KAIROS: Talking Heads and Moving Bodies for Successful Meetings

Jun-Ho Choi
Yonsei University, Incheon, Korea
idearibosome@yonsei.ac.kr

Marios Constantinides
Nokia Bell Labs, Cambridge, UK
marios.constantinides@nokia-bell-labs.com

Sagar Joglekar
Nokia Bell Labs, Cambridge, UK
sagar.joglekar@nokia-bell-labs.com

Daniele Quercia
Nokia Bell Labs, Cambridge, UK
daniele.quercia@nokia-bell-labs.com

ABSTRACT

Successful meetings create a safe environment for contribution; one that attendees feel engaged in and part of. Previous research has shown that meetings success depends not only on execution, but also on whether attendees feel psychologically safe. While this aspect is, to a great extent, partly observable through certain body cues during in-person meetings, they are often overlooked in virtual ones. To partly fix that, we developed "Kairos"–a system for multi-modal monitoring of virtual meetings that captures subtle body cues. We deployed it in 55 real-world corporate meetings and, upon six metrics for body cues, we built a model to predict a meeting's self-reported success, achieving an AUC as high as 79%. We found that certain body cues were more predictive of a meeting's success (defined as a linear combination of execution and psychological safety) than others (head movements, for example, were twice as predictive as hand movements), not least because they captured three typical meeting phases (its initiation, collective discussions, and turning points) whose presence (or absence) greatly mattered for success.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Meetings, multi-modal data, wearables, body cues

HotMobile '21, February 24–26, 2021, Virtual, United Kingdom

ACM ISBN 978-1-4503-8323-3/21/02…$15.00
https://doi.org/10.1145/3446382.3448361

ACM Reference Format:
Jun-Ho Choi, Marios Constantinides, Sagar Joglekar, and Daniele Quercia. 2021. KAIROS: Talking Heads and Moving Bodies for Successful Meetings. In *The 22nd International Workshop on Mobile Computing Systems and Applications (HotMobile '21), February 24–26, 2021, Virtual, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3446382.3448361

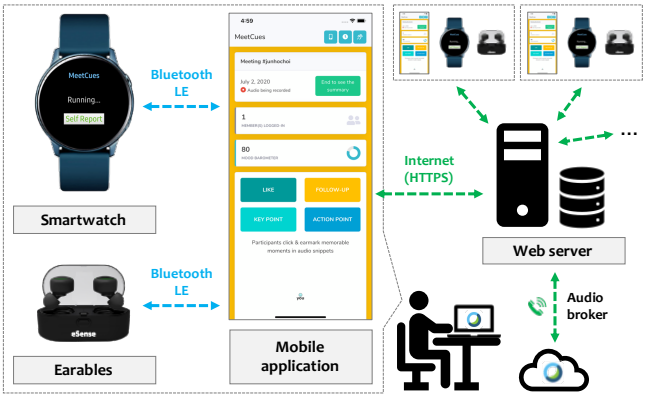


Figure 1: KAIROS monitors virtual meetings using a multi-sensory approach, and predicts their success.

1 INTRODUCTION

In any organization, whether it be a small company or a large corporation, meetings are the fuel for productivity. They enable collective work, facilitate decision-making, and foster execution [37]. To a great extent, meetings' engagement is an observable behavior, often expressed by body cues. As humans, we are attuned to these cues to *'make sense'* of an *in-person* meeting. However, the picture is slightly different in a virtual meeting. Not only visual communication cues, but also non-verbal ones such as body postures and gestures might go unnoticed [34]. Consider, for example, a virtual all-hands meeting, where one would expect a one-to-many

form of communication. One could imagine that not everyone would feel comfortable sharing their video streams in such a setting and, as such, the speaker/presenter might miss the opportunity to 'read the room' (i.e., interpret non-verbal cues expressed by meeting participants). Therefore, capturing body cues would not only bring them back to the virtual space, but also give meeting participants the opportunity to 'read the *[virtual]* room'. As the Related Work section (§2) summarizes, while previous research successfully employed techniques for analyzing audio-visual or textual information in virtual meetings [26, 41], the full spectrum of human senses is far from being captured. As the mobile-sensing technologies are growing in sophistication and ubiquitousness, we are now faced more than ever with a unique opportunity to measure human behavior that was previously impossible to measure [7, 27]. In this vein, we developed "Kairos", a system for multi-modal monitoring of virtual meetings, and made three main contributions:

- We developed a multi-modal system (mobile, watch, and earables), and deployed it in 55 real-world corporate virtual meetings for three weeks. We collected 1,968 minutes of multi-modal data, and 135 self-reported meetings success scores (§3).
- Using the collected data, we developed six body cues metrics based on the literature, and tested whether they are predictive of meetings success (§4). Our best performing model predicts success from these metrics with an AUC of 79%. We found that, among the six metrics, the most predictive one is 'head movements' (twice as important as hand movements); this metric was even more predictive than a meeting's emotional content derived from the meeting's transcript.
- Through a thematic analysis, we found that body cues were indicative of the presence (or absence) of three types of key phases in a meeting that happened to greatly impact its success: the meeting's *initiation*, *collective discussions*, and *turning points* (§4).

2 RELATED WORK

Meetings technologies: McGregor and Tang [26] developed a speech-based system that extracts "action items" from meetings' transcripts. NoteLook [10] supports note-taking through analyzing video data obtained from cameras in conference rooms, while [40] facilitates speaking time management. While textual support or audio-visual cues (e.g., prosody [9, 19]) received extensive attention, body language cues (e.g., gestures [34, 42]) have not been widely studied.

Mobile and wearables: Various technologies emerged that monitor people's psycho-physiological [21, 33] and behavioral [3] aspects. Smartphones and smartwatches [33], earables [22] and various other wearable devices are now fully equipped with sensors that make it possible to obtain multi-modal datasets [11]. For example, Gaggioli et al. [16] employed electrocardiogram sensors, wirelessly connected to smartphones, to detect stress. Mirjafari et al. [29] investigated workers' job performance from both smartphones and wearables. While such devices have been extensively used to study people's psychological and behavioral aspects, they have not been widely employed in the context of meetings.

Table 1: Data collected from our system.

Device	Sensor	Data
Phone	Accelerometer	$p\_acc = \{p\_acc_x, p\_acc_y, p\_acc_z\}$
	Gyroscope	$p\_gyr = \{p\_gyr_x, p\_gyr_y, p\_gyr_z\}$
Watch	Accelerometer	$w\_acc = \{w\_acc_x, w\_acc_y, w\_acc_z\}$
	Gyroscope	$w\_gyr = \{w\_gyr_x, w\_gyr_y, w\_gyr_z\}$
	Heart rate	w_hr
Earables	Accelerometer	$e\_acc = \{e\_acc_x, e\_acc_y, e\_acc_z\}$
	Gyroscope	$e\_gyr = \{e\_gyr_x, e\_gyr_y, e\_gyr_z\}$
_	Audio features	a_power, a_zcr
	Audio transcript	a_script

3 KAIROS SYSTEM

Our system consists of three components (Figure 1): (a) the mobile application, (b) the smartwatch and earables, and (c) the web server. Next, we describe each component.

3.1 Multi-modal monitoring of meetings

Mobile application. It is iOS and Android compatible, and was implemented using a hybrid approach that separates user interface (UI) elements from sensing. The UI was implemented using HTML5 and JavaScript, while native functionality was implemented using Swift and Kotlin for iOS and Android, respectively. In so doing, we kept the UI elements consistent on both platforms and optimized the sensor data collection for each platform independently. The application collects *motion* data (p_acc and p_gyr) at a sampling rate of 20 Hz, and transmits (acting as a local "server") the collected smartwatch's and earables' data to our web server.

Smartwatch application. We built a Tizen application that runs on Samsung Galaxy watches. It collects *motion* data (w_acc and w_gyr) from the watch's sensors at a sampling rate of 5 Hz, and heart rate (w_hr) from the device's Photoplethysmography sensor at a sampling rate of 1 Hz. All three modalities are transmitted to our mobile application via Bluetooth.

Earables. We used "eSense" [22] and collected *motion* data (e_acc and e_gyr) from its sensors. The mobile application scans nearby Bluetooth devices until it finds an "eSense" one, and broadcasts eSense's UDID. Upon registration, eSense

starts data collection, and transmits it to the mobile application every 200ms.

Web server. We developed a RESTful web server using the Python Tornado framework. It exposes an endpoint, which accepts a device's identifier, sensor type, and the raw timestamped sensor data. To prevent excessive server overloads, the mobile application transmits each sensor's data in bulk every 2 minutes (a threshold decided after testing trials), and all data is stored in a MongoDB instance. The web server also contains an "audio broker" service, which records the audio of a meeting [2]. From it, we obtained a meeting's recording and transcribed it using Google's Speech-to-Text API.

3.2 Setup and study execution

We recruited nine participants to always use our system in corporate meetings for three weeks. Given the complex "in-the-wild" nature of our study, we focused on a small-group [5], and all participants were consented in writing with strict anonymization in place.

We deployed our mobile application as an APK (Android) and via TestFlight (iOS), and distributed both watches and earables. We instructed our participants to wear both types of device, to install our mobile application in their phones, and to use it alongside their WebEx[1] meetings (the corresponding Cisco's WebEx meeting number was used to join in). Once the first user logged in, our audio broker joined WebEx via a calling-in function to obtain the meeting's audio. At the end of each meeting, our mobile application prompted a post-meeting survey comprised of two questions ($Q_{psychological}$ and $Q_{execution}$) on a 1-to-7 Likert scale, which represent our proxies for meeting success (as detailed in §4.2).

3.3 Data cleaning

In total, we collected data from 55 corporate meetings lasting 1,968 minutes, where 135 user sessions (127 sessions comprised of smartwatch data, and 71 ones of earables data) were generated in a period of three weeks. Our dataset comes from a diverse range of meetings with varying duration (), hours of day (), and days of week (). Meetings, on average, lasted for about 36 minutes with a minimum of three and a maximum of nine participants in each meeting, and all meetings were conducted during business hours (10am to 6pm, Mon-Fri). Post-meeting answers distributions were slightly skewed towards positive values ($\mu_{Q_{psychological}} = 5.80$ ), $\mu_{Q_{execution}} = 5.68$ ).

To prepare our multi-modal dataset for further analyses, we first grouped the data for each participant in each meeting and aligned their timestamps. For each sensor data (Table 1), we standardized its value by subtracting it from its average and dividing it by its standard deviation (i.e., $(x - \mu_x)/\sigma_x$), ensuring comparable values. We filled out missing data points for each feature with the average value of that feature over the whole dataset [38]; any missing data were due to participants not wearing the watch and/or earables at all times (due to the "in-the-wild" nature of our experiment).

[1]Cisco WebEx: https://www.webex.com/

4 STUDYING MEETING SUCCESS

4.1 Proxies for body cues

Using the collected dataset (Table 1), we designed six metrics based on the literature that captured body language cues [15, 21, 28, 34], and tested whether they were predictive of a meeting's success. The notation μ_f refers to the average values of a given sensor feature f. We computed the magnitude of the vector $|k_m| = \sqrt{{k_x}^2 + {k_y}^2 + {k_z}^2}$, for inertial measurement unit (IMU) sensors, which measure inertial data where $k \in \{p\_acc, p\_gyr, w\_acc, w\_gyr, e\_acc, e\_gyr\}$.

Vibrancy. Previous research found that vibrancy in speech is highly related to engagement [15]. We extracted two widely used features from audio [32], and partitioned them into 1-second windows. These are the root-mean-square power (a_power), and the zero-crossing rate (a_zcr). We computed $M_{vibrancy}$ from the averaged values of these two features ($M_{vibrancy} = \mu_{a\_power} + \mu_{a\_zcr}$). A larger μ_{a_power} suggests that the audio contains more vibrant conversations, and a larger μ_{a_zcr} value suggests a cleaner signal (i.e., a more speech-related signal) [8]. Hence, a larger $M_{vibrancy}$ value indicates a more conversational meeting.

Multi-tasking. People often use their mobile phones while taking the meeting call from another device. This might entail a positive contribution (e.g., viewing meeting-related files) or a less participatory approach (e.g., checking e-mails) [31]. To capture this, we computed $M_{multitasking}$ using the mobile IMU sensor readings, which detect the phone's movements ($M_{multitasking} = \mu_{p\_acc_m} + \mu_{p\_gyr_m}$). The larger its value, the more the phone has been moved around.

Heart rate. Past work has linked lower heart rates with focus and various states of consciousness (e.g., being awake) [18, 20, 21]. We computed M_{hr} from the average heart rate captured by the smartwatch ($M_{hr} = -\mu_{w\_hr}$). The larger its value, the lower the heart rate.

Head movement. People usually use body language to convey their (dis)agreement [34]. It is known that head gestures (e.g., nodding and shaking) are highly related to head rotations [28]. We computed M_{head} as the averaged magnitude of gyroscope readings obtained from earables ($M_{head} = \mu_{e\_gyr_m}$). The larger its value, the higher the number of head movements.

Postures. Various postures are related to people's attitudes during meetings [24]. Previous work reported that earable's acceleration data are more predictive of physical activity (i.e., postures) than its gyroscope data [28]. We defined $M_{postures}$, which captures changes in posture, from the average magnitude of earables accelerometer readings ($M_{postures} = \mu_{e\_acc_m}$). The larger its value, the more changes in posture.

Hand movement. People also express themselves with hand motions [34]. In literature, smartwatch IMU sensors are typically used to capture hand gestures [28]. We computed M_{hands} by summing the averaged magnitudes of the IMU sensor readings obtained from smartwatches ($M_{hands} = \mu_{w\_acc_m} + \mu_{w\_gyr_m}$). The larger its value, the more one has moved hands and wrists.

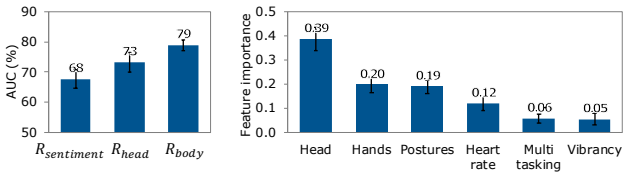
Our metrics are grounded in past work. However, they might not be exhaustive and universal as they could be influenced by the diversity of meetings, or even cultures. That is why we tested the extent to which they are predictive of success.

4.2 Self-reported success scores

We defined a meeting "success" score γ from self-reports (§3.2), and used it as the outcome variable. This score has been previously validated in a large-scale crowdsourcing study [13], and is independent from the sensor readings as it uses meeting participants' self-reported answers. In that previous study, we administered a 28-item questionnaire to 363 individuals whose answers were statistically analyzed through Principal Component Analysis (PCA). We found that two factors are sufficient to mostly capture whether a meeting is successful or not: (a) the extent to which participants felt listened during the meeting or motivated to be involved ($Q_{psychological}$), and (b) the extent to which the meeting had a clear purpose and structure ($Q_{execution}$). Using the loading factors of the first two components from the PCA analysis in [13] and the self-reports, we computed an aggregated score γ of each attendee as: $\gamma = (0.759 \cdot Q_{psychological}) + (0.673 \cdot Q_{execution})$. We binarized each γ using the median $\bar{\gamma}$ computed across all meetings' γ scores of a given attendee, and assigned them to positive ($\gamma > \bar{\gamma}$) and negative ($\gamma < \bar{\gamma}$) classes. Given each attendee's own rating, the choice of the median served as a representative way to discriminate the two classes.

4.3 Predicting success from body cues

We deployed a Random Forest (RF) model, which was the best predictive compared to SVM (69% AUC) and Logistic Regression (68% AUC) to predict our binary outcome variable (success) from the six metrics, while controlling for the hour of day (encoded using one-hot encoding) and day of the week. We refer to the RF model including all six body cues metrics as R_{body}. We measured performance using the area under curve (AUC) metric, and found the best RF model using a grid search algorithm. This algorithm searched iteratively the model's hyperparameters, employed a 5-fold cross-validation, and obtained an AUC for each classifier; we chose the RF classifier with the highest AUC.



(a) (b)

Figure 2: (a) AUCs of $R_{sentiment}$, R_{head}, and R_{body} models, and (b) R_{body} model features importance (best interpreted in a comparative way); Head movements were twice as important as hand movements.

Our best performing R_{body} model achieved 79% AUC. Inspecting its features (Figure 2b), we found that M_{head} was the most predictive feature, suggesting that head gestures were highly related to success; M_{hands}, which captured hand movements, was the next most prominent one; this was then followed by $M_{postures}$, which captured changes in posture. Surprisingly, $M_{vibrancy}$ was the least predictive metric. We speculate that, given the format of corporate meetings, not everyone might get the chance to speak up and contribute [31], and, as such, $M_{vibrancy}$ was partly compromised.

We ascertained R_{body}'s predictions by comparing it against a model predicting success from a meeting's emotional content rather than body cues. To capture that content, we defined $M_{sentiment}$, which measured the sentiment of the words used in a meeting. This choice was grounded on previous findings that established a significant relationship between positive emotions and productivity [6, 23, 25]. It is calculated as the fraction of positive (n_pos) over positive and negative (n_neg) words in the meeting's transcript (a_script): the larger its value, the more positive the meeting's sentiment. To categorize words, we used a bag-of-words technique upon the NRC Emotion Lexicon [30]. Using $M_{sentiment}$, we built a $R_{sentiment}$ model to predict success. For comparability, instead of using our R_{body} model with all six metrics, we selected its most predictive feature (M_{head}) and built a R_{head} model. This ensured that any difference between $R_{sentiment}$ and R_{head} performance would not be attributed to the number of features. We obtained 68% AUC for $R_{sentiment}$, and 73% AUC for R_{head} (Figure 2a). This translated into a 5% gain in the model trained with simply head movements compared to the model trained with emotional content, suggesting that head movements were more predictive than content.

4.4 Three typical phases in a meeting

While our best performing model reliably predicts an entire meeting's success, zooming into a meeting allows us to understand its temporal dynamics as it unfolds. To this end, we set out to explore the moments that contribute to a meeting's success, and the relevance of body cues in them.

We divided the original dataset of the 55 meetings into training (80%) and validation (20%) sets. For illustration purposes, Figure 3 reports 10 meetings from the validation set. Each row represents a meeting, and each cell represents a 5-min window[2]. For each 5-min window i in a given meeting, we defined two metrics:

$z_{success}$: We computed the probability p_i of the window contributing positively to success[3], and did so with the best performing model (R_{body}). We then transformed each p_i into its z-score $z^i_{success} = \frac{p_i - \mu_p}{\sigma_p}$, where μ_p and σ_p are the mean and variance of the probabilities p_i's across all windows. $z^i_{success}$ represents window i's *contribution to success* (depicted with a proportional color in Figure 3).

z_{body}: We computed a z-score for each *body* metric as $z^i_{body} = \frac{x^i_{body} - \mu_{body}}{\sigma_{body}}$, where $body \in \{Head, Hand, Postures, Heart\ rate, Multitasking, Vibrancy\}$, x^i_{body} is the *body*'s value in window i, and μ_{body} and σ_{body} are the mean and variance of x_{body}'s values across all windows.

To ease the interpretation of the z-score values, consider that a z-score value of 2 means that the original value was 2 standard deviations above the mean, and a z-score value of -2 means that it was 2 standard deviations below.

These two metrics work under the assumption that the model trained on an entire meeting reliably transfers on smaller temporal windows of it. That assumption appeared to be experimentally reasonable as we found that the predicted success for an entire meeting highly correlated with the predicted success averaged across all the 5-min windows of that meetings ($\rho = 0.88$).

We labeled the 5-min windows as *peaks*, where $z_{success}$ was positive, or *valleys*, where $z_{success}$ was negative. To understand what happened during the peaks and the valleys, we conducted a thematic analysis on these windows. We listened to the audio, read the transcripts, and annotated relevant statements using open coding [4]. We then examined these annotations using axial coding to identify relationships between them, and to ultimately extract themes. We reviewed the extracted themes in a recursive manner [4], with an emphasis on meeting success. We identified the three typical

[2]Spearman's rank correlation was computetd between our model's success prediction on an entire meeting and its success prediction on varying set of window sizes. The 5-min window yielded the highest correlation ($p = 0.74$).
[3]The probability of success is computed by taking the fraction of trees in the model that classify the window as positive contribution to success.

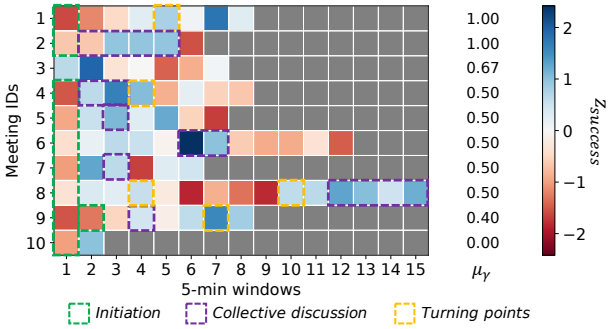


Figure 3: Heatmap of the 5-min windows contribution to success ($z_{success}$) across the ten meetings in the validation set. The z-scores greater than zero (blue) and lower than zero (red) represent positive and negative contribution, respectively. Points with no data (gray) indicate a meeting's ending time. The μ_γ denotes the meeting success score averaged across the attendees of a given meeting.

phases in a meeting that impacted its success (explained next), and named them: *(i) initiation*, *(ii) collective discussions*, and *(iii) turning points*, marked with green, purple, and yellow dotted boxes in Figure 3.

Initiation. This corresponds to the initial period, which includes actions such as setting up cameras and microphones, and periods in which participants could be muted. We observed the average $z_{success}$ (-1.03) to fall below 1 standard deviation from the mean, suggesting that a large initiation window negatively contribute to success. We also found that vibrancy and heart rate were under-expressed (both $z_{Vibrancy}$ and $z_{Heartrate}$ were < -1) in six out ten windows of this phase, indicating the presence of large periods of silence and diverse physiological states [18, 20, 21].

Collective discussions. We observed that the central parts of most meetings typically unfolded into two ways: as a *presentation* phase, or as a *discussion* phase. A presentation phase consists of one person talking, while a discussion phase consists of multiple people having a conversation. Both agreements (e.g., *"okay"*, *"sounds good"*) and disagreements (e.g., *"that's right, but I mean ..."*) were frequent in these two phases, and our model associated these *collective discussion* periods [31] with higher contribution to success (the average $z_{success}$ is 0.94). We found that vibrancy ($z_{Vibrancy} > 1$) was over-expressed in seven out of fourteen windows of this phase, and the three metrics of head ($z_{head} > 1$), hand ($z_{hand} > 1$), and postures ($z_{postures} > 1$) were over-expressed in twelve out of fourteen windows of this phase. This suggests a prolonged period of conversational and animated interactions, confirming that people's active bodily engagement contributes to success.

Turning points. Some meetings had turning points (e.g., the topic changed, or a latecomer joined). For example, at the 4th window of *Meeting ID #4*, one participant tried to end the call, but a question was immediately raised. By listening to the audio, we found that these periods stimulated attendees to (re)focus. Our model predicted the turning points windows as positively contributing to a meeting's success with an average $z_{success}$ of 0.92. We found that no body cue metric was over-expressed or under-expressed except the head movements in one out of five windows of this phase.

5 CONCLUSION

In our daily lives, body cues transmit a host of information to others, signaling our mood, attention, and emotions. Meetings are no exception to this. While we are attuned to such body signals during in-person meetings, these very same signals might go unnoticed during virtual ones. We collected a multi-modal dataset during virtual meetings, and showed that body cues are predictive of a meeting's success, even more than the meeting's emotional content, and that our six proxies for body cues essentially captured a meeting's initiation, collection discussions, and turning points; three typical meeting phases whose presence (or absence) greatly mattered for success.

From a practical standpoint, our work offers a deployable system that captures meeting success also from body cues, and generates analytics as a meeting unfolds; this is central to knowledge workers' productivity who spend a significant amount of their work time in meetings [39]. As of immediate practical use, KAIROS can be integrated with MeetCues [2], which is a companion platform for Cisco WebEx. MeetCues allows participants to engage during a meeting, and reflect on their experience through visual and interactive features. By integrating our system with MeetCues, we could improve participants' experience with three types of feedback: *i)* visual (e.g., aggregated body cues [35]), *ii)* auditory (e.g., manipulating audio features [14]), or even *iii)* haptics delivered on smartwatches [33]. Furthermore, we foresee that our system could benefit various types of meetings such as town halls or all-hands meetings. In these particular settings, as one-to-many or many-to-many conversations are natural forms of communication, our system could provide analytics to drive the course of a meeting (e.g., revise a meeting's agenda on the fly, if positive bodily engagement is lacking).

Our study looks at the problem of capturing body cues associated with meeting success, particularly in a period of the COVID-19 pandemic during which remote working is at its peak [17]. The current work has, however, limitations that call for future research efforts in the following areas. First, larger deployments would allow us to capture a wider variety of factors to control for (e.g., number of participants, topics, types of meeting), and, as such, generalize our findings. This leads us to our second limitation, which concerns the ground truth. While the meeting success score (ground truth) comes from a previous large-scale crowdsourcing study [13], future studies could explore causal relationships between the objective success of a meeting and participants' self-reported scores; this would further clarify the generalizability of our findings. The third limitation concerns the comparison of our model (R_{body}) against the emotional content ($R_{sentiment}$) of a meeting. While we used standard NLP methods (e.g., bag-of-words to extract sentiment), our ongoing work includes new NLP tools for scoring conversations in terms of types of social conversations [12, 36], and of empathy [43], which may well be used as alternative baseline models. The fourth limitation deals with privacy considerations. The analytics extracted from earables come with privacy concerns, yet KAIROS was built in a way that meeting participants could share what they felt comfortable to share. Furthermore, even if not shared, analytics based on body cues could create awareness. For example, they could be privately used by a meeting participant to reflect on how (s)he is likely to be perceived. Additionally, we foresee that our system would be used to provide analytics and interventions at a meeting/organizational level (aggregated) rather than on an individual level. Finally, our findings are based on sensed data from smartphones, smartwatches, and earable devices. Future studies could enrich these sensed modalities through camera-based head motion and facial gesture detection (bounded by users' willingness to share their video streams) [1]. Similarly, monitoring applications on work laptops could be used to track the types of task that are actually performed in real-time [21].

6 ACKNOWLEDGMENTS

We would like to thank Fahim Kawsar, Alessandro Montanari, and Chulhong Min for providing the earable devices, and Michael Eggleston for his useful feedback.

REFERENCES

[1] Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. 2015. Decoupling facial expressions and head motions in complex emotions. In *Proc. of the IEEE International Conference on Affective Computing and Intelligent Interaction*. IEEE, 274–280.

[2] Bon Adriel Aseniero, Marios Constantinides, Sagar Joglekar, Ke Zhou, and Daniele Quercia. 2020. MeetCues: Supporting online meetings experience. In *Proc. of the IEEE Visualization Conference*.

[3] Sangwon Bae, Tammy Chung, Denzil Ferreira, Anind K. Dey, and Brian Suffoletto. 2018. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive Behaviors* 83 (2018), 42–47.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[5] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 981–992.

[6] Fabio Calefato, Filippo Lanubile, Federico Maiorano, and Nicole Novielli. 2018. Sentiment polarity detection for software development. *Empirical Software Engineering* 23, 3 (2018), 1352–1382.

[7] Andrew T. Campbell et al. 2008. The rise of people-centric sensing. *IEEE Internet Computing* 12, 4 (2008), 12–21.

[8] Alexandru Caruntu, Gavril Toderean, and Alina Nica. 2005. Automatic silence/unvoiced/voiced classification of speech using a modified Teager energy feature. In *Proc. of the WSEAS Conference on Dynamical Systems and Control*. 62–65.

[9] Marcela Charfuelan and Marc Schröder. 2011. Investigating the prosody and voice quality of social signals in scenario meetings. In *Proc. of the International Conference on Affective Computing and Intelligent Interaction*. Springer, 46–56.

[10] Patrick Chiu, Ashutosh Kapuskar, Sarah Reitmeier, and Lynn Wilcox. 1999. NoteLook: Taking notes in meetings with digital video and ink. In *Proc. of the ACM International Conference on Multimedia*. 149–158.

[11] Jun-Ho Choi and Jong-Seok Lee. 2019. EmbraceNet: A robust deep learning architecture for multimodal classification. *Information Fusion* 51 (2019), 259–270.

[12] Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In *Proceedings of The Web Conference 2020*. 1514–1525.

[13] Marios Constantinides, Sanja Šćepanović, Daniele Quercia, Hongwei Li, Ugo Sassi, and Michael Eggleston. 2020. ComFeel: Productivity is a matter of the senses too. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–21.

[14] Jean Costa, Malte F. Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. 2018. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *Proc. of the ACM Conference on Human Factors in Computing Systems*.

[15] Keith Curtis, Gareth J. F. Jones, and Nick Campbell. 2015. Effects of good speaking techniques on audience engagement. In *Proc. of the ACM International Conference on Multimodal Interaction*. 35–42.

[16] Andrea Gaggioli et al. 2013. A mobile data collection platform for mental health research. *Personal and Ubiquitous Computing* 17, 2 (2013), 241–251.

[17] Arpit Gupta. 2020. Accelerating remote work after COVID-19. In *Proc. of the Covid Recovery Symposium*. 1–2.

[18] Jennifer A. Healey and Rosalind W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. on Intelligent Transportation Systems* 6, 2 (2005), 156–166.

[19] Julia Hirschberg. 2002. Communication and prosody: Functional aspects of prosody. *Speech Communication* 36, 1-2 (2002), 31–43.

[20] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology* 92, 1-2 (2004), 84–89.

[21] Harmanpreet Kaur, Alex C. Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi T. Iqbal. 2020. Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 1–15.

[22] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[23] Janice R Kelly and Sigal G Barsade. 2001. Mood and emotions in small groups and work teams. *Organizational behavior and human decision processes* 86, 1 (2001), 99–130.

[24] Nicky Kern, Bernt Schiele, Holger Junker, Paul Lukowicz, and Gerhard Tröster. 2003. Wearable sensing to annotate meeting recordings. *Personal and Ubiquitous Computing* 7, 5 (2003), 263–274.

[25] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?. In *Proc. of the ACM Conference on Mining Software Repositories*. 247–258.

[26] Moira McGregor and John C. Tang. 2017. More to meetings: Challenges in using speech-based technology to support meetings. In *Proc. of the ACM Conference on Computer Supported Cooperative Work*. 2208–2220.

[27] Geoffrey Miller. 2012. The smartphone psychology manifesto. *Perspectives on Psychological Science* 7, 3 (2012), 221–237.

[28] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Exploring audio and kinetic sensing on earable devices. In *Proc. of the ACM Workshop on Wearable Systems and Applications*. 5–10.

[29] Shayan Mirjafari et al. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–24.

[30] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[31] Karin Niemantsverdriet and Thomas Erickson. 2017. Recurring Meetings: An experiential account of repeating meetings in a large organization. *Proc. of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–17.

[32] Costas Panagiotakis and Georgios Tziritas. 2005. A speech/music discriminator based on RMS and zero-crossings. *IEEE Trans. on Multimedia* 7, 1 (2005), 155–166.

[33] Sungkyu Park, Marios Constantinides, Luca Maria Aiello, Daniele Quercia, and Paul Van Gent. 2020. WellBeat: A framework for tracking daily well-being using smartwatches. *IEEE Internet Computing* 24, 5 (2020), 10–17.

[34] Kęstutis Peleckis and Valentina Peleckienė. 2015. Nonverbal communication in business negotiations and business meetings. *International Letters of Social and Humanistic Sciences* 62 (2015), 62–72.

[35] Chao Ying Qin, Marios Constantinides, Luca Maria Aiello, and Daniele Quercia. 2020. HeartBees: Visualizing crowd affects. In *Proc. of the VIS Arts Program*. IEEE, 1–8.

[36] Alexander Robertson, Luca Maria Aiello, and Daniele Quercia. 2019. The language of dialogue is complex. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 428–439.

[37] Steven G. Rogelberg, Joseph A. Allen, Linda Shanock, Cliff Scott, and Marissa Shuffler. 2010. Employee satisfaction with meetings: A contemporary facet of job satisfaction. *Human Resource Management* 49, 2 (2010), 149–172.

[38] Peter Schmitt, Jonas Mandel, and Mickael Guedj. 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics* 6, 1 (2015), 1.

[39] Yolande Strengers. 2015. Meeting in the global workplace: Air travel, telepresence and the body. *Mobilities* 10, 4 (2015), 592–608.

[40] Diane Tam, Karon E. MacLean, Joanna McGrenere, and Katherine J. Kuchenbecker. 2013. The design and field observation of a haptic notification system for timing awareness during oral presentations. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 1689–1698.

[41] John Tang et al. 2012. Time travel proxy: Using lightweight video recordings to create asynchronous, interactive meetings. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 3111–3120.

[42] Nicole Torres and Joep Cornelissen. 2019. When you pitch an idea, gestures matter more than words. *Harvard Business Review* (2019).

[43] Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proc. of the ACM on Human-Computer Interaction* 1, CSCW (2021), 1–19.