# Multimodal interaction in online meetings: the GEHM corpus

Patrizia Paggio[1,2], Manex Agirrezabal[1], Costanza Navarretta[1] and Leo Vitasovic[1]

[1]Centre for Language Technology (CST)
NorS, University of Copenhagen
[2]Institute of Linguistics and Language Technology
University of Malta

L-Università
ta' Malta

LREC-COLING 2024

# Motivation

- The use of video conferencing for group meetings, teaching, international conference organisation, etc. has increased dramatically.
- Several studies have discussed pros and cons especially in connection with teaching (Chen et al., 2021a,b; Yarmand et al.,2021).
- Danger of fatigue and lack of engagement has been stressed (Bailenson, 2021; Fauville et al., 2021).
- Another issue is the difficulty of gauging interlocutors’ responses in large online meetings (Koh et al., 2022).

# Motivation

- ❖ In general, empirical evidence of the way gesture and speech are used in online meetings is scarce (See, however, Reverdy et al., 2022).
- ❖ We wanted to contribute to fill out this gap by creating a rich corpus of transcribed audio-visual Zoom meeting recordings annotated with visual features.
- ❖ The corpus is freely available for researchers to study multimodal group interaction in a real-life setting.

# The GEHM network

An initiative of the research network on Gesture and Head Movements in Language (GEHM) funded by the Danish Research Council.

Cooperation between nine research groups from six European countries.

UNIVERSITY OF COPENHAGEN

KTH VETENSKAP OCH KONST

upf. Universitat Pompeu Fabra Barcelona

KU LEUVEN

Linnæus University

CAU Kiel University Christian-Albrechts-Universität zu Kiel

LUND UNIVERSITY

Trinity College Dublin Coláiste na Tríonóide, Baile Átha Cliath The University of Dublin

L-Università ta' Malta

# The GEHM network

## Research themes:

- Language-specific characteristics of gesture-speech interaction
- Multimodal prominence
- Modelling of multimodal behaviour

https://cst.ku.dk/english/projects/gestures-and-head-movements-in-language-gehm/

# The GEHM online meetings

Due to COVID 19, we could not meet physically.

We decided to record our online meetings and turn them into a joint corpus we could use to provide data for at least the second and third of our themes.

**Requirements:**

- ❖ As good a sound as possible to allow for phonetic analysis.
- ❖ Extraction of visual features to allow for machine learning based on image processing.

# The corpus

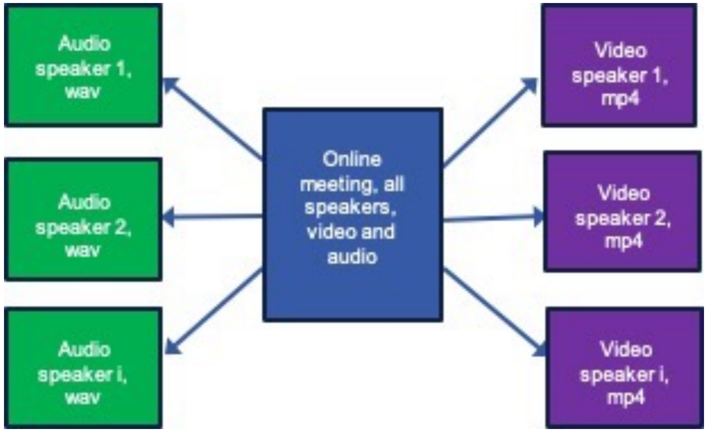
- ❖ Recordings of 12 Zoom meetings featuring 5-6 participants per meeting.
- ❖ Average duration ~40 minutes each, total ~8 hours.
- ❖ Language English (different accents).
- ❖ Native and non-native speakers.
- ❖ Researchers in the GEHM network.
- ❖ Participants have their webcams on and wear headsets.
- ❖ They have all given their consent.

# The corpus

## Preprocessing

Speakers' video and audio tracks separated manually using a video editing tool.

Separate audio files for each speaker with silences for the intervals in which they did not speak.



Every participant's video feed was exported separately, containing only their webcam's video, keeping a constant size of 1920-1080 pixels.

# The corpus

## Transcription

To derive the transcription, several automatic speech recognition systems were tested.

A manually transcribed excerpt was used as ground truth to calculate WER (word error rate).

WhisperX, based on OpenAI's Whisper,  had the lowest WER (11% on the test sample).

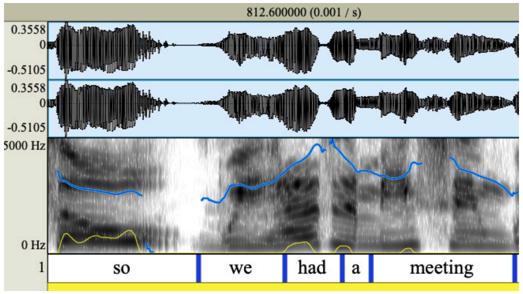
# The corpus

## Transcription

Obtained with **word-level timestamps** (start and end of every word).

Converted to a PRAAT TextGrid and corrected manually for specific types of error:

- Speaker overlaps.
- Some proper names.



Filled pauses and laughter mostly not transcribed and not added yet.

# The corpus

## Visual feature extraction

Obtained using OpenPose.

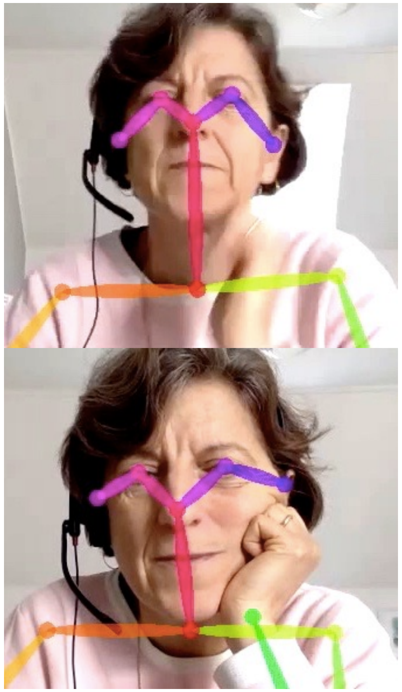
OpenPose uses a pre-trained deep learning model which, given an image (sequence) as input returns a set containing X and Y coordinates of common points found on human bodies.

Our videos included only a portion of the participant's torsos and almost always included their faces.

# The corpus

## Visual feature extraction

- We extracted position coordinates of nose, eyes, ears, neck, shoulders, elbows and wrists.
- All coordinates were saved in a series of JSON files (one file per frame per video).
- The code developed to produce transcriptions and extract visual coordinates available on GitHub.



# The corpus

## Visual feature extraction validation

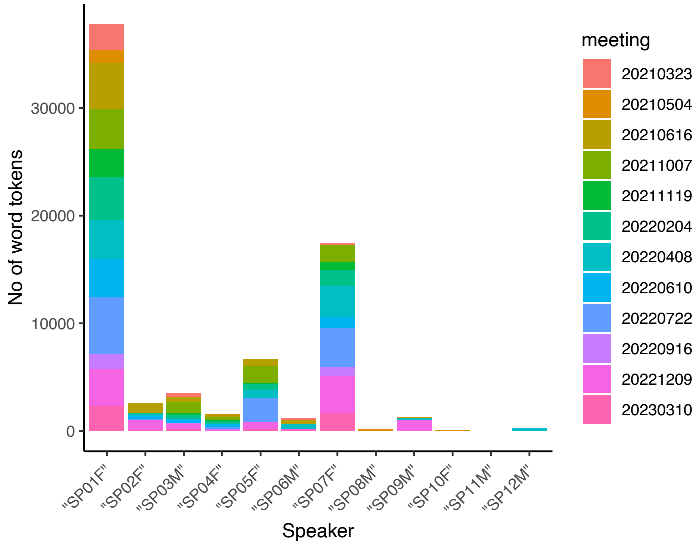
The only systematic validation was checking whether the system always identified only one person in each speaker video.

Averaged over all frames in all files, only a negligible 12 per thousand have coordinates for more than one person.

This may be due to individuals not belonging to the group of meeting participants, for instance relatives, briefly appearing in the videos.

# Statistics – words

The speech transcriptions contain a total of 72,671 word tokens and 3,785 types.

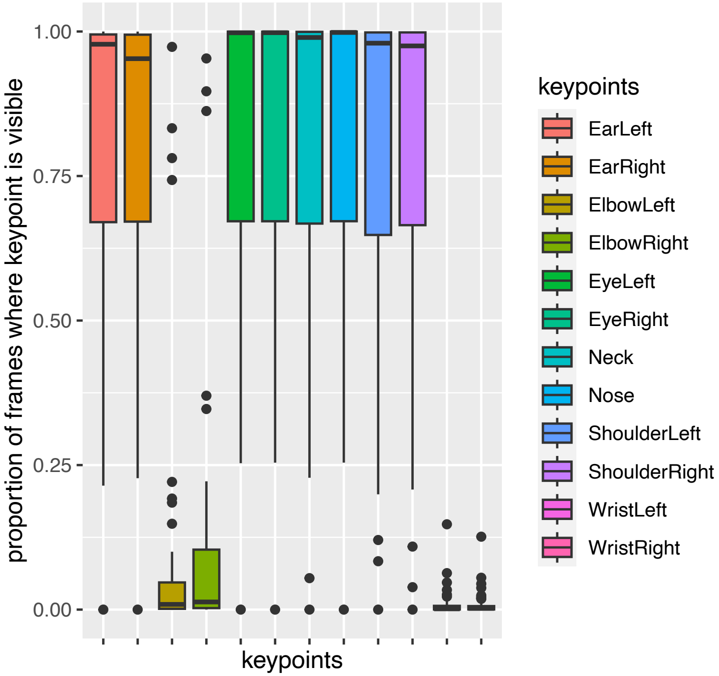


Large variation: SP01F produces 52% of the total, followed by SP07F, who produced 24% of the total.

# Statistics – visual coordinates

## What is captured?

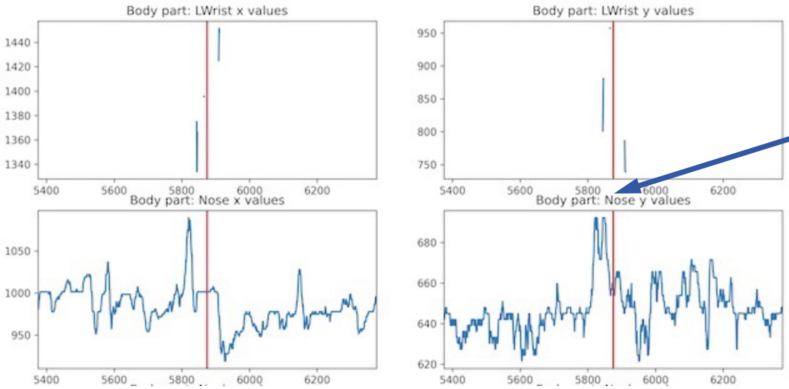
- Elbows and wrists mostly not captured. Outliers are speakers who sit a bit further from the screen.
- Face and shoulder visible most of the time, and little difference between left and right (speakers are sitting frontally).
- Visual coordinates not available while the screen is shared (esp. one meeting).



# Interpreting visual coordinates

A script was written to plot the visual coordinates of meeting participants in a specified time frame sequence to help understand movement patterns.

Just before lifting his hand, the participant had been nodding several times.



# A case study: feedback

We know from many studies that feedback is an essential component of successful communication.

How does it work in online meetings?

# A case study: feedback

We calculated the relative frequency of the most common feedback words in English, that is *yes, yeah*, *okay* and *no*.

We compared with how frequent the same words are in the naturally occurring project meetings from the AMI corpus (80,877 words).

| | GEHM | AMI |
|---|---|---|
| Positive fd words | 0.029 | 0.046 |
| Negative fd words | 0.003 | 0.005 |

# A case study: feedback

Differences between the two corpora:

- ❖ Speech production in AMI more balanced across speakers (however, meeting chair speaks more)
- ❖ Different age and nationality characteristics

The lower frequency of feedback expressions might be due to the online setting.

A similar effect was observed by Bodur et al. (2023) in a study on children’s online interaction.

# A case study: feedback

**What about gestural feedback?**

Discrete gestures have not been annotated (yet). However, continuous movement patterns can give initial indications.

# Future perspectives

- ❖ The audio-visual recordings will be made available for research purposes from a public repository but can already be obtained at request.
- ❖ Transcriptions and annotations are on GitHub.
- ❖ The size of the entire corpus is about 200 gigabytes.

# Future perspectives

The data can be further annotated if needed.

We are also considering adding more information about the participants.

Examples of relevant studies:

- ❖ How multimodal feedback is established online.
- ❖ How much speakers align to each other.
- ❖ The role of hand gestures in online interaction.
- ❖ Training of gesture recognition models (Agirrezabal et al., 2023).

# Conclusion

We hope to have shown that the GEHM Zoom meeting corpus is an interesting resource for the study of online interaction.

We release it for research purposes and further annotation by the research community.

We acknowledge all the meeting participants, and the support of the GEHM international research network, which is funded by the Danish Research Council (grant number: 9055-00004B).