American Psychological Association

Journal of Experimental Psychology: General

0096-3445/20/$12.00

2020, Vol. 149, No. 2, 391–404
http://dx.doi.org/10.1037/xge0000646

BRIEF REPORT

Gesture–Speech Physics: The Biomechanical Basis for the Emergence of Gesture–Speech Synchrony

Wim Pouw
University of Connecticut and Erasmus University Rotterdam

Steven J. Harrison and James A. Dixon
University of Connecticut

The phenomenon of gesture–speech synchrony involves tight coupling of prosodic contrasts in gesture movement (e.g., peak velocity) and speech (e.g., peaks in fundamental frequency; F0). Gesture–speech synchrony has been understood as completely governed by sophisticated neural-cognitive mechanisms. However, gesture–speech synchrony may have its original basis in the resonating forces that travel through the body. In the current preregistered study, movements with high physical impact affected phonation in line with gesture–speech synchrony as observed in natural contexts. Rhythmic beating of the arms entrained phonation acoustics (F0 and the amplitude envelope). Such effects were absent for a condition with low-impetus movements (wrist movements) and a condition without movement. Further, movement–phonation synchrony was more pronounced when participants were standing as opposed to sitting, indicating a mediating role for postural stability. We conclude that gesture–speech synchrony has a biomechanical basis, which will have implications for our cognitive, ontogenetic, and phylogenetic understanding of multimodal language.

Keywords: gesture–speech synchrony, upper limb movement, phonation acoustics, motion tracking, multimodal language

Hand gesture and speech acoustics tightly synchronize their activity (see for an example https://osf.io/29h8z/; Krivokapić, Tiede, & Tyrone, 2017; Leonard & Cummins, 2011; Shattuck-Hufnagel & Ren, 2018; for an overview see Wagner, Malisz, & Kopp, 2014). Specifically, speech prosodic contrasts, which are communicatively meaningful fluctuations in speech acoustics, structurally align with energetic contrasts in gesture. For example prosody as captured by peaks in the fundamental frequency of speech (F0; perceived as the "pitch" of speech) are often found to align with peak velocity or the point of maximum effort (Danner, Barbosa, & Goldstein, 2018; Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2011; Loehr, 2004; McClave, 1998; Pouw & Dixon, 2019).

The synchronized modulations of meaningful fluctuations in acoustics or kinematics of speech or gesture have further been found to be bidirectionally coupled (e.g., Parrell, Goldstein, Lee, & Byrd, 2014; see also Kelso, Tuller, & Harris, 1983; Zelic, Kim, & Davis, 2015). For example, Parrell and colleagues (2014) found that during a repetitive speaking-and-finger-tapping synchrony task, when a syllable or a tapping movement is given emphatic stress by lengthening the syllable duration or modulating finger-movement magnitude, such modulation of movement/tonal activity will spill over to the other modality in compulsory fashion (e.g., when stressing a syllable it will affect movement amplitude of the tapping movement).

Krahmer and Swerts (2007; Experiment 1) assessed whether hand gestures, head nods, or eyebrow raises have compulsory effects on concomitant speech acoustics. Movements were produced during either a part of the sentence that was also intended to be produced with a pitch accent (marker of prosodic stress), or during a different part of the sentence where there was no pitch accent intended. It was found that any type of gestural movement invoked two markers of pitch-accented speech, namely increased duration of phonation and a lower frequency for the second for-

This article was published Online First August 1, 2019.

Wim Pouw, Center for the Ecological Study of Perception and Action, University of Connecticut, and Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam; Steven J. Harrison and James A. Dixon, Center for the Ecological Study of Perception and Action, University of Connecticut.

Steven J. Harrison is now at the Department of Kinesiology, University of Connecticut.

Wim Pouw, Steven J. Harrison, and James A. Dixon have designed the study. Wim Pouw conducted the analyses with the supervision of Steven J. Harrison and James A. Dixon. Wim Pouw wrote the current report, as well as the preregistration, with critical revisions by Steven J. Harrison and James A. Dixon. This study has been preregistered on the Open Science Framework (OSF: https://osf.io/5aydk/) and the raw anonymized quantitative data, and analyses scripts supporting this confirmatory study are also available on the OSF. Note that parts of the current article may overlap verbatim with the preregistration. This research has been funded by the Netherlands Organization of Scientific Research (NWO; Rubicon grant "Acting on Enacted Kinematics," Grant 446-16-012; PI: Wim Pouw).

Correspondence concerning this article should be addressed to Wim Pouw, Center for the Ecological Study of Perception and Action, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020. E-mail: wimpouw@uconn.edu

mant (lower F2). These effects were found regardless of whether a pitch accent was actually intended, and the effects were similar to when a pitch accent was intended but was not accompanied by movement. These effects arose regardless of movement type, which suggests that making any burst-like body movement during speech affects speech acoustics.

Despite the pervasive phenomenon of gesture–speech coupling, a unifying theory of why these systems bind together is lacking. Extant explanations are varied (Wagner et al., 2014), and include hypotheses about communicative functions (e.g., Krauss, Chen, & Gottesman, 2000), and cognitive functions (Pouw & Dixon, 2019; Rusiewicz & Esteve-Gibert, 2018). Additionally, it has been suggested that the hand and mouth are naturally solicited to interact when bringing food to the mouth, readying opportunities for modulation of hand–mouth synchrony during social interactions (Iverson & Thelen, 1999; see also Esteve-Gibert & Guellaï, 2018).

However, all previous literature on this matter has assumed that gesture–speech synchrony must emerge from the acquisition of neural-cognitive faculties, and is completely constrained by said faculties (de Ruiter, 2000; Wagner et al., 2014). As McClave (1998, p. 96) summarized: "[coordinating] pitch and manual gesture movements is an option available to speakers, but it is not biologically mandated." In the current study, we show that there is fundamentally a biophysical basis for gesture–speech synchrony.

Gesture–Speech Synchrony and Its Medium

The fundamental frequency (F0) of speech is determined by the alveolar/subglottal (lung) air pressure and larynx muscle tonus (Lieberman, 1996). Increasing the alveolar pressure will produce more acoustic energy in the form of amplitude and will produce an increased fundamental frequency (i.e., perceived as a higher pitch). Expiratory flow is a key modulator of acoustic energy for speech. This energy for expiration is primarily delivered by the elastic recoil in the lungs.

Of further importance is that the body allows—and is dependent in its functioning on—forces that resonate through its myofascial-skeletal network (Turvey & Fonseca, 2014). Given the sensitive role of expiration-related muscles and alveolar pressure in speech, it is possible therefore that gesture movements could affect expiration-related muscles, and therefore could affect prosodic metrics of speech directly (e.g., contrasts in F0; changes in amplitude).

First, when hand gestures are performed various muscles are recruited in an anticipatory fashion, with muscle activations occurring between 100 ms before and 50 ms after onset of the limb movements so as to maintain the stability of body posture (e.g., Aruin & Latash, 1995; Bouisset & Do, 2008). In the case of arm movement, these "anticipatory postural adjustments" mobilize an interconnected set of muscles including those around the trunk (Hodges & Richardson, 1997). One of the key anticipatory postural adjustment muscles recruited for arm movements is the Rectus Abdominus (RA; i.e., "the abs"; Aruin & Latash, 1995). The trunk muscles that are recruited for anticipated postural adjustments (including the Rector Abdominus) are directly involved in the active phase of expiration (Hodges, Gandevia, & Richardson, 1997), which is the phase during which we produce speech. Anticipatory postural adjustments produce balancing reactive forces that are counteractive to the forces produced by the kinetic perturbations of moving the arms. Moving the arms faster produces more destabilizing forces and will need to be met with an equally more forceful anticipatory postural adjustment. It is finally important to note that contrary to common wisdom, the forces produced by limb movements themselves (as well as anticipatory postural adjustments) are not localized to the limb (Silva, Moreno, Mancini, Fonseca, & Turvey, 2007; for an overview see Turvey & Fonseca, 2014). Any type of muscle contraction will produce forces that travel throughout the body, and such traveling forces are essential in the effective coordination of movement that involves a synergy of components (i.e., any intentional action; Bernstein, 1966).

Now that we have established a potential route through which gestures can affect speech directly, we must wonder whether gestures really produce nontrivial forces, and whether such forces are a viable source of physical coupling. That gesture-related forces are nontrivial is indicated by the experiences of Ian Waterman, a person suffering from almost complete proprioceptive loss, who reported a need to suppress his gestures in initial stages of his disease because he was afraid of falling from the destabilizing effects of these articulations (Gallagher, 2005; McNeill, 2008). Furthermore, the forces that gestures produce and their coupling with speech prosody already seem to be entailed by the classic gesture categories that are used in gesture studies (McNeill, 2008). Namely, a common type of gesture that is identified as having the sole function of synchronizing with prosodic contrasts in speech are called *beat* or *baton* gestures (McNeill, 2008; Kendon, 2004). Such beat gestures are characterized by burst-like vertical arm movements that "beat" with the rhythm of speech (Leonard & Cummins, 2011). Importantly, beat gestures seem to possess greater physical momentum as compared to other types of gestures as they produce sudden halts (beats), and therefore possess greater potential for momentum transfers to the body. Thus, beat gestures might synchronize with speech the way they do, because they are recruited in order to produce a physical impulse on the body. Beats are classically distinguished from other gestures such as iconic or metaphoric gestures (McNeill, 2008). Compared to beat gestures, iconic and metaphoric gestures have more complex and often more fluid movement trajectories as they need to iconically present meaning. Importantly, however, although such gestures are often more variably aligned with speech acoustics, it is still often the case that iconic gestures might still have moments of emphasis wherein the perceived moment of maximum effort closely coincides with prosodic peaks in speech (Prieto, Cravotta, Kushch, Rohrer, & Vilà-Giménez, 2018; Wagner et al., 2014). The current study is about the beat-like aspects of gesture, which are necessarily present in beat gestures, but are also often present in other types of gestures such as iconic gestures (see, e.g., https://osf.io/29h8z/).

Current Study

With the current preregistered study (see https://osf.io/5aydk/), we aim to replicate and extend earlier exploratory findings (Pouw, Harrison, & Dixon, 2018), which indicated that upper limb movements with high physical impetus (one-arm beat and two-arm beat) were synchronized with peaks in F0 and the amplitude envelope during a phonation task.[1] Here we investigate the role of physical

[1] The reader can listen to audio examples of the trials together with a visual presentation of the amplitude envelope, F0, and vertical hand movement (Z movement) for the exploratory data here: https://osf.io/acmdg/.

impetus on phonation with a larger sample (10 participants; 240 trials) and assess the possible modulatory role of *postural stability*. Participants phonated the vowel [ə] (as in cinem**a**) at a steady pitch during several trials of a passive condition, wrist beat condition, one-arm beat condition, and two-arm beat condition. Participants performed these movements while sitting in a chair (sitting condition) or while standing upright (standing condition). We added this condition as it has been shown that anticipatory postural adjustments that arise when moving the upper limbs while standing are dramatically diminished when the body is in a more stable seated posture (Cordo & Nashner, 1982). If anticipatory postural adjustments are modulating effects of gestures with high physical impetus on phonation, then upper-limb-motion effects on phonation would be absent or diminished in the sitting condition relative to upper limb movement effects on phonation in the standing condition.

Method

Design

The current experiment consists of a two-factor within-subject design, with one within-subject factor (movement condition) of four levels (passive vs. wrist beat vs. one-arm beat vs. two-arm beat), and another within-subject factor (posture condition) with two levels (sitting vs. standing). Ten undergraduate students from the University of Connecticut (5 females and 5 males; 8 right-handed; M_{age} = 19.2, SD_{age} = 1.25) were asked to produce a steady voiced output of the vowel *a*: (as in cinem**a**, [ə]). The current sample size was considered appropriate as we obtained clear effects on an individual level in our exploratory study. Participants were asked to stop phonating as soon as they felt that they ran out of air and could not maintain their preferred level of pitch. For each participant we planned to perform three blocks of eight trials (total = 240 trials = 10 participants × 3 Blocks × 4 Movement Condition × 2 Posture Condition). A total of 239 trials were actually performed.[2]

The current study was approved by the IRB review board of the University of Connecticut (IRB approval #H18-176).

Procedure

Each trial consisted of the participant taking a full breath and phonating until participants felt they could not maintain their level of pitch. Participants were explicitly instructed to keep phonating at a steady state across and within trials, keeping any changes in pitch or loudness at a minimum. For half of the trials, participants were asked to sit on a chair with their feet firmly on the ground and their backs touching the backrest. The chair did not have armrests. In the other half of the trials, participants were asked to stand upright. For the passive condition, participants were asked to let their hands rest alongside their bodies during phonating when standing, or rest on their lap when sitting. For the one-arm beat condition, participants were asked to continuously move their dominant hand on the sagittal plane by lifting the hand up (via a lower arm movement) and letting it drop with a sudden complete halt (i.e., with energetic contrast, a "beat"). The beat was reached around the point where the elbow flexion angle was about 90 degrees. In the two-arm beat condition, participants made the same movement in-phase with two arms. In the wrist beat condition, participants were asked to move their dominant hand in a beat like fashion with a wrist movement (and no forearm or upper arm movements). Order of condition was randomized for each block of eight trials.

A crucial change from the exploratory study is that in the current experiment we guided the movement frequency of the participant by a visual presentation. Instead of participants moving at their own preferred frequency, participants were encouraged to move their hands at 80 beats per minute (i.e., 1.3 Hz; oscillation period = 0.77 s). This allowed us to analyze the data with a focus on a particular frequency range without having to account for individual differences in preferred moving rate. We programmed in C++ a visual presentation that takes input from the motion tracker so as to visually represent the frequency of the vertical movement to the participant. The visual presentation consisted of a bar that changed size as a function of movement frequency; participants tried to keep the size of the bar within a certain range as specified by two guide bars. The size of the guide bars corresponded to 10% faster or slower than 80 BPM. For a graphical representation of the wrist and arm movement as well as the movement guiding bars see https://osf.io/wvhyg/.

Apparatus: Motion and Audio Recording

We used a Polhemus Liberty to record movement (240 Hz), with a sensor attached to the tip of dominant hand's index finger. Since hand movements were primarily in the vertical dimension, we analyzed movement–phonation coordination and computed derivatives (i.e., velocity, acceleration, jerk) only for Z-axis movement. For derivative estimation, we applied a low-pass Butterworth filter of 33 Hz. We recorded audio using a RT20 Audio Technica Cardioid microphone (44.1 kHz).[3] We modified a C++ script made publicly available by Michael Richardson (Richardson, 2009), so as to simultaneously call and write movement as well as audio data. We modified this script to enable simultaneous recording of sound from a microphone, using toolbox SFML for C++ (https://www.sfml-dev.org/). Using a custom-made script in R (R Core Team, 2013), the data from PRAAT (Boersma, 2001) and the motion-tracking data were aggregated (code available on https://osf.io/a9hw7/).

Phonation Variables

As stated in the preregistration, our analyses focus on F0 and the amplitude envelope. These acoustic properties are key metrics for tracking prosody of speech. F0 and amplitude time series were sampled at the sampling rate of the motion-tracker (240 Hz: 1 sample per 4.16 ms).

[2] One trial was not performed because the experimenter accidentally skipped a trial. For this participant there will be one less observation, which will be statistically accounted for in terms of confidence estimates by the linear mixed regression models we employ (the types of analyses we employ do not require equal amount of observations per condition).

[3] The experimenter tried to keep participant–microphone distance equal across trials (about 2 inches). We think however there will be very slight variability in microphone–participant distance and this will change overall intensity estimates. Note that we are, however, working with a participant-scaled amplitude envelope, therefore between-subject differences in microphone distances are neutralized. Furthermore, in the current study we study temporal intensity changes as a function of moments of movement. Therefore slight overall higher or lower intensity between trials are not detrimental to our temporal analyses that are performed within trials. Finally, note that F0 should not be affected by slight changes in microphone-participant distance.

Fundamental frequency (F0; pitch). F0 time series was extracted from the audio using PRAAT (Boersma, 2001) with a range suitable for male (75–500 Hz) or female (100–500 Hz) voice range.

Amplitude envelope (ENV). A raw speech signal has both fine and gross structural changes, that is, higher and lower frequency fluctuations. The lower frequency fluctuations are important for the rhythmic structure of speech (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009) and can be captured by the Amplitude Envelope (ENV). ENV can be reconstructed from the raw audio signal using the Hilbert transform (He & Dellwo, 2017). The amplitude envelope (ENV) time series were produced by applying the PRAAT script by He and Dellwo (2017) to each audio waveform (one audio file for each block with 8 trials). ENV is scaled in Hilbert Units ranging from 0 to 1. Thus each sound recording from a participant and block is scaled from 0 (*minimum amplitude*) to 1 (*maximum amplitude*). See Figure 1 for an example of the amplitude envelope metric. For the main analysis R script see https://osf.io/q2kx8/.

Results

Descriptives

Descriptives for the fundamental frequency (F0) and the amplitude envelope (ENV) computed for each trial, and averaged across trials per condition, are provided in Table 1. Examination of Table 1 shows that phonation was less stable for the one-arm and two-arm beat trials, as standard deviations are markedly larger (especially for F0). Indeed, when listening to the audio samples it was apparent that phonation was less stable during the arm-movement conditions. The average time for each phonation trial was 8.71 s (SD = 2.69 s), with average duration for passive = 8.48s (SD = 2.65), wrist beat = 8.90 s (SD = 2.93), one-arm beat = 8.72s (SD = 2.85), two-arm beat = 8.74s (SD = 2.62), sitting = 8.58 s (SD = 2.62), and standing = 8.83s (SD = 2.76). Trial time was correlated with decreases in F0, $r = -.199$, $p < .001$ and ENV, $r = -.134$, $p < .001$, which indicates that the ability to maintain acoustic energy levels decreased as participants reached the end of their breaths. To prevent spurious effects of non-stationarity in our time series analyses, we linearly detrended the effect of time for each trial before entering into the analyses. Not surprisingly, ENV and F0 (standardized for each trial) were weakly positively correlated (average r = .22, average $p < .017$).

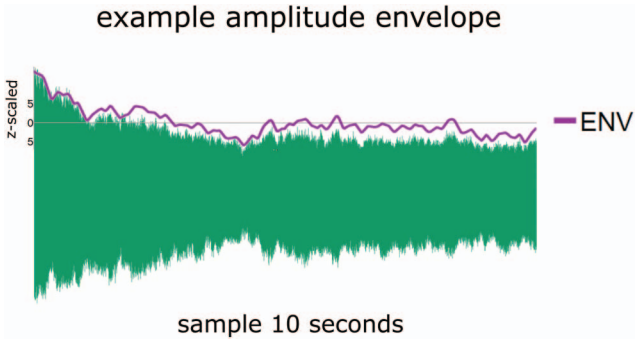

Figure 1. Example amplitude envelope. Standardized amplitude envelope (240 Hz sampling) for a sample of 10 s of phonation for the current data (trial = standing one-arm beat condition). Essentially the amplitude envelope tracks gross fluctuations of the raw audio waveform. See the online article for the color version of this figure.

Table 1
Mean and Standard Deviation of F0 and ENV per Condition

Phonation variable	Mean (average *SD*)			
	Passive	Wrist beat	One-arm beat	Two-arm beat
F0				
TOTAL	178.82 ***(1.81)***	178.82 ***(2.16)***	178.13 ***(2.67)***	176.86 ***(3.12)***
Sitting	180.89 ***(1.63)***	177.47 ***(2.45)***	178.35 ***(2.67)***	178.80 ***(3.33)***
Standing	176.47 ***(1.62)***	176.26 ***(1.88)***	179.28 ***(2.66)***	177.43 ***(2.91)***
ENV				
TOTAL	0.257 ***(.047)***	0.250 ***(0.047)***	0.251 ***(0.048)***	0.272 ***(0.056)***
Sitting	0.212 ***(.042)***	0.199 ***(0.035)***	0.191 ***(0.038)***	0.220 ***(0.045)***
Standing	0.297 ***(.051)***	0.301 ***(0.057)***	0.309 ***(0.056)***	0.325 ***(0.066)***

Note. Average standard deviations are given in boldface and italics. F0 is given in Hertz. Amplitude envelope (ENV) is given in Hilbert Units (range = 0–1). The "averaged standard deviations" are computed relative to the trial-mean, then averaged for all trials. As such the standard deviations reported here are not biased due to prominent between-subject differences in F0 between Males vs. Females.

To provide an insight into the kinematics of the current movements, we computed the maximum velocity produced during the extension phase (i.e., negative movement direction) of the vertical movement. This corresponds to the maximum velocity of the downbeat. The average maximum negative velocity is given in Table 2 for each movement condition. We also provide the average maximum vertical amplitude of the movement to provide an indication of how large the different movements were.

Exclusions for analyses. After inspection, we found for several trials that for two participants (Participants 7 and 8), PRAAT could not reliably track continuous F0 while participants were in fact phonating continuously, and also showed noisy periodicity estimates of F0 traces. We will not include these participants in our analyses for fundamental frequency.[4] The amplitude envelope could be reliably tracked and showed no anomalies.

Time series descriptives. Figure 2 shows an example of the time series for the first participant (female) of the current dataset (for 1 block).

To further summarize the (shared) periodic structure of the time series, we performed a spectral decomposition analysis with R package "spectral" (Seilmayer, 2016). This analysis used the Fast Fourier Transform (FFT) to assess periodicities in movement and phonation (see Figure 3). For the movement conditions, we expected to find periodicities in vertical movement time series around the target range of 1.3 Hz (80 BPM, period = 0.77 s). If movements are entraining phonation (as Figure 2 indicates), we would also expect to observe dominant periodicities around 1.3 Hz for the F0 and ENV time series. Visual inspection of Figure 3 indeed indicates that there are likely to be shared periodicities of movement with phonation (F0 and ENV), suggesting that high-impact movements are structurally affecting phonation.

[4] The failed F0 tracking was most likely due to too much distance between participants and the microphone (leading to a faint signal). Note that this decision does not affect our conclusions in any way; we have run the analyses with these noisy data included (and this did not affect our main conclusions).

Table 2
Descriptives Kinematics

Kinematic variable	Mean (*SD*)		
	Wrist beat	One-arm beat	Two-arm beat
Average max negative velocity			
Sitting	−75 cm/s *(24.59)*	−138.88 cm/s *(42.71)*	−141.97 cm/s *(35.70)*
Standing	−98.55 cm/s *(23.10)*	−172.28 cm/s *(31.41)*	−166.93 cm/s *(22.26)*
Total average max negative velocity	−87.21 cm/s *(26.28)*	−155.87 cm/s *(40.71)*	−154.24 cm/s *(75.86)*
Average max vertical amplitude			
Sitting	11.18 cm *(3.78)*	19.52 cm *(5.02)*	20.34 cm *(4.43)*
Standing	14.11 cm *(2.51)*	23.89 cm *(3.10)*	23.73 cm *(3.10)*
Total average max vertical amplitude	12.65 cm *(3.50)*	21.74 cm *(4.46)*	22.00 cm *(4.17)*

Note. Standard deviations (*SD*) are provided in italics. Velocity is given in centimeters per second.

Confirmatory Analyses

Coherence. To formally test whether the periodicities of movement (vertical movement; or *Z*) and phonating (F0 and ENV) were correlated, we computed coherence between the different spectral density distributions (R package "seewave"; Sueur, Aubin, & Simonis, 2008). Coherence is a measure that provides a correlation strength of the periodicities, ranging from 0 (*no correlation*) to 1 (*perfect correlation*) across a frequency range. Figure 4 provides an overview for the mean coherence per condition between (a) movement and ENV, and (b) movement and F0. Examination of Figure 4 suggests coherence levels are increased at and around 1.3 Hz for the two-arm beat and the one-arm beat conditions, for both sitting and standing conditions. The effect of movement on F0 in the standing condition seems more pronounced as compared to the sitting condition.

To directly test differences in coherence levels we performed mixed regression modeling (R package nlme: participants as random intercept) to predict coherence levels as a function of movement condition, posture condition, and the interaction of movement and posture condition; as stated in our preregistration. Coherence was assessed in a frequency band around 1.3 Hz (0.8–1.8 Hz range).

Coherence between movement and F0 was reliably predicted by movement condition, as compared to a base model that predicted the overall mean (change in $\chi^2[6] = 16.88$, $p < .001$). Adding posture condition to the model, further improved predictions for coherence as compared to the previous model, $\chi^2[7] = 6.60$, $p < .010$. Adding an interaction of movement condition × posture condition did not, however, lead to further improvement of the model, $\chi^2[10] = 1.51$, $p = .679$. The best fitting model with movement and posture condition as predictors showed that the wrist beat condition did not reliably differ from the passive condition ($b = .005$, $t(1683) = 0.360$, $p = .718$). The one-arm beat condition ($b = .038$, $t(1683) = 2.43$, $p = .015$), and the two-arm beat condition ($b = .055$, $t(1683) = 3.50$, $p < .001$), did reliably differ from the passive condition. Furthermore, we found that the standing condition resulted in increased coherence as compared to the sitting condition ($b = .028$, $t(1683) = 2.57$, $p = .010$). The one-arm beat ($b = .032$, $t(1683) = 2.12$, $p = .034$) and the two-arm beat conditions ($b = .049$, $t(1683) = 3.21$, $p = .002$) also had reliably higher coherences as compared to the wrist beat condition. These analyses show that fundamental frequency of phonation was entraining to movements with high physical impetus (one-arm beat condition and two-arm beat condition), and that standing resulted in increased coherence suggesting that postural stability is a contributing (but not a necessary) factor.

Coherence between movement and ENV was reliably predicted by movement condition, as compared to a base model (change in $\chi^2[6] = 36.43$, $p < .001$). Adding posture condition to the model did not improve model fit (change in $\chi^2[7] = 2.51$, $p = .112$), nor did adding the interaction between movement and posture condition (change in $\chi^2[10] = 2.64$, $p = .449$). The best fitting model with movement condition as a predictor revealed that the one-arm beat condition ($b = .075$, $t(2027) = 4.76$, $p < .001$) and the two-arm beat condition ($b = .074$, $t(2027) = 4.77$, $p < .001$) had reliably higher coherence levels than the passive condition. The wrist beat condition did not differ in coherence as compared to the passive condition ($b = .018$, $t(2027) = 1.13$, $p = .256$). The one-arm beat ($b = .057$, $t(2027) = 3.70$, $p < .001$) and the two-arm beat conditions ($b = .057$, $t(2027) = 3.62$, $p < .001$) also had significantly higher coherences as compared to the wrist beat condition. These analyses reveal that arm movement with high physical impetus were also entraining the amplitude envelope of phonation.[5]

Temporal dynamics. From the previous analyses we know that upper limb movements with high physical impetus (one-arm and two-arm beats) are structurally entraining phonation. The next question is how movement and phonation are locking their phases. From the time-series example in Figure 2, it can be observed that acoustic peaks are observed around the moment when the vertical movement reaches its maximum extension. To assess whether vertical movement is indeed related to phonation in this antiphase

[5] Exploratory analyses individual differences: As requested by one of the reviewers, we checked whether there were any effects of gender in the current confirmatory analyses. However, no reliable predictive value of gender was detected over and above the most predictive models of coherence between F0 and Z and coherence between ENV and Z. Recall further that effects in the current paradigm are at times very much audible (e.g., https://osf.io/acmdg/) suggesting that the effect can be manifested on the individual participant level and our analyses are able to detect such effects (Pouw, Harrison, & Dixon, 2018). However, there are nevertheless some individual differences in the current coherence effect estimates for individual participants. We performed participant-level analyses to get coherence effect estimates of the movement conditions relative to the passive condition (while controlling for fixed effects of posture condition). A graphical overview individual-level effect estimates on movement–phonation coherence relative to the passive condition can be found here: https://osf.io/xjw5p/. Despite variability it can be seen that seven out of 10 participants show increased coherence of two-arm beat condition as compared to the wrist beat condition.

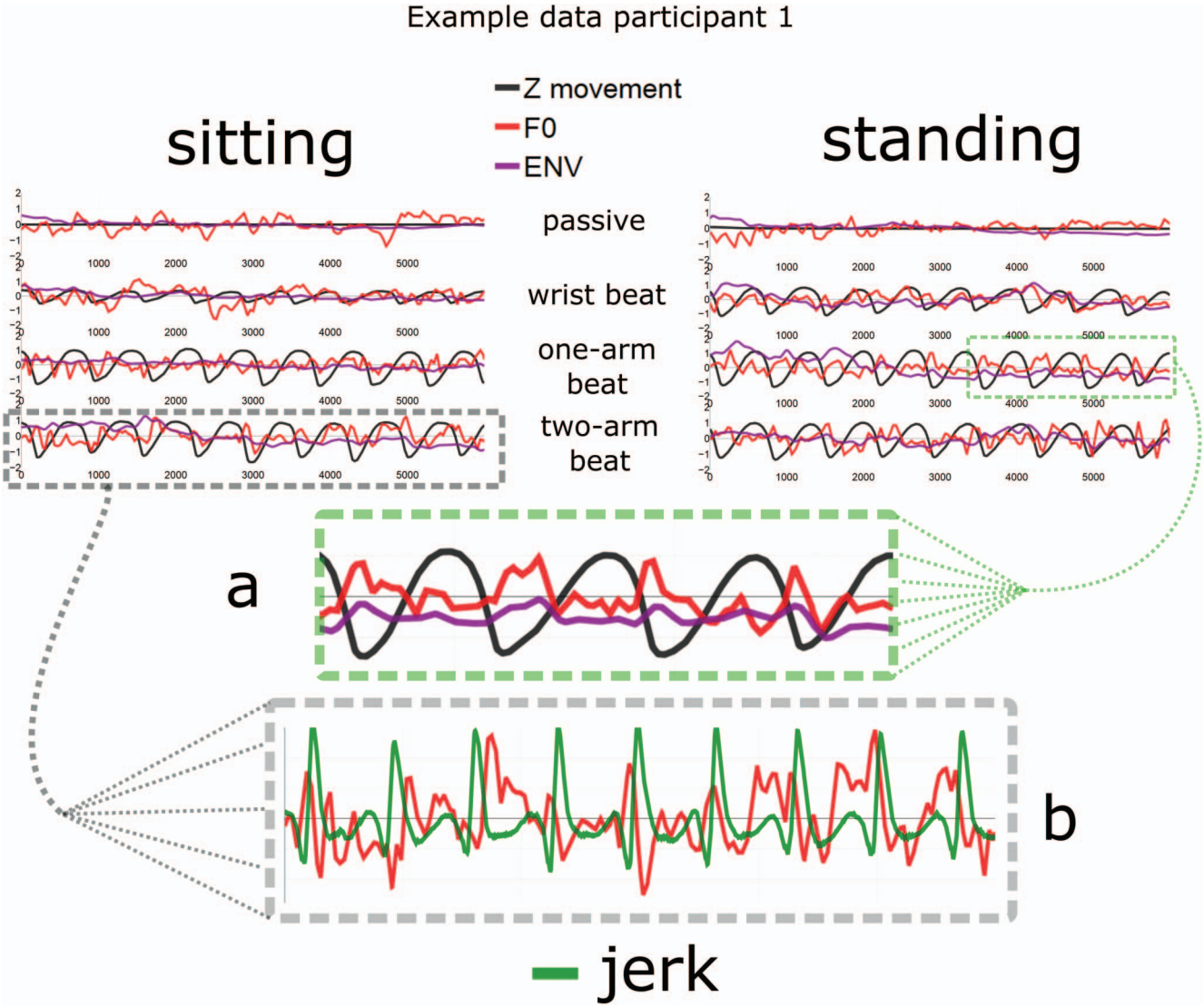

Figure 2. Example phonation and movement time series. Example time series (first 6 s) for each movement and posture condition (F0: in red [light gray]) amplitude envelope (ENV: in purple [dark gray]) and vertical movement (Z movement: in black). All measures shown are z-standardized. Panel (a) shows an enlarged section of the one-arm beat condition, where it is very clear that peaks in F0 and ENV are observed when the vertical movement reaches its maximum extension. Panel (b) shows another representation of the time series for F0 and vertical movement. Jerk is the time-derivative of acceleration, and indicates here that sudden changes in movement acceleration often co-occur with peaks in F0. See the online article for the color version of this figure.

fashion, we computed the relative phases (i.e., phi: Φ) between movement (z) and the amplitude envelope, as well as movement and F0,[6] for the one-arm and the two-arm beat conditions using cross-wavelet analyses (with R package "WaveletComp"; Rösch & Schmidbauer, 2014). Cross-wavelet analysis uses a mother wavelet as a basis for decomposing complex time series into dominant periodicities, and further allows for comparison of periodicities between time series (hence *cross*-wavelet). Our analysis used the Morlet wavelet as its daughter wavelet. We entered one-arm and two-arm movement trials into a cross-wavelet analyses (using 50 simulations to compute p values) where we assessed the relative phases of ENV with movement around the frequency 1.3 Hz (i.e., period = 0.77). Figure 5 shows a summary of the results for the cross-wavelet analyses, whereby the observed relative phases at reliability levels $p < .01$ across trial time is plotted. It can be seen that there is a reliable out-of-phase coordination.

[6] Note that in the preregistration we planned to only look at the amplitude envelope for estimating timing and relative phases, since we assumed that dynamics should be similar based on the results of the previous analyses. However, after review we have added this to the analyses (but strictly speaking relative phase results for F0 and Z belong to the exploratory analyses).

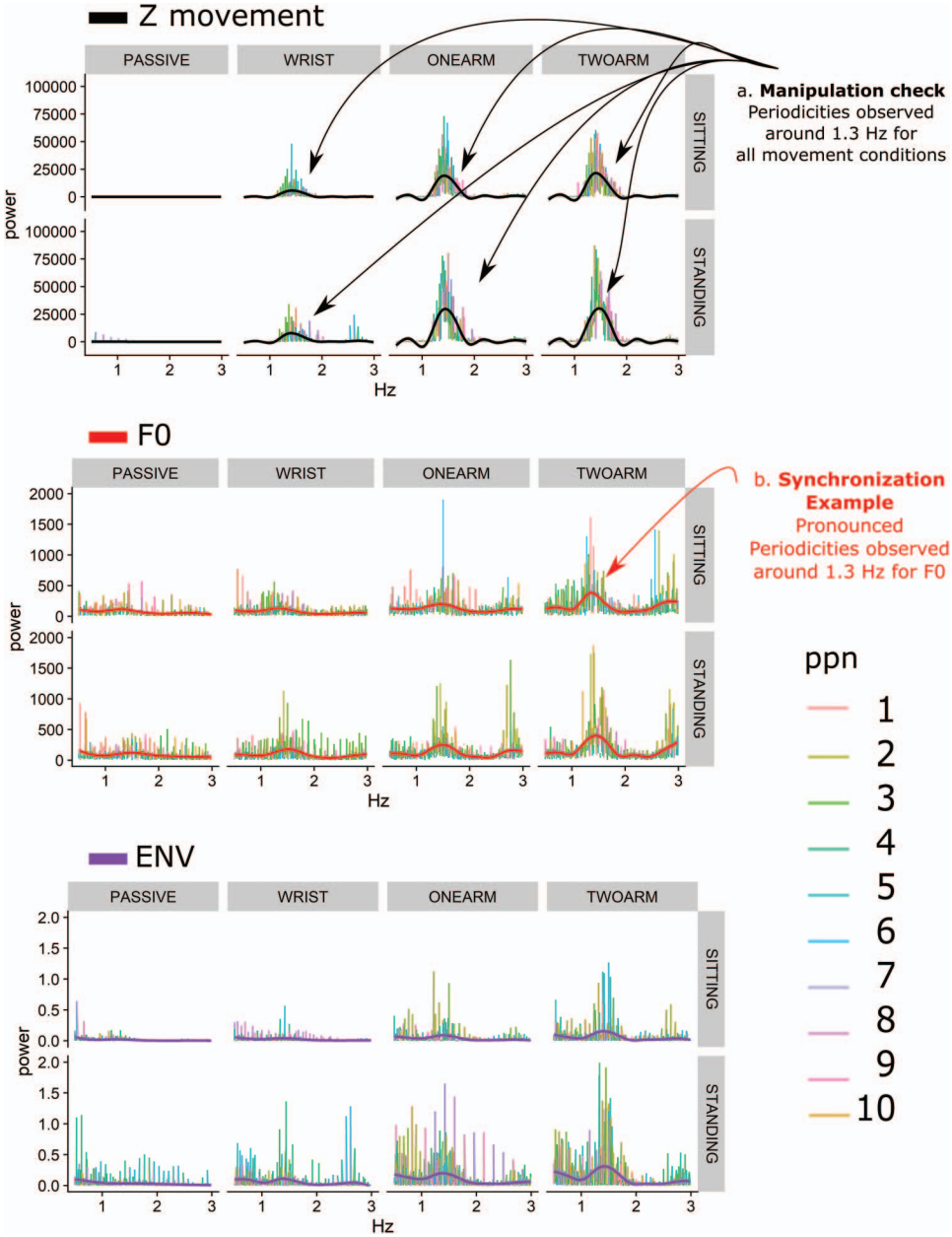

Figure 3. Spectral decomposition of periodicities of movement (black), F0 (red [light gray]), and amplitude envelope (purple [dark gray]), per condition. Solid lines (black = Z, red [light gray] = F0, purple [dark gray] = ENV) reflect smoothed mean power densities. Bumps in these solid lines indicate that the time series were defined by periodicities around that frequency. Colors reflect individual participant data. Note, at (a) that our manipulation to guide participants' movement frequency was successful as consistent periodicities are shown. Note, at (b) that peaks at the movement frequency range are found for several conditions. Note that one participant (Participant 7) was moving too fast relative to the target 1.3 Hz frequency. Interestingly though, Participant 7 also showed an increased power at this particular faster frequency range for the amplitude envelope. See the online article for the color version of this figure.

Given that we now know that there is an antiphase synchronization between movement and phonation (when movement is in the extension phase, F0 and the amplitude envelope are rising), we can estimate the time it takes for a gesture movement to reach phonation. Specifically, we can compare when a peak in amplitude envelope or peak in F0 is observed relative to the maximum extension of the beating gesture. We determined the point of maximum extension of the down-beat phase (minimum Z value) and relating this to the nearest peak in the positive rate of change of the amplitude envelope (in other words, the peak in the positive

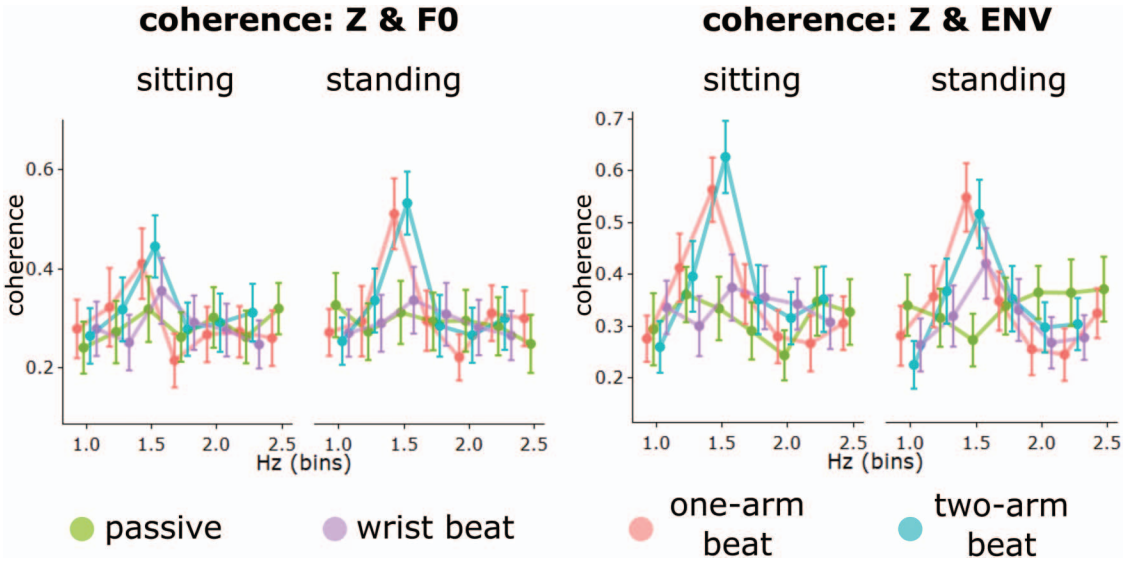

Figure 4. Coherence. This figure shows the coherence levels for each bin (of 0.25 Hz width) and condition between (a) movement (Z) and amplitude envelope (ENV), and (b) movement (Z) and fundamental frequency (F0). Data points that are directly adjacent to each other fall within a single bin. Error bars indicate 95% confidence intervals. It can be observed that the passive and wrist condition generally have lower coherence levels, confirming that movement and phonating were not coupled. Consistent with the spectral density results around the 1.3 Hz range, there are prominent peaks for the ENV and F0 for the one-arm and two-arm beat conditions, and this seems to be most pronounced for the standing condition. Note that we added the passive condition as a baseline for comparing to the other conditions. We could hence reasonably expect that movement and phonation do not have high coherence for when participants are not moving. Note that for the standing condition the coherence between Z and ENV seemed higher for the wrist condition. See the online article for the color version of this figure.

"acceleration" of the amplitude envelope). We also estimated the nearest peak in the positive rate of change of the F0, by first smoothing the F0 time series with a first-order Butterworth filter so as to reduce artifacts of fast-scale fluctuations in the signal (as observed in Figure 2). Figure 6 shows the main results. It can be obtained that slightly before (mean for negative distribution ENV = −58 ms [*SD* = 39]; mean for negative distribution F0 = −61 ms [*SD* = 34]) and slightly after (mean for positive distribution ENV = 55 ms [*SD* = 39]; mean for positive distribution F0 = 61 ms [*SD* = 37]) the point of maximum extension is reached, there is a positive acceleration in phonation (ENV and F0). These results suggest that right before the moment of maximum extension (right at the moment where *anticipatory* postural adjustments are made to brace for the impact of the beat gesture), and right after (i.e., right at the moment where the arm accelerates again for flexion) the peak in phonation is observed. That the anticipatory phonation effects are more pronounced for the standing condition is to be expected, as anticipatory postural adjustments are more pronounced when standing.

Exploratory Analyses: Higher Formants F1 and F2

One of the reviewers suggested that in addition to the amplitude envelope and F0 it would be insightful to further assess potential replication of earlier findings by Krahmer and Swerts (2007) concerning the higher formants. Krahmer and Swerts found lower levels of the second formant (F2) for gesture-accented or pitch-accented vowels, while higher levels of the first formant for pitch-accented vowels (but not for the gesture-accented vowels). We did not plan to analyze this for our confirmatory hypotheses because our study is mainly about how physical impetus affects phonation through the lower vocal tract (e.g., subglottal pressure) rather than articulatory constraints (lip rounding, tongue position relative to the palate, nasality) that are held to be primary determinants of higher formant levels. However, it is very much possible that either (a) over and above the role of physical coupling of manual movement and voice acoustics there could be informational coupling or "muscular synergies" (Krahmer & Swerts, 2007, p. 410) between manual movement and articulatory movements (see also Kelso et al., 1983; Parrell et al., 2014), or (b) that articulatory counteradjustments are made as to control phonation as interfered by physical impetus, or (c) that the direct effects of physical impetus via the lower vocal tract have cascading effects on resonant frequencies (as captured by F1 and F2). In any case, since we are interested in providing a biomechanical basis for prosodic markers (pitch accent), it is important to keep in mind that the literature in phonetics has identified higher F1 and lower F2 with pitch accented syllables (e.g., Mo, Cole, & Hasegawa-Johnson, 2009). Thus, if physical impetus affects speech in a way that is similar to pitch accents, it should produce higher F1 and lower F2 at moments of maximum physical impetus.

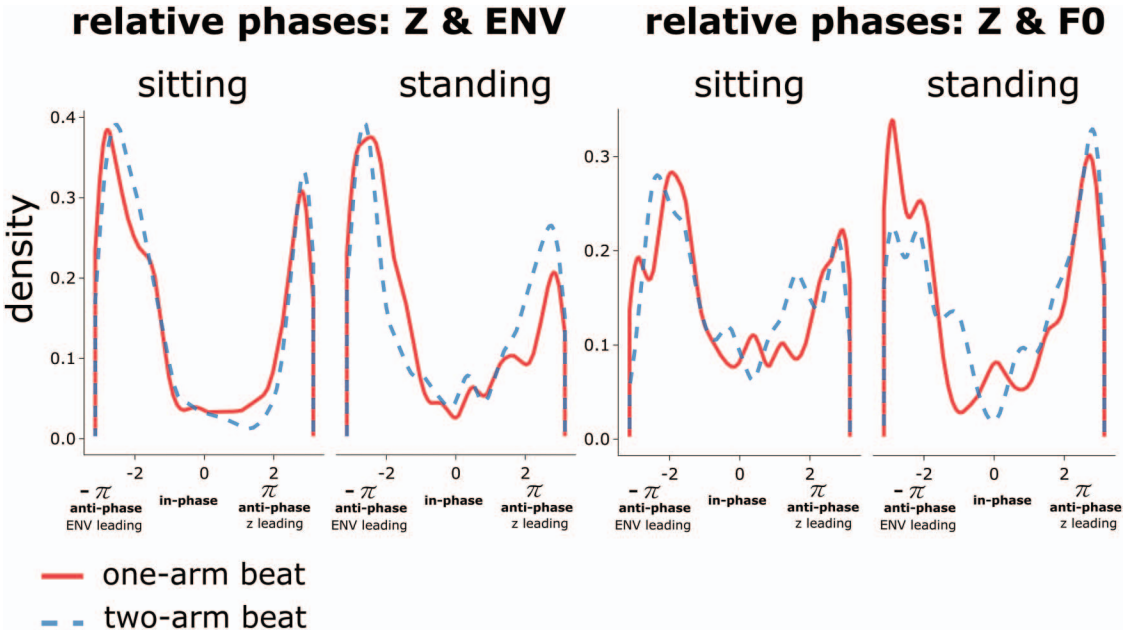

Figure 5. Density distributions of relative phases of movement (Z) versus amplitude envelope (ENV) and F0 for all movement conditions. Smoothed density distributions of relative phases (between Z and ENV, as well as Z and F0) observed at $p < .01$ for the one-arm and two-arm conditions for the 1.3 Hz range. It can be found that reliable relative phases are primarily found at the antiphase regions ($\Phi > \pi/2$ or $\Phi < -\pi/2$). Furthermore, it seems that phonation predominately leads Z time series (especially in the standing condition), and this is a likely indication that slightly before the maximum extension in the movement, the maximum physical impetus is reached due to the deceleration, which affects phonation. See the online article for the color version of this figure.

Figure 7 shows an example time series of F1, F2, and Z (movement) for the standing condition. As can be seen in the figure, at moments of maximum extension there are peaks in F1 observed (similar to our findings on amplitude envelope and the fundamental frequency). In this example time series, the relationship between F2 changes movement is not as visible, but seem to show small negative and then positive peaks after moments of physical impetus.

Figure 8 shows the coherence estimates for the movement conditions and F1 and F2.[7] It can be seen that around the relevant frequency range of 1.33 Hz there are clear elevated levels of coherence. To assess whether such inflated coherence levels were statistically reliable, we tested whether coherence in bin range containing 1.33 Hz (bin with the closed interval 1.25–1.5 Hz) was reliably higher relative to the other frequency ranges between 1 and 3 Hz. For F1, a model with movement-relevant frequency range was more reliable for predicting coherence as compared to a model predicting overall mean coherence, change in $\chi^2[1] = 103.40$, $p < .001$. There was a higher overall coherence for the movement-relevant frequency range, $b = 0.125$, $t(3095) = 11.10$, $p < .001$. This was also the case for F2, such that a model containing movement-relevant versus irrelevant frequency as predictor was a more reliable model than a model predicting overall mean coherence, $\chi^2[1] = 42.79$ $p < .001$. Again, higher coherence was found for the movement-relevant frequency, $b = 0.072$, $t(2929) = 6.563$, $p < .001$.

Having established higher formant and movement coupling, we will now assess potential differences in effects of movement conditions (wrist vs. one-arm vs. two-arm) as well as posture condition. We assessed differences in coherence levels only for the movement-relevant frequency range as a function of condition and their interactions. For F1, a model predicting overall mean coherence between F1 and movement (i.e., F1 and Z coherence) was almost surpassed in reliability by model containing movement condition, $\chi^2[1] = 5.68$, $p = .058$. Adding posture condition or its interaction with movement condition did not further improve the model, $ps > .362$. The model with movement condition showed that the two-arm beat condition had higher F1 and Z coherence levels as compared to the wrist condition, $b = 0.59$, $t(383) = 2.11$, $p = .035$. The one-arm condition did no differ on F1 and Z coherence from the wrist condition, $b = 0.13$, $t(383) = 1.86$, $p = .063$. In conclusion, we obtain exploratory evidence that physical impetus is modulating the coupling between F1 and manual movement as we observe higher coherence for the two-arm beat condition.

For F2, a model containing movement condition was more reliable than a model predicting overall mean F2 and Z coherence, $\chi^2[1] = 6.78$, $p = .033$. Adding posture condition or its interaction with movement condition did not reliably improve model fit, $ps > .643$. The most reliable model with movement condition showed

[7] Note that we used a similar procedure as in the confirmatory analyses wherein we assessed coherence levels between movement and phonation for linearly detrended phonation signals. We exclude the passive condition for this analysis and instead assess evidence for F1 and Z and F2 and Z coherence as a function of the relevant movement frequency.

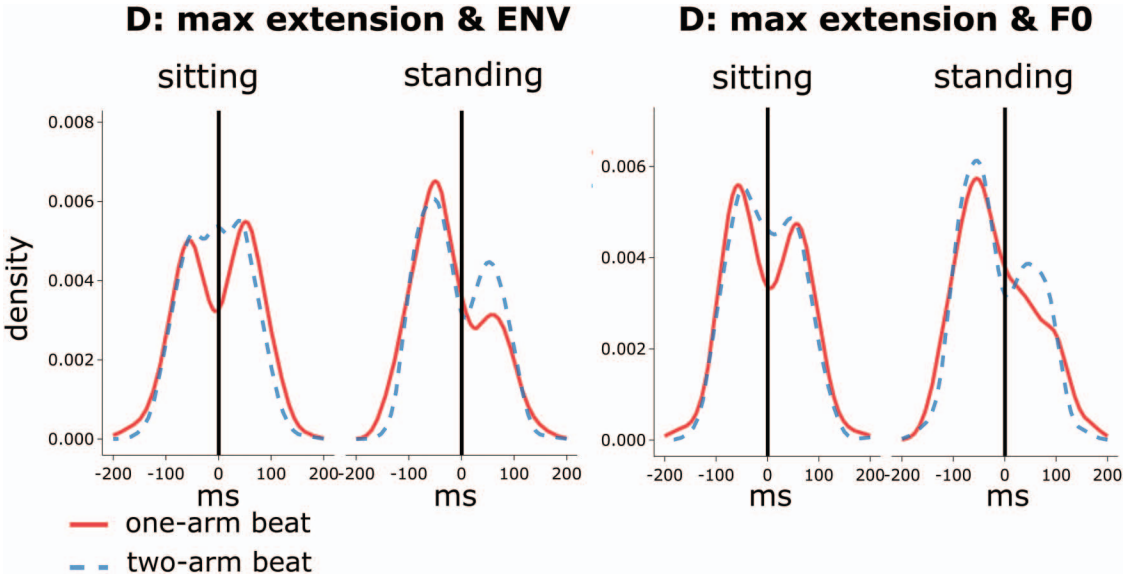

Figure 6. Timing of peak change ENV—Point of maximum extension beat. Ms (milliseconds) is the temporal distance between the nearest peak of a positive rate of change in the amplitude envelope (peak env "acceleration") versus the maximum extension of the downbeat. If temporal offset in milliseconds is negative, this indicates that peak in change of the amplitude envelope precedes the point of maximum extension. For the standing condition as compared to the sitting condition, it seems that the negative distribution becomes more peaked, likely indicating that anticipatory postural adjustments that are made before the maximum extension (and thus physical impetus) is reached are impacting phonation. See the online article for the color version of this figure.

a higher F2 and Z coherence for two-arm beat as compared to the wrist beat condition, $b = 0.072$, $t(362) = 2.60$, $p = .010$. The one-arm condition did not differ in coherence as compared to the wrist beat condition for F2, $b = 0.033$, $t(362) = 1.18$, $p = .240$.

Finally, Figure 9 provides the phase relations between the higher formants and manual movement as computed by cross-wavelet analyses (exactly the same settings were used as our previous cross-wavelet analyses). The results are not as straightforward as was the case for our previous analyses with F0 and the amplitude envelope, where we found consistent antiphase relations such that when maximum extension (low Z) was reached there was higher F0 and amplitude envelope levels (see Figure 6). Although such patterns seem to be emerging for F1, as indicated by clear peaks at the antiphase regions, there are also peaks observed more closely situated near the in-phase region (maximum flexion and higher F1). For F2 such in-phase coupling between movements is actually to be expected if it follows pitch accent patterns. Interestingly, a higher likelihood of in-phasing can indeed be observed for F2 as compared to F1, especially for the one-arm condition. This means for the one-arm condition when the maximum extension was reached negative peaks of F2 are often observed. In conclusion, there are interesting relationships with movement and the higher formant (as observed by Krahmer & Swerts, 2007) and recruitment of high intensity movements (wrist < one-arm beat < two-arm) seems to modulate this formant–movement coupling.

Discussion

We showed that repetitive arm movements (moving at 1.3 Hz) with relatively high physical impetus (one-arm and two-arm vertical beat movements) structurally entrained phonation. Repetitive arm movements with relatively low physical impetus (wrist beat condition) or making no movements at all (passive condition) did not lead to effects on phonation (as compared to the one-arm and two-arm beat condition). Movement during standing, as compared to sitting, increased the degree of phonation entrainment (but only for F0). This suggests that anticipatory postural adjustments modulate movement–phonation synchronization effects. However, since movement effects on phonation still arise when participants are seated (and posture effects were not very pronounced), we can conclude that physical impact of arm movement has direct effects on phonation. The effects were such that when about 50 ms before and after the arm movement reached its maximum extension, peaks in fundamental frequency and amplitude envelope were found. This reflects that effects on phonation arise at moments where physical impetus is highest (deceleration for stopping extension and acceleration for initiating flexion). In sum, the current study shows that merely moving the upper limbs affects phonation acoustics. Movement affects acoustics despite the fact that participants were instructed to resist any effects on their phonation.

The current results converge with previous studies showing a link between kinematic peaks and acoustic peaks such as F0 (Cravotta, Busà, & Prieto, 2018; Cravotta, Busà, & Prieto, in press; Danner, 2017; Esteve-Gibert & Prieto, 2013; Krahmer & Swerts,

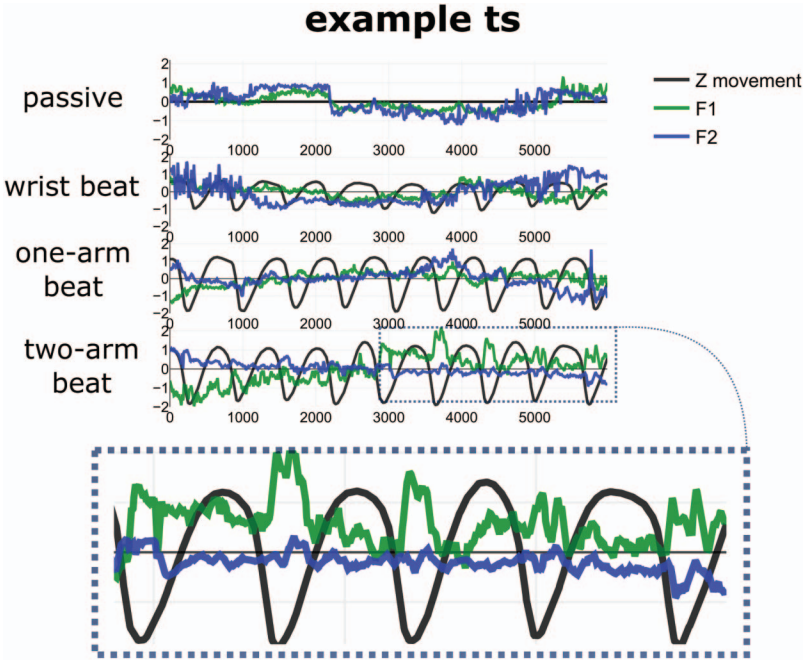

Figure 7. Example time series (standing condition) of movement, F1, and F2. An example time series is shown with the z-standardized first formant (F1 in green [light gray]) and second formant (F2 in blue [dark gray]). It can be seen for F1, especially in the two-arm beat condition, that positive peaks are observed before and after the moment of physical impetus (maximum extension). For F2 this is less readily detectable, although smaller negative peaks and then positive peaks are observed around the moment when physical impetus is reached. In general, it can be further obtained that signals show some nonlinear nonstationarities. Note further that for the exploratory analyses we smoothed the signals with a low-pass 33 Hz first-order Butterworth filter as to reduce high frequency fluctuations in the F1 and F2 estimates (for a smoothed example see: https://osf.io/49ru8/). Smoothing did not change the analyses' conclusions. See the online article for the color version of this figure.

2007; Leonard & Cummins, 2011; Loehr, 2004; McClave, 1998; Pouw & Dixon, 2019 however also see Hoetjes, Krahmer, & Swerts, 2014), but they diverge from said literature (for an overview see Wagner et al., 2014) in that a clear role of physical impetus is found.

The exploratory analysis on higher formants converge with results by Krahmer and Swerts (2007), which showed that gestural beats lead to F2 decreases in vowel acoustics, and pitch accents were associated with increases in F1 (also see Mo et al., 2009). Our results also indicated that there is clear coupling of movement and the higher formants as indicated by increased coherence between the movement and phonation signals, for F2 and especially F1. Yet, the phasing relationships between moments of observed peaks in F1 or F2 and movement phases were very variable—although we find some indication that F2 is more likely to be decreasing when movement reaches its beat (maximum extension) aligning with Krahmer and Swerts (2007). Before focusing on the implications of our confirmatory analyses, we have some speculations of why F1 and F2 show clear frequency coupling while having less clear phase coupling. First, it may be that the type of coupling between movement and the higher formants is sustained by more complex coupling relations between articulatory (e.g., mouth aperture; tongue position) and manual movement coordination. For example, when coordinating finger movements at frequencies that are comfortable to perform, spontaneous switching from in-phase to out-phase (and vice versa) is often found (Kelso & Schöner, 1988). Furthermore, it is possible that articulatory movements are recruited as a reaction to the effects of physical impetus on voice acoustics so as to keep voice integrity under perturbance. Together with the fact that changes in F0 due to physical impetus (obtained in our confirmatory analyses) will likely cascade through the resonant frequencies (i.e., F1 and F2), adding up articulatory adjustments in reaction to those changes might lead to very variable phase relationships as observed here. We will further focus on the implications of our confirmatory analyses, which focus on the role of physical impetus on the lower vocal tract (rather than articulatory dynamics).

Theoretical Implications

We think there are a couple of important implications to be drawn from the current study that will be informative for ontogenetic, phylogenetic, and cognitive theories of multimodal language.

However, we must first emphasize what the current effects do not entail. The current findings cannot explain all occurrences of prosodic gesture–speech synchrony. This is because: (a) beat gestures can be very small in their movement amplitude and still tightly synchronize with speech; (b) head movements also synchronize with speech (Krahmer & Swerts, 2007); (c) gesture and

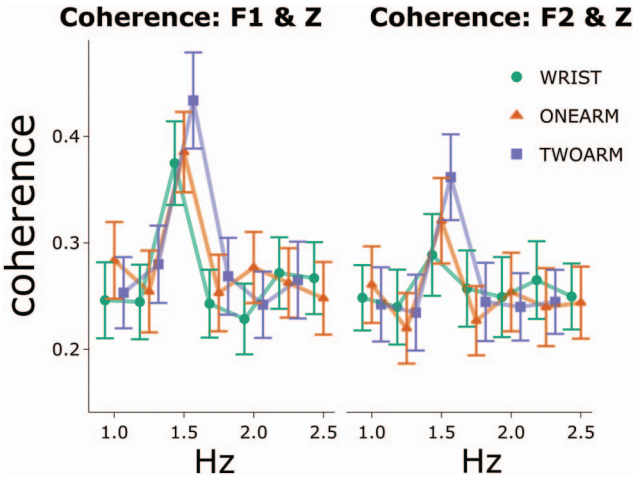

Figure 8. Coherence between movement (Z) and the higher formants (F1 and F2). Coherence levels for the first formant and movement (F1 and Z: left panel) and second formant and movement (F2 and Z: right panel). It can be clearly seen that there is increased coherence for both measures around the 1.3 Hz range, with a stepwise increase in coherence as a function of physical impetus (wrist beat < one-arm beat < two-arm beat). See the online article for the color version of this figure.

speech can be affected and their timings are often variable (Loehr, 2004; Pouw & Dixon, 2019; Rusiewicz, Shaiman, Iverson, & Szuminsky, 2014); and (d) speech prosody is not only defined by peaks in F0 and amplitude but by a myriad of other temporally dynamic features (e.g., modulation of syllable duration) that couple in interesting ways with gestural coordination that go beyond the current findings (e.g., Bernardis & Gentilucci, 2006; Shattuck-Hufnagel & Ren, 2018). Thus, to be very clear about this, bodily resonances cannot fully accommodate for gesture–speech synchrony in all contexts. What the current results do entail is that we can now directly challenge the idea that gesture–speech synchrony is "not biologically mandated" (cf. McClave, 1998).

Note that it is likely to be the case that the current biomechanical constraints play out differently in fluid communicatively intended speech. That is, although physical impetus will constrain voice acoustics by biomechanical necessity if we are to believe the current results, said effects may be counteracted, exploited, or amplified through interaction with other constraints such as communicative context. It is very likely that even within speakers there might be variability of how the current biomechanical effects are exploited or counteracted. For example, it has been shown that the movement of the chest wall to modulate alveolar pressure is used by some speakers when producing a pitch-accented syllable, while others produce an equivalent result without primary use of lower tract modulation (Petrone, Fuchs, & Koenig, 2017). The point of the current study is that we aimed to show that biomechanical factors co-constitute gesture–speech coupling which at this point of our inquiry required us to simplify the communicative context. The clear results obtained here, however, now do invite future research on how these biophysical dynamics play out in interaction with other constraints that determine gesture–speech coordination (e.g., Cravotta, Busà, & Prieto, in press).

The current study does allow for exciting speculations that prosodic gesture–speech coordination is not solely a cognitive invention, but also an *embodied innovation*. Speculating about ontogenesis, when babies perform their motor babbling together with their oral babbling, the current effects may be gradually appropriated by the infant (Iverson & Thelen, 1999; Lee, 2006). Indeed, it has been found that acoustic qualities improve, and become more speech-like, when oral babbling is concomitant with rhythmic limb motions (Ejiri & Masataka, 2001; for an overview Esteve-Gibert & Guellaï, 2018). Thus, similar to feeding-related hand-mouth linkages (Iverson & Thelen, 1999), there is a potential for gesture–speech synchrony present from the moment of birth through biomechanics, which may be further exploited through socialization processes.

On a phylogenetic level, our (primate) ancestors may have learned to harness bodily tensioning for control of the phonation system. It is, for example, well-known that our closest relative *Pan troglodytes* (the chimpanzee) lacks sufficient control of the phonation system, while there is good evidence for high control of rudimentary articulatory gestures such as lip-smacking (Ghazanfar, 2013). The current results allow for the possibility that tensioning of the myofascial-skeletal net (Turvey & Fonseca, 2014) has a possible role to play in the control of the vocal apparatus. Perhaps humans' more direct ancestors successfully developed control of the vocal apparatus in part through a range of embodied innovations (see, e.g., Hardus, Lameira, Van Schaik, & Wich, 2009; see also Blasi et al., 2019) including bodily tensioning, which might be particularly potent for the emergence of language when they co-opt an already keen sense of rhythm in the motor domain (Kotz, Ravignani, & Fitch, 2018).

Finally, the current results have implications for gesture–speech cognition. Invariably, gesture–speech synchrony is understood as controlled by the higher faculties (i.e., top-down constraints), involving some kind of prediction mechanism that times the activation of a particular gesture stress (or "maximum effort") with an acoustic marker (de Ruiter, 2000). We suggest that in principle such timing information could be based in biomechanics. Sensing and modulating the effect of physical impetus on phonation can provide a cognitively cheap (and radically embodied) opportunity

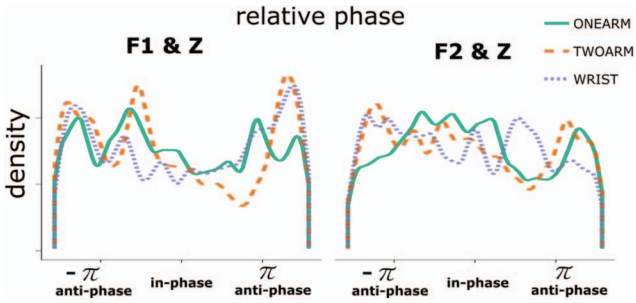

Figure 9. Observed relative phases between movement and higher formants. Frequency density distributions of reliable ($p < .05$) relative phase estimates as obtained from cross-wavelet analyses. It can be seen that there are primarily antiphase peaks for first formant and movement (F1 and Z: left panel), while there also seem to be in-phase coupling for the second formant and movement (F2 and Z: right panel), especially for the one-arm beat. In general the phasing relations are much more variable as compared to the consistent antiphasing relationship of movement versus F0/amplitude envelope. See discussion for possible explanation of this variable phasing. See the online article for the color version of this figure.

for what can be intentionally modulated and exploited based on contextual communicative requirements. Indeed, there is a promising indication that the current effects must be to some degree in play, as it is now widely accepted that the beat-like, force-producing aspect of gestures is not only present in baton or beat gestures, but superimposed on most types of gestures, including those that reserve degrees of freedom for referential expression (Prieto et al., 2018; Shattuck-Hufnagel & Ren, 2018). Finally, the current forces that are producing the acoustic effects are present in many gestures observed in the wild (see, e.g., https://osf.io/29h8z/).

References

Aruin, A. S., & Latash, M. L. (1995). Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements. *Experimental Brain Research, 103,* 323–332. http://dx.doi.org/10.1007/BF00231718

Bernardis, P., & Gentilucci, M. (2006). Speech and gesture share the same communication system. *Neuropsychologia, 44,* 178–190. http://dx.doi.org/10.1016/j.neuropsychologia.2005.05.007

Bernstein, N. (1966). *The co-ordination and regulation of movements*. London, UK: Pergamon Press.

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science, 363*(6432), eaav3218. http://dx.doi.org/10.1126/science.aav3218

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

Bouisset, S., & Do, M. C. (2008). Posture, dynamic stability, and voluntary movement. *Clinical Neurophysiology, 38,* 345–362. http://dx.doi.org/10.1016/j.neucli.2008.10.001

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

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

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

Cravotta, A., Busà, M. G., & Prieto, P. (in press). Effects of Encouraging the Use of Gestures on Speech. *In Journal of Speech, Language, and Hearing Research*.

Danner, G. M. (2017). *Effects of speech context on characteristics of manual gesture*. Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

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

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 248–311). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620850.018

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

Esteve-Gibert, N., & Guellaï, B. (2018). Prosody in the auditory and visual domains: A developmental perspective. *Frontiers in Psychology, 9,* 338. http://dx.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,* 850–864. http://dx.doi.org/10.1044/1092-4388(2012/12-0049)

Gallagher, S. (2005). *How the body shapes the mind*. New York, NY: Oxford University Press.

Ghazanfar, A. A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behavioral Ecology and Sociobiology, 67,* 1441–1448. http://dx.doi.org/10.1007/s00265-013-1491-z

Hardus, M. E., Lameira, A. R., Van Schaik, C. P., & Wich, S. A. (2009). Tool use in wild orangutans modifies sound production: A functionally deceptive innovation? *Proceedings of the Royal Society B: Biological Sciences, 276,* 3689–3694.

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

Hodges, P. W., Gandevia, S. C., & Richardson, C. A. (1997). Contractions of specific abdominal muscles in postural tasks are affected by respiratory maneuvers. *Journal of Applied Physiology, 83,* 753–760. http://dx.doi.org/10.1152/jappl.1997.83.3.753

Hodges, P. W., & Richardson, C. A. (1997). Relationship between limb movement speed and associated contraction of the trunk muscles. *Ergonomics, 40,* 1220–1230. http://dx.doi.org/10.1080/001401397187469

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

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. A. S., & Schöner, G. (1988). Self-organization of coordinative movement patterns. *Human Movement Science, 7,* 27–46. http://dx.doi.org/10.1016/0167-9457(88)90003-6

Kelso, J. A. S., Tuller, B., & Harris, K. (1983). A "dynamic pattern" perspective on the control and coordination of movement. In P. MacNeilage (Ed.), *The production of speech* (pp. 138–173). New York, NY: Springer-Verlag.

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

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

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

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp. 261–283). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511620850.017

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,* 3. http://dx.doi.org/10.5334/labphon.75

Lee, D. (2006). *How movement is guided*. Unpublished manuscript. Retrieved from http://www.pmarc.ed.ac.uk/ideas/pdf/HowMovtGuided100311.pdf

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

Lieberman, P. (1996). Some biological constraints on the analysis of prosody. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (pp. 67–78). Mahwah, NJ: Erlbaum.

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

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27,* 69–89. http://dx.doi.org/10.1023/A:1023274823974

McNeill, D. (2008). *Gesture and thought*. Chicago: University of Chicago press.

Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: Evidence from formant structure. *Proceedings of Interspeech 2009* (pp. 2535–2538). Brighton, UK: ISCA.

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

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,* 1715–1725. http://dx.doi.org/10.1121/1.4976073

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

Pouw, W., Harrison, S. J., & Dixon, J. A. (2018). *The physical basis of gesture–speech synchrony*. Unpublished preprint. http://dx.doi.org/10.31234/osf.io/9fzsv

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

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, M. J. (2009). *Polhemus applications and example code*. Retrieved from http://xkiwilabs.com/software-toolboxes/

Rösch, A., & Schmidbauer, H. (2014). WaveletComp: Computational wavelet analysis (R package version 1.0). Retrieved from https://cran.r-project.org/package=WaveletComp

Rusiewicz, H. L., & Esteve-Gibert, N. (2018). Temporal coordination of prosody and gesture in the development of spoken language production. In P. Prieto & N. Esteve-Gibert (Eds.), *The development of prosody in first language acquisition* (pp. 103–124). Amsterdam, the Netherlands: John Benjamins. http://dx.doi.org/10.1075/tilar.23.06rus

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. http://dx.doi.org/10.1016/j.specom.2013.06.004

Seilmayer, M. (2016). Common methods of spectral data analysis: Package "spectral" (Computer software). Retrieved from https://cran.r-project.org/web/packages/spectral/spectral.pdf

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. http://dx.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,* 64–68. http://dx.doi.org/10.1016/j.neulet.2007.09.066

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics, 18,* 213–226. http://dx.doi.org/10.1080/09524622.2008.9753600

Turvey, M. T., & Fonseca, S. T. (2014). The medium of haptic perception: A tensegrity hypothesis. *Journal of Motor Behavior, 46,* 143–187. http://dx.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. http://dx.doi.org/10.1016/j.specom.2013.09.008

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

Received December 17, 2018
Revision received April 10, 2019
Accepted May 25, 2019 ■