Hand-Mouth Coordination in a Pointing Task Requiring Manual Precision

Aleksandra Ćwiek[1,2], Susanne Fuchs[1]

[1] *Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany*
[2] *Humboldt Universität zu Berlin, Berlin, Germany*
cwiek@leibniz-zas.de, fuchs@leibniz-zas.de

Abstract

In daily life, articulatory movements and pointing gestures are tightly coupled. Nevertheless, the two motor systems governing the movements of the articulators and hands differ in their dynamics: the articulators are fast and much lighter than the limbs, which are slower due to their mass. We investigated the timely coordination of those motor systems in a pointing task requiring manual precision. In our experiment, the initial segment was always [p], allowing the participants for early articulatory preparation. Most importantly, we found that the hand gesture onset precedes the onset of the articulatory gesture. We also found that some speakers begin the articulatory movement only after reaching the hand gesture target. Overall, our data reveal that when the articulatory movement is not audible, as it is the case of [p], speakers are very flexible in the coordination between hand and mouth.

Keywords: coordination, pointing task, motor system, anticipation, motion capture

1. Introduction

The synchronization between pointing gesture and speech is an important step in development (Iverson and Thelen 1999) and is a frequent phenomenon in our daily life. Various authors have investigated the interplay between pointing gestures and speech in detail and found tight coordination between the two systems. For example, Rochet-Capellan et al. (2008) provided evidence that while the movement of the hand always starts earlier than the movement of the jaw, both are temporally coordinated during the production of a stressed syllable in a bisyllabic word. Changing the stress from the first to the second syllable of a word also leads to a temporal shift in coupling from the first to the second syllable.

Krivokapić et al. (2017) explored the hand–mouth coordination in varying prosodic structures, i.e., prominent syllables, and prosodic boundaries. Their findings reveal lengthening for both, manual gestures and speech gestures at boundaries, and under prominence. Pouw and Dixon (2019), as well as Chu and Hagoort (2014), showed that, when a perturbation is applied to either the hand or the mouth, the other system is also affected, emphasizing the functional link between the two.

A recent spontaneous speech study (Bekke et al. 2020) discussed the earlier onset of hand gestures concerning the predictive power of hand motion in language processing. The authors hypothesized that the hand–mouth asynchrony at the onset of the movement might be used for making predictions about upcoming words and lead to faster response times to a question. However, they could not find evidence for this claim. Aspects that are often missing in the discussion of the hand–mouth coordination are the dynamic properties of the two motor systems. Speech articulators consist to a large extent of soft tissue, which is relatively light in mass, very fast in speed, and has a large number of degrees of freedom (Grimme et al. 2011). In contrast, hands and arms consist largely of joints and bones, which are heavier in mass, slower, and have more limited degrees of freedom than speech articulators. Thus, the motor systems with vastly different dynamic properties need to be coordinated in speaking and gesturing.

These different dynamical properties may affect the synchronization behavior. While it is principally possible that both motor systems adapt to one another, it may be more economic than the slower motor system starts earlier than the faster system, so that both reach the target at the required time point. The faster system is thanks to its properties more flexible in adjusting to the slower system. If the faster system would start together with the slower system, it would have to slow down at some point or wait at the target.

The importance of the interaction between motor systems with different agility has been pointed out in the speech motor control literature, e.g., regarding the tongue–jaw coordination. Most notably within the "Frame–Content theory", and the acquisition of speech (MacNeilage 1998). Tongue–jaw coordination has certain similarities with the hand–mouth coordination, because the jaw is a heavy articulator, due to its bone structure. It is restricted in speed, in comparison to the faster and more flexible tongue. During babbling, babies cannot control their tongue and jaw motions independently (MacNeilage 1998). The cyclic movement of the jaw, with the tongue on top of it, leads to the production of specific syllables. At a later stage in speech acquisition, the tongue can be controlled independently of the jaw, and this freedom allows children to acquire the rich phonemic inventory of their respective language.

Similar approaches are rare in the literature on hand–speech coordination (but see Stoltmann and Fuchs 2017), while empirical data for synchronization are rich (e.g., Chu and Hagoort 2014; Esteve-Gibert and Prieto 2013; Habets et al. 2011; Krivokapić et al. 2017; Levelt et al. 1985; Pouw et al. 2020; Rochet-Capellan et al. 2008).

Our exploratory study aims at investigating the timing between hand and mouth in a speech–pointing task requiring large arm movements and a high degree of manual precision. In contrast to previous work, we did not vary the prosodic parameters. The design of the study allows for large flexibility in the hand–mouth coordination because the initial speech sound is a voiceless bilabial stop [p]. Since this sound is acoustically silent during the closure phase, speakers can prepare the lip closure at any time during the arm movement without disturbing the acoustic speech output.

Given the dynamic properties of the two motor systems and the large number of studies reporting an early hand gesture onset in comparison to the speech gesture, we assume that such a pattern will also be found in our data. Alternatively, the flexibility allowed by the task may lead to individual behavior, so that some participants may prepare their speech earlier than others.

2. Methods

2.1. Procedure and Participants

The experimental task for the participants was to "shoot" cans projected onto the wall in front of them. They were asked to point at the can with a laser pointer and say the word that was written on the can (either *piff* or *paff*, which are German onomatopoeias for shooting). The participants stood approx. 1 m from the wall with their hands down. They first saw a blank screen. Then, a can was shown and the participants were asked to point to it and say the word. Afterward, an animation of the can falling was played. After a short blank screen, a new can in a different position appeared (for further details, see Ćwiek and Fuchs 2019). Thirty-one female German speakers were recorded. Our preliminary analysis is based on a subset of the data from seven speakers.

2.2. Data Annotation

The motion data, as well as the acoustic data, were recorded simultaneously. We used an Optitrack motion capture system with 12 cameras for recording the movements (Prime 13, with the Motive software, ver. 1.9.0), and a Sennheiser ME 64 cardioid microphone for recording the acoustics. The sampling frequency was 120 Hz for the motion data, and 44.1 kHz for the acoustics. Several markers were placed on the participant's body. Here we will focus on the hand wrist marker of the pointing arm, and the upper lip and jaw markers that were used to calculate the lip distance during speech production.

An example of the annotation is shown in Figure 1. For the acoustics, we focused on the speech onset and offset. Lip distance is calculated as the Euclidian distance for the x, y, and z coordinates between the upper lip and jaw marker. There, the lip closing gesture onset and offset were annotated. In some cases, it was impossible to define a clear lip closing gesture onset, because some speakers also moved their lips during the pause before moving the arm.

The hand movement is labeled on the velocity signal of the wrist marker, using a 20% threshold criterion in MVIEW software (Tiede 2005). We annotated the onset and the target of the hand gesture.

2.3. Data Processing and Analysis

With the time stamps mentioned above, we were able to test the temporal coordination between hand and lip motions. To do so, we calculated two intervals, as shown in Figure 1 (7 and 8). The first interval of interest refers to the difference between the time points of the hand and lip gesture onsets. It was calculated by subtracting the hand gesture onset from the lip closing gesture onset. The second interval shows the difference between the lip gesture onset and the hand gesture target. It was calculated by subtracting the hand gesture target from the lip closing gesture onset.

If the two events – e.g., the lip closing gesture onset (cf. 3 in Figure 1) and the hand gesture onset (cf. 5 in Figure 1), or the lip closing onset (cf. 3 in Figure 1) and the hand gesture target (cf. 6 in Figure 1) – would occur simultaneously, the difference between them would amount to 0. Nonetheless, because of our prediction that the slower motor system should start earlier, we expect the values of interval 7 (cf. Figure 1) to be positive.

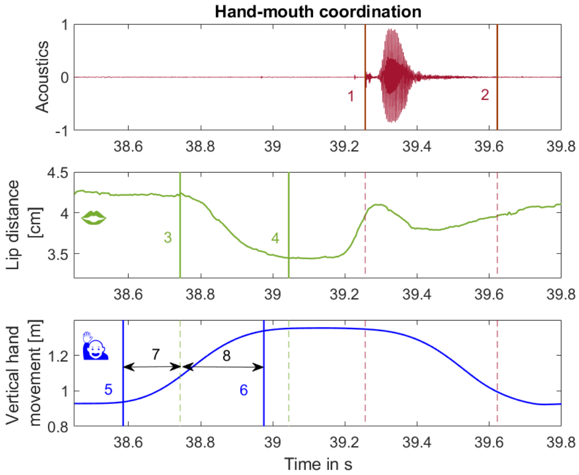

Figure 1: *The plots show the annotations of the data. In the acoustic signal, the speech onset (1), and the speech offset (2) were annotated. As for the lip distance data, we annotated the lip closing gesture onset (3), and the lip closing gesture offset (4). Lastly, we annotated the hand gesture onset (5), and the hand gesture target (6). We subsequently calculated two intervals: hand gesture onset to lip closing onset (7), and lip closing onset to hand gesture target (8).*

As for interval 8 (cf. Figure 1), which tests the coordination between hand gesture target and lip closing gesture onset, we expect speaker behavior to be variable. Also, we wanted to explore the difference between both interval durations.

The data was not normally distributed, therefore we used paired Wilcoxon signed-rank tests to analyze the data. The NA values were excluded leaving a total of $N = 582$ pairs. The values reported further as M refer to the medians. The p-values were corrected for multiple comparisons using Bonferroni correction. We calculated the effect size by dividing the z-score by the square root of the pairs ($N = 582$). All analyses were performed in R (R Core Team 2019), with the `tidyverse` package for data wrangling (Wickham 2017), and `ggplot2` to generate the plots (Wickham 2016).

3. Results

First, the time point of the hand gesture onset was compared with the time point of the lip closing gesture onset. Overall, the hand gesture onset occurred earlier ($M = 69.17$) than the lip closing gesture onset ($M = 69.85$). A Wilcoxon signed-rank test indicated that the difference was statistically significant with $p < .001, z = -20.86$, and the effect size $r = 0.86$. The results for individual speakers are depicted in Figure 2. It shows the duration of the interval between the hand gesture onset and lip closing onset, calculated by subtracting the hand gesture onset from the lip closing onset. It is visible that the values are positive – the lip closing gesture onset has a greater value, thus, occurs only after the hand already started to move. Most notably, speakers 9, 3, and 2 begin the lip closing gesture shortly after beginning hand motion.

Then, the time point of the lip closing onset was com-

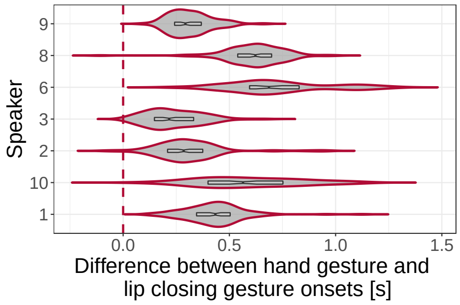

Figure 2: *The difference between the time points of the hand gesture onset and lip closing onset for each speaker; calculated by subtracting the hand gesture onset from the lip closing onset.*

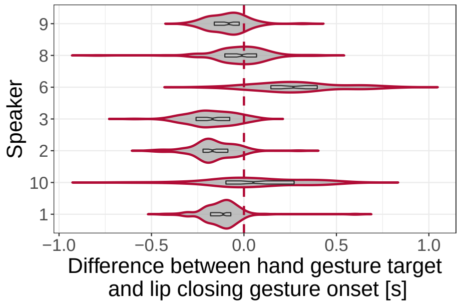

Figure 3: *The difference between the lip closing onset and the hand gesture target for each speaker; calculated by subtracting the hand gesture target from the lip closing onset.*

pared with the time point of the hand gesture reaching its target. The lip closing onset occurred earlier ($M = 69.85$) than the hand gesture target ($M = 69.92$). The Wilcoxon signed-rank test showed a statistically significant difference with $p < .001, z = -10.14$, and the effect size equal to $r = 0.42$. The results for individual speakers are given in Figure 3, which shows greater variability in the speakers' behavior. While speakers 9, 3, 2, and 1 tend to start the articulatory gesture before reaching the hand gesture target, speakers 8, 6, and 10 are more variable and may start the articulatory gesture only after having reached the hand gesture target.

Lastly, the difference between the two intervals described above was calculated: the absolute difference of hand gesture onset to the lip closing onset (cf. 7 in Figure 1), and the absolute difference of hand gesture target to the lip closing onset (cf. 8 in Figure 1). The median for the interval 7 was equal to $M = 0.43$ and for the interval 8 $M = 0.16$. The difference was significant, according to the Wilcoxon signed-rank test, with $p < .001, z = -17.57$, and the effect size of $r = 0.87$. Figure 4 demonstrates the difference between the two intervals. Despite the high inter-speaker variance, the interval 7 is generally longer. Most notably and reliably, it can be seen with speaker 6.

4. Discussion

In the current study, we investigated the coordination of hand and articulatory movements in a task requiring manual precision. Our experimental design allowed for large flexibility. Similar to previous studies, we found that the slower motor system, i.e., the hand motion starts earlier than the motion of the lips. In their study on spontaneous speech, Bekke et al. (2020) reported a delay of 215 ms on average, which is comparable to our data, with speaker-specific variations. Individual differences cannot be attributed to the prediction of upcoming words in our study, because the speech material only consisted of two ideophones that were repeated several times. Three of seven speakers showed a tendency to early anticipation (9, 3, and 2 in Figure 2). Those speakers began the articulatory gesture shortly after beginning the hand gesture. The other speakers started the articulatory gesture later, most notably speakers 8, and 6.

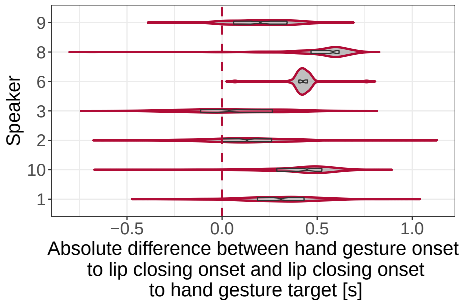

Figure 4: *The difference between the two intervals: the hand gesture onset to the lip closing onset (cf. 7 in Figure 1), and the hand gesture target to the lip closing onset (cf. 8 in Figure 1) for each speaker; calculated by subtracting the absolute difference of 7 from 8.*

These variations in onset delay can be either a result of the task, which allows for the temporal flexibility, or they could also be the consequence of individual differences in the amplitude of the hand motion. Some speakers may elevate their arm/hand to a larger extent in comparison to others, since we did not restrict our participants in any way. Larger movement amplitudes are commonly produced with higher velocity and still take longer (Ostry and Munhall 1985). Thus, speakers who elevate their arms more vigorously may also have to start their hand motion earlier.

The second result of our study revealed that the lip closing gesture for the initial [p] occurred before the hand reached the target in four out of seven speakers (cf. Figure 3). While speakers 9, 3, 2, and 1 tended to start the lip closing gesture before reaching the hand gesture target, other speakers behaved more variably. For instance, speaker 6 tends to produce the task sequentially, i.e., she started to produce the lip closing gesture only after having reached the hand gesture target. Interestingly, the speakers who anticipated the speech gesture early in the analysis shown in Figure 2, are also the ones who begin with the lip closing gesture onset before hitting the target with the hand gesture. The only difference is speaker 1, who, despite a slightly later start of the articulatory gesture, nevertheless managed to begin the lip closing gesture before she reached the target with her hand.

It has to be noted that this behavior might not apply in all contexts. Many scholars studying the hand–mouth coordination use sonorants as initial segments in the data. In our study, the initial segment was a voiceless bilabial stop [p], which allows for inaudible articulatory preparation. Our data reveals that some speakers make use of this possibility and prepare the articulatory gesture in advance.

Finally, our findings show that the interval from the hand gesture onset to the lip closing gesture onset (cf. 7 in Figure 1) is generally longer than the interval from the lip closing gesture onset to reaching hand gesture target (cf. 8 in Figure 1). This suggests closer coordination of the lip closing onset with reaching the target of the hand movement than with the beginning of the hand movement.

The analyses so far are limited as we were only able to analyze a subset of the data (i.e., seven out of 31 speakers). Since in some cases the marker on the laser pointer was hidden, we focused on the marker glued on the dominant arm wrist, which limits the analysis of precise hand movements. However, most studies on arm or hand motions have not decoupled one motion from the other. As a next step, we will analyze the whole data set, and consider body height, as well as the individual amplitude of the hand motion, as additional factors that might explain speaker-specific coordination behavior (i.e., the onset delays).

Any discussions about the hand–mouth coordination, should not only rely on representational aspects but, more importantly, should consider the hand–mouth coordination as the coordination of a slow and a fast motor system. Kinematic properties, like the amplitude of the motion itself, might also be of importance. Thus, the dynamical properties should be an integral part of the discussions about hand–mouth coordination, predictions, and gesture research.

5. Acknowledgements

This work was funded by the German Research Council as a part of the XPrag.de project PSIMS: The Pragmatic Status of Iconic Meaning in Spoken Communication (FU 791/6-1).

6. References

Bekke, Marlijn ter, Linda Drijvers, and Judith Holler (June 2020). "The predictive potential of hand gestures during conversation: An investigation of the timing of gestures in relation to speech". en. In: *Preprint*. DOI: 10.31234/osf.io/b5zq7. URL: https://osf.io/b5zq7 (visited on 01/28/2021).

Chu, Mingyuan and Peter Hagoort (2014). "Synchronization of speech and gesture: Evidence for interaction in action". In: *Journal of Experimental Psychology: General* 143.4. Place: US Publisher: American Psychological Association, pp. 1726–1741. DOI: 10.1037/a0036281.

Ćwiek, Aleksandra and Susanne Fuchs (2019). "Iconic Prosody is Rooted in Sensori-Motor Properties: Fundamental Frequency and the Vertical Space". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 41. Montreal, Canada: Cognitive Science Society, pp. 1572–1578. URL: https://cogsci.mindmodeling.org/2019/papers/0282/0282.pdf (visited on 05/14/2020).

Esteve-Gibert and Pilar Prieto (June 2013). "Prosodic Structure Shapes the Temporal Realization of Intonation and Manual Gesture Movements". In: *Journal of Speech, Language, and Hearing Research* 56.3. Publisher: American Speech-Language-Hearing Association, pp. 850–864. DOI: 10.1044/1092-4388(2012/12-0049). URL: https://pubs.asha.org/doi/abs/10.1044/1092-4388%282012/12-0049%29 (visited on 01/28/2021).

Grimme, Britta, Susanne Fuchs, Pascal Perrier, and Gregor Schöner (Jan. 2011). "Limb versus Speech Motor Control: A Conceptual Review". In: *Motor Control* 15.1, pp. 5–33. DOI: 10.1123/mcj.15.1.5. URL: https://journals.humankinetics.com/view/journals/mcj/15/1/article-p5.xml (visited on 01/28/2021).

Habets, Boukje, Sotaro Kita, Zeshu Shao, Asli Özyurek, and Peter Hagoort (Aug. 2011). "The Role of Synchrony and Ambiguity in Speech–Gesture Integration during Comprehension". In: *Journal of Cognitive Neuroscience* 23.8, pp. 1845–1854. DOI: 10.1162/jocn.2010.21462.

Iverson, Jana M and Esther Thelen (1999). "Hand, mouth and brain. The dynamic emergence of speech and gesture". en. In: *Journal of Consciousness Studies* 6.11-12, pp. 19–40.

Krivokapić, Jelena, Mark K. Tiede, and Martha E. Tyrone (2017). "A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection". In: *Laboratory phonology* 8.1. DOI: 10.5334/labphon.75. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5472837/ (visited on 08/26/2020).

Levelt, Willem J.M, Graham Richardson, and Wido La Heij (Apr. 1985). "Pointing and voicing in deictic expressions". en. In: *Journal of Memory and Language* 24.2, pp. 133–164. DOI: 10.1016/0749-596X(85)90021-X. URL: https://linkinghub.elsevier.com/retrieve/pii/0749596X8590021X (visited on 01/28/2021).

MacNeilage, Peter F. (Aug. 1998). "The frame/content theory of evolution of speech production". en. In: *Behavioral and Brain Sciences* 21.4, pp. 499–511. DOI: 10.1017/S0140525X98001265. URL: https://www.cambridge.org/core/product/identifier/S0140525X98001265/type/journal_article (visited on 01/28/2021).

Ostry, David J. and Kevin G. Munhall (Feb. 1985). "Control of rate and duration of speech movements". In: *The Journal of the Acoustical Society of America* 77.2. Publisher: Acoustical Society of America, pp. 640–648. DOI: 10.1121/1.391882. URL: https://asa.scitation.org/doi/abs/10.1121/1.391882 (visited on 01/30/2021).

Pouw, Wim and James A. Dixon (2019). "Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback". en. In: *Cognitive Science* 43.3, e12721. DOI: 10.1111/cogs.12721. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12721 (visited on 03/21/2019).

Pouw, Wim, Steven J. Harrison, and James A. Dixon (Feb. 2020). "Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony." en. In: *Journal of Experimental Psychology: General* 149.2, pp. 391–404. DOI: 10.1037/xge0000646. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000646 (visited on 01/28/2021).

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rochet-Capellan, Amélie, Rafael Laboissière, Arturo Galván, and Jean-Luc Schwartz (Dec. 2008). "The Speech Focus Position Effect on Jaw–Finger Coordination in a Pointing Task". In: *Journal of Speech, Language, and Hearing Research* 51.6. Publisher: American Speech-Language-Hearing Association, pp. 1507–1521. DOI: 10.1044/1092-4388(2008/07-0173). URL: https://pubs.asha.org/doi/full/10.1044/1092-4388%282008/07-0173%29 (visited on 08/26/2020).

Stoltmann, Katarzyna and Susanne Fuchs (2017). "Syllable-pointing gesture coordination in Polish counting out rhymes: The effect of speech rate". en. In: *Journal of Multimodal Communication Studies. Special issue: Gesture and Speech in Interaction* 4.1. Ed. by Silva Bonacchi and Maciej Karpiński, pp. 63–68.

Tiede, Mark (2005). *MVIEW: software for visualization and analysis of concurrently recorded movement data*. New Haven, CT.

Wickham, Hadley (June 2016). *ggplot2: Elegant Graphics for Data Analysis*. en. Google-Books-ID: XgFkDAAAQBAJ. Springer.

Wickham, Hadley (2017). *tidyverse: Easily install and load the "tidyverse."* https://CRAN.R-project.org/package=tidyverse.