Effects of delayed auditory feedback on gesture-speech synchrony: Pre-registration and exploratory study

VERSION 1.1: 14-4-2018

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

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

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

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

Open data & Pre-registration: This experiment code, (raw) anonymized quantitative data, and analyses scripts supporting this pre-registration and exploratory data report are available on the Open Science Framework (https://osf.io/pcde3/). De-identified raw video data are available upon request.

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).

Abstract

Research on co-speech gestures has shown that gesture-speech synchrony is stable when hand movement or speech is disrupted by a delayed feedback manipulation, suggesting strong bidirectional feedback between gesture and speech. Yet, it has also been argued from case studies in perceptuo-motor pathology, that hand gestures are a special kind of action that do not require closed-loop control mechanisms (e.g., efferent feedback) to function properly in synchrony with speech. In the current pilot study utilizing motion-tracking methods, we reassessed gesture-speech synchrony under conditions of delayed auditory feedback (DAF; 130-150 ms delay) leading to speech disruption. The pilot study indicated that gesture-speech synchrony was indeed stable under DAF, even to a higher degree than the control condition. We also find a promising indication that gesture-speech dynamics does entrain to the external auditory delay as indicated by a consistent shift in gesture-speech synchrony offsets. This pilot study forms the basis for the current pre-registration of a larger-scale study.

Keywords: co-speech gesture, motion tracking, synchrony, speech perturbation, delayed auditory feedback

Introduction

Research combining motion-tracking methodology and speech analyses is beginning to confirm that gesture and speech are tightly synchronizing their energetic patterns (Chu & Hagoort, 2014; Krivokapic, Tiede, Tyrone, & Goldenberg, 2016; Krivokapic, Tiede, Tyrone, 2017; Leonard & Cummins, 2010; Pouw & Dixon, under review; Rochet-Capellan, Laboissier, Galvan, & Schwartz, 2008; Rusiewicz, Shaiman, Iverson, Szumisky, 2014). The general finding from these studies is that the timing of movements in gesture (e.g., gesture end phase/apex; peak velocity) are structurally related to prosodic contrasts (e.g., stressed syllable; peak pitch) in speech. Although such findings are primarily based on pointing gestures produced on command (but see Pouw & Dixon, under review), these findings seem generalizable given the host of pioneering studies that have studied gesture-speech synchrony on the basis of careful (but subjective) analysis of video recordings of spontaneous gesturing (e.g., McClave, 1994; Loehr, 2004, 2012; Mcneill, 2008; for an overview see Wagner, Malisz, Kopp, 2014). Given this increasing evidence for the entrainment of gesture-speech rhythms, the question arises of *how* and *why* gestures are so closely controlled with respect to speech (Iverson & Thelen, 1999; Rusciewicz, 2011; Esteve-Gibert & Guellai, 2018).

A remarkable case study that has left considerable theoretical imprint on how gesture researchers think about the perceptuo-motor control of hand gestures, is the gesturing ability of Ian Waterman (IW) (Gallagher, 2005; McNeill, 2008; McNeill, Quaeghebeur, Duncan, 2008). IW suffers since early adulthood on from the absence of proprioception from the neck down, which makes instrumental actions (e.g., picking up objects) practically impossible without continuous visual feedback. Without visual control,

IW simply does not know where his limbs are located, let alone whether a grasp is successfully unfolding. Yet, IW produces typical looking gestures when his vision is blocked to the hands, sometimes without any intention and awareness of doing so. Researchers studying IW have concluded that his non-visually guided gestures are impaired when topo-kinetic accuracy is required (e.g., tracing out an imagined triangle in the air), but are otherwise largely unaffected (but see McNeill et al., 2008; Gallagher, 2005 for a more detailed description). Most important, these researchers also concluded that IW's gestures are produced in synchrony with his speech (see Dawson & Cole, 2010 for video example of IW gesticulation). This finding has lead these researchers to conclude that gestures are controlled in a different way than instrumental actions (Gallagher, 2005; McNeill, 2008), which invokes the idea that the gesture system does not require so-called closed-loop control; it does not require a continuous causal cycle of perception (where are my hands now/what effects do my actions have) and action (where do my hands go) to maintain gesture-speech synchrony. The idea that gestures are somehow different from instrumental actions has further been argued for based on research showing that pantomimes are sensitive to visual illusions but instrumental actions are not (e.g., Westwood, Heath, & Roy, 2000), as well as by case studies of subjects with congenital phantom limbs who report gesturing with otherwise passive phantom limbs (e.g., phantom arms do not swing during walking; Ramachandran, Blakeslee, & Shah, 1998).

Yet it has also been shown that gesture-speech synchrony is relatively stable under the perturbance of hand movements or speech production, suggesting continuous bidirectional coupling of gesture and speech (McNeill, 1992; Chu & Hagoort, 2014; Rusiewicz, Shaiman, Iverson, Szumisky, 2014). Chu & Hagoort (2014) found that when

visual feedback of a pointing gesture is disrupted, speech will halt as to synchronize with the perturbed (and therefore delayed) pointing movement. Specifically, when the visual feedback of the pointing gesture was delayed in the virtual environment between 117-417 ms (experiment 1), or when it was suddenly horizontally replaced to the right or to the left (experiment 2), or put to full halt while the gesture being in actual movement (experiment 3), or when visual feedback of the hand movement was removed altogether while changing the target's location (experiment 4), gesture execution time was delayed and so was speech onset time. Gesture-speech (re)synchronization was thus maintained. Importantly, perturbing the gestures, affected speech even in very late phases before speech onset (as short as an estimated 99 ms), suggesting that interaction between speech and gesture does not become impossible while the gesture is in its execution, i.e., gesture and speech do not become "ballistic" at some point (see e.g., de Ruiter, 1998). In the final fifth experiment speech was perturbed, by changing the color of the to-be-referenced light disrupting the speech intention (e.g., "this blue light" to "this yellow light"). Equivalent to the previous experiments perturbations were administered at early and late phases in the gesture execution. Speech perturbation indeed delayed gesture execution times in early and late phases (although this was not enough to reach complete re-synchronization of speech and gesture). This study shows that, at least in pointing gestures that are controlled with respect to the external environment, continuous visual feedback is utilized to maintain gesture-speech synchrony. This weakens the case that gesture-speech synchrony is not regulated by perceptual feedback of actions in typical populations (cf. IW's case).

McNeill (1992) reported on participants' gesture-speech synchrony when speech was disrupted by Delayed Auditory Feedback manipulation (DAF). When auditory feedback

from speech is delayed by about 75-200 milliseconds, speech becomes noticeably disfluent and slurred, and more frequent speech errors (e.g., repetition of phonemes) and slower speech rates are observed (Stuart, Kainowski, Rastatter, Lynch, 2002; see demonstration DAF effect from current study https://osf.io/5h3bx/). However, despite obtaining a classic DAF effect when participants retold a cartoon they had just watched, McNeill (1992) reported that gesture-speech synchrony looked completely unaffected. In a second experiment however, McNeill noted that gesture-speech synchrony was noticeably affected when participants had to recite memorized sentences and gesture movements, suggesting that there might be important role for spontaneity in gesture-speech synchrony.

Rusiewicz and colleagues (2014) used a common pointing and verbalizing task where participants point and label targets under DAF. In half of the trials participants heard their own speech with a 200ms delay, which indeed resulted in elongated spoken responses. It was further found that the time between gesture launch midpoint and the vowel to vowel midpoint of the referenced target word was increased as compared to when speech was not perturbed by DAF, although gestures were lengthened, signaling an attempt to correct movement so as to reach gesture-speech synchrony. However, all results concerning the effect of DAF were not statistically reliable. Thus gesture potentially adjusted to the elongation of speech in the perturbed conditions, but at least not to the extent of full compensation (c.f., Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). Rusiewicz et al., (p. 293) carefully conclude that the "*Qualitative changes in total gesture time and gesture launch time suggest that there is potentially some interaction and feedback between the two motor systems.*", but also leave open the possibility that once speech and manual gestures are planned they cease their interactions and become "ballistic" (de

Ruiter, 1998), therefore leading to gesture-speech de-synchronization when speech is interrupted and is slacking behind due to the DAF.

Summary

There are still several open questions about the stability of gesture-speech synchrony under conditions of perturbance and the precise role that perceptual feedback plays in maintaining stability. Firstly, perturbation research with relatively high kinematic detail, has solely focused on gesture-speech synchrony of pointing gestures (Chu & Hagoort, 2014; Rusziewisc et al., 2014). These types of gestures are an exotic member of the family of gesture as they are exogenously controlled with respect to an external target, and are thus also closely related to instrumental actions in this respect. Therefore, it is not certain that more fluid and spontaneous gestures (e.g., Beat and Iconic gestures) produced during narration show similar dynamics. Secondly, even if the previous effects are generalizable, they do not converge in their findings. While Chu & Hagoort (2014) provides strong evidence for stability of gesture-speech synchrony when *visual feedback* is disrupted, Rusiewicz and colleagues (2014) do not obtain reliable evidence for stability of gesture when *auditory feedback* is disrupted. It could even be that the gesture-speech system is asymmetrically coupled, such that speech follows gesture, given that only strong evidence for synchrony stability is obtained for visual feedback perturbations rather than speech perturbations (Rusiewicz et al., 2014; Chu & Hagoort, 2014, exp 5). Yet, McNeill's (1992) original descriptive report is that gesture-speech synchrony is maintained under DAF for spontaneous gestures produced during narration. However, the descriptions of this study, although promising, lack empirical detail and therefore cannot yet fully support the

assertion that gesture-speech synchrony is not affected by DAF for spontaneous gesturing (as also argued by Rusiewicz et al., 2014).

The theoretical import of these mixed findings is that it is still possible that some kind of motor plan and a speech plan are synchronized in the case of fluid spontaneous gestures that are *not* exogenously controlled, in contrast to pointing gestures (de Ruiter, 1998). This may account for IW's gesture-speech synchrony without apparent perceptual feedback of his movements. Simply put, IW's motor intentions and speech production are planned in synchrony, and under non-perturbed circumstances this will result in gesture-speech synchrony. However, the question is whether this account could equally explain gesture-speech synchrony under DAF (McNeill, 1992). Note, that it has been argued from a dynamical systems perspective of gesture (Rusiewicz et al., 2014; see also Rusiewicz et al., 2011) that the stability of gesture-speech synchrony under DAF would signal that a system is dynamically entrained and flexibly organizes into equivalent functional sensorimotor solutions through tight bidirectional coupling of the systems. This stability is likened to the classic finding by Kelso and colleagues (1984) who showed that when the jaw is locked in the midst of articulating a syllable, the lips *or* tongue spontaneously jump into place to successfully complete the syllable (depending on whether the perturbation is successfully resolved by a lip or tongue intervention). Yet, in the case of gesture-speech synchrony under DAF, it could still be that stability is maintained *because* the delayed auditory feedback only affects speech *and not gesture*. That is, *if* the gesture system merely follows speech production and is not attuning to perceptual auditory feedback, it will always be produced in synchrony with speech. Stability, under such a view, is maintained because the gesture system is ignorant with respect to the perceptual feedback of speech, it may merely

follow speech wherever it goes (i.e., gesture will be disrupted when speech is disrupted, maintaining synchrony).

We think there is another way a dynamically coupled gesture-speech system would behave under DAF. Namely, in line with the entrainment effect as described by Rusziewisc and colleagues (2014; see also Rusziewisc, 2011; Iverson & Thelen, 1999), coupled oscillators will diverge from their individually preferred rate of oscillation and will form a new joint preferred oscillation pattern. This joint preferred oscillation pattern will function as a stable point attractor, such that the system will return to this stable state even when one of the sub systems is perturbed (i.e., stability of gesture-speech synchrony). We can maintain that in ordinary gesture-speech synchrony that *two* oscillators are dynamically entrained, but under DAF there is a third oscillator in the form of a delayed feedback. Although it is certainly likely that that a dynamically coupled gesture-speech system should be able to flexibly modulate its activity to resist perturbations of the third oscillator, it is also likely that the DAF signal will serve as an attractor that draws in any system that is attuning to it. Thus, from a dynamic systems perspective we might argue that gesture and speech operate like a coupled oscillator and that its coupling strength might be intentionally modulated (e.g., Amazeen, Amazeen, Treffner, & Turvey, 1997), but *also* that gesture-speech synchrony (operating as a joint oscillator) will be attracted to a third oscillator. Furthermore, given that that speech and gesture have their own preferred rate of oscillation, it is also likely that DAF will affect speech and gesture differently. That is, only if gesture is also tuned to the perceptual feedback of speech. Note that this possibility is not far-fetched given that it is well known that humans naturally and involuntarily tune their actions (e.g., finger tapping) to auditory rhythms (e.g., metronome; for an overview see

Port, 2003). In sum, stability under perturbation is a necessary but not sufficient condition for arguing for the bidirectional coupling of gesture and speech, as an account that posits that gesture couples with speech production (and not speech feedback) would equally predict gestures-speech synchrony under DAF. Rather, entrainment to the DAF signal is also to be predicted if the gesture-speech system is dynamically attuned to it.

Current pilot study and pre-registration

In the current exploratory study, participants narrated a cartoon they had previously watched (see McNeill, 1992). We used motion-tracking of the dominant hand (240 Hz) to record hand movements as to identify energetic peaks during each gesture event (peak velocity, peak acceleration, peak deceleration). Gesture events and gesture types (Beat and Iconic) were identified using ELAN and the motion-tracking time series (Crasborn, SLoetjes, Auer, & Wittenburg, 2006; Lausberg & Sloetjes, 2009). We extracted pitch (F0) from the audio (using PRAAT, Boersma, 2001) so as to and identify peaks of pitch within relevant gesture-speech events. In a previous study, using a similar cartoon narration paradigm, we obtained promising results that gesture's energetic peaks, such as peak velocity and peak deceleration were structurally synchronized (within ~45ms) with peak in pitch (Pouw & Dixon, under review). As such, in the current exploratory pilot study, we will explore the effect of DAF on gesture-speech synchrony based on the coordination of energetic peaks in gesture and peak pitch.

Method Exploratory Study

Participants and design

Four right-handed females were tested at the University of Connecticut. Two participants were native speakers- and two participants were non-native speakers of American English with high proficiency in spoken and written English. The current study entails a within-subjects design with one factor with 2 levels: DAF vs. no DAF. The exploratory study consisted of 16.5 minutes of narration in total for all participants combined (6.7 min DAF vs. 9.8 min NO DAF).

Apparatus

Motion tracking. To track hand movements we used a Polhemus Liberty (Polhemus Corporation, Colchester, VT, USA) collecting 3D position data at 240Hz (~0.13 mm spatial resolution) with a single motion-sensor attached to the top of the index finger of the dominant hand. Thus the position data is determined by movements of the arms, wrists and fingers. We recorded the motion of the dominant hand to simplify interpretations regarding gesture-speech synchrony.

Audio. We obtained speech data by using a RT20 Audio Technica Cardioid microphone (44.1kHz) which suppresses surrounding noises including any experimenter comments.

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

Camera. Participants were video recorded (29.97 fps) using Sony Digital HD Camera HDR-XR5504 Recorder.

DAF. For the DAF manipulation we used a second microphone and wireless headphone connected to an additional PC. We used open software called pitchbox 2.0.2 (Juillerat, 2007) to delay feedback with a microphone. This software is originally developed to modify acoustics such as shifting the pitch of voice, but in the "normal" mode it will produce latency between voice production and auditory feedback. This degree of latency is dependent on the computer system's hardware specifications. We pre-tested the latency between microphone input and audio output that was produced when running it on a laptop, and we obtained that the latency was consistently between 130-150 ms for this system. This range for testing the auditory feedback delay phenomenon lies well above the lowest DAF manipulations that have been known to induce a DAF effect (e.g., Stuart et al., 2002). The effect of DAF was indeed very noticeable (see sample of speech and gesture under DAF here https://osf.io/5h3bx/).

Procedure

Participants were equipped with a glove for the dominant hand which allowed attachment of Polhemus motion sensor with Velcro tape. In a previous study we already obtained that this glove did not restrict spontaneous gesturing rates (Pouw & Dixon, under review). The glove was attached before watching the video so that subjects got used to wearing it. In the current experiments participants first watched the cartoon "Canary Row" lasting about 350 seconds. Participants were informed they would retell the cartoon narrative to the experimenter later on. After watching the cartoon, participants were familiarized with the DAF manipulation. The experimenter further explained that during

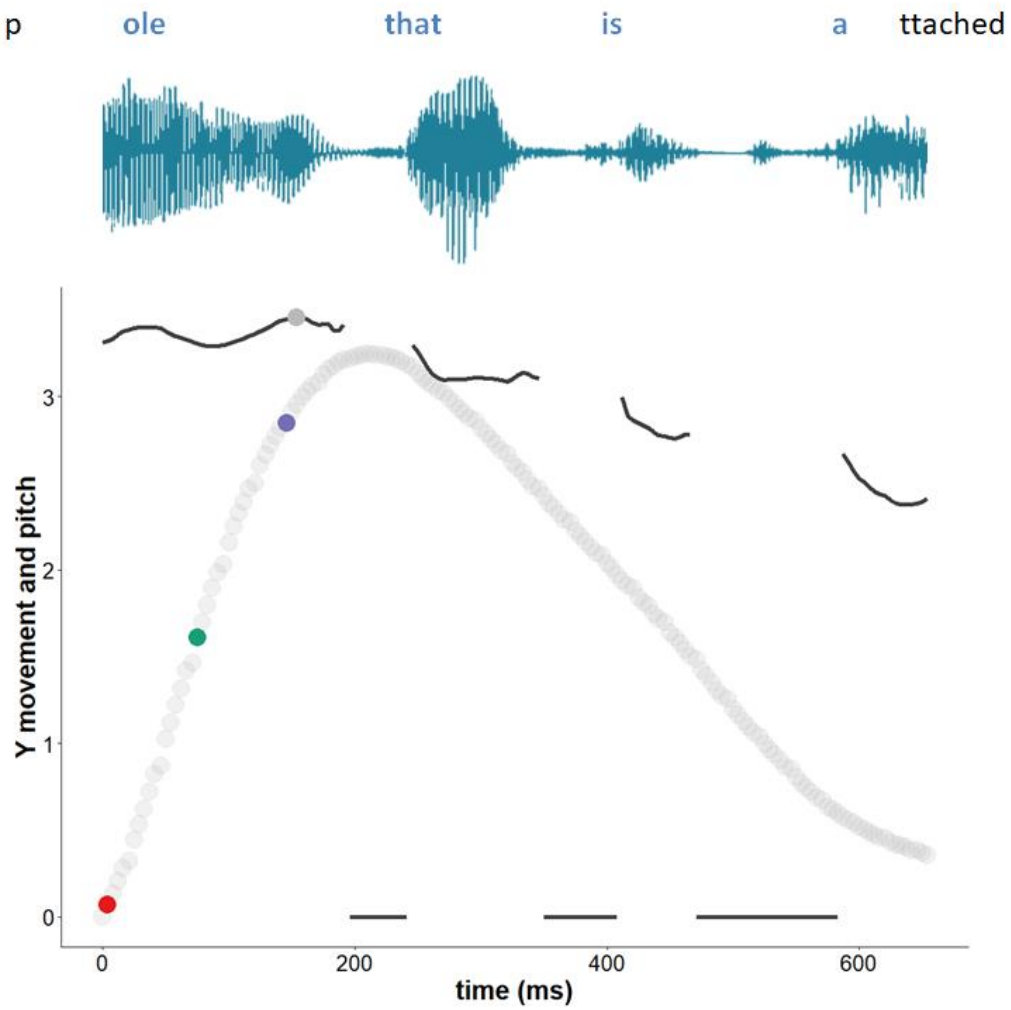
narration the DAF would sometimes be turned on, but that participants would need to keep narrating as best as they could. In the narration phase, for the first 30 seconds participants always narrated without DAF manipulation (warm up phase). After the warm-up phase the DAF was turned on, or turned off (depending on counterbalanced condition) every alternating minute for a duration of 60 seconds. We counterbalanced DAF timing as it is possible that otherwise DAF would occur for a particular segment of the cartoon narrative. No instructions were given about whether to use hand gestures.

Data preparation

Using the same method as Pouw and Dixon (under review), the first author transcribed speech and identified gesture events using the annotation software ELAN (Lausberg & Sloetjes, 2009). We also loaded into ELAN the motion-tracking time series which was used to determine the onset and end phase of the gestures (see Crasborn et al., 2006), whereas the video data were used to determine the type of gestures (Beat vs. Iconic vs. Undefined gestures). Each gesture was marked as ending at the point at which the gesture completed its main stroke. Thus, we did not include a post-stroke hold, nor a retraction-phase if present (see Kita, van Gijn, & van de Hulst, 1998).

Similar to the procedure used by Pouw and Dixon (under review), peak velocity, peak acceleration, and peak deceleration were determined with respect to peak pitch (see code online). We applied a low-pass Butterworth filter to the position velocity traces with a cut-off of 10Hz.

Figure 1. Example gesture and peak finding results



Note. Example of an Iconic gesture produced under NO DAF (see example videoclip https://osf.io/ax48y/). The pitch track (rescaled for this example) reflects the opening of the vocal folds for the voiced parts of the speech segments. The participant traces out the outline of a pole while saying "pole that is attached to the building", wherein the gesture overlaps with the blue segment of speech. In this example, it is clear that energetic peaks are situated in the beginning phase of the gesture which coincides with the peak and pitch (and the semantically relevant part "pole").

Speech Pitch. We extracted pitch time series using PRAAT with a range suitable for female voice range, 100-500 Hz[1] (Boersma, 2001). We matched the sampling rate of pitch with that of the motion tracker (1 sample per 4.16 milliseconds).

Data aggregation and analysis. Using a custom-made code in R (R core Team 2013) ELAN, PRAAT, and motion-tracking data were aggregated into a single time series dataset as to compare movement properties and pitch time series. We temporally aligned movement data with the pitch data using alignment functions in *R* (code available on https://osf.io/pcde3/). Smoothed distribution plots are produced with the ggplot2 "geom_density" function. This function draws on the "1d Kernel Density Estimate" function called "stat_density".

[1] We will include male participants in the proposed study. In such cases we will extract pitch series within a range of 75-500 Hz.

Results

Descriptives

A total of 275 gesture events were observed (Beat = 149, Iconic = 113, Undefined = 13). Average time for gesture events for the NO DAF condition was 644 ms (*SD* = 320 ms, 95% CI [586, 702]) versus 733 ms (*SD* = 468 ms, 95% CI [659, 807]) for the DAF condition; beat gestures (M_{NODAF} = 592, SD_{NODAF} = 266, M_{DAF} = 549, SD_{DAF} = 268), Iconic gesture (M_{NODAF} = 882, SD_{NODAF} = 596, M_{DAF} = 738, SD_{DAF} = 315). Table 1 provides an overview of the production rates of the different gestures, as well as speech rate (spoken words per minute narration). These findings indicate that there are no prominent differences in gesture rates depending on condition. Further, speech rate was considerably slowed by about 23% for the DAF condition as compared to the NO DAF condition, confirming the qualitative observations of pronounced speech impairment for all participants.

Table 1. Gesture and speech rates

Condition	**Beat p/m**	**Iconic p/m**	**Undefined p/m**	**Pitch (F0) Mean *SD***	**Speech rate p/m**	**Time of Narration**
DAF	9.9	7.65	0.3	193.04 *(38.98)*	100.78	400 s
NO DAF	8.44	6.30	1.12	188.48 *(42.84)*	131.10	590 s

Note. The gesture and speech rates (words spoken) are given for frequency *per minute* (p/m).

Gesture kinematics

Qualitatively gestures looked less pronounced under the DAF condition, with some gestures having a jerk-like motion. Table 2 provides some estimates for two kinematic properties per condition and gesture types, which do not however reveal reliable differences as function of condition in terms of peak velocity or average jerk.

Table 2. Gesture kinematics

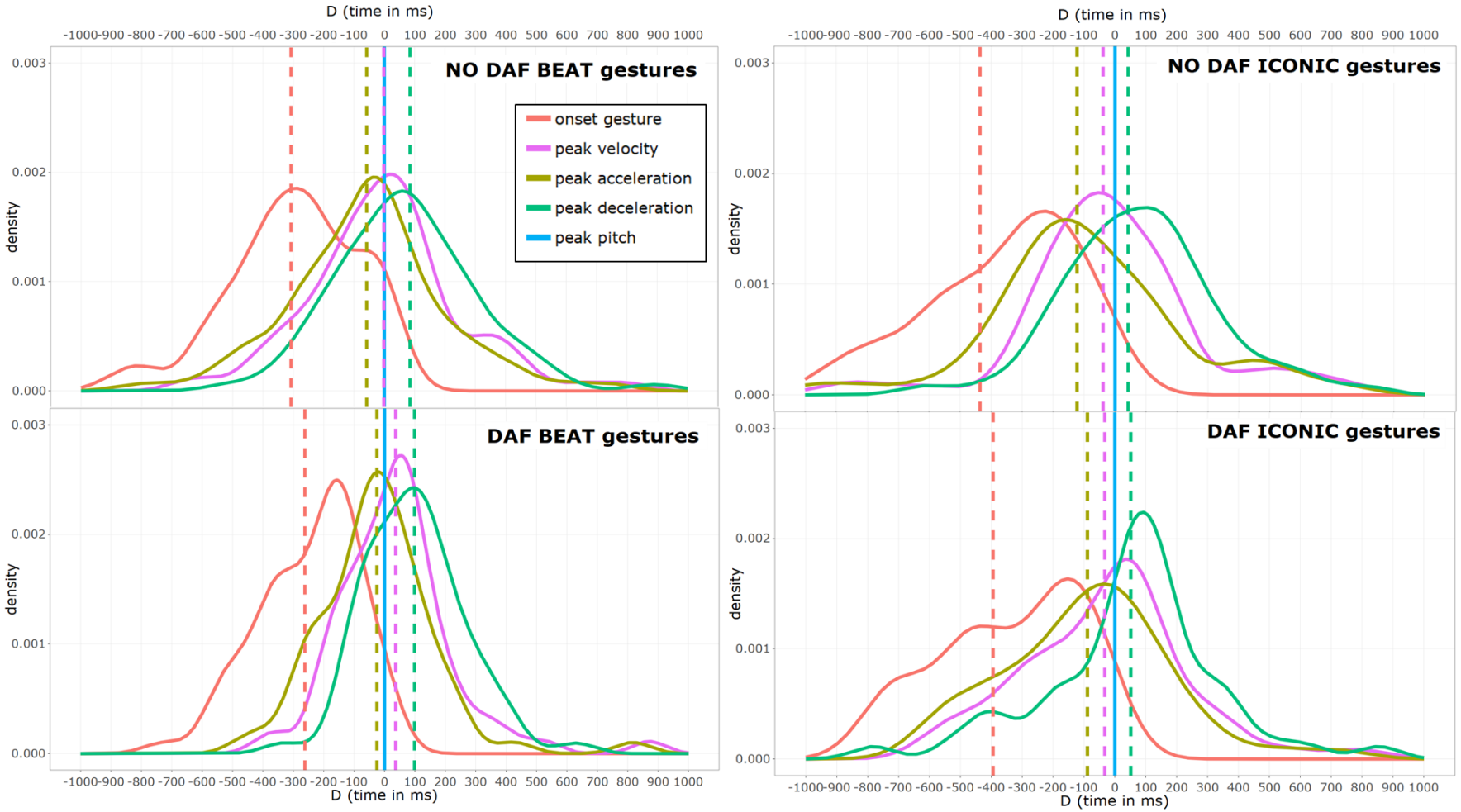
Condition		**BEAT**		**ICONIC**	
		Peak Velocity	**Average Jerk *z*-score**	**Peak Velocity**	**Average Jerk**
DAF	M *(SD)*	0.16 *(0.11)*	-.065 *(.07)*	0.44 *(1.58)*	0.27 *(2.96)*
	95CI% [lower, upper]	[0.13, 0.18]	[-0.08, -0.05]	[0.00, 0.88]	[-0.37, 0.91]
NODAF	M *(SD)*	0.15 *(0.08)*	-.075 *(.06)*	0.21 *(0.09)*	-0.04 *(0.08)*
	95CI% [lower, upper]	[0.13, 0.16]	[-0.09, -0.06]	[0.18, 0.23]	[-0.06, -0.02]

Note. Average Jerk is rescaled (standardized) as values are very small.

Main Analyses: Gesture-speech synchrony

The main analyses concern the investigation of gesture-speech synchrony under DAF versus NO DAF. Figure 3 show gesture-speech synchrony distributions for Beat gestures and Iconic gestures, and Table 3 shows the summary statistics for these distributions. Here the temporal differences (*D*) in milliseconds are reported for particular kinematic property (e.g., peak velocity) occurring relative to peak in pitch. There are several observations that can be made from the figures and tables.

Figure 2. Distributions D per condition and gesture type



Note. Distributions of *D's* are shown, where blue line at time = 0 reflects the peak pitch in speech. Note that for DAF gestures there seems to be consistent positive shift of all D distributions.

Table 3. Mean Difference *D* milliseconds (peak pitch - gesture property) per condition

Kinematic property	BEAT		ICONIC	
	NO DAF	DAF	NO DAF	DAF
Onset				
M (SD)	-308 *(234)*	-261 *(193)*	-436 *(419)*	-349 *(321)*
95% CI [lower, upper]	[-359, -256]	[309, -214]	[-542, -330]	[-484, -304]
Peak acceleration				
M (SD)	-58 *(259)*	-25 *(194)*	-123 (391)	-88 *(270)*
95% CI [lower, upper]	[-115, -2]	[-73, 23]	[-222, -23]	[-165, -13]
Peak velocity				
M (SD)	-2 *(250)*	37 *(194)*	-39 (*432*)	-32 *(268)*
95% CI [lower, upper]	[-56, 53]	[-11, 84]	[-149, 71]	[-108, 43]
Peak deceleration				
M (SD)	84 *(253)*	99 *(163)*	43 *(388)*	51 *(285)*
95% CI [lower, upper]	[30, 140]	[58, 139]	[-56, 141]	[-29, 132]

Firstly, similar to a previous study (Pouw & Dixon, under review), we find clear coordination of energetic peaks in gesture relative to peak pitch, as indicated by the single peaked distributions of *D*. Peak pitch and the structure of gesture movements are thus not arbitrarily related which would have resulted in flatter or multi-peaked distributions. Furthermore, for NO DAF gestures, peak velocity is a reliable anchor point for gesture-speech synchrony, especially for Beat gestures. For Iconic gestures peak velocity and peak deceleration are both closely situated near peak pitch. This replicates our previous findings (Pouw & Dixon, under review).

However, there are two more important observations for present purposes which can be obtained. Firstly, the standard deviations for the distributions seem to be consistently smaller for gestures produced under DAF as compared to NO DAF, and this is

especially pronounced for Beat gestures (clearly demonstrated by the sharper peaks of D for the DAF Beat gestures in Figure 2). This is an indication that gestures are more tightly coordinated with peak pitch under DAF as compared to NO DAF. To assess the reliability of this finding we performed a linear mixed effects analyses (nlme version 3.1-131), wherein we assessed the absolute deviance from peak pitch for peak velocity; firstly for Beat gestures. We used maximum likelihood estimation with a random intercept for participant. A model containing condition had added predictive value as compared to a model predicting the overall mean (Chi-square change [4] = 6.55, p = .01). Model estimates indicated that gestures produced under DAF had smaller mean deviances from peak pitch (~40 ms) for *D* peak velocity, b = -39.23, $t(717)$ = 2.57, p = .01. We also ran these analyses for Iconic gestures yielding qualitatively similar, but statistically unreliable, results.

A second important observation is that there is a consistent temporal shift for all kinematic properties of gestures, but again especially pronounced for Beat gestures. However, this shift is not in a classically predicted direction for the hypothesis that gesture and speech are decoupled at some point. Under such a hypothesis gestures' energetic peaks would be completed earlier as speech is slowed due to DAF, which would lead to a negative shift for gesture properties with respect to prosodic patterns in speech. Instead in the current sample, we find that gestures produced under DAF have energetic peaks that are positively shifted. We assessed the reliability of this shift in a mixed effects linear model (with random intercept for participant), wherein we assessed the D's relative to peak pitch as a function of kinematic properties (model 1), as well as condition (model 2; interaction effects were not statistically reliable). Both the model including kinematic properties as predictor and the model with condition as additional predictor were more reliable than a

base model predicting the overall mean (Chi-square change [6-7] > 202.12, *ps* < .001). Although, the model with condition and kinematic properties was not a more reliable model as compared to the model which only included kinematic properties (Chi-square change [7] = 3.16, $p = .076$), the direction and consistency of the effect of condition was promising and will serve as guiding hypothesis for the proposed larger scale replication study. Namely, after accounting for the variance attributed to the different kinematic properties that determine *D* (*ps* < .001), DAF condition had an estimated main effect of about 33 ms ($b = 32.62$, t (588) = 1.77, $p = .077$). Thus, regardless of kinematic property (gesture onset, peak -acceleration, -velocity, -deceleration) DAF gestures had positively shifted D's of about 33 ms with respect to peak pitch.

Brief Discussion

The current exploratory study served as a proof of concept for a kinematic study of the stability of gesture-speech synchrony for fluid spontaneous gestures produced under Delayed Auditory Feedback (DAF) which we will replicate in a large scale study (see next section "pre-registration"). We will briefly summarize the current results, but note that a detailed discussion of the results go beyond the purposes for the current pre-registration report.

In the current exploratory study, we provide promising preliminary results that gestures are remarkably stable under conditions where speech is perturbed due to DAF (conceptually replicating McNeill [1992] and Rusciewicz and colleagues' [2014] original predictions). Moreover, Beat gestures were reliably more synchronized with respect to peak velocity and peak pitch when produced under DAF as compared to NO DAF, as

indicated by smaller absolute temporal deviations of peak velocity from peak pitch. We think this is an indication that speakers can intentionally modulate the coupling strength of gesture and speech as to resist entrainment to an external rhythm that interferes with its functioning (i.e., delayed auditory feedback). As such the current finding may indicate a special functional role of gesture-speech synchrony, such that the intentional modulation of coupling strength of gesture-speech systems might allow for a higher degree of stability under perturbance (e.g., Amazeen, et al., 1997; Richardson et al., 2007).

Secondly, we obtain a promising indication that delayed auditory feedback may attract gestures' energetic peaks to the speech delay as opposed to simply following speech production. We speculate that the current finding that DAF potentially leads to a positive consistent shift of kinematic properties relative to peak pitch is due to the entrainment of gesture to this third oscillator. We think this is a viable hypothesis given that speech under DAF leads to lengthening and slurred production. Therefore, if bidirectional feedback between gesture and speech is non-existent at late stages of execution, then gesture's kinematic properties should be completed *earlier* with respect to prosodic peaks which slack behind under DAF. Instead we find that gestures' energetic peaks are happening later with respect to peak pitch when produced under DAF as compared to NO DAF. We speculate that this result indicates that gestures are entraining to the external auditory loop, similarly as rhythmic tapping would be affected by a rhythmic auditory pulse (e.g., Port, 2003). We aim to test these hypotheses in a larger scale study introduced in the next section.

Pre-registration

The method section and results section of the pilot study are the basis of this pre-registration as this approach will be *directly* replicated. Data manipulation code and analyses code that will be used for this replication study are available at: https://osf.io/pcde3/. Next we will explicitly state the confirmatory analyses, exploratory analyses, sample size justification, procedures for data exclusions, as well as some small amendments to the gesture annotation method.

Confirmatory Analyses

For this study, we will exactly replicate the two confirmatory analyses as indicated in the "main analyses" result section of the pilot study (see R analyses code; https://osf.io/pcde3/). For all the mixed linear models tested, we will set participant as a random intercept. Specifically, we predict that Beat gestures are more tightly coupled to speech under DAF as compared to NO DAF, as indicated by a lower absolute deviation of peak velocity from peak pitch for DAF (as compared to NO DAF). Secondly, we predict a positive shift of D's for all kinematic properties of Beat DAF gestures relative to NO DAF Beat gestures, suggesting that gestures attune to the auditory delay. We will only focus on Beat gestures as these gestures are a) the most tightly coupled to speech prosody (providing the least noisy estimates for our hypotheses), and b) have peak velocity as a reliable anchor point which allows for a straightforward hypothesis test of whether gesture is more strongly coupled to speech under DAF. Since we will perform two analyses we will restrict our alpha level to .025 (Bonferroni correction); i.e., any result will be deemed statistically reliable if $p < .025$.

Exploratory analyses

We will perform a range of exploratory analyses that will supplement the final report. For example, we aim to provide a more detailed description of the kinematics of gestures as well as articulatory processes in speech under DAF and NO DAF conditions. We are also planning to apply non-linear measures of the structure of gesture's kinematics (e.g., Recurrence Quantification Analysis). All non-confirmatory analyses will be reported as exploratory in the final report.

Participants & Sample Size Justification

Every participant narrates about 2 minutes per condition, and produces about 10 Beat gestures per minute. Thus per participant we have about 20 observations per condition. We plan to test 10 participants, which will produce 200 samples per condition. Note that, a conservative estimate from G*Power (version 3.1.9.2) provides for a within-subject design with one two-level factor, alpha of .025, 80% power and a small to medium effect size (Cohen's d = 0.3), a total number of 109 observations (~54 observations per condition). The current sample should therefore have enough statistical power.

Participants (N = 10) will receive a small monetary reward or course credit for their participation.

Data exclusions

We foresee no circumstances for participant or data exclusions. Any post-hoc exclusions will be reported and their effect on results will be made explicit with extra analyses assessing model fit with and without excluded data points.

Gesture analysis

For 20% of the participants (i.e., 2 participants), Beat, Iconic and Undefined gestures will be annotated by a second rater. To assess interrater reliability, we will compute a modified Cohen's Kappa (Holle & Rein, 2013) as provided by the ELAN Annotation software. Following common practices, if reliability is lower than .75 annotation procedures will be reevaluated and adjusted, and we will recode the video data to reach a higher agreement between raters.

Data availability

All quantitative data (excluding the video data) are made publically available at the Open Science Framework (https://osf.io/pcde3/). Due to privacy restrictions we cannot publicly share the raw sound and video data from the participants, but we will share all quantitative data.

References

Amazeen, E. L., Amazeen, P. G., Treffner, P. J., & Turvey, M. T. (1997). Attention and handedness in bimanual coordination dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(5), 1552.

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

De Ruiter, J. P. A. (1998). Gesture and speech production (Unpublished doctoral dissertation). University of Nijmegen, Nijmegen, the Netherlands.

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

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.

Dawson A., & Cole, J. (2010). http://www.thearticulatehand.com/ian.html

Esteve-Gibert, N., & Guellaï, B. (2018). Prosody in the Auditory and Visual Domains: A Developmental Perspective. *Frontiers in Psychology*, *9*, 338. doi: 10.3389/fpsyg.2018.00338

Holle, H., & Rein, R. (2013). The modified Cohen's kappa: Calculating interrater agreement for segmentation and annotation. In H. Lausberg (Ed.), *Understanding Body*

Movement. A Guide to Empirical Research on Nonverbal Behaviour. With an Introduction to the NEUROGES Coding System. (New York: Peter Lang).

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.

Kelso, J. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(6), 812.

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

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

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

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

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

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

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

McNeill, D., Quaeghebeur, L., & Duncan, S. (2010). IW-"The Man Who Lost His Body". In S. Gallagher & D. Schmicking (eds.), *Handbook of Phenomenology and Cognitive Science*, (pp. 519-543). Dordrecht: Springer.

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.

Port, R.F. (2003). Meter and speech. *Journal of Phonetics*, 31, 599–611.

Ramachandran, V. S., Blakeslee, S., & Shah, N. (1998). *Phantoms in the brain: Probing the mysteries of the human mind* (pp. 224-25). New York: William Morrow.

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

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R., & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, *26*(6), 867-891. doi: 10.1016/j.humov.2007.07.002

Rusiewicz, H. L. (2011). Synchronization of prosodic stress and gesture: A dynamic systems perspective. Proceedings of the 2nd Conference on Gesture and Speech in Interaction (GESPIN 2011), Bielefeld, Germany.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, *57*, 283-300. doi: 10.1016/j.specom.2013.06.004.

Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, 111(Pt. 1), 2237–2241. doi: 10.1121/1.1466868

Pouw, W. & Dixon, J. A. (under review). Quantifying gesture-speech synchrony: Exploratory study and pre-registration. Pre-print available at: https://psyarxiv.com/983b5

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.

Westwood, D. A., Heath, M., & Roy, E. A. (2000). The effect of a pictorial illusion on closed-loop and open-loop prehension. *Experimental Brain Research*, *134*(4), 456–463. doi: 10.1007/s002210000489