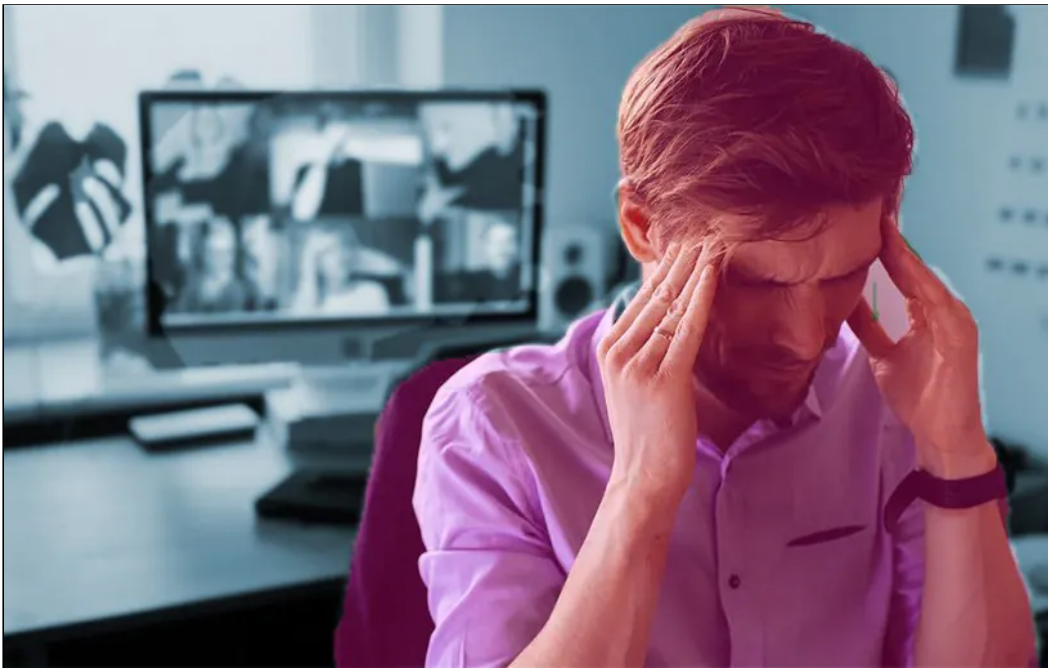


AI Helps Nervous Speakers to ‘Read the Room’ During Videoconferences

Originally published at <https://www.unite.ai/ai-helps-nervous-speakers-to-read-the-room-during-videoconferences/>



In 2013, a poll on common phobias determined that the prospect of public speaking was [worse than the prospect of death](#) for the majority of respondents. The syndrome is known as [glossophobia](#).

The COVID-driven [migration](#) from ‘in person’ meetings to online zoom conferences on platforms such as Zoom and Google Spaces has, surprisingly, not improved the situation. Where the meeting contains a large number of participants, our natural threat assessment abilities are impaired by the low-resolution rows and icons of participants, and the difficulty in reading subtle visual signals of facial expression and body language. Skype, for instance, has been found to be a poor platform for conveying non-verbal cues.

The effects on public speaking performance of perceived interest and responsiveness are well-documented by now, and intuitively obvious to most of us. Opaque audience response can cause speakers to hesitate and fall back to [filler speech](#), unaware of whether their arguments are meeting with agreement, disdain or disinterest, often making for an uncomfortable experience for both the speaker and their listeners.

Under pressure from the unexpected shift towards online videoconferencing inspired by COVID restrictions and precautions, the problem is arguably getting worse, and a number of ameliorative audience feedback schemes have been suggested in the computer vision and affect research communities over the last couple of years.

Hardware-Focused Solutions

Most of these, however, involve additional equipment or complex software that can raise privacy or logistics issues – relatively high-cost or otherwise resource-constrained approach styles that predate the pandemic. In 2001, MIT proposed the [Galvactivator](#), a hand-worn device that infers the emotional state of the audience participant, tested during a day-long symposium.



From 2001, MIT's Galvactivator, which measured skin conductivity response in an attempt to understand audience sentiment and engagement. Source: <https://dam-prod.media.mit.edu/x/files/pub/tech-reports/TR-542.pdf>

A great deal of academic energy has also been devoted to the possible [deployment of 'clickers'](#) as an Audience Response System (ARS), a measure to increase active participation by audiences (which automatically increases engagement, since it forces the viewer into the role of an active feedback node), but which has also been envisaged as a means of speaker encouragement.

Other attempts to 'connect' speaker and audience have included [heart-rate monitoring](#), the use of complex body-worn equipment to leverage electroencephalography, 'cheer meters', computer-vision-based [emotion recognition](#) for desk-bound workers, and the use of audience-sent [emoticons](#) during the speaker's oration.



From 2017, the EngageMeter, a joint academic research project from LMU Munich and the University of Stuttgart. Source: http://www.mariamhassib.net/pubs/hassib2017CHI_3/hassib2017CHI_3.pdf

As a sub-pursuit of the lucrative area of audience analytics, the private sector has taken a particular interest in gaze estimation and tracking – systems where each audience member (who may in their turn eventually have to speak), is subject to [ocular tracking](#) as an index of engagement and approbation.

All of these methods are fairly high-friction. Many of them require bespoke hardware, laboratory environments, specialized and custom-made software frameworks, and subscription to expensive commercial APIs – or any combination of these restrictive factors.

Therefore the development of minimalist systems based on little more than common tools for videoconferencing has become of interest over the last 18 months.

Reporting Audience Approbation Discreetly

To this end, a new research collaboration between the University of Tokyo and Carnegie Mellon University offers a novel system that can piggy-back onto standard videoconferencing tools (such as Zoom) using only a web-cam-enabled website on which lightweight gaze and pose estimation software is running. In this way even the need for local browser plugins is avoided.

The user's nods and estimated eye-attention are translated into representative data that is visualized back to the speaker, allowing for a 'live' litmus test of the extent to which the content is engaging the audience – and also at least a vague indicator of periods of discourse where the speaker may be losing audience interest.



With *CalmResponses*, user attention and nodding is added to a pool of audience feedback and translated into a visual representation that can benefit the speaker. See embedded video at end of article for more detail and examples. Source: https://www.youtube.com/watch?v=J_PhB4FCzk0

In many academic situations, such as online lectures, students may be entirely unseen by the speaker, since they have not turned their cameras on because of self-consciousness about their background or current appearance. *CalmResponses* can address this otherwise thorny obstacle to speaker feedback by reporting what it knows about how the speaker is looking at the content, and if they are nodding, without any need for the viewer to activate their camera.

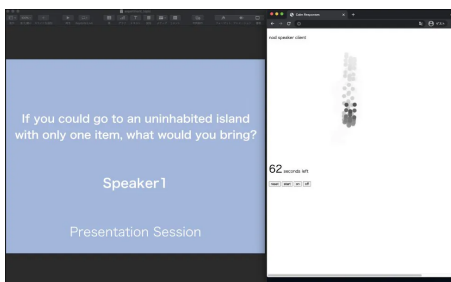
The [paper](#) is titled *CalmResponses: Displaying Collective Audience Reactions in Remote Communication*, and is a joint work between two researchers from UoT and one from Carnegie Mellon.

The authors offer a live web-based demo, and have released the [source code at GitHub](#).

The *CalmResponses* Framework

CalmResponses' interest in nodding, as opposed to other possible dispositions of the head, is based on research (some of it hailing [back to the era of Darwin](#)) that indicates that more than 80% of all listeners' head movements [are comprised of nodding](#) (even when they are [expressing disagreement](#)). At the same time, eye gaze movements have been shown [over numerous studies](#) to be a reliable index of interest or engagement.

CalmResponses is implemented with HTML, CSS, and JavaScript, and comprises three subsystems: an audience client, a speaker client, and a server. The audience clients pass eye gaze or head movement data from the user's webcam via WebSockets over the cloud application platform Heroku.

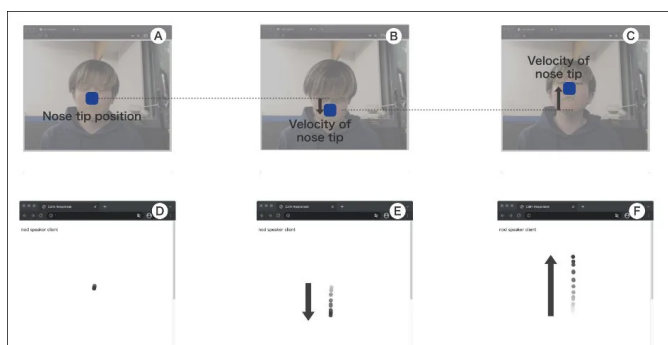


*Audience nodding visualized on the right in an animated movement under *CalmResponses*. In this case the movement visualization is available not only to the speaker, but to the entire audience.* Source: <https://arxiv.org/pdf/2204.02308.pdf>

For the eye-tracking section of the project, the researchers used [WebGazer](#), a lightweight, JavaScript-based browser-based eye-tracking framework that can run with low latency directly from a website (see link above for the researchers' own web-based implementation).

Since the need for simple implementation and rough, aggregate response recognition outweighs the need for high accuracy in gaze and pose estimation, the input pose data is smoothed according to mean values before being considered for the overall response estimation.

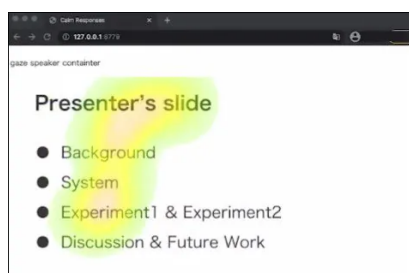
The nodding action is evaluated via the JavaScript library [clmtrackr](#), which fits facial models to detected faces in images or videos through [regularized landmark mean-shift](#). For purposes of economy and low-latency, only the detected landmark for the nose is actively monitored in the authors' implementation, since this is enough to track nodding actions.



The movement of the user's nose tip position creates a trail that contributes to the pool of audience response related to nodding, visualized in an aggregate manner to all participants.

Heat Map

While the nodding activity is represented by dynamic moving dots (see images above and video at end), visual attention is reported in terms of a heat map that shows the speaker and audience where the general locus of attention is focused on the shared presentation screen or videoconference environment.



All participants can see where general user attention is focused. The paper makes no mention of whether this functionality is available when the user can see a 'gallery' of other participants, which could reveal specious focus on one particular participant, for various reasons.

Tests

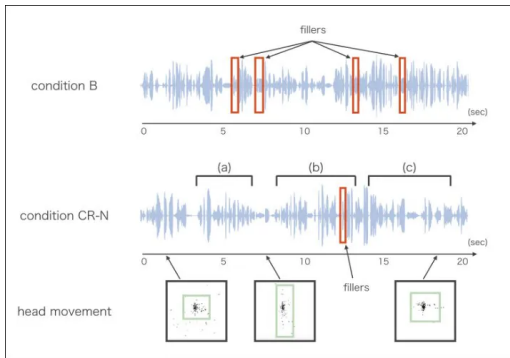
Two test environments were formulated for CalmResponses in the form of a tacit ablation study, using three varied sets of circumstances: in 'Condition B' (baseline), the authors replicated a typical online student lecture, where the majority of students keep their webcams turned off, and the speaker has no ability to see the faces of the audience; in 'Condition CR-E', the speaker could see gaze feedback (heat maps); in 'Condition CR-N', the speaker could see both the nodding and gaze activity from the audience.

The first experimental scenario comprised condition B and condition CR-E; the second comprised condition B and condition CR-N. Feedback was obtained from both the speakers and the audience.

In each experiment, three factors were evaluated: objective and subjective evaluation of the presentation (including a self-reported questionnaire from the speaker regarding their feelings about how the presentation went); the number of events of 'filler' speech, indicative of momentary insecurity and prevarication; and qualitative comments. These criteria are [common estimators](#) of speech quality and speaker anxiety.

The test pool consisted of 38 people aged 19-44, comprising 29 males and nine females with an average age of 24.7, all Japanese or Chinese, and all fluent in Japanese. They were randomly split into five groups of 6-7 participants, and none of the subjects knew each other personally.

The tests were conducted on Zoom, with five speakers giving presentations in the first experiment and six in the second.



Filler conditions marked as orange boxes. In general, filler content fell in reasonable proportion to increased audience feedback from the system.

The researchers note that one speaker's fillers reduced notably, and that in 'Condition CR-N', the speaker rarely uttered filler phrases. See the paper for the very detailed and granular results reported; however, the most marked results were in subjective evaluation from the speakers and audience participants.

Comments from the audience included:

'I felt that I was involved in the presentations' [AN2], "I was not sure the speakers' speeches were improved, but I felt a sense of unity from others' head movements visualization." [AN6]

'I was not sure the speakers' speeches were improved, but I felt a sense of unity from others' head movements visualization.'

The researchers note that the system introduces a new kind of artificial pause into the speaker's presentation, since the speaker is inclined to refer to the visual system to assess audience feedback before proceeding further.

They also note a kind of 'white coat effect', difficult to avoid in experimental circumstances, where some participants felt constrained by the possible security implications of being monitored for biometric data.

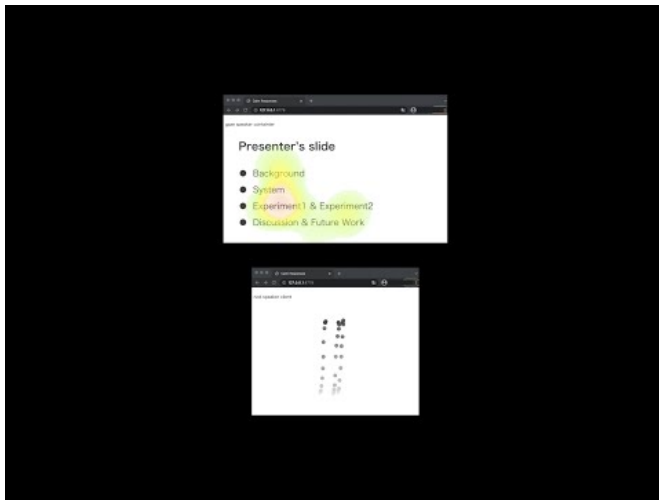
Conclusion

One notable advantage in a system like this is all the non-standard adjunct technologies needed for such an approach completely disappear after their usage is over. There are no residual browser plugins to be uninstalled, or to cast doubts in the minds of participants as to whether they should remain on their respective systems; and there is no need to guide users through the process of installation (though the web-based framework does require a minute or two of initial calibration by the user), or to navigate the possibility of users not having adequate permissions to install local software, including browser-based add-ons and extensions.

Though the evaluated facial and ocular movements are not as precise as they might be in circumstances where dedicated local machine learning frameworks (such as the YOLO series) might be used, this almost frictionless approach to audience evaluation provides adequate accuracy for broad sentiment and stance analysis in typical videoconference scenarios. Above all else, it's very cheap.

Check out the associated project video below for further details and examples.

CalmResponses: Displaying Collective Audience Reactions in Remote Communication



◻

First published 11th April 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More