Multimodal Entrainment and Feedback in Online Group Meetings

Patrizia Paggio[1], **Manex Agirrezabal**[1], **Giulia Di Cristina**[2], **Bart Jongejan**[1] **and Costanza Navarretta**[1]
[1]University of Copenhagen
[2]University of Turin
{paggio,manex.aguirrezabal,bartj,costanza}@hum.ku.dk
giulia.dicristina@edu.unito.it

Abstract

This paper presents the results of a study on multimodal speaker behaviour in a corpus of online Zoom meetings. We investigate two questions: i) whether speakers display a higher degree of head movement when they exchange verbal feedback than when they don't, as would be expected if verbal and gestural feedback reinforce one other, and ii) whether they move more or less similarly under the same conditions. Several linear mixed models were fitted to test the difference in head movement values in target and control intervals of two different durations. The results indicate that speakers indeed entrain by moving their heads more in target intervals where verbal feedback is present. This result confirms our expectations. However, speakers also appear to move in less similar ways in the same target intervals. This dissimilarity can be explained by the fact that not all speakers give the same type of gestural feedback, but also by noise created by non-communicative movements in which speakers adjust their positions or reach out for objects during the meeting.

Keywords: speaker entrainment, feedback behaviour, head movement

1. Introduction

In their seminal paper, Pickering and Garrod (2004) proposed an interactive alignment account of dialogue according to which speakers align their linguistic behaviours as well as the underlying representations at many levels as a consequence of a need to simplify language processing while interacting. Since then, a body of research from linguistics and gesture studies have investigated the phenomenon of temporal coordination across individuals, for which other terms have been used including mimicry, interpersonal coordination, synchronisation and entrainment (see Rasenberg et al., 2020 for an overview).

In fact, temporal coordination may concern not only communicative behaviours in the speech and gestural modalities, but also physiological functions such as heart rate, electrodermal activity and pupillary response, as well as coordination of neural activity (Gordon et al., 2025). We know that the type of communication, in particular whether a specific task is involved, has an effect on the amount of temporal coordination during face-to-face communication. For instance, gazing towards the same region of interest, or repeating words and gestures, may be necessary to ground the interaction by establishing joint focus of attention and common reference (Rasenberg et al., 2022). In general, successful synchronisation with others is crucial for people to establish and carry out social interaction.

To our knowledge, however, the topic has not been investigated in the context of online interaction, in particular virtual group meetings, in spite of the fact that group meetings since the COVID-19 pandemic have become a very common way for people to interact in work-related as well as social contexts. But since corpora of online group meetings have begun to be available (Reverdy et al., 2022; Paggio et al., 2024), the time seems ripe to start addressing the question to what degree speakers entrain in such a context.

In this study, we investigate head movement entrainment between speakers in a collection of online Zoom meetings. The velocity of a number of six different keypoints in the head are used to predict the overall energy among participants and the differences between them.

This article is structured as follows. We review relevant literature in the next section. After that, we present the methods employed in our experimentation. Then, results are presented and discussed in the next two sections. At the end, we conclude the paper and we suggest some possible future directions.

2. Related work

Temporal coordination across speakers has been investigated from different perspectives looking not only at speech, but also at gestural and facial behaviour. Investigated aspects include the effect of interaction type (free vs. task-oriented) on lexical and syntactic alignment (Dideriksen et al., 2019), entrainment at the level of prosodic features in dyads (Levitan and Hirschberg, 2011) and groups (Litman et al., 2016), mirroring of facial expressions in speaker dyads (Navarretta, 2016), co-speech gesture mimicry (Holler and Wilkin, 2011), self-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9139–9149
11–16 May 2026.

alignment (Bergmann and Kopp, 2012) and eye blinking entrainment between speaker and listener (Nakano and Kitazawa, 2010).

There are two fundamentally different ways of looking at temporal coordination. The approach that seems to dominate in linguistics-oriented studies investigates the repetition of discrete elements (words, gestures, head movements, etc.) between or within speakers over temporal sequences (Louwerse et al., 2012). In contrast, in studies based on the automatic extraction of features from video-recorded interaction, the phenomenon is modelled in terms of continuous variables, for example using prosodic features (Levitan and Hirschberg, 2011; Litman et al., 2016) or, if movement behaviour is studied, motion capture measurements (Béres et al., 2026) and motion energy values (Khosrobeigi et al., 2025). The term *entrainment* has been used in the latter approaches to refer to the convergence of patterns of behaviour across speakers. In this sense, which is the one we embrace in this study, entrainment refers to a kind of interpersonal synchrony driven by social interaction as opposed to similarity of responses to an external stimulus – a different sense of the term used in neurocognition (Hamilton et al., 2025).

Investigations of entrainment employ a range of methods to model convergence across speakers and make different predictions about which factors affect inter-speaker coordination. Litman et al. (2016), which was the initial source of inspiration for the present study, analysed entrainment in task-oriented group dialogues. Prosodic features were extracted from the audio recordings and used to compute group-level partner differences in various phases of the interaction (of 3-7 minutes duration). The study found that the difference across partners decreased for some of the features, in other words, entrainment could be observed in the final part of the interactions, probably in parallel with the group converging on a task solution.

Khosrobeigi et al. (2025) analysed the temporal flow of motion energy (ME) in dialogues by using alternating lagged correlation tests on consecutive 0.3 s windows and Granger causality tests. The goal was to investigate whether speaker dominance affected entrainment. The study showed that dominant speakers are in fact more likely to lead motion dynamics, in other words to drive the coordination, though spans of mutual influence also occur.

Trujillo et al. (2023) investigated entrainment at various linguistic and kinematic levels in affiliative and task-oriented Danish and Norwegian conversations. They used turns as units and applied dynamic time warping to determine the entrainment of head and hand movements. Kinematic entrainment was highest in task-oriented conversations.

Studying temporal coordination implies of course some kind of time series analysis since the goal is to find recurrent patterns or correlations of values in temporal windows. Hamilton et al. (2025) gives an overview of the different timescales at which different phenomena can be observed to entrain. For example, latencies of 0.6-1.5 s have been reported for coordinated nods or facial expressions between two speakers (Hale et al., 2020; Louwerse et al., 2012) while coordination of more complex behaviours like deictic gestures or certain speech acts reach peak lags of 25 s (Louwerse et al., 2012). An argument for the usefulness of considering longer units can also be found in Inden et al. (2013), who investigated multimodal backchanneling and found that feedback head movements tended to occur at the end of a speaker's utterance and extended over the subsequent utterance pause.

In this study of multimodal entrainment in online meetings, we build on Litman et al. (2016)'s method by looking at entrainment between speakers in a group. Rather than looking at whether the group entrains as the conversation proceeds, however, we focus on specific spans in which speakers are coordinating their verbal behaviour by exchanging feedback words. This is motivated by the fact that naturally occurring meetings like the ones we analyse in this paper, do not necessarily have a primary task which may drive participant entrainment. They are instead organised around a sequence of discussion items such that participants may agree or disagree with each other, and possibly show entrainment, at several points during the interaction. We hypothesise that entrainment, with specific focus on head movement, will be observed in the visual modality in these spans as opposed to other parts of the interaction in which no feedback is exchanged, in other words we expect verbal and non-verbal behaviour to contribute together to inter-speaker entrainment.

3. Methods

3.1. The corpus

We perform our experimentation on the GEHM Zoom meeting corpus (Paggio et al., 2024).[1] This corpus comprises a set of 12 video recordings of meetings held on Zoom, with participants of different nationalities. For each meeting, the independent video and sound recording is available for each participant, together with the transcription of their speech. These meetings have an average duration of 40 minutes and the number of participants ranges between 5 and 9. The language of the meetings is English, which some people speak as their native and others as a second language. Two

[1] https://archive.org/details/GEHM_meeting_corpus

of the meetings were excluded from the analysis for this study because of extensive screen sharing during the interaction.

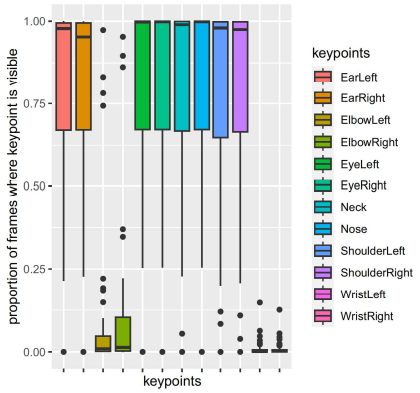

Figure 1: Distribution of visual keypoint values in the GEHM corpus, reproduced from Paggio et al. (2024)

The corpus is publicly available[2] complete with speech transcriptions and position coordinates of body keypoints, i.e. nose, eyes, ears, neck, shoulders, elbows and wrists obtained using OpenPose (Cao et al., 2017). Figure 1, which is reproduced from Paggio et al. (2024), shows the distribution of keypoint values in the corpus.

3.2. Units of analysis

To model the effect of exchanged verbal feedback on movement entrainment, we compare speakers' motion energy in intervals where feedback words occur (target intervals) against those in which they don't (control intervals). To be more precise, we define as target intervals those in which at least two speakers utter a feedback word. The list of feedback words includes the following: *yes, yeah, okay, no, ok, hmm, mm, mhm, oh, right, fine, sure, good, nice,* and *great*. We experiment with target and control intervals of two different durations, i.e. 15 and 30 s. The decision to experiment with these durations was reached based on empirical observations to maximise the number of target intervals in which 2-4 speakers exchanged verbal feedback. In addition, we looked at latencies between feedback words to gather more evidence of the validity of the chosen spans.

Feedback word latency was defined as the time between the onset of the first feedback word and the onset of the last subsequent feedback word, with and without intervening non-feedback words. In the first case, latency is between 0.12 and 12.45 s. In the second case, it ranges from 0.1 to 7.49 s. For comparison, we also measured the latency of feedback words in six face-to-face conversations involving four participants from the AMI corpus (Carletta et al., 2005) using the same methodology. In those data, the latency of feedback words with intervening non-feedback words ranges from 0.12 to 10.32 s, while the latency with non intervening words is between 0.1 and 5.86 s.

Given these numbers, and considering lags between gestural behaviour and any co-occurring words especially if several speakers are involved (Louwerse et al., 2012; Inden et al., 2013), it does not seem unreasonable to choose intervals of 15 s corresponding to 3-4 subsequent feedback words with intervening latencies of about 5 s. Doubling this duration to 30 s gives us a sufficiently different term of comparison to validate our analyses.

3.3. Movement features

Head movement entrainment between speakers is investigated by means of a subset of the visual keypoint values available in the corpus. We work with continuous variables for two reasons: i) we wanted to apply to visual coordinates a method similar to that used by Litman et al. (2016) for prosodic analysis (theoretically, both prosody and gesturing are suprasegmental phenomena); ii) there are no discrete labels (annotated head movements) in our dataset.

For each speaker and each video frame, we consider x and y coordinate values for six different visual keypoints relating to head movements (Nose, Neck, Left and Right Eye, Left and Right Ear), we compute the first derivative (velocity) of these values and square it to obtain *energy* values for each keypoint coordinate. From these energy values, two different kinds of measure are then calculated for each interval (both target and control). The first measure is the average energy per speaker for each keypoint coordinate. The other is the average difference across all speaker pairs.

In sum, we have four different datasets with movement values referring to either individual speakers or speaker pairs. Note that the total number of data points examined varies depending on the duration of the intervals as well as whether the averages being compared refer to individual speakers (for energy levels) or speaker pairs (for speaker differences). Note also that the 10 meetings have different lengths and different number of speakers involved. The total data points for each dataset (and type of analysis) are displayed in Table 1.

The two types of measures we consider allow us to look at potential entrainment in the target intervals in two different ways. In the case of average energy per speaker, if it is true that verbal and gestural feedback go hand in hand, we would expect

[2]See https://archive.org/details/GEHM_meeting_corpus.

Dataset	Object	Data points
Energy level 15 s	Individual speakers	92,304
Energy level 30 s	Individual speakers	46,032
Energy diff 15 s	Speaker pairs	112,752
Energy diff 30 s	Speaker pairs	59,628

Table 1: Total data points examined in the four datasets used

the overall motion energy displayed in target intervals to be higher than in the controls. In the case of motion energy difference across speakers, however, we can only expect diminished differences in the target intervals if all speakers move similarly.

3.4. Statistical methods

We used *R* (R Core Team, 2024) and the *lme4* package (Bates et al., 2015) to create two sets of linear mixed effect (LME) models. In the first set, we test the effect of a number of different variables on the total energy levels in video intervals while considering the two interval durations we have defined. In the second set of models we test the effect of a number of variables on the differences across speaker pairs, again in intervals of two different durations. In both sets of models, the movement values used (either speaker-specific motion energy values or values reflecting the difference in energy across speakers) are log-transformed to approximate normality (Curran-Everett, 2018). Maximum likelihood was used to fit the models. Significance was established in all cases by comparing the full model with a model without the predictor under scrutiny (Winter, 2013).

The variables used in all the models are summarised in Table 2. Most variable names should be self explanatory with two possible exceptions. In the datasets used to predict energy, the variable *fb_speakers* refers to the number of speakers uttering feedback words in a given interval. Similarly, *fb* is used by the models predicting energy differences to refer to the number of speakers uttering feedback words in a given speaker pair.

Dependent		log_energy
Fixed effects	Model1:	interval, keypoint
	Model2:	fb_speakers, keypoint
	Model3:	fb_speakers, direction
Random effects		meeting, speaker
Dependent		log_diff
Fixed effects	Model1:	interval, keypoint
	Model2:	fb, keypoint
	Model3:	fb, direction
Random effects		meeting, pair

Table 2: Predictors used in the mixed linear models

Datasets and statistical code are available from Open Science Framework (OSF).

4. Results

Table 3 shows for the two types of interval considered, of 30 and 15 s respectively, the total number of meetings, the total number of target intervals and control targets, as well as the mean (sd) number of speakers engaged in verbal feedback in the target intervals.

Duration	No. meetings	No. targets (FB+)	No. controls (FB-)	Mean (sd) FB sp.
30 s	10	484	85	2.22 (0.45)
15 s	10	559	469	2.12 (0.34)

Table 3: Analysed intervals of different durations

Table 4 shows, again for intervals of 15 and 30 s, mean and sd values for motion energy levels and energy difference across speakers in target and control intervals.

	15 s		30 s	
	FB+	FB-	FB+	FB-
Energy (mean)	0.960	0.855	0.967	0.806
Energy (sd)	1.902	1.797	1.634	1.477
Energy diff (mean)	0.018	0.015	0.018	0.014
Energy diff (sd)	0.033	0.031	0.028	0.024

Table 4: motion energy levels and differences in target (FB+) and control (FB-) intervals of two durations

As expected, the overall energy level is higher in the target intervals than in the control ones. When we consider energy differences across all speakers, we see that this difference is slightly higher in the target intervals than the control ones. For both types of measure, the difference between target and control spans is slightly higher in the 30 s intervals although the figures for the targets are quite similar for both durations.

We tested the differences between target and control intervals for intervals of both durations and obtained very similar results. For brevity, we report here only the results referring to the shorter, 15 s duration.

4.1. Predicting energy values

The linear mixed model we fitted to predict energy from interval and keypoint, with speaker and meet-

ing as random effects, showed a significance effect of interval. This main effect is visualised in Figure 2.

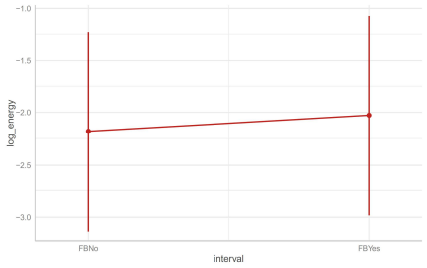

Figure 2: Predicted motion energy values from interval

Most of the keypoint values added significantly to the effect achieved by the presence of verbal feedback. The neck, however, diminished this effect, as did horizontal movement of the right ear. A plot of these effects is available in the Appendix (plot at the top of Figure 8). All effect plots were generated by means of the *sjPlot* (Lüdecke, 2025) and *ggeffects* packages (Lüdecke, 2018). Comparing this model with one without the interval predictor using an anova test confirmed the significant result (χ^2 84.56, *df* 1, $p < 0.0001$).

The second model we fitted predicts energy from the number of speakers exchanging feedback words (*fb_speakers*), which ranges from none to four, as well as keypoint. Speaker and meeting are still treated as random effects.

We observe in Figure 3 (upper plot) that each additional speaker engaging in verbal feedback adds a significant effect. As in the first model, all keypoints add to these effects with the exception of the neck and the horizontal movement of the right ear (the effects are shown in the Appendix, plot in the middle of Figure 8). Comparing this model with one without the *fb_speakers* predictor using an anova test confirmed the significant result (χ^2 150.39, *df* 4, $p < 0.0001$).

To get a more general impression of the way movement keypoints contribute to the differences between target and control intervals, we fitted a third model in which keypoint values are summarised in terms of horizontal and vertical movement. Thus, energy is now predicted based on fb_speakers and movement direction (as well as their interaction). The random effects are kept unchanged. The interaction between the two predictors is shown in the bottom plot of Figure 3.

We see that the highest levels of energy are consistently provided by speakers moving along the horizontal axis. A more detailed plot is again shown in the Appendix (plot at the bottom in Figure 8). A comparison of this model with one only using direction but not fb_speakers as predictors showed significance difference (χ^2 148.03, *df* 8, $p < 0.0001$).

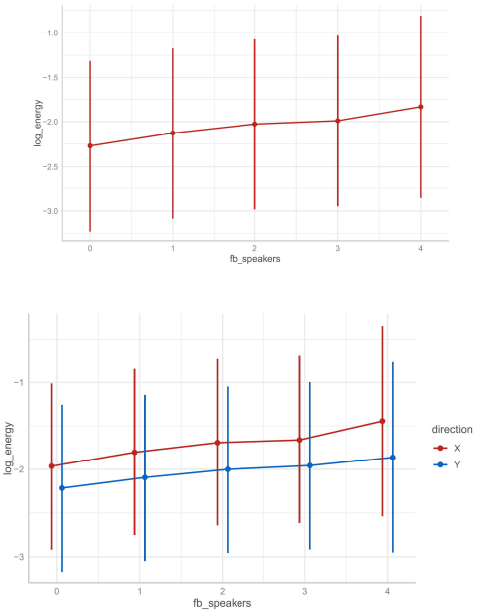

Figure 3: Predicted motion energy values from number of speakers uttering feedback words. The graph at the bottom shows the interaction with movement direction, where x=horizontal and y=vertical.

4.2. Predicting energy differences

Similarly to what done to predict energy levels, three models were fitted to test the effect of interval type on average motion energy differences between speakers. In the first one, the difference in energy across speaker pairs is predicted given the interval. We know already that the motion energy level is increased in target intervals. However, we also see that the difference in energy between speakers increases significantly, as shown in Figure 4. The effects of the various keypoints for each of the three models are available in the Appendix (Figure 9).

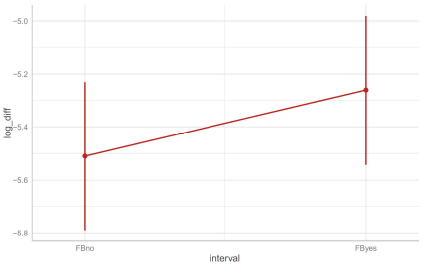

Figure 4: Predicted motion energy differences from interval

The significance of the effect provided by the interval type was confirmed by comparing this model with a similar one without the interval predictor (χ^2 695.42, *df* 1, $p < 0.0001$).

In the second model we test the effect of the number of speakers uttering feedback words (*fb*) on energy differences. In this dataset, however, we are considering speaker pairs, so the relevant values are 0-2. This effect is significant, and grows with the number of speakers, as visualised in the top plot of Figure 5. Significance was confirmed by the comparison with a model without the *fb* predictor (χ^2 1362.1, *df* 2, $p < 0.0001$).

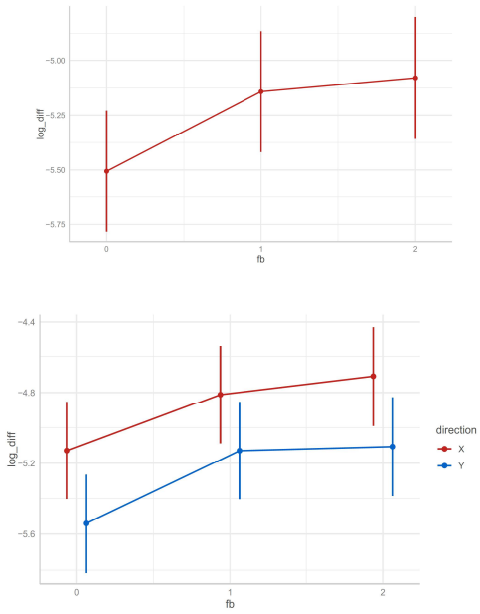

Figure 5: Predicted motion energy differences from number of speakers uttering feedback words in each speaker pair. The graph at the bottom shows the interaction with movement direction, where x=horizontal and y=vertical.

In the third model we look at the combined effect of number of speakers uttering feedback words in each speaker pair and the direction of the movement, which is shown in the bottom plot of Figure 5. We saw for level of energy that horizontal movement provided increased values. Here we see that it also provides increased differences across the speakers. This model was compared with *direction* as the only predictor, and the difference between the two confirmed the significance of the effect created by *fb* (χ^2 1308.7, *df* 4, $p < 0.0001$).

5. Discussion

Based on the results of our models, it would seem that our initial expectation holds true. Speakers engaged in verbal feedback show motion entrainment in the sense that, on average, they tend to move more than if they are not giving each other feedback. We see examples of this in several intervals in which several speakers show agreement with the speaker by nodding simultaneously while feedback words are exchanged. On closer scrutiny, we see that part of this movement is due to the meeting participants adjusting their focus to look at the speaker providing verbal feedback, thus increasing the overall energy for a short stretch of time.

Our results also show that the difference between speakers increases in the target intervals. To understand what lies behind these figures, we analysed qualitatively the most representative intervals, in other words those showing the most and the least difference between the speakers.

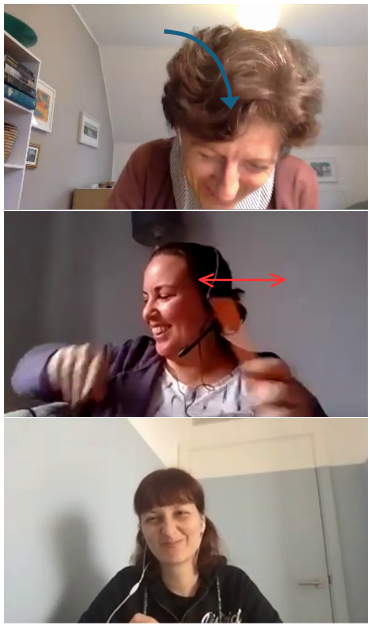

Figure 6: Different movements in the same target interval

In some cases, differences are due to the fact that speakers in a target interval show feedback in different ways. In a study on multimodal feedback behaviour on data from 14 participants, Blomsma et al. (2024) also found that there is a large variability in the number and type of feedback signals provided by the participants to the same speaker's utterance. A relevant example from our dataset is shown in Figure 6: The speaker at the top suggests to change the date of the next meeting while laughing and producing a forward vertical head movement. Simultaneously, another speaker moves her head horizontally. Both movements are shown by superimposed arrows in the images. The third participant at the bottom remains stationary and reacts just by laughing. Feedback words are uttered.

However, increased differences in motion are not necessarily related to the feedback mechanism. A large difference between speakers can also be caused by strong movements with no communicative intent (as when somebody moves laterally to adjust their position or to reach out for an object during a meeting). An example is shown in Figure 7.

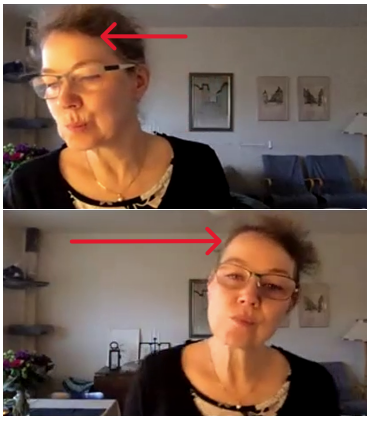

Figure 7: A speaker moving horizontally to adjust her position

A case like this affects the energy level as well as the potential difference between speakers. However, it may happen in a target interval without having anything to do with feedback. This kind of behaviour may also be at least partly responsible for the additive effect of horizontal movement direction shown in two of our models. Ideally, such behaviour should be filtered out not to affect the analysis of target intervals.

Conversely, small differences are not necessarily due to the fact that speakers are moving similarly, but are also seen in situations where people move very little, especially in the control intervals. Such a case is not problematic for the models predicting energy levels since lack and presence of movement are clearly distinguished. For the models predicting motion energy difference, however, there is no way to discern situations where speakers are not moving, or moving very little, from those in which they are moving in similar ways: in both situations, the difference will be small.

6. Conclusion

In this paper we analysed head movement entrainment between participants in a set of Zoom meetings. We did this by predicting the amount of energy across the participants and the differences between these energy levels in speaker pairs in video intervals where verbal feedback is being uttered as opposed to control intervals in which there is no verbal feedback. Based on our experiment results, we argued that speakers tend to move their heads more when they exchange verbal feedback than when they don't. We also observed that in the target intervals, the differences between the speakers increase: they move in different ways.

As mentioned above, participants in meetings tend to perform involuntary physical movements not necessarily related to feedback or other interaction behaviour, and these movements may lead to noise in our analysis. In order to model the communicative nature of the movements in a cleaner way, a possible future direction would be to develop methods to filter out these movements, e.g. based on their amplitude. There is evidence that involuntary, non-communicative movements should be recognisable from kinematic clues only (Kendon et al., 1980; Trujillo et al., 2018; Derchi et al., 2023) or from eye-gaze patterns (Trujillo et al., 2018). In our videos, it seems that amplitude and duration of the movements, in combination with head pose, could be used to detect head movements where a speaker is busy doing something which is not directly related to the meeting, e.g. looking at or reaching out for something on their desk as may be the case for the movement shown in Figure 7.

The experimentation in this work was done by splitting the videos in fixed-time bins (15 or 30 s). We attempted to capture the gestures using continuous values obtained from OpenPose. It may have happened, though, that some gestures were split between two different bins. We would like to expand this study to using sliding windows, which would also increase the number of target and control data points.

We would also like in future to add prosody measurements to the analysis in order to model speaker entrainment in terms of not only words and visual movement features, but also including the effect of features relating to intensity and pitch.

Finally, our analysis method was applied to a relatively small dataset. Therefore, in order to validate its usefulness, we would like to apply it to other data collections from online (Reverdy et al., 2022) as well as in-person meetings (Koutsombogera and Vogel, 2018). In fact, although we were particularly interested here in analysing online interaction, we see no a priori reason why our methodology should not also be applicable to other group meeting data.

7. References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Luca Béres, Péter Nagy, Tibor Pólya, Béla Weiss,

Ádám Boncz, and István Winkler. 2026. Interpersonal coordination in communication: Effects of alignment in multiple modalities on objective and subjective task outcomes. *Frontiers in Psychology*, 17:1655164.

Kirsten Bergmann and Stefan Kopp. 2012. Gestural alignment in natural dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34.

Peter Blomsma, Julija Vaitonyté, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16(4):1158–1181.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Douglas Curran-Everett. 2018. Explorations in statistics: the log transformation. *Advances in Physiology Education*, 42(2):343–347. PMID: 29761718.

CC Derchi, E Mikulan, A Mazza, S Casarotto, A Comanducci, M Fecchio, J Navarro, G Devalle, M Massimini, and C Sinigaglia. 2023. Distinguishing intentional from nonintentional actions through EEG and kinematic markers. *Scientific Reports*, 13(1):8496.

Christina Dideriksen, Riccardo Fusaroli, Kristian Tylén, Mark Dingemanse, and Morten H Christiansen. 2019. Contextualizing conversational strategies: backchannel, repair and linguistic alignment in spontaneous and task-oriented conversations. In *CogSci'19*, pages 261–267. Cognitive Science Society.

Ilanit Gordon, Alon Tomashin, and Oded Mayo. 2025. A theory of flexible multimodal synchrony. *Psychological Review*, 132(3):680–718.

Joanna Hale, Jamie A. Ward, Francesco Buccheri, Dominic Oliver, and Antonia F. de C. Hamilton. 2020. Are You on My Wavelength? Interpersonal Coordination in Dyadic Conversations. *Journal of Nonverbal Behavior*, 44(1):63–83.

Antonia Hamilton, Victoria Southgate, Kamilla Miskowiak, Anne J Bjertrup, Arianna S Lomoriello, Sara De Felice, Rui Liu, and Ivana Konvalinka. 2025. Rhythms of interaction – the timescales of social synchrony and why they matter.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35:133–153.

Benjamin Inden, Zofia Malisz, Petra Wagner, and Ipke Wachsmuth. 2013. Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, page 181–188, New York, NY, USA. Association for Computing Machinery.

Adam Kendon et al. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication*, 25(1980):207–227.

Zohreh Khosrobeigi, Maria Koutsombogera, and Carl Vogel. 2025. Methods and findings in the analysis of alignment of bodily motion in cooperative dyadic dialogue. *Multimodal Technologies and Interaction*, 9(6).

Maria Koutsombogera and Carl Vogel. 2018. Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech*.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The TEAMS corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

Max M Louwerse, Rick Dale, Ellen G Bard, and Patrick Jeuniaux. 2012. Behavior matching in multimodal communication is synchronized. *Cognitive science*, 36(8):1404–1426.

Daniel Lüdecke. 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26):772.

Daniel Lüdecke. 2025. *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.9.0.

Tamami Nakano and Shigeru Kitazawa. 2010. Eyeblink entrainment at breakpoints of speech. *Experimental brain research*, 205:577–581.

Costanza Navarretta. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 469–474.

Patrizia Paggio, Manex Agirrezabal, Costanza Navarretta, and Leo Vitasovic. 2024. Multimodal behaviour in an online environment: The GEHM Zoom corpus collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11890–11900, Torino, Italia. ELRA and ICCL.

Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science*, 44(11):e12911.

Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. *Discourse Processes*, 59(3):209–236.

Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R. Cowan, and Naomi Harte. 2022. RoomReader: A multimodal corpus of online multiparty conversational interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2517–2527, Marseille, France. European Language Resources Association.

James P Trujillo, Christina Dideriksen, Kristian Tylén, Morten H Christiansen, and Riccardo Fusaroli. 2023. The dynamic interplay of kinetic and linguistic coordination in Danish and Norwegian conversation. *Cognitive Science*, 47(6):e13298.

James P Trujillo, Irina Simanova, Harold Bekkering, and Asli Özyürek. 2018. Communicative intent modulates production and comprehension of actions and gestures: A kinect study. *Cognition*, 180:38–51.

Bodo Winter. 2013. Linear models and linear mixed effects models in R with linguistic applications.

Appendix

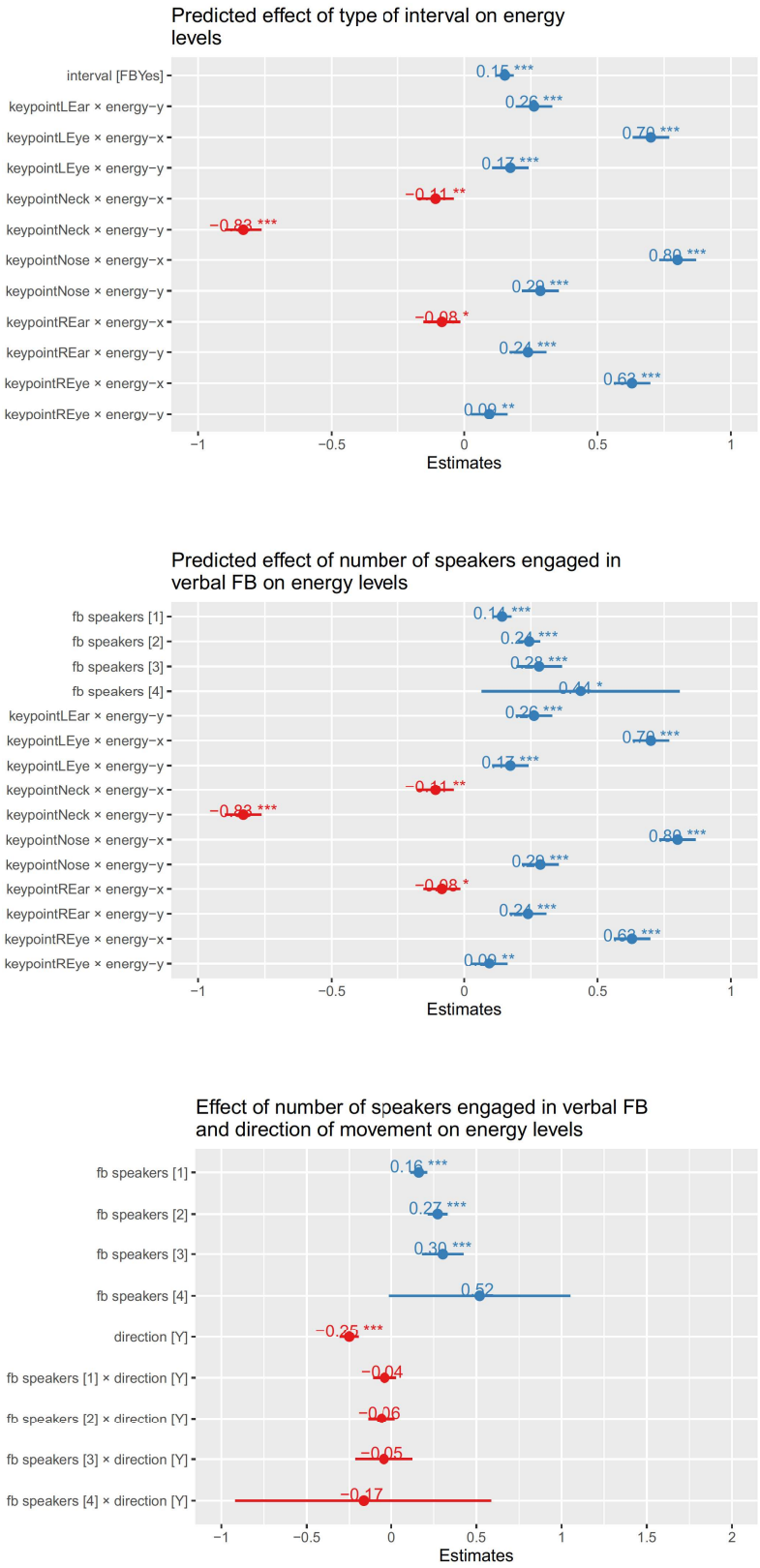

Figure 8: Visualisation of effects for models predicting speakers' overall energy levels

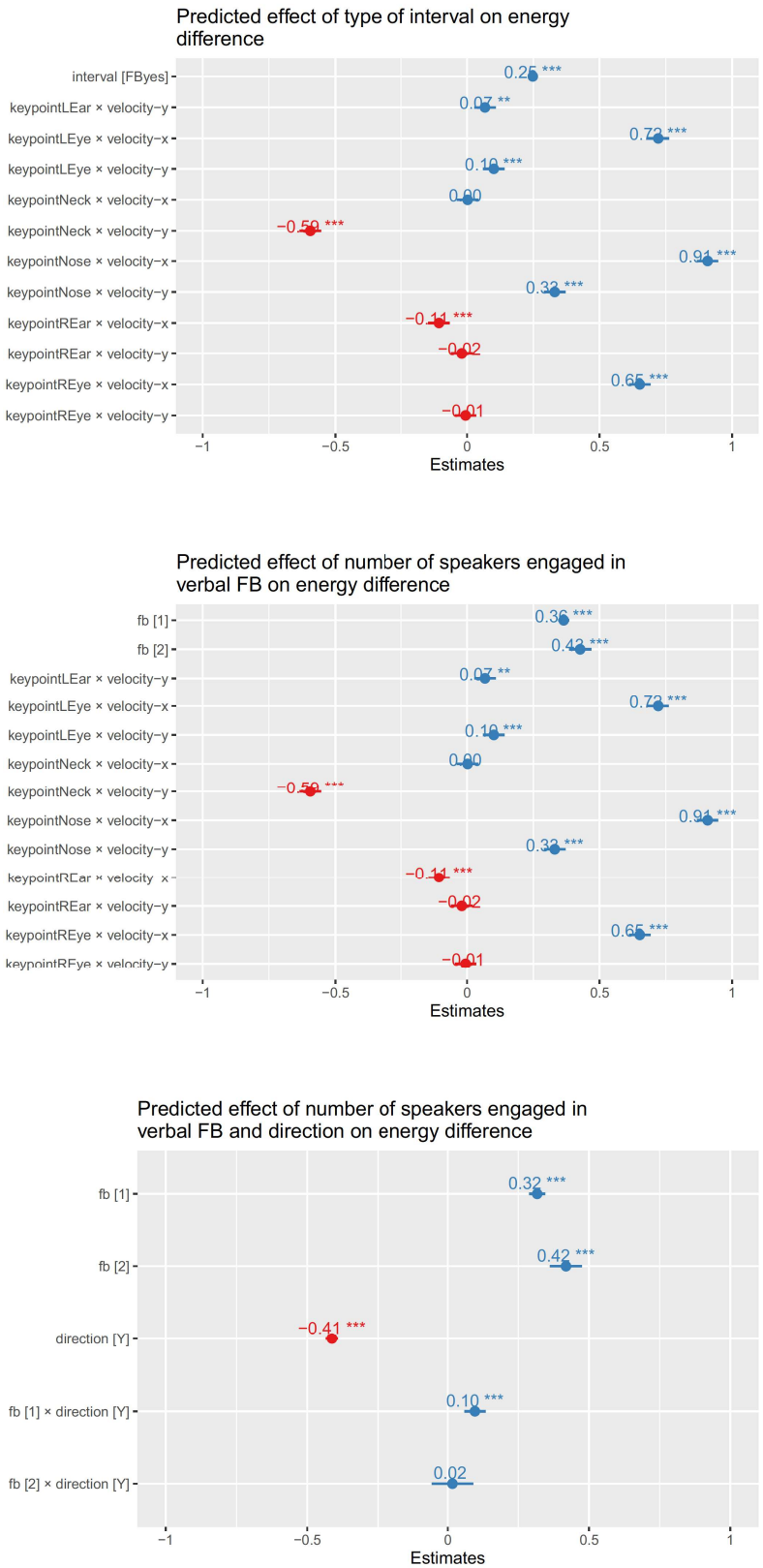

Figure 9: Visualisation of effects for models predicting difference of energy level across speakers