UNIVERSITY OF COPENHAGEN
DEPARTMENT OF NORDIC STUDIES AND LINGUISTICS

Visualising motion energy in videorecorded online meetings

Get this poster in PDF format

Bart Jongejan, Manex Agirrezabal, Costanza Navarretta and Patrizia Paggio
{bartj,manex.aguirrezabal,costanza,paggio}@hum.ku.dk

Abstract

Conversation participants align at different levels [4], both in terms of repetition of discrete behaviours, like nods [6], and entrainment of continuous behaviours [5], e.g. motion energy [2], across speakers.

Recently, [1] found that speakers display higher motion energy levels when verbal feedback (words like yes and no) is being exchanged but could not distinguish communicative gestures from involuntary physical movement unrelated to the meeting interaction.

Here, we present an implemented methodology to visualise motion energy in video-recorded online meetings as an intermediate step to study entrainment, but also as a tool to aid gesture annotation.

The visualisations of motion energy can also be used to analyse specific gestures, to inspect more general patterns of interaction, e.g. engagement and entrainment.

Motivation

We have made many attempts to automatically discern communicative gestures in meetings. In the current study the data consists of recordings of Zoom meetings. For each movie frame the positional data is provided by the OpenPose software.

We translated the raw **OpenPose** data (a JSON file for each movie frame) to multiple timelines, one **timeline** for each **speaker**. The timelines run across the movie, showing a few seconds of future frames and several seconds of past frames. A timeline condenses a lot of frame information in a simple colour coded vertical line. Most prominent are the bars hanging down. **Red** bars indicate **horizontal** movements, and **blue** bars indicate **vertical** ones. Bars can be a mix of red and blue, indicating movements that go in a slanted fashion. The length of each **coloured** bar indicates the amount of **kinetic energy**. Sometimes a **green** line can be seen above these bars, indicating that the person in question is **speaking**. The words themselves are not shown, except if the speaker says one of the **feed back words** that we have defined. Another useful information is coded as a pair of black and white horizontal lines. They are shown where the OpenPose software did not provide nose tip data or the data for the left or right ear was missing.

It was soon clear that a **head gesture**, even a non-repeated one, **almost never** consisted of a single sequence of **consecutive**, coloured bars, and that many such sequences were part of non-communicative head movements like momentarily looking down at the (invisible) table or taking a sip from a mug. There were almost always **breaks**, because in many gestures there are moments where there is no movement, for example when somebody is nodding and the head has turned maximally up or down.

We came up with the idea to model some types of **head movements** as though the head were fixed to a helical spring **coil** the other end of which would be kept in a fixed position. If the head were pulled or pushed a little bit to the side and then slipped free, the head would start an **oscillating** movement, at some points moving fast and at other points virtually standing still. If we ignore damping, then the **total energy** in such a system had to stay **constant** over many back-and-forth swings, and we would have to take potential energy into account as well as kinetic energy, the coil temporarily storing potential energy where the kinetic energy was low or zero to make the sum of potential and kinetic energy constant. The expectation was that also plotting the **potential energy** would **connect** otherwise **isolated** stretches of coloured **bars** that all belonged to the same head movement.

Movement: kinetic+potential energy

We compute the levels of kinetic and potential energy from velocity $v(t)$, acceleration $a(t)$ and jerk $j(t)$ of head movements, which in turn are computed from position values that OpenPose has provided for each video frame. Whereas kinetic energy at any time can be estimated as $E_{kin}(t) = v^2(t)$ (setting mass $m = 1$), potential energy $E_{pot}(t)$ is a more theory-laden concept.

We set $E_{pot}(t) = 0$ when there is no head gesture. In our idealized model, we call a head movement a head gesture if the head seems to make a linearly oscillating movement with frequency ω and amplitude r under the influence of an elastic force towards a point halfway between the extremes of the position. A movement is considered linearly oscillating only if the velocity and jerk vectors $\vec{v}(t)$ and $\vec{j}(t)$ are opposite each other to within some threshold angle, e.g. 26° ($\arccos 0.9$) or 37° ($\arccos 0.8$) from 180°.
From the known values of

$$v(t) = r\omega\cos(\omega t),$$
$$a(t) = -r\omega^2\sin(\omega t)$$
$$j(t) = -r\omega^3\cos(\omega t)$$

we compute the potential energy during head movements as follows. Since

$$E_{kin}(t) = v^2(t) = r^2\omega^2\cos^2(\omega t)$$

and

$$E = E_{kin}(t) + E_{pot}(t)$$

is conserved during an undamped oscillating motion, we make the approximation

$$E_{pot}(t) = r^2\omega^2\sin^2(\omega t) + C,$$

thereby ensuring that

$$E = r^2\omega^2 + C$$

is constant. One can check that the potential energy term, up to an arbitrary constant C, can be computed if we know the velocity, acceleration and jerk, because

$$E_{pot}(t) = r^2\omega^2\sin^2(\omega t) + C = -\frac{a(t)^2 v(t)}{j(t)} + C$$

In the above computations, we have treated $v(t)$, $a(t)$ and $j(t)$ as quantities in one dimension. In reality, these quantities live in three dimensions, of which OpenPose provides two. In our annotation software, we correct for the misalignment of $\vec{v}(t)$ and $\vec{j}(t)$ by multiplying the first term by $\frac{\vec{v}(t).\vec{j}(t)}{\sqrt{|\vec{v}(t)|.|\vec{j}(t)|}}$, but we overestimate the potential energy somewhat by taking the norm of the acceleration, disregarding misalignment with the other two quantities.
In this way, head gestures can be recognized as contiguous series of frames where the sum of the energies $E = E_{kin}(t) + E_{pot}(t)$stays far from zero, even though $v(t)$, $a(t)$ and $j(t)$ periodically come very close.

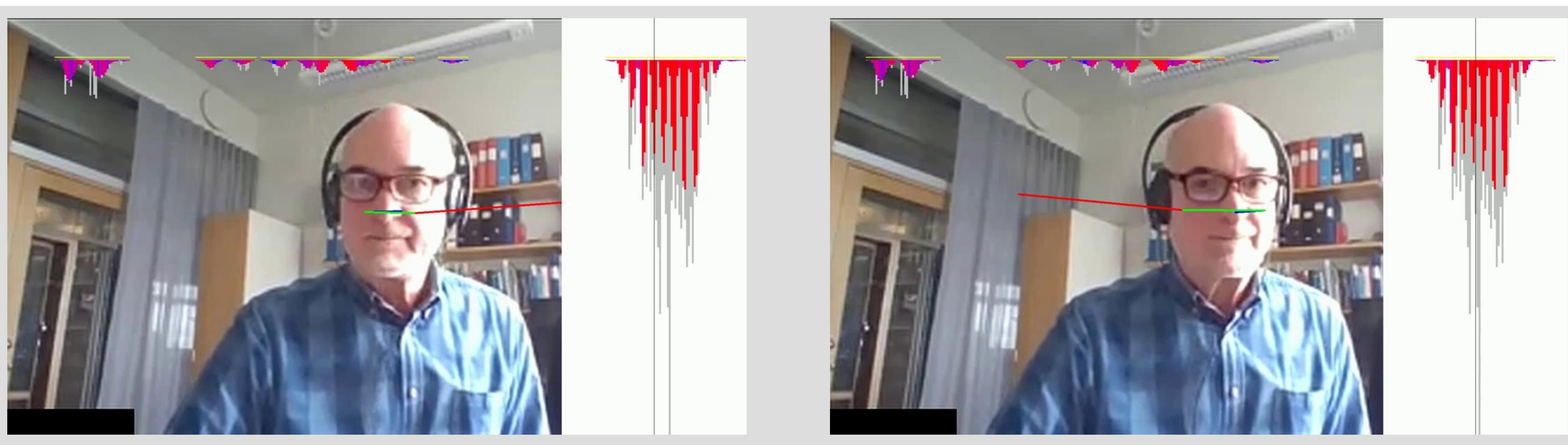

Figure 3: Anatomy of a repeated shake. When the software is run with OpenPose data for a single person, it blanks out one quarter of the movie frame to make an undisturbed view of the energy bars in the near future (right) and past (left) around the currently shown frame. It also draws the velocity, acceleration and jerk vectors. The velocity vector (blue) emanates from the nose tip, the acceleration vector from the velocity vector and the jerk vector from the acceleration vector. When viewing the movie, one clearly sees how jerk changes acceleration and how acceleration changes velocity.
Notice that some of the long red bars are separate from the rest. Those separations indicate that the kinetic energy was practically zero. Further notice that the 'free spaces' (except one) between all red bars are populated with grey bars, indicating that the algorithm decided that the condition for an oscillating head gesture was fulfilled, namely that velocity and jerk must be opposite each other.

The energy levels are visualised as continuous bar plots on top of the original video.

Figures 1 and 2 show screenshots from one of the GEHM videos [3] with energy bar plots and ELAN annotations for two subsequent intervals.

Five continuous bar plots show the motion energy displayed by the participants.

Vertical bars indicate motion energy during a frame. Grey bars are a measure of potential energy, coloured bars are kinetic energy, red for horizontal and blue for vertical movement.

In the first interval, the bar plot at the top, corresponding to the participant in the middle of the top row (SP01F), shows a succession of horizontal movements (the red bars): the participant shakes her head repeatedly while talking about something she did not understand.

The participant at the top left corner (SP05F) responds with a repeated nod to acknowledge SP01F. This is shown in the second interval, where the highlighted grey and blue bars correspond to that participant's vertical motion energy.

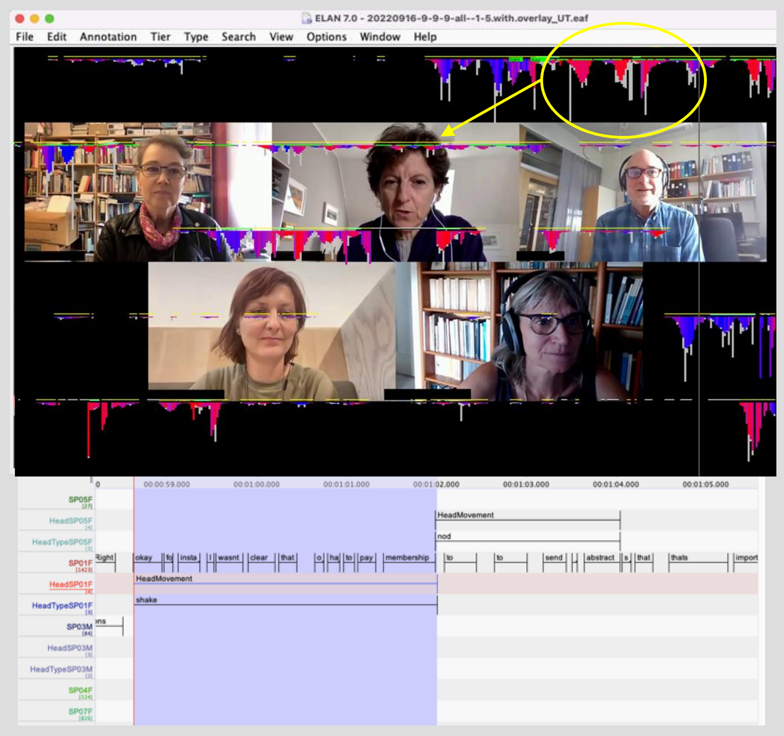

Figure 1: Example of repeated shake by one speaker (SP01F). The red bars circled in yellow correspond to the horizontal energy. The yellow circle highlights a significant part of the bar plots. The arrow points at the relevant speaker.

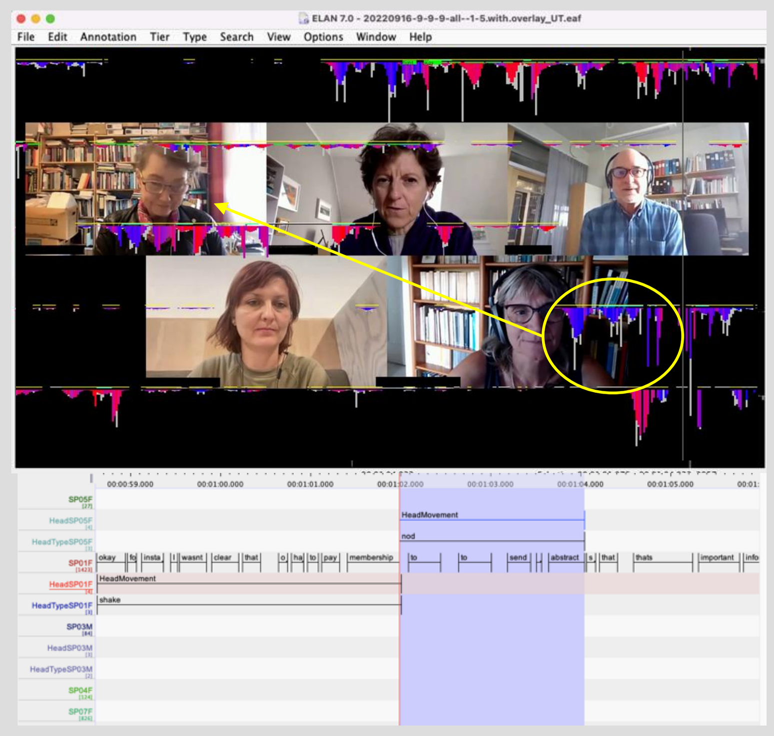

Figure 2: Example of nodding by one speaker (SP05F). The grey and blue bars circled in yellow correspond to the vertical energy. The yellow circle highlights a significant part of the bar plots. The arrow points at the relevant speaker.

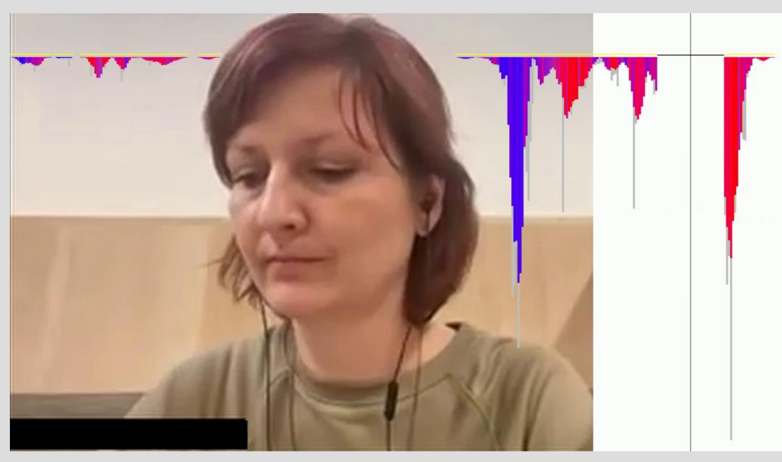

Figure 4: Missing (enough) OpenPose data. To estimate a scaling factor, the software needs the OpenPose positions for *both* ears. But the gap (here indicated with the black horizontal line) is itself also indicative of a head gesture: head turn far to the side. See also the red bars flanking the data gap.

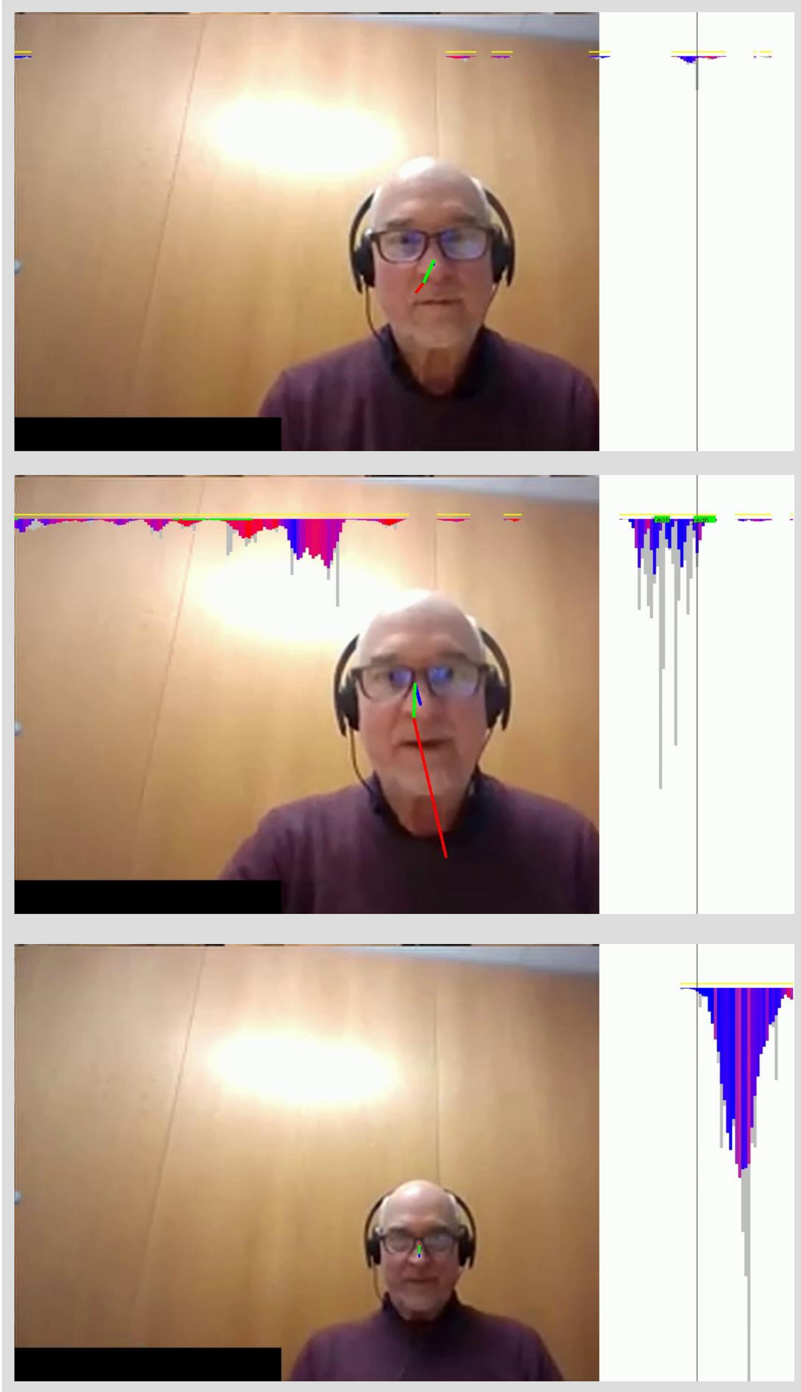

Figure 5. (to the right) The amount of energy in a head movement is only a rough indicator of the communicative quality of the movement. The pattern of the movement is at least as important. Some very small movements are head gestures, such as the repeated nod (Top). Some very large movements, such as getting closer to the screen, are not (bottom). And some movements are both large and communicative (middle).

References
[1] Paggio, P., Agirrezabal, M., Cristina, G. D., Jongejan, B., & Navarretta, C. (2026). Multimodal Entrainment and Feedback in Online Group Meetings. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026) (pp. 9139–9149). European Language Resources Association (ELRA).
[2] Khosrobeigi, Z. Koutsombogera, M. and Vogel, C. (2025) Methods and Findings in the Analysis of Alignment of Bodily Motion in Cooperative Dyadic Dialogue, *Multimodal Technologies and Interaction*, 9, (6).
[3] Paggio, P., Agirrezabal, M., Navarretta, C. and Vitasovic, L. (2024) (Multimodal Behaviour in an Online Environment: The GEHM Zoom Corpus Collection - https://aclanthology.org/2024.lrec-main.1038/). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 11890–11900, Torino, Italia. ELRA and ICCL.
[4] Pickering, M.J. and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.
[5] Litman, D., Paletz, S., Rahimi, Z., Allegretti, S. and Rice, C. 2016. The Teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1421–1431.
[6] Louwerse, M.M., Dale, R., Bard, E.G. and Jeuniaux, P. (2012) Behavior matching in multimodal communication is synchronized. *Cognitive Science*, 36(8):1404–1426.