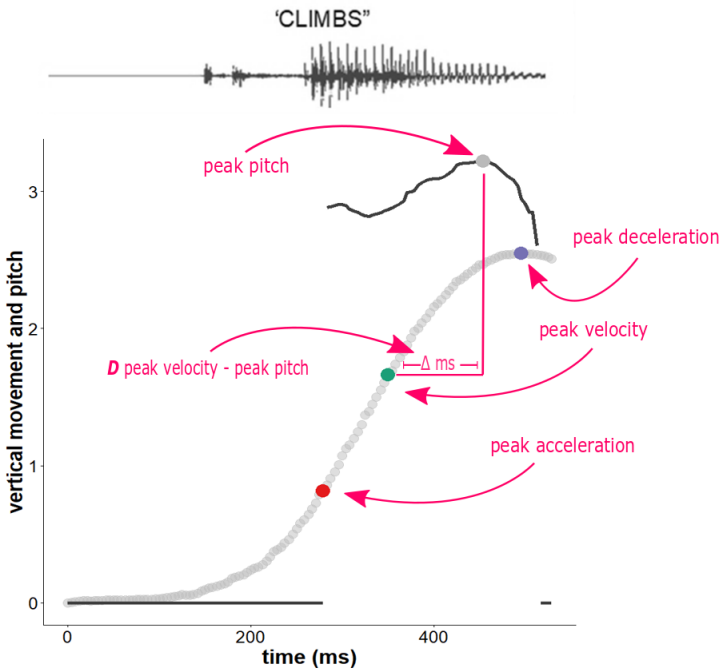
The procedure of marking a gesture in the current dataset was as follows. Gesture onset was identified by spotting a gesture in the video, categorizing it as either a beat, iconic, or pointing gesture (based on gesture categorization guidelines by McNeill, 2005). In cases where the gesture was not of a clear nature, it was categorized as "undefined"; we also categorized "abandoned" gestures that were not completed (see e.g., Kita, Alibali, & Chu, 2017). After having spotted a gesture, the experimenter would go back to beginning of the gesture event and seek out the onset of the gesture (first fluent change from static position), on the basis of the time series of the kinematic data (with the use of x and y axis, and velocity trace). The gesture event was marked as ending at the place where the gesture completed its main stroke, thus not including a possible post-stroke hold, and not including a retraction phase. Excluding these optional end-phases of gesture allowed us to ensure that our peak-finding functions do not pick out possible energetic peaks in the retraction phase (which is generally known not to coordinate meaningfully with speech).
Speech Pitch. We extracted pitch time series of the audio recording using PRAAT with default range suitable for males 75-500 Hz (Boersma, 2001). We matched the sampling rate of pitch with that of the motion tracker (1 sample per 4.16 milliseconds).
Speech content. For exploratory purposes, also using ELAN, speech was transcribed and lexical affiliates of iconic gesture were identified if possible, but not when gestures did not clearly refer to what was mentioned in text.

Data aggregation and analysis. We wrote a custom code in R (R core Team 2013) to aggregate the ELAN, PRAAT, and motion tracking data. We interpolated the movement data to match the pitch data with an interpolation function in *R* (code available on the Open Science Framework; https://osf.io/5ja6y). Using a custom-made function, we automatically read in ELAN gesture and speech annotation files so that these events were marked in the movement and pitch time series.

For each gesture event, the peak velocity, peak acceleration, and peak deceleration were extracted by a custom-written function in R. Since our peak-finding function could be sensitive to small but significant jumps in position data due to noise, we applied a low-pass Butterworth filter to the position velocity and accuracy traces with a cut-off of 10Hz (e.g., Leonard & Cummins, 20).
Data Availability. All (raw) data, pitch data (PRAAT), annotation data (ELAN), experiment code (C++), data preparation code (R), & analyses code(R) generated for this exploratory study are publicly available on the Open Science Framework (https://osf.io/5ja6y).

3.4. Descriptive Results

A total of 231 gesture events were observed (beat = 152, iconic = 44, pointing = 31, undefined/abandon = 4). Average time for gesture events was 829 ms (*SD* = 602 ms); beat gesture *M* = 739 (*SD* = 398), iconic gesture *M* = 947 (SD = 789), pointing gesture *M* = 667 (*SD* = 443). Table A (see here: https://osf.io/3n79f/) provides an overview of the production rates of the different gestures, as well as speech rate (spoken words per minute narration). It is important to note that these gesture ratios are very comparable to other studies that have used the same retelling of cartoon procedure (see e.g., McNeill, 2005, p. 42, where a comparable 41% of iconic gestures was found). This serves as evidence that in the current sample the glove and measuring apparatus did not seem to greatly alter spontaneous gesture tendencies.



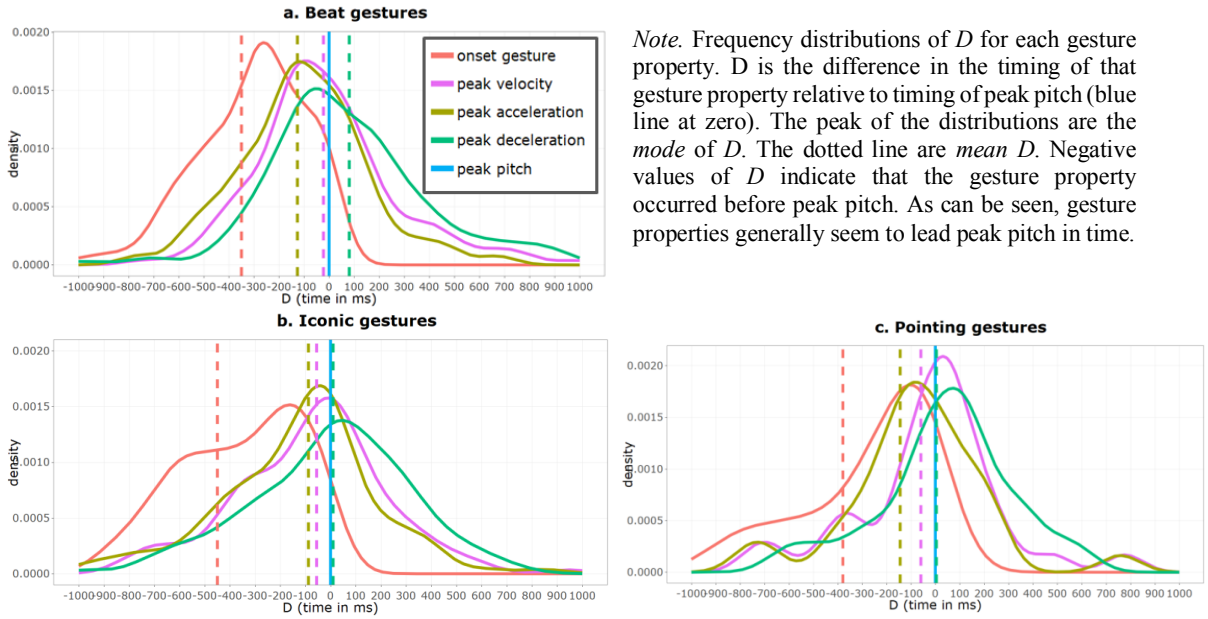
Note. Example of change y-axis position (grey) and pitch track (black) over time (ms; centered and scaled) for the "CLIMBS the wall" gesture. Red dot = peak acceleration, green dot = peak velocity, purple dot = peak deceleration, solid grey dot = peak pitch. These gesture properties were extracted using the custom-written function in R. Further note that speech starts around onset of gesture. We have super imposed the raw sound waveform in red above. The pitch (F0) reflects the vocal fold opening at pronouncing the "I" in "climbs".

Figure 1. Visual example peak extraction method.

3.5. Gestures and Peak Pitch

We first assessed the temporal relation between speech (peak pitch) and properties of gesture. Table B (see here: https://osf.io/c7qbm/) shows the mean difference in milliseconds, *D*, between the peak pitch and the different kinematic properties of gesture - gesture onset, peak velocity, peak acceleration and peak deceleration, for each gesture type separately. Figure 2 shows the relative frequency distributions of *D* for these gesture properties relative to peak pitch (which defines the zero point).

A flat distribution curve of *D* would be an indication of a random occurrence of a kinematic property of gesture with regards to peak pitch. We obtain a clearly non-uniform distribution of *D* for beat, iconic, and pointing gestures, showing an impressive temporal coupling between gesture and speech prosody. Furthermore, the data show that gesture's peak velocity, peak acceleration and gesture onset, all lead peak pitch in time (and is followed by peak deceleration). Gesture onset and peak acceleration are clearly not the point at which gestures synchronize with peak pitch. For each gesture property separately (i.e., onset, peak velocity, peak acceleration, peak deceleration), we performed a within-subjects ANOVA to assess differences in *D* between each gesture type (3 levels: beat vs. iconic vs. pointing gesture events; see Table B https://osf.io/c7qbm/). In the current sample, we did not find statistically significant differences between gesture types for *D*. This suggests that all the gestures types addressed here in this exploratory sample are roughly comparable in the degree to which they synchronize with peak pitch. The Bayesian Analyses further show that the observed data were 3 times or more likely under the null-hypothesis (absence of effect of gesture type) for gesture onset, peak velocity, peak acceleration. However, for peak deceleration we did not find substantial evidence for the null-model, suggesting that peak deceleration may differ in *D* between gesture types (when tested with larger samples).



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

Figure 2. Distribution of D's: Gesture properties relative to peak pitch.

A further question that arises is whether there is one particular gesture property that is most closely coordinated with peak pitch in speech. Since we did not find reliable statistical differences in *D* between gesture types, we collapsed all beat, iconic, and pointing gesture events for the following analyses. With this combined data, we performed a within-subjects ANOVA with gesture property (peak velocity vs. peak acceleration vs. peak deceleration) as a within-subjects variable, and *D* as the dependent variable.

We found that these gesture properties differed reliably in their *D*'s, $F(2, 6) = 17.54$, $p < .001$. Paired post-hoc comparisons (p-values Bonferroni corrected) revealed that peak velocity shortly led peak pitch (M_D = -39, SD_D = 454, 95%CI[-90: 11]), as compared to peak deceleration which followed peak pitch ($p < .001$; $M_D = 44$, $SD_D = 424$, 95%CI[-3 : 92]). Peak acceleration was furthest from peak pitch (M_D = -113, SD_D = 494, 95%CI[-168 : -58]), and was statistically different from peak velocity and peak deceleration (*ps* < .001). As can be seen, both peak velocity and peak deceleration have 0 in their confidence intervals, suggesting that both closely synchronize with peak pitch, with peak velocity shortly leading (39 ms), and peak deceleration shortly following (44 ms) peak pitch.

4. Discussion

This exploratory study has provided the following preliminary implications with regards to classic questions in gesture research. These implications should be regarded as tentative.

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

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

Future research is needed to ensure that our findings can be reproduced in a more comprehensive experiment. Also, we may wonder whether findings can be reproduced when movement of the non-dominant hand can be made as well.

References

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

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

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, 71, 268–283.

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

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

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.

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

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.

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.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement Phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.), *Gesture and Sign Language in Human-Computer Interaction*, Proceedings International Gesture Workshop Bielefeld, Germany, September 17-19, 1997.

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

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

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

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

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

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

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

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

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

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, *57*, 283-300.

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.* https://doi.org/10.31234/osf.io/jm3hk

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