Gesture-Speech Physics in Fluent Speech and Rhythmic Upper Limb Movements

Wim Pouw1,2,3, Lisette de Jonge-Hoekstra1,5, Steven J. Harrison1,6, Alexandra Paxton1,7, & James A. Dixon1,7

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

2 Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen

3 Institute for Psycholinguistics, Max Planck Nijmegen

4 Faculty of Behavioral and Social Sciences, University of Groningen

5 Royal Dutch Kentalis, Sint-Michielsgestel

6 Department of Kinesiology, University of Connecticut

7 Department of Psychological Sciences, University of Connecticut

Author note

All anonymised data and analysis code are available at the Open Science Framework (<https://osf.io/tgbmw/>). This manuscript has been written with Rmarkdown - for the code-embedded reproducible version of this manuscript please see the Rmarkdown (.Rmd) file available at the OSF page. 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).  
Some sections of this paper have been submitted (in verbatim) as 2-page abstract to GESPIN2020. Acknowledgement: We would like to thank Jenny Michlich for pointing us to relevant bioacoustic literature. We thank Susanne Fuchs for valuable comments on this work.

Correspondence concerning this article should be addressed to Wim Pouw, Donders Institute for Brain, Cognition and Behaviour, Heyendaalseweg 135, 6525 AJ Nijmegen. E-mail: [w.pouw@psych.ru.nl](mailto:w.pouw@psych.ru.nl)

Abstract

Communicative hand gestures are often coordinated with prosodic aspects of speech, and salient moments of gestural movement (e.g., quick changes in speed) often co-occur with salient moments in speech (e.g., near peaks in fundamental frequency and intensity). A common understanding is that such gesture and speech coordination is culturally and cognitively acquired, rather than having a biological basis. Recently, however, the biomechanical physical coupling of arm movements to speech movements has been identified as a potentially important factor in understanding the emergence of gesture-speech coordination. Specifically, in the case of steady-state vocalization and mono-syllable utterances, forces produced during gesturing are transferred onto the tensioned body, leading to changes in respiratory-related activity and thereby affecting vocalization F0 and intensity. In the current experiment (N = 37), we extend this previous line of work to show that gesture-speech physics impacts fluent speech, too. Compared with non-movement, participants who are producing fluent self-formulated speech, while rhythmically moving their limbs, demonstrate heightened F0 and amplitude envelope, and such effects are more pronounced for higher-impulse arm versus lower-impulse wrist movement. We replicate that acoustic peaks arise especially during moments of peak-impulse (i.e., the beat) of the movement, namely around deceleration phases of the movement. Finally, higher deceleration rates of higher-mass arm movements were related to higher peaks in acoustics. These results confirm a role for physical-impulses of gesture affecting the speech system.We discuss the implications of gesture-speech physics for understanding of the emergence of communicative gesture, both ontogenetically and phylogenetically.

*Keywords:* hand gesture, speech production, speech acoustics, biomechanics, entrainment

*Word count:* X

Gesture-Speech Physics in Fluent Speech and Rhythmic Upper Limb Movements

Communicative hand gestures are ubiquitous across human cultures. Gestures aid communication by seamlessly interweaving relevant pragmatic, iconic and symbolic expressions of the hands together with speech (Feyereisen, 2017; Holler & Levinson, 2019; Streeck, 2008). For such multi-articulatory utterances to do their communicative work, gesture and speech must be tightly temporally coordinated to form a sensible speech-gesture whole. In fact, gestures’ salient moments are often timed with emphatic stress made in speech, no matter what the hands depict (Shattuck-Hufnagel & Prieto, 2019; Wagner, Malisz, & Kopp, 2014). For such gesture-speech coordination to get off the ground, the system must functionally constrain its degrees of freedom (Turvey, 1990); in doing so it will have to utilize (or otherwise account for) intrinsic dynamics arising from the bio-physics of speaking and moving at the same time. Here we provide evidence that movement of the upper limbs constrain fluent self-generated speech acoustics through biomechanics.

## The gesture-speech prosody link

The tight coordination of prosodic aspects of speech with the kinematics of gesture has been long appreciated and is classically referred to as the beat-like quality of co-speech gesture (McNeill, 1992). As obtained from video analysis, gesture apices are often found to align with *pitch accents*— accents which are acoustically predominately defined by positive excursions in the fundamental frequency (FO), lowering of the second formant, longer vowel duration, and increased intensity (Loehr, 2012; McClave, 1998; Mendoza-Denton & Jannedy, 2011). Pitch accents can be perceptually differentiated by sudden lowering of F0 as well, but gestures do not seem to align with those events quite as much (Im & Baumann, 2020).  
In more recent motion-tracking studies, gesture-speech prosody correlations have been obtained as well. For example, gestures’ peak velocity often co-occurs near peaks in F0, even when such gestures are depicting something (Danner, Barbosa, & Goldstein, 2018; Krivokapić, 2014; Leonard & Cummins, 2011; Pouw & Dixon, 2019a, 2019b). In pointing gestures, stressed syllables align neatly with the maximum extension of the pointing movement, such that the hand movement terminates at the first syllable utterance in strong-weak stressed “PApa” and terminates later during the second syllable utterance in the weak-strong “paPA” (Esteve-Gibert & Prieto, 2013; Rochet-Capellan, Laboissière, Galván, & Jean-Luc, 2008). During finger-tapping and mono-syllabic utterances, when participants are instructed to alternate prominence in their utterances (“pa, PA, pa, PA”), the tapping action spontaneously aligns with the syllable pattern, such that larger movements are made during stressed syllables (Parrell, Goldstein, Lee, & Byrd, 2014). Conversely, if participants are instructed to alternate stress in finger tapping (strong-weak-strong-weak force production), speech will follow, with larger oral-labial apertures for stressed vs. unstressed tapping movements.  
Even when people do not intend to change the stress patterning of an uttered sentence, gesturing concurrently affects speech acoustics in a way that makes it seem intentionally stressed, inducing an increase in vocalization duration and a lowering of the second formant of co-occurrent speech (Krahmer & Swerts, 2007). Further, gesture and speech cycle rates seem to be attracted towards particular (polyrhythmic) stabilities: In-phase speech-tapping is preferred over anti-phase coordination, and 2:1 speech-to-tapping ratios are preferred over more complex integer ratios such as 5:2 (Kelso & Tuller, 1984; Stoltmann & Fuchs, 2017; Treffner & Peter, 2002; Zelic, Kim, & Davis, 2015). These previous results indicate that gesture and speech naturally couple their activity, like many living and non-living oscillatory systems (Pikovsky, Kurths, & Rosenblum, 2001), requiring further study on the exact nature of this coupling.

## Gesture-speech physics

The current mainstream understanding of the gesture-prosody link is that is not “biologically mandated” (p.69 ; McClave, 1998; Shattuck-Hufnagel & Ren, 2018), requiring neural-cognitive timing mechanisms (De Ruiter, 2017; Ruiter, 2000) that appear only after about 16 months of age (Iverson & Thelen, 2005; see also Esteve-Gibert & Guellai, 2018). Recently, however, a potential physical coupling of arm movements with speech via myofascial tissue biomechanics was investigated, where it was found that hand gesturing physically impacts steady-state vocalizations and mono-syllabic consonant-vowel utterances (Pouw et al., 2019a, 2019b; Pouw, Harrison, Esteve-Gibert, & Dixon, n.d.; Pouw, Paxton, Harrison, & Dixon, 2020). Specifically, hand and arm movements can transfer a force (a physical impulse) onto the musculoskeletal system, thereby modulating respiration-related muscle activity, leading to changes in vocalization’s intensity. If vocal fold adjustments do not accommodate for gesture-induced impulses, the fundamental frequency (F0) of vocalizations is affected as well. Higher-impulse arm movements or two-handed movements will induce more pronounced effects on F0 and intensity than lower-impulse wrist movements or one-handed movements. This is because the mass of the “object” in motion is higher for arm versus wrist movements, thereby changing the momentum of the effector (everything else—such as effector speed—being equal, as momentum effector = effctor mass x effector velocity). The change in momentum is the physical impulse, and physical impulse is highest when the change in velocity (i.e., acceleration) is highest (everything else—such as effector mass—being constant).  
How physical impulses are absorbed by the respiratory system is likely complex and not a simple linear function (Levin, 2006). However, a complete understanding will involve an appreciation of the body as a pre-stressed system (Bernstein, 1967; Profeta & Turvey, 2018), forming an interconnected tensioned network of compressive (e,g., bones) and tensile elements (e.g., fascia, muscles) through which forces may reverberate nonlinearly (Silva, Moreno, Mancini, Fonseca, & Turvey, 2007; Turvey & Fonseca, 2014). Specifically, the upper limb movements are controlled by stabilizing musculoskeletal actions of the scapula and shoulder joint, which directly implicate accessory expiratory muscles that also stabilize scapula and shoulder joint actions (e.g., the serratus anterior inferior; see Pouw et al., n.d. for an overview).  
Peripheral actions also play a role, as performing an upper limb movement recruits a whole kinetic chain of muscle activity around the trunk (e.g., rector abdominus) to maintain posture (Hodges & Richardson, 1997). Indeed, when people are standing vs. sitting, for example, the effects of peak physical impulse of gestures onto vocalization acoustics are more pronounced (Pouw et al., 2019a). We reasoned that this is because standing involves more forceful anticipatory postural counter adjustments (Cordo & Nashner, 1982), which reach the respiratory system via accessory expiratory muscles also implicated in keeping postural integrity. Recently, more direct evidence has been found for the gesture-respiratory-speech link: Respiratory-related activity (measured with a respiratory belt) was enhanced during moments of peak-impetus of gesture as opposed to other phases in the gesture movement, and respiratory-related activity itself was predictive of the gesture-related intensity modulations of mono-syllable utterances (Pouw et al., n.d.).  
The evidence reviewed so far has been based on experiments on continuous vocalizations or monosyllabic utterances. Such results cannot directly generalize to fluent, self-generated, full-sentenced speech. However, recent work suggest that gesture-speech physics does generalize to fluent speech. For example, Cravotta, Busà, and Prieto (2019) found that encouraging participants to gesture during cartoon-narration versus giving no instructions lead to 22Hz increase in observation of max F0 and to greater F0 ranges of speech and intensity. Furthermore, computational modelers have reported on interesting successes in synthesizing gesture kinematics based on speech acoustics alone (Ginosar et al., 2019; Kucherenko, Hasegawa, Henter, Kaneko, & Kjellström, 2019), indicating that information about body movements inhabits the speech signal. Such results do not necessitate a role for biomechanics but only suggests a strong connection between gesture and speech.

## Current experiment

The current experiment was conducted as a simple test of the constraints of upper limb movement on fluent speech acoustics. Participants were asked to retell a cartoon scene that they had just watched, while either not moving, vertically moving their wrist, or vertically moving their arm at a tempo of 80 beats per minute (1.33Hz). Participants were asked to give a stress or beat in the downward motion with a sudden stop at maximum extension (i.e., sudden deceleration). Participants were asked to not allow movements to affect their speaking performance in any way. Similar to previous experiments (e.g., Pouw et al., n.d.), we assessed the following to conclude that gesture-speech physics is present: \* 1) Does rhythmic co-speech movement change acoustic markers of prosody (i.e., F0 and amplitude envelope)?  
\* 2) At what moments of co-speech movement is change in acoustics observed? \* 3) Does degree of physical impulse (as measured by effector mass or changes in speed) predict acoustic variation?

# Method

## Participants & Design

A total of 37 undergraduate students at the University of Connecticut were recruited as participants (*M* age = 18.76, *SD* age = 0.95, %cis-gender female = 67.57, %cis-gender male = 32.43, %right-handed = 94.59).  
The current design was fully-within subject, with a three-level movement manipulation (passive vs. wrist-movement vs. arm-movement condition). Movement condition was randomly assigned per trial. Taken together, participants performed 419 trials, each lasting about 40 seconds. The study design was approved by the IRB committee of the University of Connecticut (#H18-227).

## Material & Equipment

### Cartoon vignettes.

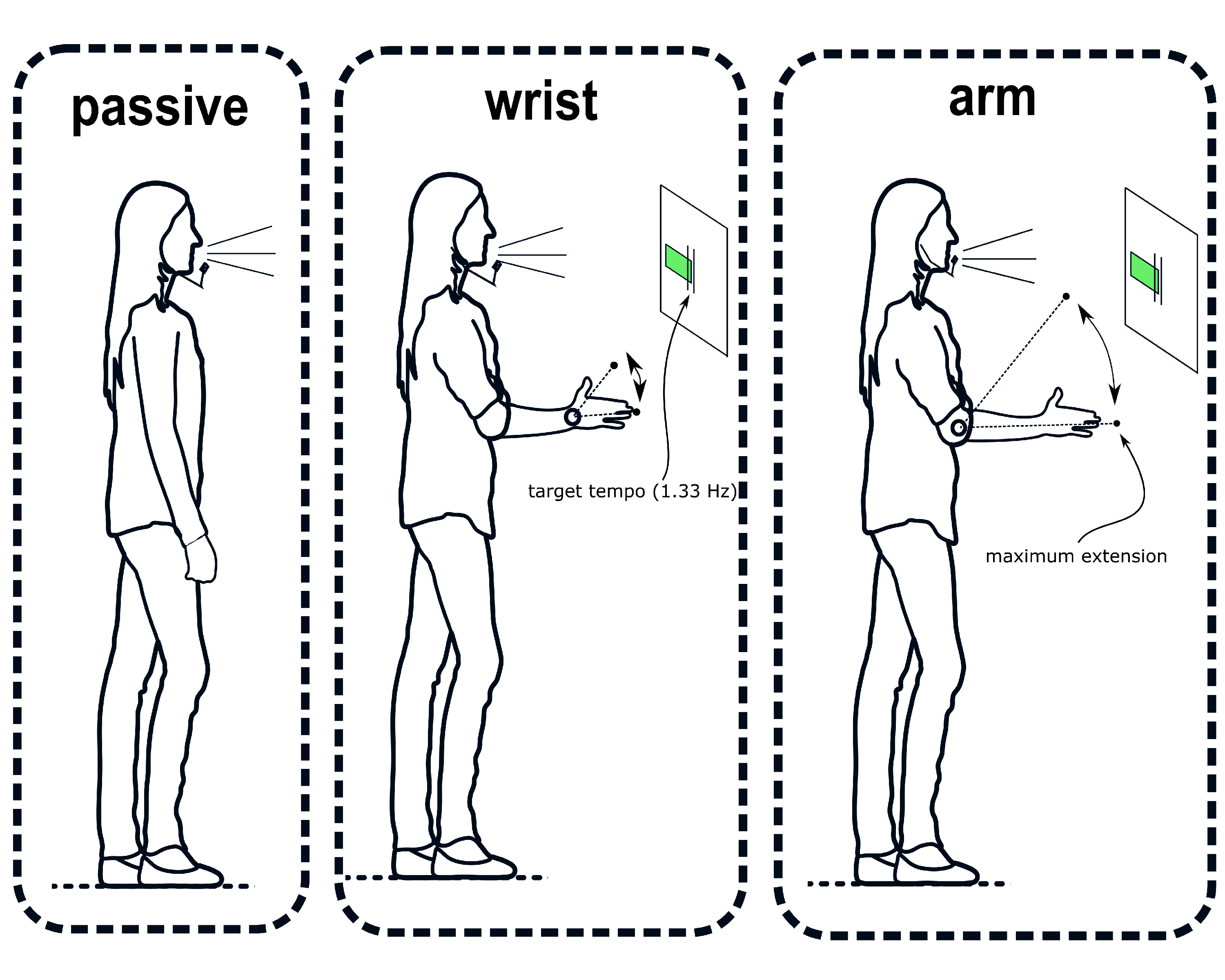
Twelve cartoon vignettes were created from the “Canary Row” and “Snow Business” Tweety and Sylvester cartoons (M vignette duration = 59.42seconds (*SD* = 32.11). These cartoons are often used in gesture research (McNeill, 1992). The videos can be accessed here: <https://osf.io/rfj5x/>.

### Audio and Motion Tracking.

A MicroMic C520 cardioid condenser microphone headset (AKG, Inc.) was used to record audio at 44.1Khz. The microphone was plugged into a computer that handled the recording via a c++ script. Also plugged into this computer was a Polhemus Liberty motion tracking system (Polhemus, Inc.) which tracked position of the participant’s index finger of the dominant hand, sampling with one 6D sensor at 240 Hz. We applied a first-order Butterworth filter at 30 Hz for the vertical position (z) traces and its derivatives.

## Procedure

Upon arrival, participants were briefed that this 30-minute experiment entailed retelling cartoon scenes while standing and performing upper-limb movements. A motion sensor was attached to their tip of the index finger of the dominant hand, and a microphone headset was put on. Participants were asked to stand upright and were introduced to three movement conditions (see Figure 1). In the passive condition, participants did not move and kept their arm resting alongside the body. In the wrist-movement condition, participants were asked to continuously move the hand vertically at the wrist joint while keeping the elbow joint at 90 degrees. In the arm-movement Condition, participants moved their arm vertically at the elbow joint, without wrist movement. Similar to previous studies (e.g., Pouw et al., 2020), participants were asked to give emphasis in the downward motion of the movement with a sudden halt—in other words, a beat—at the maximum extension of their movement.  
After introduction of the movements, participants were told that they were to move at a particular tempo, indicated by visual feedback system. The feedback system consisted of a horizontal bar that continually updated to report on the participant’s movement speed in the previous movement cycle. The participant was to keep the horizontal bar between the lower and higher boundaries (a 20% region, [72-88] BPM) of the 1.33-Hz target tempo (i.e., 80 BPM). Participants briefly practiced moving at the target rate before starting the experiment.  
Critically, the participants were not exposed to an external rhythmic signal, like a visual metronome. Subsequently, participants were instructed that they would watch and then retell cartoon clips while making one of the instructed movements (or making no movements). Participants were asked to keep their speech as normal as possible while making the movements (or no movement). In the conditions requiring movement, participants were to keep their movement tempo within the target range. Twelve cartoon vignettes were readied to be shown before each trial. The experiment ended when the participant saw and retold all 12 vignettes or when the total experiment time reached 30 minutes. To ensure that all movement conditions would be performed at least once within that time, we set the maximum time per trial at 1 minute. In other words, when participants were still retelling the same scene after 60 seconds, the experimenter would terminate the trial and move to the next trial. Mean retelling time was, however, well below 1 minute (*M* = 26.00 seconds, *SD* = 7.06).

Figure 1. Graphical overview of movement conditions  *Note*. Movement conditions are shown. Each participant performed all conditions (i.e., within-subjects). To ensure that movement tempo remained relatively constant, participants were shown a moving green bar, which indicated whether they moved too fast or too slow relative to a 20% target region of 1.33Hz. Participants were instructed to have an emphasis in the downbeat with an abrupt stop (i.e., beat) at the maximum extension.

## Preprocessing

### Speech acoustics.

The fundamental frequency was extracted with sex-appropriate preset ranges (male = 50-400Hz; female = 80-640Hz). We used a previously written R script (<https://osf.io/m43qy/>; Pouw & Trujillo, 2019) utilizing the R package ‘wrassp’ (Winkelmann, Bombien, & Scheffers, 2018), which applies a K. Schaefer-Vincent algorithm. It should be noted that F0 tracking is always susceptible to noisy estimation. We have however checked for multiple participants of the data if there were gross mistrackings of F0 algorithm (e.g., sudden jumps to higher harmonics) and we did not find any. However, we did not hand-check the F0 track for all the data given the current sample size, therefore we must accept a certain range of noise that is common to F0 tracking.  
We also extracted a smoothed (5-Hz Hann window) amplitude envelope using a previously written custom-written R script (<https://osf.io/uvkj6/>, which reimplements in R a procedure from He & Dellwo, 2017). The amplitude envelope was calculated by applying a Hilbert transformation to the sound waveform, yielding a complex-valued analytic signal from which we take the complex modulus. After smoothing and downsampling to 240Hz, this gives a one-dimensional time series referred to as the amplitude envelope tracing the extrema’s (i.e., envelope) of the sound waveform as shown in Figure 2.

### Data and Exclusions.

We collected 189.70 minutes of continuous data (passive condition = 63.45, wrist-movement condition = 63.56, arm-movement = 62.69). However, a C++ memory allocation error resulted in the loss of the precise timing information of the sampling of the motion tracker after a certain period (1 million milliseconds), thus for a subset of the experiment data. Full data was therefore obtained for the first 16 minutes and 40 seconds of each trial for all participants. We limit our analyses to this complete data set. This dataset consists of 124.49 minutes of continuous speech and movement data (passive = 40.08, wrist-movement condition = 42.32, arm-movement condition = 42.10).

## Baseline

For gesture-speech analysis we also created a surrogate condition. We randomly paired the speech of the passive condition trials of participant x with motion-tracking data from the movement conditions for that participant x (without scrambling the order of the speech and motion time series extracted in these falsely paired trials). This surrogate randomly paired condition allowed us to exclude the possibility that any effects of movement were due to chance correlations inherent to the structure of speech and movement, rather than the correlations arising out of the coupling of speech and movement. We only use this surrogate control condition as a contrast when we are performing analysis on the temporal relation between speech and movement.  
We computed the following measures to check whether our movement manipulation was successful and whether speech rates were comparable across conditions. Figure 2 shows a summary of the results for key manipulation check measures.

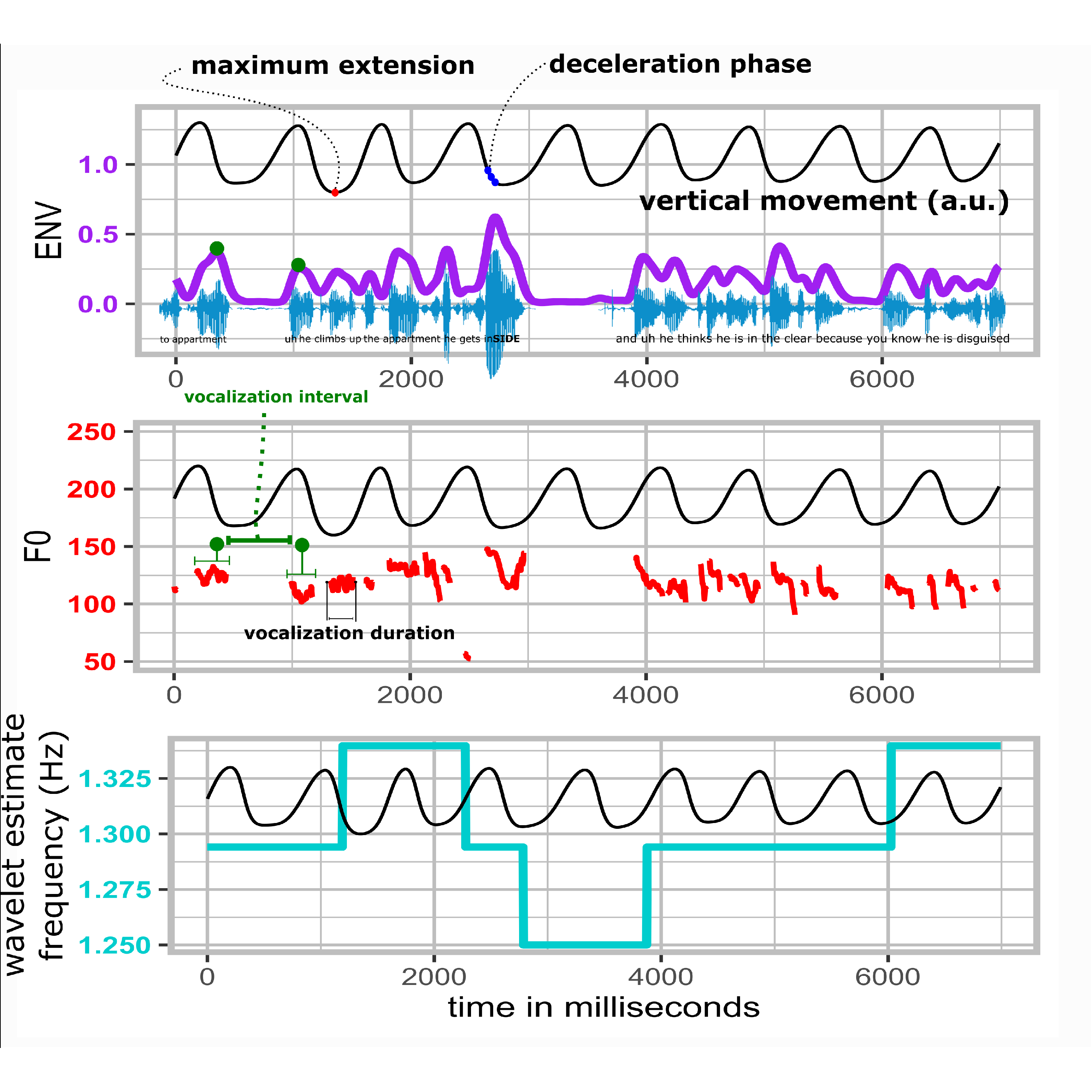
## Manipulation Checks

### Movement Frequency.

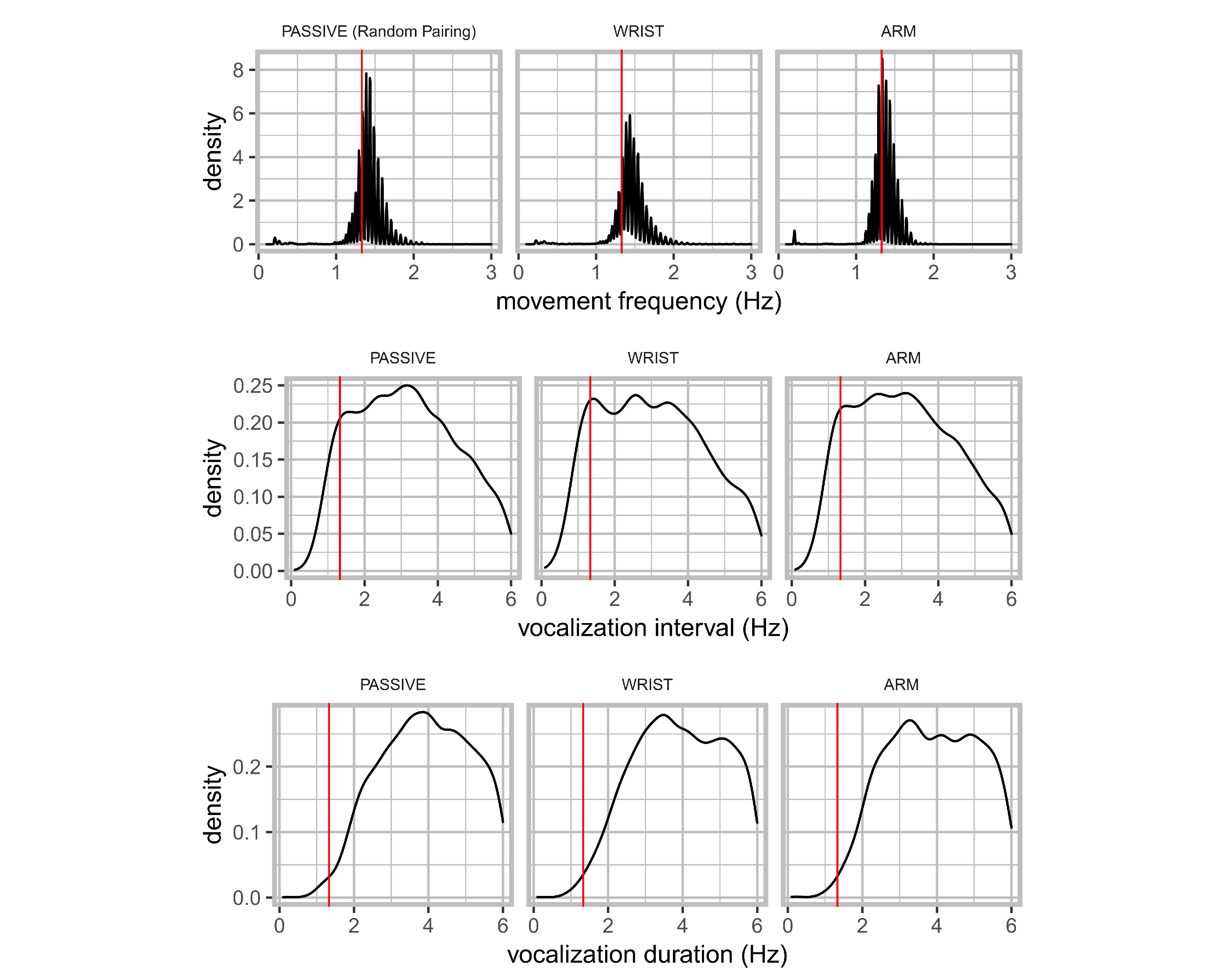
To ascertain whether participants moved their limbs within the target 1.33-Hz range, we performed a wavelet-based analysis (using R package “WaveletComp”; Rosch & Schmidbauer, 2014). Wrist movements were performed at slightly faster rates (*M* = 1.44 Hz, *SD* = 0.24) than arm movements (*M* = 1.36 Hz, *SD* = 0.19), but in both cases the movements were distributed over the target range. This confirms that our movement manipulation was successful. For our surrogate control condition, the mean frequency of the artificially paired movement time series fell between both arm- and wrist-movement condition frequency distributions (*M* = 1.41 Hz, *SD* = 0.22).

### Speech Rate.

To provide a description about speech rate we calculated two measures, namely vocalization duration and vocalization interval (see Fig. 2 for examples), which are measures derived from information in the F0 track, as well as the amplitude envelope for the interval calculation. The vocalization duration was defined as the ms length of an uninterrupted run of F0 observations. The vocalization interval was determined by identifying two consecutive runs of F0 observations (i.e., vocalization events) and determining the peak amplitude envelope of each of those vocalization events as to compare the relative timing between those peaks. This way we have a single time point for each vocalization event that we can compare with the next vocalization event’s time point (i.e., the vocalization interval).  
Figure 3 shows relatively uniform distributions for these specific speech measures. No clear 1:1 frequency couplings of movement and vocalization duration or vocalization interval nor any other clear signs of polyrhythmic coupling of movement and speech are observed (see e.g., Zelic et al., 2015; Stoltmann & Fuchs, 2017). Note though that there are other possible (acoustically defined) units of speech that might entrain to movements that we do not further pursue here (Lin & Rathcke, 2020). We restrict ourselves for the current report to speech vocalization acoustics rather than speech-movement cycle dynamics, as the former is the confirmatory research topic of the current study.  
To compare vocalization rates to movement, we computed the average vocalization duration and interval for each trial by tracking the time of uninterrupted runs of F0 observations and then converting the time in milliseconds to Hz. For the passive condition, the average vocalization duration was *M* = 6.28 Hz, *SD* = 6.03, and the vocalization interval was *M* = 5.17 Hz, *SD* = 6.94.. For the wrist-movement condition the vocalization duration was *M* = 6.24 Hz, *SD* = 5.96, and the vocalization interval was the vocalization interval was *M* = 5.02 Hz, *SD* = 6.86. For the arm-movement condition, the vocalization duration was *M* = 6.08 Hz, *SD* = 5.83 and the vocalization interval was *M* = 4.86 Hz, *SD* = 5.76.

Figure 2. Example movement-, amplitude envelope-, F0- time series, and time-dependent movement frequency estimates 

*Note figure 2.* A sample of about 10 seconds is shown. With the participant’s permission the speech sample is available at <https://osf.io/2qbc6/>. The smoothed amplitude envelope in purple traces the waveform maxima. The F0 traces show the concomitant vocalizations in Hz, with an example of vocalization interval and vocalization duration (which was calculated for all vocalizations). The bottom panel shows the continuously estimated movement frequency in cyan, which hovers around 1.33 Hz. In all these panels, the co-occurring movement is plotted in arbitrary units (a.u.) to show the temporal relation of movement phases and the amplitude envelope, F0, and the movement frequency estimate. In our analysis, we refer to the maximum extension and deceleration phases as relevant moments for speech modulations. In this example, a particularly dramatic acoustic excursion occurs during a moment of deceleration of the arm movement, possibly an example of gesture-speech physics.

Figure 3. Summary of movement-frequency, vocalization duration and vocalization interval  


*Note Figure 3*. Density distributions of movement frequencies, vocalization interval, and vocalization duration are shown. There was no movement for the passive condition, but we display the randomly paired movement time series in the surrogate baseline pairing for which frequency information is shown. The red vertical line indicates the target movement frequency at 1.3 Hz.

# Results

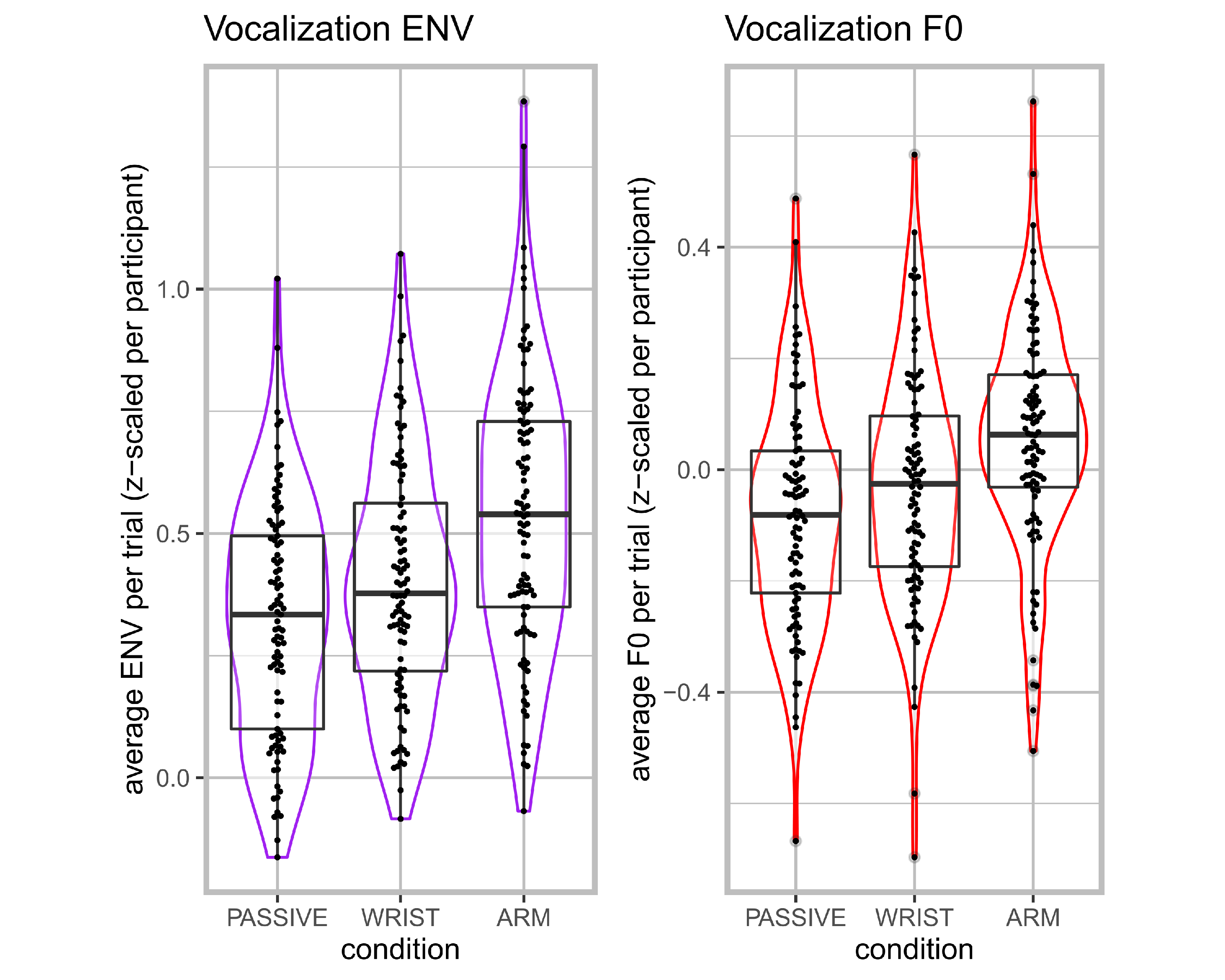
## Overview of analyses

We report three main analyses to show that gesture-speech physics is present in fluent speech. Firstly, we assess overall effects of movement condition on vocalization acoustics (F0 and the amplitude envelope); these would support our hypothesis that upper limb movement—and, especially, high-impulse movement—constrains fluent speech acoustics. Secondly, we assess whether vocalization acoustic modulations are observed at particular phases of the movement cycle, which gesture-speech physics holds should occur at moments of peaks in deceleration. Thirdly, we assess whether a continuous estimate of upper limb physical impulse through deceleration rate predicts vocalization acoustic peaks, which would support the gesture-speech physics hypothesis that physical impulses are transferred onto the vocalization system.  
The following generally applies to all analyses. For hypothesis testing, we performed mixed linear regression models (using R-package “nlme”; Pinheiro, Bates, DebRoy, Sarkar, & R Team, 2019), and non-linear generalized additive modeling or GAM (using R-package “gam”; Hastie, 2019) with random intercept for participants by default.

## Acousic correlates of movement condition

Figure 4 shows the average F0 and amplitude envelope (z-scaled for participants) per trial per condition. The passive condition had generally lower levels of F0 and amplitude envelope as compared to the arm- and wrist-movement conditions. Furthermore, the higher-impulse arm-movement condition generally had higher levels of F0 and amplitude envelope as compared to lower-impulse wrist-movement condition.  
Table 1 shows the results of mixed linear regression analysis. For the amplitude envelope, the passive condition had a lower average amplitude envelope as compared to the the wrist-movement condition, as well as the arm-movement condition. After accounting for differences in F0 for sex (males had generally 73 Hz lower F0), wrist-movement condition had about 1.6 Hz increase in average as compared to the passive condition, but this was not statistically reliable. Further, the arm-movement condition increased in F0 by 3.5 Hz over the passive Condition.

Figure 4. Average F0 and amplitude envelope (ENV) per trial per condition

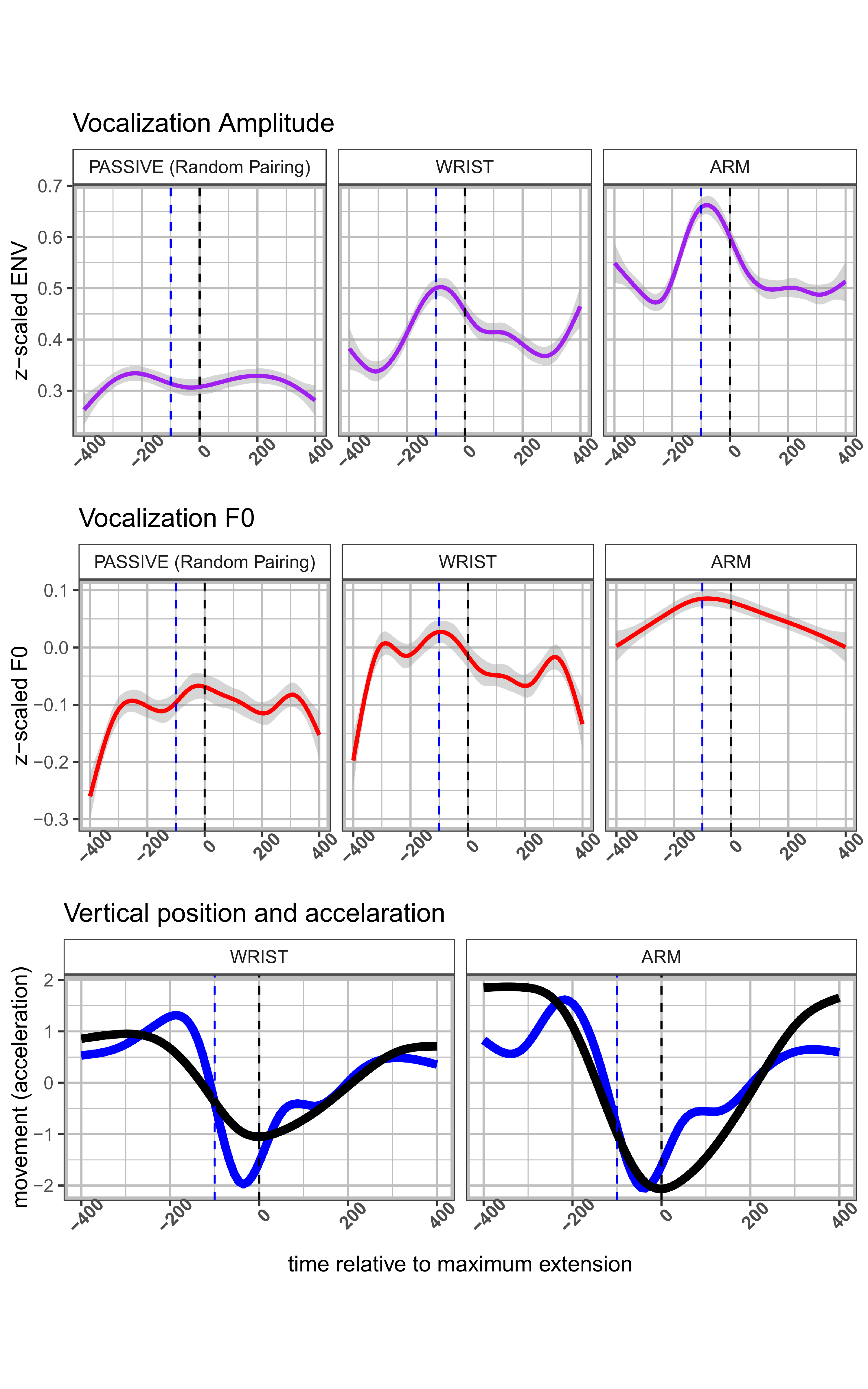


*Note Figure 4*. Violin and box plots are shown for average F0 (Hz) and amplitude envelope (*z*-scaled) per trial (jitters show observation).

Table 1. Linear mixed effects for effects of condition on F0 and amplitude envelope (ENV)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | contrast | b | SE | df | p |
| ENV (z-scaled) | intercept | 0.32 | 0.036 | 251 | < .0001 |
|  | Wrist vs. Passive | 0.094 | 0.028 | 251 | 0.001 |
|  | Arm vs. Passive | 0.215 | 0.028 | 251 | < .0001 |
|  |  |  |  |  |  |
| F0 (Hz) | intercept | 186.577 | 3.22 | 251 | < .0001 |
|  | Male vs. Female | -73.268 | 5.437 | 33 | < .0001 |
|  | Wrist vs. Passive | 1.603 | 0.845 | 251 | 0.0588 |
|  | Arm vs. Passive | 3.504 | 0.828 | 251 | < .0001 |

## Coupling of vocalization duration and movement

Having ascertained in the previous analysis that acoustics were modulated for movement versus no movement, we further need to confirm that such modulations occur at particular moments in the movement cycle. Figure 5 shows the main results for all data, where we model over time the acoustic patterning in vocalizations around the maximum extension of the movement cycle, for all movement cycles that occurred. If there are particular moments in the movement cycle where vocalization is affected—for example, the moment when the hand starts decelerating (estimated from the data as shown in Figure 5)—we would expect acoustic modulations (peaks) at such moments of the movement cycle.  
Just before the moment of maximum extension, the observed amplitude envelope shows a clear peak, most dramatically for the arm-movement condition, but also for the wrist-movement condition. For speech in the randomly paired movement- and passive condition, this was not the case; this provides evidence that the results observed in the arm- and wrist-movement conditions are not due to mere chance. For F0, the pattern is somewhat less clear, but positive peaks still occur just before the maximum extension. These findings replicate our earlier work on steady-state vocalization and mono-syllabic utterances, showing that moments of peak deceleration also show peaks in acoustics (Pouw et al., 2019a, n.d.).  
To test whether trajectories are indeed non-linear and are reliably different from the passive condition, we performed Generalized Additive Modeling or GAM, a type of non-linear mixed effects procedure. GAM is a popular time-series analysis in phonetics and allows the automatic modeling of more (and less) complex non-linear patterns by combining a set of smooth basis functions. Furthermore, GAM allows for testing whether those non-linear trajectories are modulated depending on some grouping of the data (see e.g., Wieling, 2018). We assessed with GAM the trajectory of acoustics around 800 milliseconds of the maximum extension of the movement. We chose 800 milliseconds (-400, 400) as this is about the time for a 1.33Hz cycle (1000/1.33Hz = 752 ms) with an added margin of error of about 50ms. The model results with random slopes and intercept for participant are shown in Table 2.  
Firstly, for all models tests for non-linearity of the trajectories were statistically reliable (*p*’s < .0001), meaning that there were peaks or valleys in acoustics over the movement cycle rather than a flat linear trend (Figure 6). As shown in Table 2, our results replicate the general finding that the wrist movements condition led to reliably different non-linear peaks in acoustics as compared to the passive condition (*p* < .001). Moreover, this effect—relative to the passive condition—is even more extreme for the arm movement condition (*p* < .001). Figure 6 provides the fitted trajectories for the GAM models.  
For readers interested in individual differences in trajectories, we have created interactive graphs for each participant’s average amplitude envelope trajectories (<https://osf.io/a423h/>) and F0 trajectories (<https://osf.io/fdzwj/>).  
 Figure 5. Average observed vocalization acoustics relative to the moment of maximum extension  


*Note Figure 5*. For the upper two panels the average acoustic trajectory is shown around the moment of maximum extension (t = 0, dashed black line). In the lower panel, we have plotted the *z*-scaled average vertical displacement of the hand and the *z*-scaled acceleration trace. The blue dashed vertical line marks the moment where the deceleration phase starts, which aligns with peaks in acoustics.

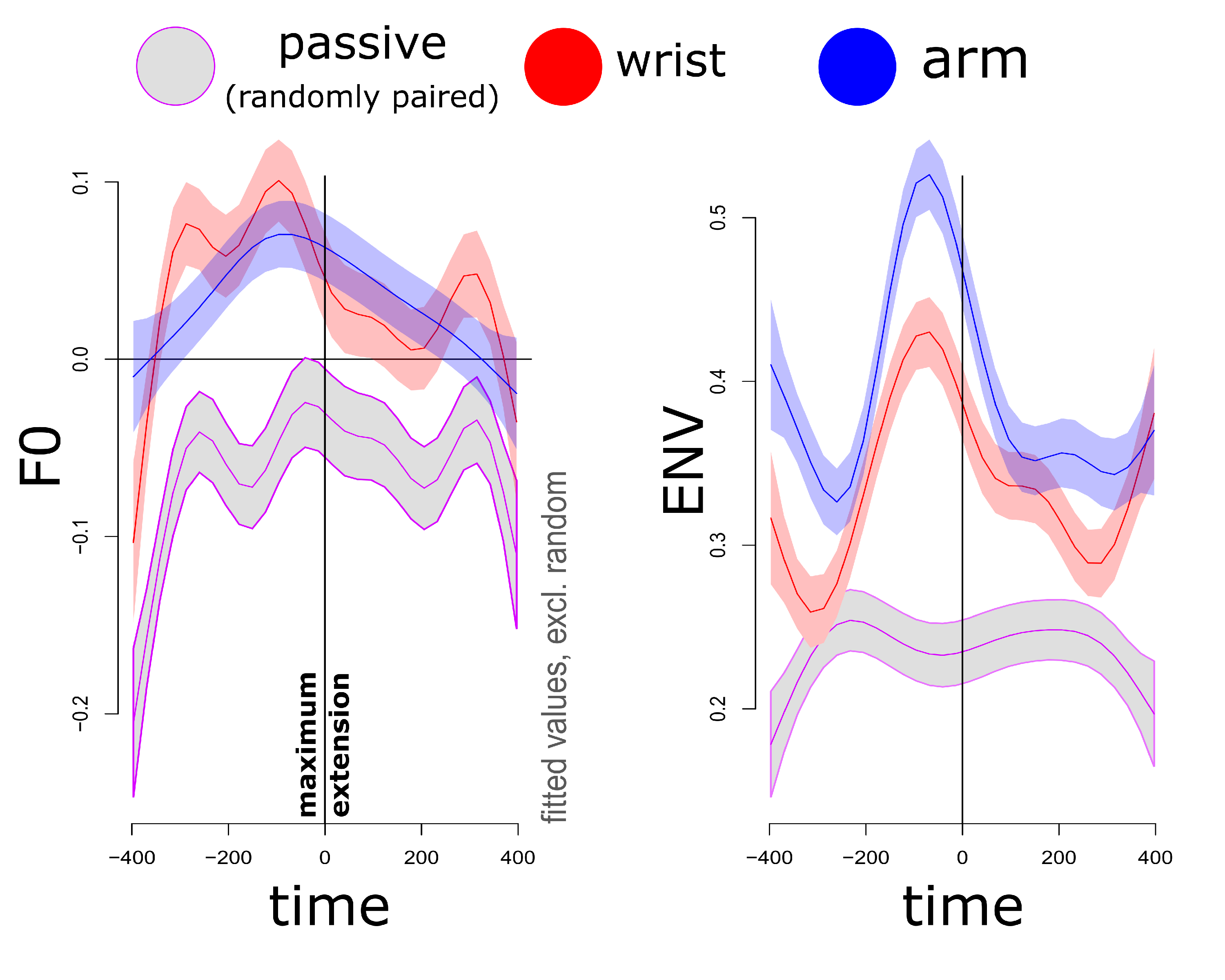
Figure 6. Fitted trajectories GAM 

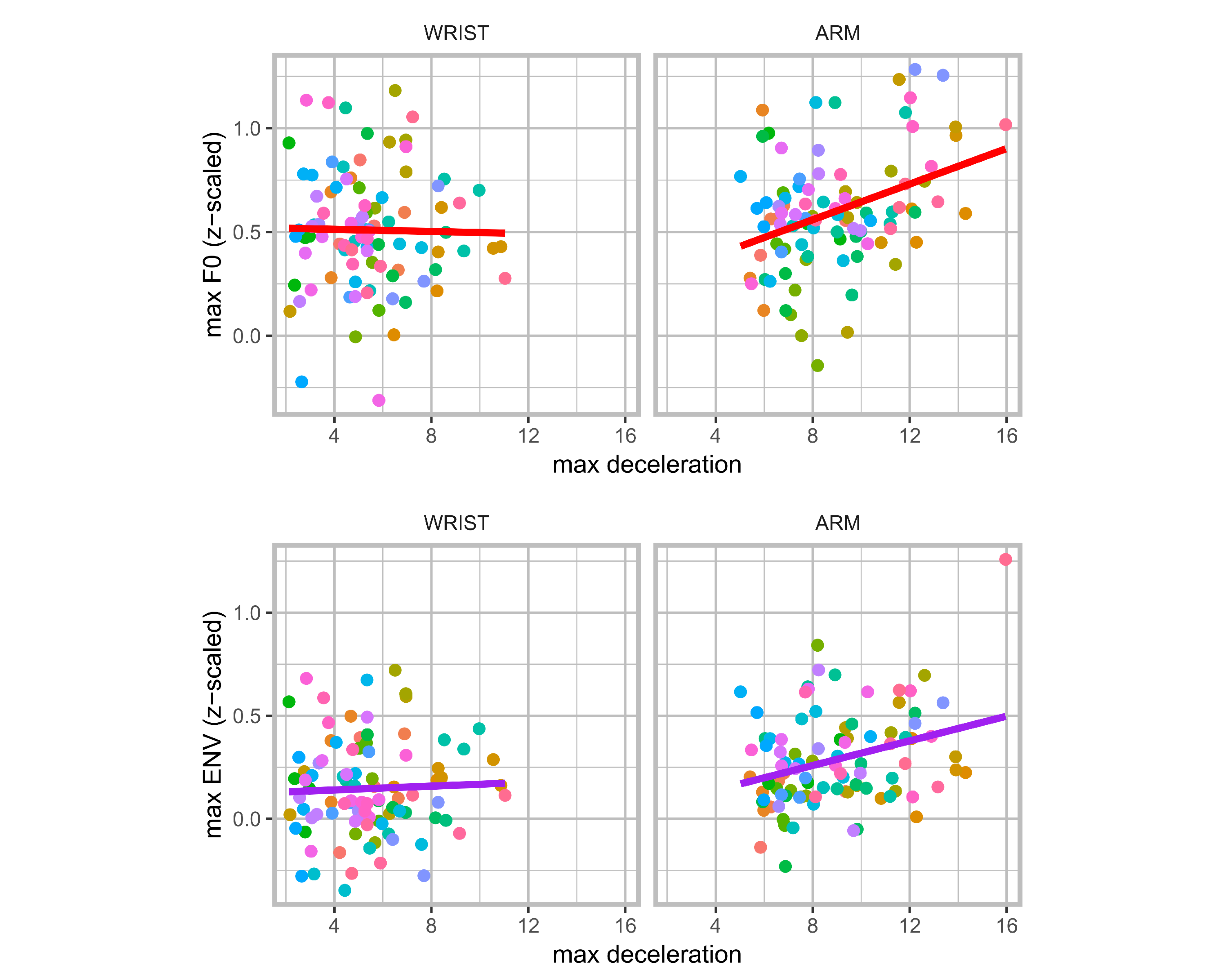
Table 2. Model results for GAM analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | contrast | b | SE | df | p |
| ENV (z-scaled) | intercept | 0.237 | 0.006 | 36.923 | < .0001 |
|  | Wrist vs. Passive | 0.096 | 0.009 | 10.579 | < .0001 |
|  | Arm vs. Passive | 0.152 | 0.009 | 16.862 | < .0001 |
|  |  |  |  |  |  |
| F0 (Hz) | intercept | -0.061 | 0.006 | -8.35 | < .0001 |
|  | Male vs. Female | -0.019 | 0.009 | -4.29 | < .0001 |
|  | Wrist vs. Passive | 0.101 | 0.009 | 10.222 | < .0001 |
|  | Arm vs. Passive | 0.094 | 0.103 | 9.546 | < .0001 |

*Note*. Model results are shown for the amplitude envelope (ENV; *z*-scaled) and F0 (Hz). For F0, we accounted for sex differences when estimating independent effects of condition.

## Degree of physical impetus and acoustic peaks

We have confirmed that speech acoustics are modulated around moments of the deceleration phase, about 0-200 ms before the maximum extension. The effect of gesture-speech physics can be further examined by looking at how the forces produced by the upper limb movement predict acoustic peaks. Therefore, for all vocalizations that occurred between 200–0ms before the maximum extension, we assessed whether acoustic peak (i.e., maximum F0 or maximum amplitude envelope) was predicted by the maximum deceleration value (i.e., minimum acceleration observation) observed in that 200 ms window. In previous research, we found that higher deceleration was related to higher amplitude envelope observations but not F0 (Pouw et al., n.d.).  
Figure 7 shows the general pattern of the results for the wrist- and arm-movement condition. Here we averaged per trial the maximum deceleration values of max F0 and max ENV for each vocalization event. Table 3 shows the model results of linear mixed effects model with random intercept and slopes for participants, in which we regressed the trial-averaged max observed deceleration against the co-occurring trial-averaged vocalization acoustic peaks for amplitude envelope and F0 (separately). Higher deceleration indeed predicted higher amplitude envelope. This was also the case for F0, but only for arm movements (as opposed to wrist movement), as indicated by a statistically reliable interaction between condition and max deceleration effect (*p*’s < .05). Together, these demonstrate the roles of both acceleration and effector mass in producing physical impulses.

Figure 7. Relation max deceleration and height acoustic peak 

*Note Figure 7*. The x-axis shows the average maximum deceleration per trial (absolutized negative acceleration value), where 0 indicates no deceleration (absolutized) and positive values indicate higher deceleration rates in cm/seconds squared. Each point contains trial averaged values. It can be seen that deceleration rates are more extreme for the arm versus the wrist condition. On the y-axis, we have the average maximum observed amplitude envelope (lower panel) and F0 (upper panel) for those moments of deceleration. Higher decelerations co-occur with higher peaks in acoustics for arm movements (but not or less so for wrist movements).

Table 4. Linear mixed effects of deceleration and acoustic peaks

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | contrast | b | SE | df | p |
| ENV (z-scaled) | Intercept | 0.003 | 0.06 | 153 | 0.9597 |
|  | Max Deceleration | 0.029 | 0.007 | 153 | 1e-04 |
|  |  |  |  |  |  |
| F0 (z-scaled) | intercept | 0.512 | 0.086 | 151 | < .0001 |
|  | Arm vs. Wrist | -0.284 | 0.134 | 151 | 0.0349 |
|  | Max Deceleration | -0.001 | 0.015 | 151 | 0.9603 |
|  | Arm x Max Deceleration | 0.042 | 0.018 | 151 | 0.0205 |

# Discussion

In the current study, we demonstrated biomechanical effects of flexion-extension upper limb movements on speech by replicating in fluent speech effects obtained in steady-state vocalization and mono-syllabic utterances. We showed that rhythmically moving the wrist or arm affects vocalization acoustics by heightening F0 and amplitude envelope of speech vocalizations, as compared to a passive control and statistical control conditions. We finally show that higher deceleration rates observed within 200 milliseconds before the moment of the maximum extension of the arm movement materializes into more extreme acoustic peaks, demonstrating a role for acceleration and effector mass for gesture’s effect onto speech (i.e., an effect of physical impulse). Indeed, in all analyses, we observe that higher-mass arm versus wrist movements affect speech more clearly.  
Thus, stabilities in speaking may arise out of gesture-speech biomechanics in fluent speech as well as more simplified speech sounds. This does not mean that speech prosody necessarily requires gesture for reaching prosodic targets. Indeed, other sensorimotor solutions are available for modulating F0 and intensity (e.g., vocal-fold tensioning, respiratory actions; Perrier & Fuchs, 2015). Furthermore, F0 is across the board less, if at all, affected in line with our previous work (Pouw et al., n.d.), and work on the variable and often negligible role of respiratory actions in F0 modulations (Petrone, Fuchs, & Koenig, 2017). However, we think on the basis of present work we can argue that the biomechanical coupling of gesture and speech provides a ‘smart’ mechanism for ‘timing’ acoustic and movement expressions, and provides way in to the phylogenetic origin of gesture.  
We should wonder still whether the current effects of upper limb movement can be produced due to the attentional guidance to move (in the sense of “I must stop my wrist here and move up”), rather than the physical impulses produced by moving. In the previous studies we provided additional evidence with a respiration belt that tensioning around the trunk is involved in gesture-induced effects on vocal acoustics (Pouw et al. (n.d.)), or that postural stability moderates said effects (Pouw et al., 2019a). The additional evidential strength of these previous studies for gesture-speech physics lies in part in that a cognitive control account does a) not readily predict that trunk tensioning is involved in synchronizing upper limb movement and speech, and b) equally does not predict that standing or sitting matters for synchronizing two speech and gesture trajectories. It should be noted here, that one could in principle explain trunk tensioning and postural control effects with some new cognitive control account, but it does not seem parsimonious to do so in light of a gesture-speech physics alternative.  
This reasoning from parsimony extends to the basic kinematic-acoustic analysis of the current study, too. We should therefore ask in the current context: Does a cognitive control account predict that arm motion versus wrist motion should lead to heightened acoustic effects? Does a cognitive control account predict that acoustic peaks arise around the deceleration phase rather at the maximum extension phase? Does a cognitive control account predict that the degree to which a limb in motion decelerates scales with the acoustic peak that ensues? It is wholly possible that a particular cognitive control account may still account for all these effects, or more likely a subset of these effects. But to do so one needs to invoke some new hypothesis how this cognitive control system produced these observables. This comes at the cost of parsimony, as we are invoking new unobservable mechanisms to explain these observables, especially if there is a more parsimonious explanation available that explains these effects.  
To be clear, this does not mean that we can fully exclude cognitive control. Not in principle, but also more forcefully, not in degree. In fact, it is likely that complex interactions arise between on the one hand biophysical constraints arising out of moving your upper limb and vocalizing at the same time, and a speech system organizing meaningful speech in the context of those constraints. It is likely then, that in contrast to previous studies with non-meaningful speech, that there are interactions either of amplification or counteraction of gesture-speech physics that are likely bidirectional with lexical, syntactics, and prosodic speech organization. For example, one might speed up the occurrence of a physical impulse as then it will occur during a part of speech where there is a lexical stress. Or one might counteract an F0 effect of a physical impulse laryngeally, as its acoustic effect would lead to an inappropriate acoustic marker in the syntactic context of the sentence. These potential interactions between gesture-speech physics and meaningful speech organization must be studied in controlled experiments, but we do think they are likely to be there in the current context.  
Future research may thus consist of controlled experiments on syntactic, lexical and prosodic interactions with biophysical constraints. However, on the other hand more research is needed on truly spontaneous speech. In the current study, participants are retelling a cartoon, which is a very different context to speak from as compared to say a conversation. In part because there is some cognitive load of having to retell something accurately from recent memory, which too can have contextualizing effects.

## Wider implications

Gesture-speech physics holds promise for revising our understanding of the emergence of communicative gesture in anatomically modern humans, both ontogenetically and phylogenetically.  
It is well known that infants produce concurrent vocal-motor babblings. Furthermore, increased rhythmicity or frequency of motor babbling predicts speech-like maturation of vocalization (Ejiri, 1998; Ejiri & Masataka, 2001). Rather than a primarily neural development that instantiates gesture-speech synchrony (Iverson & Thelen, 2005), we suggest that during such vocal-motor babblings gesture-speech physics is discovered; this could provide the basis for infants to develop novel stable sensorimotor solutions for communication, such as a synchronized pointing gesture with a vocalization. Such sensorimotor solutions are, of course, likely solicited and practiced through support of caretakers, yet without the biomorphological scaffolding, gesture-speech synchrony would not get off the ground ontogenetically.  
Phylogenetic accounts have been central in discussions of the drivers of the depiction and referential function of gesture (Fröhlich, Sievers, Townsend, Gruber, & Schaik, 2019; Kendon, 2017; Tomasello, 2008). However, the current work supports the view that peripheral body movements may have served as a control parameter of an under-evolved vocal system. Previous work has proposed that the vocal system may have been evolutionarily exapted from rhythmic abilities in the locomotor domain (Larsson, Richter, & Ravignani, 2019; Ravignani et al., 2019), and viewing upper limb movements as constraints on the vocal system’s evolution fits neatly in such views. When our species became bipedal, the respiratory system was thereby liberated from upper-limb locomotary perturbations. We know that breathing (and vocalization) cycles often rigidly couple 1:1 with locomotion cycles in quadrupeds (Carrier et al., 1984), rigidly limiting what can be done (or communicated) in one breath. For example, vocalization acoustics of flying bats are synchronized with their wing beats (Lancaster, Henson, & Keating, 1995). Bipedalism, however, did not only free respiration from locomotion; it freed the upper limbs, too, allowing these highly skilled articulators to modulate a possibly less skilled respiratory-vocal system. Gestures, then, may have played a role in the complexification of the respiratory system in our species, which has been attributed to have occurred to serve speech evolution (MacLarnon & Hewitt, 1999).  
Thus, gesture-speech physics is not culturally specific, as animals can do it, too [e.g., bats; (Lancaster et al., 1995). It can further be related to other species, such as orangutangs who deepen their vocalizations by cupping their hands in front of their mouth (Hardus, Lameira, Schaik, & Wich, 2009). Other animals have been found to be sensitive to body-related information in sound in that body size and strength can be detected from vocalizations alone (Ghazanfar et al., 2007; Pisanski et al., 2016a), and humans are able to do this with some accuracy as well (Pisanski, Fraccaro, Tigue, O’Connor, & Feinberg, 2014), even when they are blind from birth (Pisanski et al., 2016b). In a recent experiment, we found that listeners are exquisitely sensitive to gesture-modulated acoustics: Listeners can synchronize their own upper limb movement by simply listening to a vocalizer producing a steady-state vocalization while rhythmically moving their wrist or arm (Pouw et al., 2019b, 2020). Thus, bodily dynamics can imprint the (human) voice, and this can be informative for listeners.  
To conclude, gesture-speech physics opens up the possibility that gesture may have evolved as a control parameter on vocal actions. This ecological revision (Kugler & Turvey, 1987; Turvey & Fonseca, 2014) of gesture-speech coupling provides a solid phylogenetic basis for an evolution of multimodal behavior, whereby peripheral bodily tensioning naturally formed coalitions with sound-producing organs that were still very much under development.

# References

Bernstein, N. (1967). *The Co-ordination and Regulations of Movements* ([1st English ed.] edition). Pergamon Press.

Carrier, D. R., Kapoor, A. K., Kimura, T., Nickels, M. K., Satwanti, Scott, E. C., … Trinkaus, E. (1984). The Energetic Paradox of Human Running and Hominid Evolution [and Comments and Reply]. *Current Anthropology*, *25*(4), 483–495. Retrieved from <http://www.jstor.org/stable/2742907>

Cordo, P. J., & Nashner, L. M. (1982). Properties of postural adjustments associated with rapid arm movements. *Journal of Neurophysiology*, *47*(2), 287–302. <https://doi.org/10.1152/jn.1982.47.2.287>

Cravotta, A., Busà, M. G., & Prieto, P. (2019). Effects of Encouraging the Use of Gestures on Speech. *Journal of Speech, Language, and Hearing Research*. <https://doi.org/10.21437/SpeechProsody.2018-42>

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

De Ruiter, J. P. (2017). The asymmetric redundancy of gesture and speech. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins Publishing Company.

Ejiri, K. (1998). Relationship between Rhythmic Behavior and Canonical Babbling in Infant Vocal Development. *Phonetica*, *55*(4), 226–237. <https://doi.org/10.1159/000028434>

Ejiri, K., & Masataka, N. (2001). Co-occurences of preverbal vocal behavior and motor action in early infancy. *Developmental Science*, *4*(1), 40–48. <https://doi.org/10.1111/1467-7687.00147>

Esteve-Gibert, N., & Guellai, B. (2018). Prosody in the Auditory and Visual Domains: A Developmental Perspective. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2018.00338>

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. <https://doi.org/10.1044/1092-4388(2012/12-0049)>

Feyereisen, P. (2017). *The Cognitive Psychology of Speech-Related Gesture*. New York: Routledge. Retrieved from <http://books.google.com?id=nJguDwAAQBAJ>

Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., & Schaik, C. P. van. (2019). Multimodal communication and language origins: Integrating gestures and vocalizations. *Biological Reviews*, *94*(5), 1809–1829. <https://doi.org/10.1111/brv.12535>

Ghazanfar, A. A., Turesson, H. K., Maier, J. X., van Dinther, R., Patterson, R. D., & Logothetis, N. K. (2007). Vocal-tract resonances as indexical cues in rhesus monkeys. *Current Biology*, *17*(5-2), 425–430. <https://doi.org/10.1016/j.cub.2007.01.029>

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., & Malik, J. (2019). Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3497–3506). Retrieved from <https://arxiv.org/abs/1906.04160>

Hardus, M. E., Lameira, A. R., Schaik, C. S., & Wich, S. A. (2009). Tool use in wild orang-utans modifies sound production: A functionally deceptive innovation? *Proceedings of the Royal Society B: Biological Sciences*, *276*(1673), 3689–3694. <https://doi.org/10.1098/rspb.2009.1027>

Hastie, T. (2019). Gam: Generalized Additive Models (Version 1.16.1). Retrieved from <https://CRAN.R-project.org/package=gam>

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

Hodges, P. W., & Richardson, C. A. (1997). Feedforward contraction of transversus abdominis is not influenced by the direction of arm movement. *Experimental Brain Research*, *114*(2), 362–370. <https://doi.org/10.1007/pl00005644>

Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, *23*(8), 639–652. <https://doi.org/10.1016/j.tics.2019.05.006>

Im, S., & Baumann, S. (2020). Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America*, *5*(1, 1), 685–697. <https://doi.org/10.3765/plsa.v5i1.4755>

Iverson, J. M., & Thelen, E. (2005). Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, 22.

Kelso, J. A., & Tuller, B. (1984). Converging evidence in support of common dynamical principles for speech and movement coordination. *The American Journal of Physiology*, *246*(6 Pt 2), R928–935. <https://doi.org/10.1152/ajpregu.1984.246.6.R928>

Kendon, A. (2017). Reflections on the “gesture-first” hypothesis of language origins. *Psychonomic Bulletin & Review*, *24*(1), 163–170. <https://doi.org/10.3758/s13423-016-1117-3>

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. <https://doi.org/10.1016/j.jml.2007.06.005>

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

Kucherenko, T., Hasegawa, D., Henter, G. E., Kaneko, N., & Kjellström, H. (2019). Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents - IVA ’19* (pp. 97–104). Paris, France: ACM Press. <https://doi.org/10.1145/3308532.3329472>

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Lawrence Erlbaum Associates, Inc.

Lancaster, W. C., Henson, O. W., & Keating, A. W. (1995). Respiratory muscle activity in relation to vocalization in flying bats. *Journal of Experimental Biology*, *198*(1), 175–191. Retrieved from <https://jeb.biologists.org/content/198/1/175>

Larsson, M., Richter, J., & Ravignani, A. (2019). Bipedal Steps in the Development of Rhythmic Behavior in Humans. *Music & Science*, *2*, 2059204319892617. <https://doi.org/10.1177/2059204319892617>

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

Levin, S. M. (2006). Tensegrity: The new biomechanics. In M. Hutson & R. Ellis (Eds.), *Textbook of muscularskeletal medicine* (pp. 69–80). Oxford, England: Oxford University Press.

Lin, C.-Y., & Rathcke, T. (2020). How to hit that beat: Testing acoustic anchors of rhythmic movement with speech. In *10th International Conference on Speech Prosody 2020* (pp. 1–5). ISCA. <https://doi.org/10.21437/SpeechProsody.2020-1>

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

MacLarnon, A. M., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology*, *109*(3), 341–363. [https://doi.org/10.1002/(SICI)1096-8644(199907)109:3<341::AID-AJPA5>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1096-8644(199907)109:3%3C341::AID-AJPA5%3E3.0.CO;2-2)

McClave, E. (1998). Pitch and Manual Gestures. *Journal of Psycholinguistic Research*, *27*(2), 69–89. <https://doi.org/https://doi.org/10.1023/A:1023274823974>

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University Of Chicago Press.

Mendoza-Denton, N., & Jannedy, S. (2011). Semiotic Layering through Gesture and Intonation: A Case Study of Complementary and Supplementary Multimodality in Political Speech. *Journal of English Linguistics*, *39*(3), 265–299. <https://doi.org/10.1177/0075424211405941>

Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, *42*, 1–11. <https://doi.org/10.1016/j.wocn.2013.11.002>

Perrier, P., & Fuchs, S. (2015). Motor Equivalence in Speech Production. In *The Handbook of Speech Production* (pp. 223–247). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118584156.ch11>

Petrone, C., Fuchs, S., & Koenig, L. L. (2017). Relations among subglottal pressure, breathing, and acoustic parameters of sentence-level prominence in German. *The Journal of the Acoustical Society of America*, *141*(3), 1715–1725. <https://doi.org/10.1121/1.4976073>

Pikovsky, A., Kurths, J., & Rosenblum, M. (2001). *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge, Mass: Cambridge University Press. Retrieved from <https://books.google.nl/books?hl=en&lr=&id=FuIv845q3QUC&oi=fnd&pg=PP1&dq=pikovsky+synchronization&ots=RM-tIgHgU7&sig=PHOyMBVgtLIao8rdkP5bCTVKwo0#v=onepage&q=pikovsky%20synchronization&f=false>

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Team, R. C. (2019). *Nlme: Linear and nonlinear mixed effects models*.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016a). Voice modulation: A window into the origins of guman vocal control? *Trends in Cognitive Sciences*, *20*(4), 304–318. <https://doi.org/10.1016/j.tics.2016.01.002>

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O’Connor, J. J. M., & Feinberg, D. R. (2014). Return to Oz: Voice pitch facilitates assessments of men’s body size. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(4), 1316–1331. <https://doi.org/10.1037/a0036956>

Pisanski, K., Oleszkiewicz, A., & Sorokowska, A. (2016b). Can blind persons accurately assess body size from the voice? *Biology Letters*, *12*(4), 20160063. <https://doi.org/10.1098/rsbl.2016.0063>

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

Pouw, W., & Dixon, J. A. (2019b). Quantifying gesture-speech synchrony. In *Proceedings of the 6th meeting of Gesture and Speech in Interaction* (pp. 68–74). Paderborn: Universitaetsbibliothek Paderborn. <https://doi.org/10.17619/UNIPB/1-812>

Pouw, W., Harrison, S. H., & Dixon, J. A. (2019a). Gesture-speech physics: The biomechanical basis of the emergence of gesture-speech synchrony. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0000646>

Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (n.d.). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*. <https://doi.org/10.1121/10.0001730>

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2019b). Acoustic specification of upper limb movement in voicing. In *Proceedings of the 6th meeting of Gesture and Speech in Interaction* (pp. 75–80). Paderborn, Germany: Universitaetsbibliothek Paderborn. <https://doi.org/10.17619/UNIPB/1-812>

Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences*, *117*(12), 11364–11367. <https://doi.org/10.1073/pnas.2004163117>

Pouw, W., & Trujillo, J. P. (2019). *Materials Tutorial Gespin2019 - Using video-based motion tracking to quantify speech-gesture synchrony*. Retrieved from <10.17605/OSF.IO/RXB8J>

Profeta, V. L. S., & Turvey, M. T. (2018). Bernstein’s levels of movement construction: A contemporary perspective. *Human Movement Science*, *57*, 111–133. <https://doi.org/10.1016/j.humov.2017.11.013>

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

Rochet-Capellan, A., Laboissière, R., Galván, A., & Jean-Luc, S. (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. <https://doi.org/10.1044/1092-4388(2008/07-0173)>

Rosch, A., & Schmidbauer, H. (2014). WaveletComp 1.1: A guided tour through the R package, 59.

Ruiter, J. P. de. (2000, August). The production of gesture and speech. <https://doi.org/10.1017/CBO9780511620850.018>

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

Shattuck-Hufnagel, S., & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, *9*(1514). <https://doi.org/https://doi.org/10.3389/fpsyg.2018.01514>

Silva, P., Moreno, M., Mancini, M., Fonseca, S., & Turvey, M. T. (2007). Steady-state stress at one hand magnifies the amplitude, stiffness, and non-linearity of oscillatory behavior at the other hand. *Neuroscience Letters*, *429*(1), 64–68. <https://doi.org/10.1016/j.neulet.2007.09.066>

Stoltmann, K., & Fuchs, S. (2017). Syllable-pointing gesture coordination in Polish counting out rhymes: The effect of speech rate. *Journal of Multimodal Communication Studies*, *4*(1-2), 63–68.

Streeck, J. (2008). Depicting by gesture. *Gesture*, *8*(3), 285–301. <https://doi.org/10.1075/gest.8.3.02str>

Tomasello, M. (2008). *The origins of human communication*. Cambdride, MA: MIT press.

Treffner, P., & Peter, M. (2002). Intentional and attentional dynamics of speech–hand coordination. *Human Movement Science*, *21*(5-6), 641–697. <https://doi.org/10.1016/S0167-9457(02)00178-1>

Turvey, M. T. (1990). Coordination. *American Psychologist*, *45*(8), 938–953. <https://doi.org/10.1037/0003-066X.45.8.938>

Turvey, M. T., & Fonseca, S. T. (2014). The Medium of Haptic Perception: A Tensegrity Hypothesis. *Journal of Motor Behavior*, *46*(3), 143–187. <https://doi.org/10.1080/00222895.2013.798252>

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

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. <https://doi.org/10.1016/j.wocn.2018.03.002>

Winkelmann, R., Bombien, L., & Scheffers, M. (2018). Wrassp: Interface to the ’ASSP’ Library (Version 0.1.8). Retrieved from <https://CRAN.R-project.org/package=wrassp>

Zelic, G., Kim, J., & Davis, C. (2015). Articulatory constraints on spontaneous entrainment between speech and manual gesture. *Human Movement Science*, *42*, 232–245. <https://doi.org/10.1016/j.humov.2015.05.009>