Multimodal behaviour in an online environment: The GEHM Zoom corpus collection

Patrizia Paggio[1,2], **Manex Agirrezabal**[1], **Costanza Navarretta**[1], **Leo Vitasovic**[1]
[1]University of Copenhagen, Centre for Language Technology
[2]University of Malta, Institute of Linguistics and Language Technology
{manex.aguirrezabal,paggio,costanza}@hum.ku.dk
leo.vitasovic@icloud.com

Abstract

This paper introduces a novel multimodal corpus consisting of 12 video recordings of Zoom meetings held in English by an international group of researchers from September 2021 to March 2023. The meetings have an average duration of about 40 minutes each, for a total of 8 hours. The number of participants varies from 5 to 9 per meeting. The participants' speech was transcribed automatically using WhisperX, while visual coordinates of several keypoints of the participants' head, their shoulders and wrists, were extracted using OpenPose. The audio-visual recordings will be distributed together with the orthographic transcription as well as the visual coordinates. In the paper we describe the way the corpus was collected, transcribed and enriched with the visual coordinates, we give descriptive statistics concerning both the speech transcription and the visual keypoint values and we present and discuss visualisations of these values. Finally, we carry out a short preliminary analysis of the role of feedback in the meetings, and show how visualising the coordinates extracted via OpenPose can be used to see how gestural behaviour supports the use of feedback words during the interaction.

Keywords: multimodal corpus, Zoom meetings, visual keypoints

1. Introduction

This paper describes a new multimodal corpus for the study of online group communication in a real-life setting developed under the auspices of the international network on Gesture and Head Movement in Language (GEHM). The network is a cooperation among leading researchers from six European countries working in the area of gesture and language at nine different European universities and research bodies.

Due to the restrictions on social interaction imposed by the COVID-19 pandemic worldwide, and to the necessity of cutting CO2 emissions deriving from travelling, we have seen an increase in the use of video conferencing for group meetings, international conference organisation and teaching (Pratama et al., 2020). In fact, the meetings of our network also had to be held online for almost the entire duration of the research collaboration.

While the practical and environmental advantages of holding meetings online are evident, there are also disadvantages with interaction and communication taking place through videoconferencing. Several studies have discussed pros and cons especially, but not exclusively, in connection with teaching (Chen et al., 2021a,b; Yarmand et al., 2021). Fatigue and lack of engagement, for example, are certainly possible consequences (Bailenson, 2021; Fauville et al., 2021). Another issue is the difficulty of gauging interlocutors' responses in large online meetings (Koh et al., 2022).

From a more general perspective, it is not clear that communication takes place in the same way and with the same ease online as in physical presence. In particular from our perspective, empirical evidence of the way gesture and speech are used in online meetings is scarce[1]. For instance, we need to have a better understanding of the way speakers establish common ground (Clark, 1996), regulate turn taking (Sacks et al., 1978), react to each other to give and receive feedback (Allwood et al., 1992), and coordinate their behaviours (Pickering and Garrod, 2004) when interacting online, to name a few important aspects of communication that have been studied extensively in face-to-face communication.

Since little is understood about the nature of online vs in-person interactions, we believe we need data showing online interaction in different settings and different communicative situations. Online meetings, as we argue in the paper, are a specific type of communicative setting, which is different from other types of online interaction, e.g. in tutorials. This is what makes the corpus novel and different from the examples discussed below in the related works section.

Therefore, it seemed a natural step for the network to collect and process the recordings of our meetings to create a multimodal corpus of Zoom meetings[2]. The corpus consists of audio-visual recordings of actually occurred, non-scripted on-

[1] See, however, Reverdy et al. (2022).

[2] Zoom, version: 5.14.10 (19202), `https://zoom.us`.

line meetings enriched with transcriptions and extracted visual coordinates. The purpose of the collection is to facilitate future studies of the way gesture and speech are used in online communication and the way this interaction can be modelled computationally. The data will be made available to the community for educational and dissemination purposes from an open repository, e.g. the CLARIN infrastructure www.clarin.eu, under an appropriate license. However, it is already possible to download transcriptions, as well as the scripts that were used to extract visual features and visualise them, from GitHub repositories[3].

After discussing related work, in the rest of the paper we describe the way the corpus was collected, transcribed and enriched with the visual coordinates, we give descriptive statistics concerning both the speech transcription and the visual keypoint values and we present and discuss visualisations of these values. Finally, we carry out a short preliminary analysis of the role of feedback in the meetings, and show how visualising the coordinates extracted via OpenPose can be used to see how gestural behaviour supports the use of feedback words during the interaction.

2. Related work

In what follows we review previous work on meeting corpora collected in face-to-face settings as well as data collected from different types of online interaction, but also research dealing with the augmentation of multimodal datasets with visual features.

Perhaps the most influential multimodal corpus of group interaction to this day is the publicly available AMI corpus (Carletta, 2006), a multimodal dataset covering 100 hours of meeting recordings[4]. While most of the corpus was recorded by having participants play different roles, around one-third of the data were recorded from real meetings. The corpus was transcribed and annotated with labels for a number of phenomena such as dialogue acts, topic segmentation, meeting summaries, named entities, communicative head and hand gestures we well as gaze direction. Since the development of AMI, several other audio and audiovisual recordings of meetings and group conversations have become available. Some only consist of audio data (Janin et al., 2003; Tardy et al., 2020; Nedoluzhko et al., 2022). Others are audiovisual (Koutsombogera and Vogel, 2018), enriched with eye-tracking signals (Brône and Oben, 2015), body motion coordinates obtained through motion capture (Edlund et al., 2010), and physiological measures such as breathing, perspiration, electrodermal activity and heart rate (Hennig et al., 2014).

To the best of our knowledge, no attempt has been made so far to collect interaction meeting data in a videoconferencing setting with the intention of making them publicly available.

Several data collections have been made, however, in connection with online task-oriented interaction. Ringeval et al. (2013), for example, recorded participant dyads during a video conference while completing a task requiring collaboration. The dataset includes audio, video, ECG and EDA data, and was used in Kantharaju et al. (2018), where OpenFace features were added to it for the recognition of affective laughter. The RoomReader corpus described in Reverdy et al. (2022) RoomReader, consists of Zoom video and audio recordings in a collaborative student-tutor scenario designed to elicit spontaneous speech. The data are annotated with labels for student engagement, and enriched with accompanying personality test scores and behavioural assessments to investigate student engagement in online tutorials. Cappellini et al. (2023) also developed a multimodal corpus of teacher-trainee videoconference interactions with the aim of investigating the pedagogical aspects of this interaction. The goal of Bodur et al. (2023), in contrast, is to investigate the use of multimodal backchanneling behaviour (Yngve, 1970) by children compared to adults. The study is based on a corpus of Zoom video chats involving 6-12 years old children and their caregivers structured around a weakly structured word-guessing game. Interestingly, the authors claim that compared to other datasets of spontaneous child-caregiver interactions in the wild, the Zoom recordings provide much clearer access to the speakers' facial expressions and head gestures, and are therefore a more reliable empirical source. The study finds that children use backchanneling roughly as often as adults. It is also mentioned, however, that the backchanneling behaviour is overall lower in their data than what has been reported previously for interaction in physical presence. We do not know whether the same difference which was observed between online and offline children's behaviour holds for adults. In general, in order to build knowledge about the way communication takes place in online interaction, we need to collect and study interactional data collected in different online scenarios. The GEHM corpus will allow us to study the behaviour of adults in an online meeting situation.

An important component of research in multimodal interaction analysis, whether face-to-face or online, concerns the way in which the visual modality is represented. In multimodal cor-

[3] https://github.com/kuhumcst/GEHM_zoom_corpus and https://github.com/goodPointP/GEHM_dataset/.

[4] https://groups.inf.ed.ac.uk/ami/corpus/.

pora, video-recordings can be enriched with extracted visual features. There is a substantial body of work on sign language, for example, in which video data are enriched with body position information extracted using OpenPose (Bauer et al., 2023; Vahdani et al., 2021). In Provè (2022), the authors used OpenPose to extract visual coordinates of singing people, and analysed the head movements of different sound patterns. Researchers working with multimodal sentiment analysis and emotion detection (Zadeh et al., 2016; Bagher Zadeh et al., 2018) have also constructed multimodal datasets based on audiovisual online monologue data. These datasets are made available together with automatically extracted visual information, e.g. the facial action units based on Facial Action Coding System (FACS) (Ekman et al., 1980). Other researchers also used visual features obtained through OpenPose for the detection of specific postures among students in classrooms (Chen and Gerritsen, 2021), or for Human Activity Recognition (Ali et al., 2023). In general, a common characteristics of the work summarised here is that visual analysis is conducted automatically, whereas human annotators are used for higher-level annotation such as the annotation of emotion, sentiment or pragmatic categories such as feedback.

3. Data collection

Given the duration of the network and the foreseen frequency of online meetings, the expectations were that we would be able to collect about 10 hours of recorded meeting time. This seemed an acceptable size worth striving for. If we compare with data collected in physical presence, a corpus of this size would be somewhat in between smaller annotated multimodal collections such as NOMCO (Paggio and Navarretta, 2017) and Multisimo (Koutsombogera and Vogel, 2018) (about one hour of dyadic interaction and three hours of group interaction, respectively) and larger corpora such as Spontal (Edlund et al., 2010) or AMI (Carletta, 2006) (50 hours and 100 hours of group meetings, respectively). It is also comparable to the nine hours of online interaction data covered by the RoomReader dataset (Reverdy et al., 2022).

A number of requirements were agreed upon prior to starting the collection. To strive for as high as possible an audio quality, it was decided that participants should wear headsets, and avoid bluetooth equipment to minimise time lag between video and audio signals. The quality of the video recordings, in turn, depends on computer screen and internet connection. No specific requirements could be enforced concerning those two aspects. However, participants were encouraged if possible to sit in front of a light background and in good light. They were also asked to be sitting and visible at the centre of the screen in such a way that part of the torso and not just the face would also be visible. We decided not to enforce stricter requirements in an effort not to affect the naturalness of the recorded meetings, the primary goal of which was not to provide data for the corpus, but to advance the network's research agenda with discussions and planning.

A series of meetings were then recorded from September 2021 to March 2023 through the Zoom accounts of two of the universities involved in the research network. All the meetings were hosted by the same researcher, who was the network coordinator, and followed a predefined agenda. There was, however, no preconceived script or scenario.

3.1. Ethical considerations

All meeting participants have signed a consent form in which they give permission for the use, distribution and visualisation of the recordings and their transcriptions for research, education and dissemination purposes. It is also specified in the form that the participants will be visible and therefore recognisable, and their names may be mentioned during the recordings. The complete consent form that was used can be inspected at `https://cst.dk/patrizia/gehm_consent/form.pdf`.

4. The corpus

The corpus consists of 12 video recordings of meetings held on Zoom by the network researchers. The meetings have an average duration of about 40 minutes each, for a total of 8 hours. The number of participants is 5-9 per meeting. The language used is English. Participants are native and non-native speakers, all of them researchers active in the network.

The audio-visual recordings will be distributed together with an orthographic transcription as well as face and body position coordinates. The size of the corpus, including audio and video files, as well as transcription and visual coordinate files, is 195 gigabytes.

4.1. Preprocessing

The recordings were done by the meeting chair using the Zoom recording functionality. Speakers' video and audio tracks were separated manually from the Zoom meeting recordings using a video editing tool. Separate audio files for each speaker were then created with silences for the intervals in which they did not speak. The audio was exported as a lossless waveform to maintain maximum audio quality for future use. Every participant's video feed was exported separately, con-

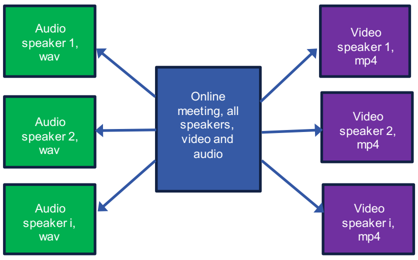

Figure 1: Preprocessing of the Zoom audio-video recordings

taining only their webcam's video. Note that the space taken up by the individual speakers in the videos varies due to different numbers of participants in each meeting. It was decided to keep a constant size of 1920-1080 pixels in the extracted single videos. In all the file names, speakers' names were replaced by unique identifiers. Participant IDs are kept constant throughout the meetings. The process is illustrated in Figure 1.

4.2. Orthographic transcription

The orthographic transcription automatically created by the Zoom software, unfortunately, is not word-aligned. Therefore, we needed a different method to transcribe the speech signal in an automatic way. We started by performing an initial evaluation of a number of models from the Google speech-to-text API[5] by measuring their error rate on a manually transcribed extract. We considered models for British and American English, given the fact that the meeting participants speak different varieties of English, as well as models trained on telephone and video interactions. Halfway through this process, WhisperX[6] (Bain et al., 2022), which is based on OpenAI's Whisper[7], became available to perform the same task, so we decided to include it in the evaluation of possible models. WhisperX produced the lowest word error rate (11%) compared with the best of the Google API models (15.89%) and was therefore chosen to transcribe the entire corpus. Each speaker's speech output in each video was consequently transcribed using WhisperX.

The output was converted into the Praat TextGrid format (Boersma and Weenink, 1992), where the spoken contributions of all meeting participants are transcribed with a separate tier for each participant and time aligned with the video through time stamps before and after each word. An example is shown in Figure 2. In the end, the automatic transcription underwent a rough manual revision especially focusing on overlaps between participants, where we had noticed that the automatic system sometimes could not separate the output of the two speakers correctly. The transcription of some of the recurrent proper nouns was also checked and corrected. Conversely, although we noticed that the system does not transcribe filled pauses and laughter, at least not in a systematic way, we chose not to add the relevant tokens and instead to leave the issue for future work.

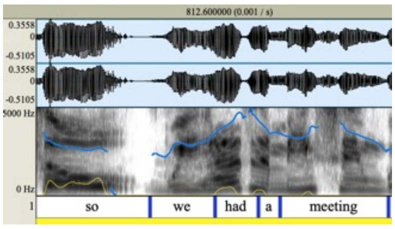

Figure 2: Sample transcription shown in the Praat interface

4.3. Visual coordinates

OpenPose (Cao et al., 2017) was run on each individual speaker video to extract position coordinates of nose, eyes, ears, neck, shoulders, elbows and wrists. The software uses a pre-trained deep learning model which, given an image (sequence) as input returns a set containing X and Y coordinates of common so-called *keypoints* found on human bodies. Our videos include only a portion of the participant's torsos. They always include their faces and sometimes their arms and hands. The positional keypoint values were saved in JSON files, one file per video frame per speaker. The visual coordinate extraction process is quite demanding (it was run on high-performance NVIDIA GPUs with 40GB of VRAM). Therefore, we believe that making the results of this processing step available will be of great service to the community and ultimately avoid unnecessary energy consumption.

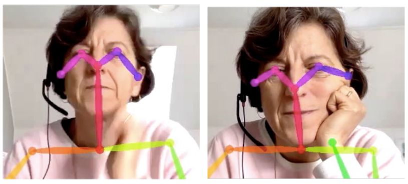

Figure 3: OpenPose skeletons in two frames from the GEHM Zoom corpus

The keypoint values found by OpenPose in two video frames are visualised by means of a skeleton outline in Figure 3.

[5] https://cloud.google.com/speech-to-text.
[6] https://github.com/m-bain/whisperX.
[7] https://openai.com/research/whisper.

4.4. Statistics

4.4.1. Words

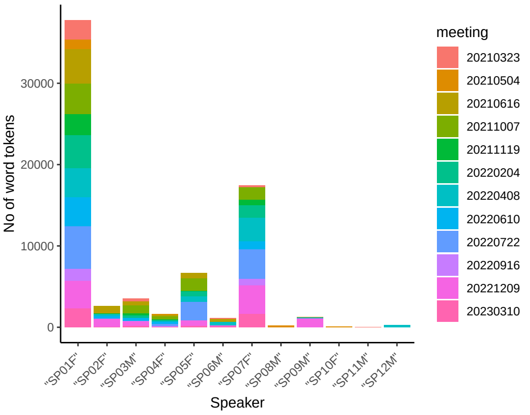

Figure 4: Number of word tokens produced by each speaker

Meeting \ Speaker	SP01F	SP02F	SP03M	SP04F	SP05F	SP06M	SP07F	SP08M	SP09M	SP10F	SP11M	SP12M
20210323	1	0	1	0	0	1	1	0	0	0	1	0
20210504	1	0	0	1	0	0	0	1	1	1	0	0
20210616	1	1	1	1	1	1	0	0	0	0	0	0
20211007	1	0	1	1	1	0	1	0	0	0	0	0
20211119	1	1	1	1	1	0	1	0	0	0	0	0
20220204	1	1	1	1	1	0	1	0	0	0	0	0
20220408	1	1	1	1	1	1	1	0	1	0	0	1
20220610	1	1	1	1	0	0	1	0	0	0	0	0
20220722	1	1	0	1	1	0	1	0	0	0	0	0
20220916	1	0	1	1	1	0	1	0	0	0	0	0
20221209	1	1	1	1	1	1	1	0	1	0	0	0
20230310	1	1	1	1	1	1	1	0	0	0	0	0

Table 1: Speaker participation in each meeting: blue indicates presence

The speech transcriptions contain a total of 72,671 word tokens and 3,785 types. There is a large variation in the number of words uttered by each speaker, with SP01F producing 37,776 words, i.e. 52% of the total, followed by SP07F, who produced 17,452, i.e. 24% of the total. This disparity is due to the fact that speaker SP01F was always the meeting chair, but also to the fact that not all speakers participated in all meetings. The number of words produced by each speaker in total and in the various meetings is visualised in Figure 4, while Table 1 shows which speaker was present in which meeting.

To get an impression of the topics discussed in each of the meetings, the ten most distinctive words were extracted from each meeting transcription using the tf-idf weighting scheme. The results are displayed in Table 2.

4.4.2. Visual coordinates

The plot in Figure 5 shows the distribution of values found for the various keypoints over video frames in the corpus. The values on the y axis correspond to the proportion of frames in each speaker video where the keypoint is visible.

Meeting	Words
m01	sensitive, recordings, data, research, delay, behavior, influence, sitting, hoc, objection
m02	Clarin, license, internet, we will, ClarinDK, ACCA, licenses, compare, tracks, Malta
m03	online, September, summer, July, phonetic, mid, travel, quality, wear, planning
m04	eyebrows, gestures, gallery, prominence, frontiers, eyebrow, 19th, nods, structure, linguistic
m05	music, February, replace, mature, Clarissa, ironic, stance, masks, Patrick's, off
m06	March, enjoyed, 11th, L2, shared, Google, coherence, tutorials, framed, ambitious
m07	proceedings, coffee, reviews, ISGS, tickets, charged, fees, reviewers, committee, scientific
m08	July, 30th, action, countries, June, 16th, Belgium, bureaucratic, takes, application
m09	references, figures, papers, pages, template, degrees, version, chair, acronym, PDF
m10	10th, 14th, 11th, Santa, bye, assignments, urgent, assignment, 13th, plane
m11	dinner, weak, zone, registration, price, oral, accept, reject, reviewers, acceptance
m12	doctoral, dance, book, future, Kendon. merchandising, sticks, history, rooms, sandwich

Table 2: Distinctive words in the corpus meetings

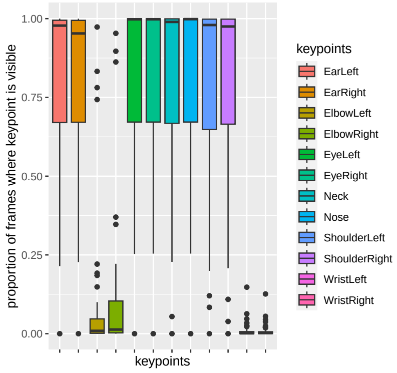

Figure 5: Average number of frames for each individual speaker video in which various visual keypoints are visible

The most obvious observation is that the elbows and even more the wrists, most of the time are not captured, probably because the arms and hands are outside the field of view of the camera. There are, however, a few speakers who fall outside the main distribution and constitute outliers with respect to this general tendency. These speakers sit a bit further away from the screen than the others such that a larger part of their torso is visible. As mentioned earlier, although speakers were asked to sit in similar positions, the requirement was not enforced in a rigid manner in order not to affect the naturalness of the interaction.

The second observation is that for the face and shoulder keypoints, all of which are visible most of the time, there is little difference between left and right. In fact, we would not expect much difference since the speakers are sitting frontally and relatively centred in front of the camera.

During the meetings, sometimes one of the speakers shares their screen through Zoom. The effect of this is that visual keypoints of the speakers are not available while the screen is shared. This happens especially in meeting 20220408, where the screen is shared for approximately 40 minutes, and for which the highest proportion of frames where visual keypoints are available is only 0.25.

A systematic validation of the correctness of the visual coordinate extraction was not attempted. We did, however, check whether the system always identified only one person in each speaker video. Averaged over all frames in all files, only 12 per thousand seem to have coordinates for more than one person – a proportion that seems negligible. This may be due to the fact that on a few occasions, individuals that do not belong to the group of meeting participants, for instance relatives, briefly appeared in the videos.

5. Interpreting visual coordinates

A script was written to help understand patterns of visual coordinates in the meetings. The script reads all keypoint value files from Openpose, and can plot the visual coordinates of a meeting participant in a specified time frame sequence. These visualisations can be used to get an impression of the way a speaker moves their head, torso and partially their hands in the course of an entire meeting, but also in specific time intervals of a given meeting. In what follows we give a few examples of the way these visualisations can be interpreted.

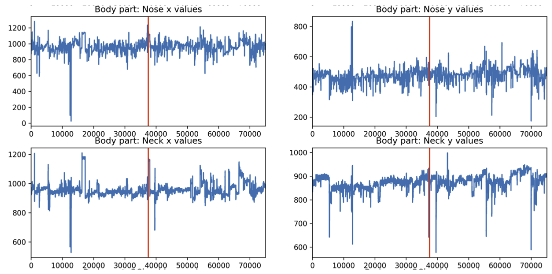

Figure 6: Visualisation of the nose (first row) and neck (second row) coordinates for one participant over the course of an entire meeting. In the left column we can see the positions on the horizontal axis and in the right column those on the vertical axis (x and y values, respectively).

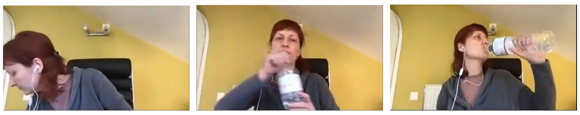

Figure 7: Moment in which a participant moves their torso to the left of the screen to get a water bottle and drink out of it.

We visualise the movements of the nose and neck of one of the participants over the course of a meeting in Figure 6. The plots can give an impression of different movements made by the participant. If we look at one relatively pronounced change that happens at the area around frame number 12,000, we can see that the x values of both nose and neck (the plots on the left) hit a low point. The y values, shown in the plots on the right (nose and neck values, respectively) display a similar change, but the nose seems to move upwards. In the video, at this moment, the participant turns the torso to the left of the screen to grab a water bottle and after coming back to the initial position, she ingests the water, which is shown with a raising nose position in the y values, while the neck stays in a relatively similar position. The sequence can be seen in the frames in Figure 7.

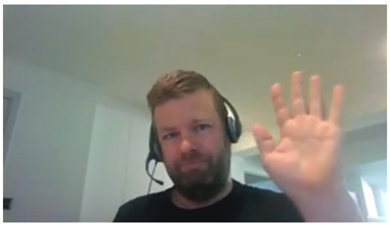

Figure 8: A speaker asking to take the floor

Another interesting example is shown in the frame in Figure 8, where we see one of the speakers lift his hand to ask for the floor. Just before doing this, he had been nodding repeatedly in response to what the meeting chair was explaining.

Figure 9 displays x and y coordinate measures relative to the left wrist and nose of the speaker 20 frames before and after the stroke of the hand movement (marked by the red line in the plots).

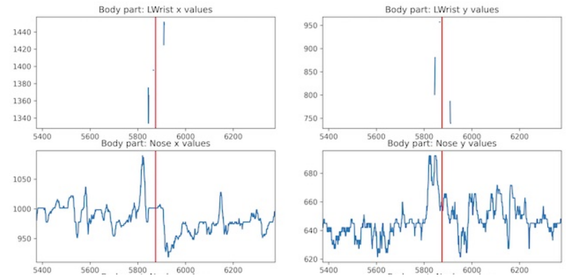

Figure 9: Measures of x and y coordinates for left wrist (upper part of the graph) and nose (lower part) in connection with hand movement.

The curves corresponding to the nose in the lower part of the graph show how the head moves up and down just before the wrist is lifted. To sum up, the script offers the possibility of visualising keypoint values in sequences of varying length, and is meant as tool to help the researcher get an insight into co-articulation patterns, compare the way different speakers move, or inspect how gestural behaviour evolves at specific moments during a meeting.

6. A case study: Feedback

One of the characteristics of human communication is the continuous production of feedback signals by the conversation participants to indicate to each other whether they perceive, understand and accept what the interlocutor is communicating (Allwood et al., 1992). Feedback can be given with speech or gestures, and it is often expressed multimodally (Kendon, 2004). Feedback giving signals are often referred to as backchannels (Yngve, 1970). Both the giving and the eliciting of feedback are pervasive in face to face communication. Feedback is in fact essential to successful communication, as we know from studies of the phenomenon in several languages and different types of communication (Cerrato, 2007; Navarretta et al., 2012), and can be considered a specific dialogue act type (Bunt et al., 2010). We don't have much empirical evidence, however, of the way feedback is used in online group meetings. In this section, we present some descriptive statistics of feedback given through speech in our corpus, and point at ways gestural feedback behaviour can be taken into account in future analyses of this phenomenon.

The most common feedback words in English are the positive *yes*, *yeah*, and *okay*, as well as the negative *no*. They were used to analyse the occurrence of feedback in the GEHM Zoom corpus transcriptions[8].

The relative frequency of the positive feedback words in the corpus is 0.029, while that of *no* is 0.003. To see whether participants other than the meeting chair produced more feedback behaviour than the chair herself, who is also the most active speaker, we repeated the counting excluding the spoken contributions of the meeting chair. The relative frequency of the positive feedback words is in this case 0.031, while that of *no* is 0.004. In other words, there is almost no difference.

To get a sense of whether the amount of feedback we identified matched what could be expected from similar interactions, we compared with counts from a portion of the AMI's naturally occurring project meetings[9]. We calculated the relative frequency of the same positive and negative feedback words as for the Zoom corpus. The relative frequency of the positive feedback words in the AMI's dialogues under consideration is 0.046, while for the negative feedback word the frequency is 0.005.

We must be cautious in directly comparing the two datasets since not only the setting of the meetings (virtual vs. physical), but also other factors differ such as the participants' native languages, their age, and the type of project discussed. Moreover, the amount of speech produced by the various participants in the AMI meetings is more balanced than in the GEHM ones, even though also the AMI meeting chairs speak more than the other participants. Having said that, we note that the relative frequency of common feedback words is higher in the AMI corpus than in the GEHM Zoom meeting corpus, especially as concerns positive feedback words, which are also in general the most frequent type of feedback in both corpora. This difference is similar to the one noted in the study by Bodur et al. (2023) mentioned earlier, where it was found that backchanneling (feedback giving in our terminology) occurs less frequently in online interaction than in physical presence. Whether this tendency also holds for gestural and multimodal behaviour, we leave for future research.

Head movements, and less frequently hand gestures, are used in the Zoom meetings alone or together with feedback words, to signal or elicit feedback. These movements can be visualised through the visual keypoints. The plot in Figure 10, for example, illustrates the way two different speakers move their nose (and therefore their head) in the course of an entire Zoom meeting. The plots in the left part of the graph show the x coordinates, and the ones in the right part the y coordinates. The difference between the two speak-

[8] Different transcriptions of these words, e.g. okay and OK, were normalised.

[9] The dialogues we included are all the naturally occurring meetings from Edinburgh in the two series EN2001 and EN2002. They consisted of 80,877 words.

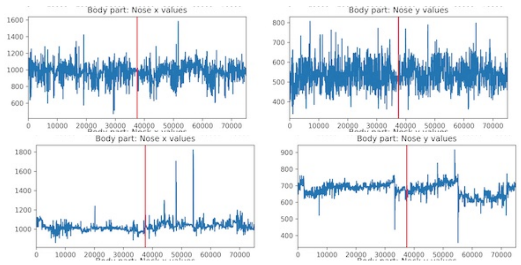

Figure 10: Nose movements by two different speakers during an entire meeting: x coordinates are on the left, and y coordinates on the right.

ers is quite evident. The speaker whose movements are depicted at the top moves the head continually both horizontally and especially vertically. The speaker depicted at the bottom, conversely, does not move the head much in either directions although there are some movement peaks at specific points.

Not all the movements reflected in this graph, of course, are associated with feedback. Nor does gestural feedback necessarily co-occur with feedback words. We can see from inspecting the videos, for example, that the speaker whose head movements are displayed in the top part of the graph often nods while another participant is speaking. Therefore, a more careful analysis must be carried out to characterise the use of multimodal and gestural feedback in the meetings. Inspecting visual coordinates in this way, however, can be a first step before a detailed analysis.

An interesting aspect of the way feedback is given, and therefore consensus reached, in the meetings, relates to the way speakers align their gestural behaviours. A relevant example is shown in the frame sequence in Figure 11. First one of the speakers shows her agreement with a thumbs-up gesture, then a second speaker does the same, and finally a third speaker also aligns her feedback, although with a little delay.

7. Conclusion and Future Work

In this paper we have introduced a new corpus of Zoom meetings which will soon be made available to the research community in its entirety (and which is already partly downloadable). We have explained the way the corpus was collected, transcribed and enriched with visual coordinates relating to various keypoints of the head, torso and wrists of the meeting participants. While the corpus was not collected with a specific research question in mind, we believe it is a rich and useful dataset to study real life interaction in online group interaction.

Our immediate plans for further developing and analysing the corpus data include, first of all, using the visual coordinates to derive features that can be used to detect specific head movement types such as nods or shakes. In order to do that, we would like to adapt the models for head movement detection described in Agirrezabal et al. (2023). We see head movement detection as a necessary step to study the way both speech and gestural behaviour are used to express feedback in the online meetings.

Additional future work will be devoted to quantitative and qualitative studies of other interaction phenomena, in particular turn taking, overlapping and aligning between speakers.

Finally, we would like to enrich the resource further by applying automatic detection of facial action units, e.g. using OpenFace (Baltrusaitis et al., 2018), to make it possible to analyse the role of facial expressions.

8. Acknowledgements

We acknowledge all the researchers who participated in the Zoom meetings. We also acknowledge the support of the international research network GEstures and Head Movements in language (GEHM), which is funded by the Danish Research Council (grant number: 9055-00004B).

9. Bibliographical References

Manex Agirrezabal, Patrizia Paggio, Costanza Navarretta, and Bart Jongejan. 2023. Multimodal detection and classification of head movements in face-to-face conversations: Exploring models, features and their interaction. In *Gesture and Speech in Interaction Conference (GeSpIn 2023)*.

Mohammed Abduljabbar Ali, Abir Jaafar Hussain, and Ahmed T Sadiq. 2023. Estimating human running indoor based on the speed of human detection by using OpenPose. In *Proceedings of Data Analytics and Management: ICDAM 2022*, pages 749–761. Springer.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

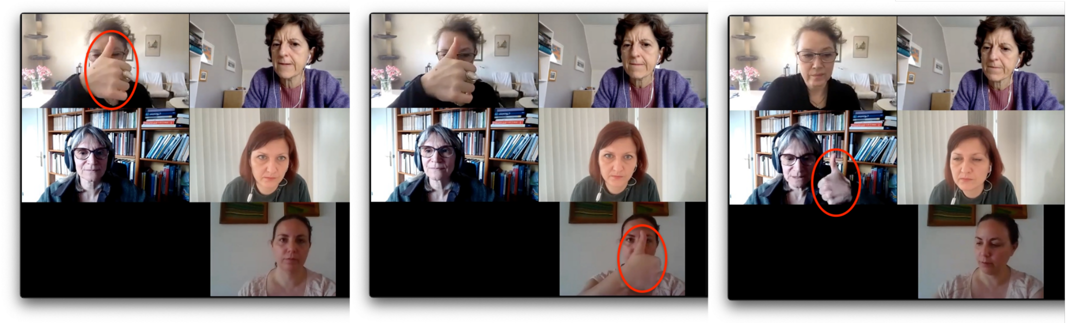

Figure 11: Three speakers aligning their feedback behaviours. Red circles were superimposed to show the okay gesture on which they align

Jeremy N Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, 2(1).

M Bain, J Huh, T Han, and A Zisserman. 2022. Whisperx: Time-accurate speech transcription of long-form audio. *Advances in Neural Information Processing Systems*, 35.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Anastasia Bauer, Anna Kuder, and Marc Schulder. 2023. Phonetics of head nods in German sign language (DGS). In *GESPIN 2023*.

Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. 2023. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, 5:1088752.

Paul Boersma and David Weenink. 1992. 2022. Praat: Doing phonetics by computer [computer program]. version 6.0. 43.

Geert Brône and Bert Oben. 2015. Insight interaction: A multimodal and multifocal dialogue corpus. *Language resources and evaluation*, 49:195–214.

H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. Chengyu Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings 7th international conference on language resources and evaluation (LREC 2010)*, pages 2548–2555.

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.

Marco Cappellini, Benjamin Holt, Brigitte Bigi, Marion Tellier, and Christelle Zielinski. 2023. A multimodal corpus to study videoconference interactions for techno-pedagogical competence in second language acquisition and teacher education. *Corpus*, (24).

Jean Carletta. 2006. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5.

Loredana Cerrato. 2007. *Investigating Communicative Feedback Phenomena across Languages and Modalities*. Ph.D. thesis, Stockholm, KTH, Speech and Music Communication.

L. Chen and David Gerritsen. 2021. Building interpretable descriptors for student posture analysis in a physical classroom. In *22nd International Conference on Artificial Intelligence in Education AIED*.

Xinyue Chen, Si Chen, Xu Wang, and Yun Huang. 2021a. "I was afraid, but now I enjoy being a streamer!" understanding the challenges and prospects of using live streaming for online education. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–32.

Zhilong Chen, Hancheng Cao, Yuting Deng, Xuan Gao, Jinghua Piao, Fengli Xu, Yu Zhang, and Yong Li. 2021b. Learning from home: A mixed-methods analysis of live streaming based remote education experience in Chinese colleges during the COVID-19 pandemic. In *Proceedings of the 2021 CHI Conference on human factors in computing systems*, pages 1–16.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David

House. 2010. Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In *LREC*, pages 2992–2995.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Geraldine Fauville, Mufan Luo, Anna Carolina Muller Queiroz, Jeremy N Bailenson, and Jeff Hancock. 2021. Nonverbal mechanisms predict Zoom fatigue and explain why women experience higher levels than men. *Available at SSRN 3820035*.

Shannon Hennig, Ryad Chellali, and Nick Campbell. 2014. The D-ANS corpus: the Dublin-autonomous nervous system corpus of biosignal and multimodal recordings of conversational speech. In *LREC*, pages 3438–3443.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1. IEEE.

Reshmashree B Kantharaju, Fabien Ringeval, and Laurent Besacier. 2018. Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 220–228.

Adam Kendon. 2004. *Gesture - Visible Action as Utterance* . Cambridge University Press.

Jung In Koh, Samantha Ray, Josh Cherian, Paul Taele, and Tracy Hammond. 2022. Show of hands: Leveraging hand gestural cues in virtual meetings for intelligent impromptu polling interactions. In *27th International Conference on Intelligent User Interfaces*, pages 292–309.

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).

C. Navarretta, E. Ahlsén, J. Allwood, K. Jokinen, and P. Paggio. 2012. Feedback in Nordic First-Encounters: A Comparative Study. In *Proceedings of LREC 2012*, pages 2494–2499, Istanbul Turkey.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182.

Patrizia Paggio and Costanza Navarretta. 2017. The Danish NOMCO corpus: Multimodal interaction in first acquaintance conversations. *Language Resources and Evaluation*, 51:463–494.

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

Hendri Pratama, Mohamed Nor Azhari Azman, Gulzhaina K. Kassymova, and Shakizat S. Duisenbayeva. 2020. The trend in using online meeting applications for learning during the period of pandemic COVID-19: A literature review. *Journal of Innovation in Educational and Cultural Research*, 1(2):58–68.

Valentijn Prové. 2022. Measuring embodied conceptualizations of pitch in singing performances: Insights from an OpenPose study. *Frontiers in Communication*, 7:957987.

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.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then summarize: Automatic alignment methods for summarization corpus creation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6718–6724, Marseille, France. European Language Resources Association.

Elahe Vahdani, Longlong Jing, Yingli Tian, and Matt Huenerfauth. 2021. Recognizing American sign language nonmanual signal grammar errors in continuous videos. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1–8. IEEE.

Matin Yarmand, Jaemarie Solyst, Scott Klemmer, and Nadir Weibel. 2021. “it feels like i am talking into a void”: Understanding interaction gaps in synchronous online classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos.