IS GESTURE-SPEECH PHYSICS AT WORK IN RHYTHMIC POINTING? EVIDENCE FROM POLISH COUNTING-OUT RHYMES

Šárka Kadavá[1,3,*], Aleksandra Ćwiek[1], Katarzyna Stoltmann[2], Susanne Fuchs[1], Wim Pouw[3]

[1]Leibniz Center General Linguistics, Berlin, Germany [2]adesso SE, Berlin, Germany [3]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

[*] corresponding author: kadava@leibniz-zas.de

ABSTRACT

'Gesture-speech physics' refers to a possible biomechanical coupling between manual gesture and speech. According to this thesis, rapid gesturing leaves a direct imprint on acoustics (intensity, F0), as gesture accelerations/decelerations increase expiratory forces and therefore subglottal pressure, leading to higher amplitude envelope peaks and higher F0 values. This acoustic effect has been reported in lab experiments, spontaneous speech, clinical studies, and professional vocal performers. The current study investigates this phenomenon in Polish counting-out rhymes, using motion capture data and acoustic recordings from 11 native Polish speakers. Following the gesture-speech physics thesis, we expect acceleration/deceleration peaks to be correlated with speech intensity/F0. Through Bayesian analyses, we obtained a weak but reliable coupling of deceleration of the pointing hand and the nearest peak in the smoothed amplitude envelope.

Keywords: pointing gestures, motion tracking, poetry, prosody, coupling

1. INTRODUCTION

The evidence that gesture-speech coordination on the prosodic level arises out of basic properties of physiology and motor control is increasing [1, 2, 3, 4, 5, 6]. This contrasts with the argument that gesture is a sophisticated cognitive achievement, proliferating due to cultural conventionalization [7].

While not downplaying either of those constraints, according to the gesture-speech physics account (see [6]), there is a biomechanical nudge for aligning peaks in F0 and amplitude envelope with the peak of the physical impulse. As such, the human voice receives an 'imprint' due to the gestural activation of expiration-related muscles. Specifically, upper limb acceleration and deceleration affect rib-cage movement and thus subglottal pressure. That forceful gestures imprint the voice is in line with machine learning studies, showing that neural networks trained on acoustics and body kinematics can come to predict the presence of gesture or kinematic properties of gestures [8, 9, 10].

Although gesture-speech physics seems robust in some tasks, a recent study on leg and arm biking suggests that acceleration may need to reach a certain threshold to affect speech acoustics [11]. This is in line with previous research showing that body parts with lower mass (hand vs. arm) have much weaker effects on speech (e.g., [12]), if at all [13].

The reason why the biomechanical gesture-speech coupling is weak is likely because there must remain the flexibility to speak in certain ways when gesturing. The larynx should indeed be flexible to resist the effect of motion at times it is appropriate to do so. After all, the primary function of the larynx is to act quickly and protect the lungs from inhaling foreign bodies [14].

This study replicates the basic kinematic-acoustic coupling findings from previous research. Our dataset consists of motion data recorded while performing Polish counting-out rhymes involving pointing movement. During a counting-out rhyme game, one person speaks a rhyme while rhythmically moving their index finger between themselves and another person. Having clear turning points, these childhood poems are a valid paradigm to investigate speech-gesture physics, as appreciated by previous studies [15].

We extend previous work by studying forward and backward pointing movement and speech rate as additional factors. So far, only flexion-extension movements have been studied [16] and it is possible that different movements will have different respiratory interactions. Looking at a faster rate is motivated by the fact that this may go hand in hand with larger accelerations/forces. However, the rate may also change the complexity of gesture-speech physics, as different coupling strengths and muscular stiffness are involved.

Following earlier work by amongst

others [16, 17], we expect that (a) higher acceleration/deceleration peaks scale with higher amplitude envelope peaks, and higher F0 values (but less strongly than amplitude), and (b) that the strength of this correlation is dependent on the body motion (forward vs. backward) as it might lead to different effects on airflow due to using antagonistic muscle units. Though we manipulated whether the task was performed with the right or left hand, we ignore this variable as we have no strong hypotheses about handedness in the context of gesture-speech physics.

2. METHODOLOGY

2.1. Experimental set-up and procedures

Participants were informed about the experimental procedures and experimental equipment and signed a consent form. They wore an OptiTrack jacket and a headband; 14 markers were placed on various joints and body parts. The marker relevant to the current analyses was located at the wrist. The participants were instructed to play a counting-out rhyme game with a teddy which was placed on a chair about 1.5 m distance from the participant.

Movements were recorded in 3D space with an OptiTrack system (Motive Version 1.9.0) with 12 cameras (Prime 13), at a 200 Hz sampling rate. Acoustic data were simultaneously recorded with a Sennheiser cardioid microphone, at a 44.1 kHz sampling rate.

Each rhyme was produced in two different speech rates (normal vs. fast) and either the right or left hand for pointing. There was also a control reading condition without any pointing. The conditions were fully crossed, resulting in five different blocks: (1) left hand, normal rate; (2) left hand, fast rate; (3) right hand, normal rate; (4) right hand, fast rate; and (5) reading.

In each block, all rhymes were produced. The order of the blocks was the same for all participants (reading, normal rate, fast rate), but the order of counting-out rhymes was randomized, as well as whether they began with the dominant (right) or non-dominant (left) hand.

2.2. Participants

Eleven Polish native speakers (8 female, 3 male, mean age = 24.1, range = 21–27) took part in the experiment. All participants were right-handed according to the Edinburgh handedness scale [18].

2.3. Data processing

2.3.1. Audio

We applied the Hilbert transform to the sound signals. Taking the complex modulus of the complex-valued transformed signal yields a 1D amplitude envelope of each signal. To smooth the envelopes, we used a Hanning window of 10 Hz. We then downsampled the smoothed envelopes to 200 Hz. To extract the F0 traces of the sound signals, we applied a K. Schaefer-Vincent periodicity detection algorithm (using the `R`-package `wrassp` [19]). Based on sex, the F0 range was limited to 100–450 Hz (female) or 70–300 Hz (male).

2.3.2. Motion tracking

To remove noise-related jitter in motion tracking, a zero-phase 2nd order Butterworth low-pass filter with a cut-off of 30 Hz was applied to the position traces. We also differentiated the signals with respect to time to retrieve the 3D speed, and the next derivative of speed, 3D acceleration. After differentiation, we also apply the same Butterworth filter again. Gesture phases during the ongoing speech were automatically annotated by marking turning points and the movement direction (forward vs. backward).

2.3.3. Aggregation of acoustics and kinematics

Acoustics and kinematics were merged based on their timestamps, and misaligned recordings were approximated using linear interpolation (based on time). Then we resampled all signals at exactly 200 Hz. An example of time series can be seen in Figure 1.

2.3.4. Peak datasets

For the analysis, we constructed a dataset containing all the relevant data points from the time series – namely the peaks from acoustics (envelope, F0) and kinematics (acceleration, deceleration for wrist), using `findpeaks` from the `pracma` `R`-package [20]. The kinematic peaks were extracted for each gesture phase (i.e., movement from one turning point to another). Acoustic peaks were identified that were closest in time to those kinematic peaks. The peaks under study are displayed in Figure 1.

For the final datasets used for modeling, paired kinematic-acoustic peaks that occurred at more than ± 80 ms difference were excluded. This is because peaks that are too far apart are

not possible candidates for mechanical coupling. These estimates are further based on the time of anticipatory or reactionary muscles that happen before or after a deceleration peak [21]. In the datasets with envelope peaks, we omitted those data points in which the envelope peaks were detected either in a pause within the rhyme or in moments without voicing (F0 = 0). Furthermore, some trials needed to be excluded since the wrist marker was not visible and caused tracking problems. Finally, outliers of each variable within each dataset were detected and removed using Tukey's 1.5 IQR rule.

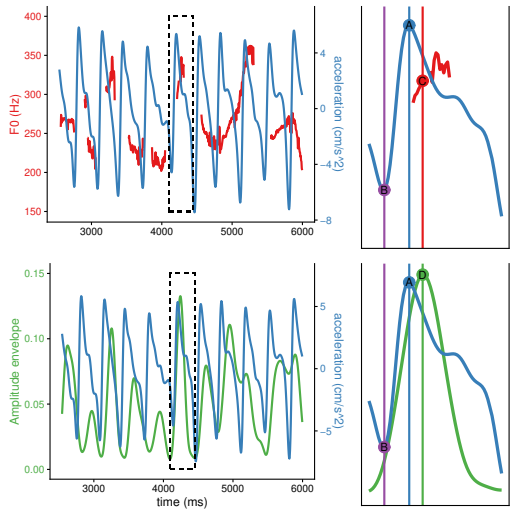

Figure 1: Example of amplitude envelope, F0 and acceleration time series (first 3.4 seconds from the onset of speech). Plots on the right side show enlarged sections with peaks under study: A – acceleration peak, B – deceleration peak, C – F0 peak, D - amplitude envelope peak.

2.4. Data analyses

All analyses were conducted in `R` version 4.2.1 [22]. We used `BayesFactor` [23] to compute Bayes Factor, `brms` [24] and `cmdstanr` [25] for modeling, and `shinystan` [26] for model diagnostics.

Four Bayesian mixed effects models were fitted for the investigated dependent variables: (1) F0 peak nearest to acceleration peak (*F0 acc*), (2) F0 peak nearest to deceleration peak (*F0 dec*), (3) amplitude envelope peak nearest to acceleration peak (*ENV acc*), and (4) amplitude envelope peak nearest to deceleration peak (*ENV dec*).

All models included *speech rate* (fast vs. normal), *movement direction* (forward vs. backward), and *sex* (female vs. male) as contrast-coded predictors. Models for the acceleration (*F0 acc* and *ENV acc*) also included the *wrist acceleration peak* and the interaction of the *wrist acceleration peak* with *movement direction* as predictors. Respectively, models for the deceleration (*F0 dec* and *ENV dec*) included the *wrist deceleration peak* and its interaction with *movement direction*. All models included random intercepts for *speaker* and *rhyme* with random slopes corresponding to all of the factors (except for *sex* in *speaker*). The models included weakly informative priors (i.e., unbiased with respect to H0/H1); the intercept prior for the models was truncated to only include positive values.

Prior to modeling, we controlled for the correlation between speech rate and the respective acceleration or deceleration peak of the wrist in order to assess the possible collinearity of the predictors. *Speech rate* (contrast-coded) and *wrist peak* (acceleration or deceleration, respectively) were correlated in all cases. This, however, did not cause any issues for the model to converge. All processed data, scripts and models are available in the OSF repository: https://osf.io/2abtd/.

3. RESULTS

Table 1 presents the results for parameters that were identified as reliably affecting the outcome variables; for reference, it also includes the intercepts of all models. For complete model outputs see here. Both F0-related models, i.e., *F0 acc* and *F0 dec*, show that *sex* and *speech rate* have an effect on F0 peaks, such that women and faster speech rate exhibit higher F0 peaks. Other parameters did not turn out to be reliable predictors of F0 peaks.

As for the envelope-related models, there were no reliable predictors in the *ENV acc* model. In the *ENV dec*, we found an effect of *wrist deceleration peak* on the amplitude envelope peak ($\beta = -0.02[-0.04; -0.00]$) with the posterior probability $\beta < 0$ equal to 0.99. The magnitude of the effect is small, as per the mean posterior estimate (-0.02). It is, however, stable and reliably distributed, as, given the priors, the data, and the model, there is a 99% posterior probability that there is an effect of wrist deceleration on the amplitude envelope, such that when deceleration (negatively) increases, the nearest envelope peak is higher.

Apart from that, there were two further parameters in the *ENV dec* model, which reached 94% posterior probability (i.e., almost the common cut-off of 95%), but both encompassed 0 in the 95% CrI. These were *speech rate* ($\beta = -0.10[-0.23; 0.03]$) and *movement direction* ($\beta = -0.08[-0.18; 0.02]$). Given our priors, data, and the model, the effect of these parameters

Model	Parameter	Estimate [95% CrI]	Pr($\beta <$ or > 0)
F0 acc	Intercept	185.58[172.59;198.94]	
	Speech rate	8.16[1.70;14.51]	0.99
	Sex	104.07[77.81;129.97]	1.00
F0 dec	Intercept	177.12[162.24;191.94]	
	Speech rate	5.04[0.20;9.90]	0.98
	Sex	100.84[71.03;130.46]	1.00
ENV acc	Intercept	0.82[0.72;0.93]	
ENV dec	Intercept	0.70[0.54;0.87]	
	Wrist dec. peak	−0.02[−0.04;−0.00]	0.99

Table 1: The table lists out all intercepts, as well as the parameters with a reliable effect on the outcome variables of the four models, with posterior means and the 95% CrI. The rightmost column is the posterior probability of the effect to be below or above 0, depending on the direction.

on amplitude deceleration peak is not reliable, however, since the task was not heavily controlled, we acknowledge that those relationships should be further studied in the future. As can be seen in Figure 2, the main effect of *wrist deceleration peak* on the amplitude envelope peak we reported above differs when it is within backward vs. forward movement. Nevertheless, given the priors, the data, and the model, there was no interaction effect of the two parameters ($\beta = -0.00[-0.02;0.01]$).

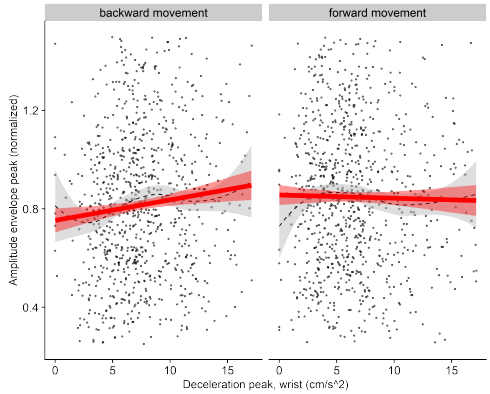

Figure 2: The linear relationship between deceleration peak and amplitude envelope peak. Dashed non-linear 'loess' line reflects possible non-linearities. Note that the deceleration values have been absolutized.

4. DISCUSSION

The current study goes beyond previous research on gesture-speech physics by assessing statistical coupling in (1) multi-directional (2) pointing movements in (3) the Polish language. Our findings suggest that deceleration peaks scale to their nearest amplitude envelope peak, rather than F0. This scaling effect was small, but reliable. Acceleration peaks did not display significant scaling with either envelope or F0 peaks.

Why did the rapidity by which participants halted a pointing movement (i.e., deceleration) not scale to the nearest F0 peak? The gesture-speech physics thesis proposes that there is a mechanical interaction between an upper limb and the body during acceleration or deceleration. The physical impulse of a upper-limb movement produces a mechanical loading onto the rib cage, which limits its movement and impacts subglottal pressures necessary for voice production. Subglottal pressures are primarily linked with affecting intensity, and only secondarily F0, which is under more flexible laryngeal control.

That there is a coupling of deceleration rather than acceleration might look like a counterargument for gesture-speech physics. However, comparing the absolute raw values of deceleration and acceleration peaks, we found 20% lower magnitudes for acceleration than deceleration. In line with [11], we suppose that a certain threshold needs to be reached before a significant effect of physics arises.

As for the deceleration effect alone, it is known that speakers coordinate their emphasis in speech with the moment when the limb movement reaches its destination [27]. Thus, emphasis is generally not located at the initial stage of pointing; rather, it occurs when reaching the intended target.

Further, we did not find that kinematic peaks affected speech differently depending on the direction of movement. This means forward and backward movements along the sagittal plane likely perturb vocalization by increasing subglottal pressure, much like flexion-extension movements along the frontal plane [16].

For future research, potential alternative hypotheses should be investigated. For example, other kinematic variables (e.g., speed) need to be assessed for speech coupling. Our study is also limited in the number of participants, cautioning generalizability, but it has a large number of trials and events that have been analyzed, increasing the reliability of the reported effects within our sample. Moreover, since we do not directly measure muscle activity in relation to respiratory-vocal states, it is always possible to maintain that the current kinematic-acoustic effect is solely a neurally controlled achievement. Such an explanation requires an auxiliary hypothesis about why the brain would monitor acceleration peaks and couple them to vocalization. While we deem it possible that the brain is tuned like this, it would be precisely because there is a weak biomechanical coupling to begin with. Gesture-speech coordination, then, is a 'smart' combination of brain and brawn [6].

5. ACKNOWLEDGEMENTS

This work has been supported by DFG grants FU791/9-1, CW 10/1-1 and PO 2841/1-1. WP is funded by a VENI grant (VI.Veni 0.201G.047: PI Wim Pouw). We would like to thank the participants of the study.

6. REFERENCES

[1] M. Chu and P. Hagoort, "Synchronization of speech and gesture: Evidence for interaction in action," *Journal of Experimental Psychology: General*, vol. 143, no. 3, 2014.

[2] J. M. Iverson and E. Thelen, "Hand, mouth and brain: The dynamic emergence of speech and gesture," *Journal of Consciousness Studies*, vol. 6, no. 11-12, pp. 19–40, 1999.

[3] B. Parrell, L. Goldstein, S. Lee, and D. Byrd, "Spatiotemporal coupling between speech and manual motor actions," *Journal of Phonetics*, vol. 42, pp. 1–11, 2014.

[4] G. Zelic, J. Kim, and C. Davis, "Articulatory constraints on spontaneous entrainment between speech and manual gesture," *Human Movement Science*, vol. 42, pp. 232–245, 2015.

[5] W. Pouw and J. A. Dixon, "Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback," *Cognitive Science*, vol. 43, no. 3, 2019.

[6] W. Pouw and S. Fuchs, "Origins of vocal-entangled gesture," *Neuroscience & Biobehavioral Reviews*, vol. 141, p. 104836, 2022.

[7] E. McClave, "Pitch and manual gestures," *Journal of Psycholinguistic Research*, vol. 27, no. 2, pp. 69–89, 1998.

[8] F. Yunus, C. Clavel, and C. Pelachaud, "Sequence-to-sequence predictive model: From prosody to communicative gestures," *arXiv:2008.07643 [cs, eess]*, 2020.

[9] T. Kucherenko, R. Nagy, M. Neff, H. Kjellström, and G. E. Henter, "Multimodal analysis of the predictability of hand-gesture properties," *arXiv:2108.05762 [cs]*, 2022, arXiv: 2108.05762.

[10] Y. Ferstl, M. Neff, and R. McDonnell, "Understanding the Predictability of Gesture Parameters from Speech and their Perceptual Importance," in *Proceedings of the 20th ACM international Conference on Intelligent Virtual Agents*. Scotland, UK: ACM New York, 2020, pp. 1–8.

[11] H. Serré, M. Dohen, S. Fuchs, S. Gerber, and A. Rochet-Capellan, "Leg movements affect speech intensity," *Journal of Neurophysiology*, 2022.

[12] W. Pouw, A. Paxton, S. J. Harrison, and J. A. Dixon, "Acoustic information about upper limb movement in voicing," *Proceedings of the National Academy of Sciences*, vol. 117, no. 12, pp. 11 364–11 367, May 2020.

[13] W. Pouw, S. J. Harrison, and J. A. Dixon, "Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony." *Journal of Experimental Psychology: General*, vol. 149, no. 2, pp. 391–404, 2020.

[14] J. J. Ohala, "Respiratory Activity in Speech," in *Speech Production and Speech Modelling*, ser. NATO ASI Series, W. J. Hardcastle and A. Marchal, Eds. Dordrecht: Springer Netherlands, 1990, pp. 23–53.

[15] S. Fuchs and U. D. Reichel, "On the relationship between pointing gestures and speech production in German counting out rhymes: Evidence from motion capture data and speech acoustics," in *Tagungsband der 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, München, Germany, 2016, p. 5.

[16] W. Pouw, L. de Jonge-Hoekstra, S. J. Harrison, A. Paxton, and J. A. Dixon, "Gesture-speech physics in fluent speech and rhythmic upper limb movements," *Annals of the New York Academy of Sciences*, vol. 1491, no. 1, pp. 89–105, 2020.

[17] W. Pouw, S. J. Harrison, N. Esteve-Gibert, and J. A. Dixon, "Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1231–1247, 2020.

[18] R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, 1971.

[19] R. Winkelmann, L. Bombien, M. Scheffers, and M. Jochim, *wrassp: Interface to the 'ASSP' Library*, 2023, R package version 1.0.4.

[20] H. W. Borchers, *pracma: Practical Numerical Math Functions*, 2022.

[21] A. S. Aruin and M. L. Latash, "Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements," *Experimental Brain Research*, vol. 103, no. 2, pp. 323–332, 1995.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[23] R. D. Morey, J. N. Rouder, T. Jamil, S. Urbanek, K. Forner, and A. Ly, *BayesFactor: Computation of Bayes Factors for Common Designs*, 2018, R package version 0.9.12-4.4.

[24] P.-C. Bürkner, "Advanced Bayesian Multilevel Modeling with the R Package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.

[25] J. Gabry and R. Češnovar, *cmdstanr: R Interface to 'CmdStan'*, 2020, R package version 0.5.2.

[26] Stan Development Team, *shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models.*, 2017, R package version 2.4.0.

[27] Esteve-Gibert Núria and Prieto Pilar, "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 850–864, 2013.