This is the Pre-Published Version.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in MM '23: Proceedings of the 31st ACM International Conference on Multimedia, https://doi.org/10.1145/3581783.3612873.

MultiMediate 2023: Engagementment Level Detection using Audio and Video Features

ABSTRACT

Participants in conversations may either actively engage in and enjoy the process, or become disinterested and allow their minds to wander. Measuring the engagement level of the participants can be quantified and used for analyzing numerous conversation scenarios such as business negotiations and remote learning, where the engagement level may indicate the success of negotiations or reflect the quality of the learning process. This paper presents a novel approach to engagement estimation using a versatile, resource-efficient, end-to-end training network. The network leverages body motion and audio characteristics, addressing challenges in data quality and model interpretability. Validated using the Noxi database and tested in the MultiMediate'23 competition, our approach achieved state-of-the-art performance, improving the baseline model's concordance coefficients correlation from 59% to 70% on the test set.

KEYWORDS

engagement, machine learning, neural networks

ACM Reference Format:

. 2023. MultiMediate 2023: Engagementment Level Detection using Audio and Video Features. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In dialogue, participants interact with each other, exchanging opinions and sharing stories. The level of participation and involvement of each participant in the interaction is referred to as "engagement." It is also defined as "the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" [14]. Estimating the value of engagement plays a vital role across various domains including education [18], healthcare [6] and UX optimization [2].

Traditional methods for evaluating engagement have predominantly relied on questionnaires and human observations, providing subjective measures of engagement [1, 17]. However, with the rapid advancement of machine learning (ML) techniques in recent years, a shift towards more objective and automated measures of engagement has been observed. The direct application of ML techniques to analyze engagement levels has emerged as a promising approach, offering the potential for real-time and accurate measurements of engagement[3, 16]. Although the potential of utilizing ML and deep learning techniques to conduct engagement evaluation is recognized by researchers, it also presents several key challenges.

Firstly, the quality and quantity of data are crucial for the effective training of machine learning models. In the context of engagement estimation, this implies the need for multi-modal records of the target subjects including facial expressions, user interactions and other physiological data in a non-intrusive way[7]. Current open datasets about engagement level focusing mainly on one or two modalities [5]. For example, DAiSEE dataset provides with facial expression [9]. Others might use wearable and external sensors to collect ambient environmental data for predictive model [8]. The lack of high-quality multi-modal training data, coupled with the bias induced by intrusive devices, poses significant challenges to the development of a reliable machine learning model that can be generally applied across various human interaction scenarios[4].

Moreover, the interpretability of the trained ML models has hindered researchers to correlate the performance of the model with the recorded physiological features. While these models can make accurate predictions, understanding why they made a specific prediction can be difficult. This lack of interpretability can be a barrier to trust and acceptance of these models[11].

In order to address the challenges identified, we propose a training network that is capable of addressing a wide range of situations and is easy to implement. This network utilizes both body motion and audio characteristics of individuals involved in an conversation. The selection of these features was based on the Noxi database [13], which is a novel database consisting of natural interactions between novice and expert individuals in multiple languages. These interactions were conducted through screens and focused on exchanging and retrieving information. The database includes records of both audio and video channels, which were captured using Microsoft Kinect2 [12]. Additionally, the dataset contains annotations for each frame indicating the level of engagement, which serves as the ground-truth labels for the MultiMediate'23 engagement estimation competition [13]. To validate our approach, we participated in this competition and achieved the state-of-the-art performance. Our approach resulted in a significant enhancement, increasing the concordence coefficients correlation of the baseline model from 59% to 70% evaluated on the test set [13].

2 METHODOLOGY

In this section, we will delve into the specifics of our methodology, showcasing how we proposed to conduct feature engineering based on the findings made from the Noxi database [13], and provide a detailed explanation of our model's design and its implementation process.

2.1 Data Observation

We first conduct data visualization using the session recordings to identify potential features that could be significant for our task. From each session in the training set, we extract a sequence of 125 consecutive frames (equivalent to 5 seconds) that exhibit the lowest or highest average engagement value. This allows us to discern any common features that might indicate useful attributes.

Our findings suggest that participant engagement peaks when they are trying hard to convey specific information to the listener. This can occur in situations where an expert is attempting to explain something to a novice, or when a novice expresses their

Conference'17, July 2017, Washington, DC, USA
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

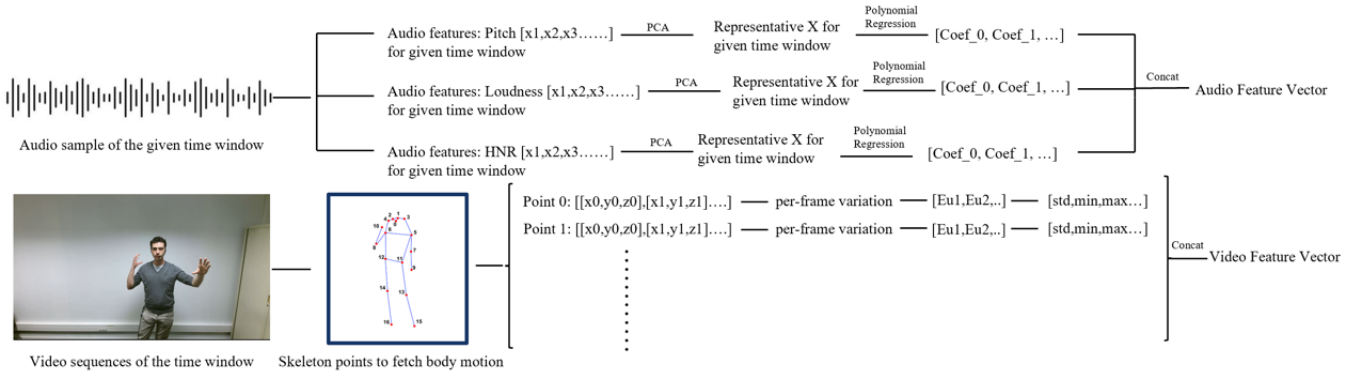

Figure 1: The workflow of feature construction of a single subject, this process will be repeated for both participants involved in a interaction session, then the four feature vectors belong to the two individuals will be concatanated to form the final feature vector.

understanding and poses questions. Such states often accompanied by extensive body gestures and a tone rich in emotion.

We have also noted that instances of minimal engagement typically correspond to intentionally introduced disruptions. For example, the engagement level of the participant is normally low when the other person involved in the conversation is engaged in irrelevant tasks such as taking a phone call or adjusting experimental equipment. Participants anticipating these interruptions generally exhibit subdued body language and maintain silence in such situations.

Based on above observations, we propose a multi-input model that leverages features from audio signals to represent speaking characteristics, and video features to describe the participant's body movements. The feature of both modalities would be builded exploiting the past information to predict the engagement status of the current moment, and a session-based min-max normalization is applied to each selected feature. The feature engineering process will be conducted for each frame and the overall workflow of our proposed approach is presented in Fig 1

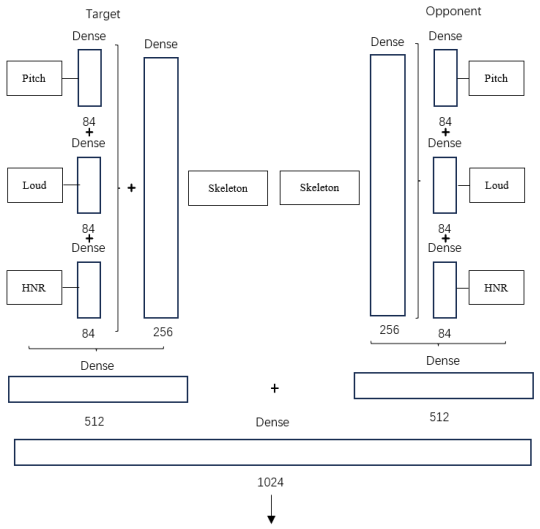

Figure 2: The proposed prediction model

2.2 Audio Feature Engineering

Our goal is to construct the audio feature in a way that it can fully capture the key audio characteristics of the speech. This includes looking at important factors like pitch, loudness, and the Harmonics-to-Noise Ratio (HNR) within a set time window. It's worth noting that pitch and loudness can often show the speaker's emotions [10]. At the same time, HNR is key parameter to indicate the clarity of a speech. This could be really important for keeping the listener's attention during a conversation [15], which will significantly impact the engagement level.

We aim to devise a method for capturing essential audio characteristics of speech, focusing on three key factors: pitch, loudness, and the Harmonics-to-Noise Ratio (HNR). It is well-documented that pitch and loudness can significantly influence the perception of the speaker's emotions [10], while HNR is a crucial parameter for speech clarity, contributing to the listener's attention and engagement during conversation [15]. Each of these factors are normally represented by a set of related parameters. To simplify this representation and reduce the dimensionality for further processes, we will employ Principle Component Analysis (PCA), through which we can represent each factor by a single value. Hence, each of these factors within the local time window would be represented by a single value across the time series. We then propose applying polynomial regression to these time-series, allowing us to model and capture complex, potentially non-linear relationships with time. By choosing a higher degree polynomial, we aim to capture more complex features from these series. The coefficients from these regressions then form the final feature vector, creating a comprehensive, multi-dimensional representation of the sound sample within the given time window.

2.3 Video Feature Engineering

We utilize the Noxi database's data collected using Microsoft's Kinect2 [13], a device that records 'skeleton points' or specific locations on a person's body that indicate body motion. Kinect2 captures up to 25 unique skeleton points, such as the head, neck, shoulders, elbows, hands, knees, and feet, offering a comprehensive view of a person's posture, movement, and interactions with their surroundings.

For each skeleton point, we extract the 3D position and calculate the frame-by-frame variations using Euclidean distances. This enables us to quantify the amount of movement between each frame. After normalizing these variation values, we represent them using a six-dimensional vector (refer to Table 1). This vector provides a comprehensive summary of the per-frame variations, capturing the range and distribution of movement across time, thus is well-suited for subsequent predictive modeling.

Feature	Description
min	Minimum value of variation
max	Maximum value of variation
std	Standard error of variation values
mean	Average value of variation
qt15	Value at the 15% quantile of variation range
qt75	Value at the 75% quantile of variation range

Table 1: Illustration of the builded feature vector

2.4 Model Design

In our research, we have conducted extensive feature engineering to prepare our data for the prediction model. This comprehensive process allows us to primarily utilize fully-connected layers in the formation of our model. The use of fully-connected layers is advantageous as it increases the flexibility and robustness of the model. This design choice also provides a straightforward path for the model to adapt to changes in input modalities, which is crucial in a dynamic data environment. The overall structure of the proposed network is shown in Fig

As illustrated in Figure 1, each modality of input is transformed into a one-dimensional feature vector. This transformation process is essential as it standardizes the input data, making it easier for the model to process. Each of these feature vectors is then individually processed through a separate block of fully connected layers. This step results in an embedding vector of the same length for each input modality. This is because we believe that each modality carries unique and valuable information, and we do not want to introduce any bias by giving more importance to one modality over another.

Once the embedding vectors are generated, they are concatenated to create a comprehensive representation of both the speaker and listener in a given time window. This combined vector encapsulates the information from all modalities, providing a holistic view of the input data.

Finally, the final feature vector is fed into the prediction block to generate predictions, which is also composed of fully connected layers. We utilize Adam as the optimizer and employ the mean-squared error (mae) as the loss function. Additionally, the learning rate of the network is set to 1e-6 to mitigate potential overfitting issues.

We intend to train a pair of distinct networks, each mirroring the other in structure, to specifically predict the engagement value of the two participants in an interaction session. For clarity, we will designate the subject for whom we are currently predicting the engagement value as the 'target,' while the other participant will be referred to as the 'opponent.'

3 EXPERIMENTAL RESULTS

In this section, we will illustrate the methodology we take to conduct evaluation, aiming at justifying our choice of design from the results of the ablation study and further identifying the key impact factors during fine-tuning our model. We will then present the final result of our model by participating in the engagement estimation of the multimediate challenge 2023 utilizing the Noxi dataset [13] where we achieved the state-of-the-art performance.

The engagement estimation challenge called for per-frame predictions for both participants in an interaction and employ the Concordance Correlation Coefficient (CCC) as evaluation metric. The CCC, considering both variability between and correlation among measurements, furnished a comprehensive evaluation of our model's performance.

3.1 Ablation Study

An ablation study was conducted using the Noxi validation set to understand the contribution of different parts of our model and guide the fine-tuning process.

Used Feature	Feature Engineering	Val CCC
Audio only	**raw data PR coef**	**0.623**
Audio only	variation data PR coef	0.358
Audio only	per-frame var data summary	0.423
Audio only	raw data summary	0.541

Table 2: Validation performance of different audio feature engineering, where PR coef is polynomial regression coefficient and Val CCC is the validation concordance correlation coefficient

3.1.1 Choice of audio feature engineering. In the first part of the ablation study, we examined the impact of different strategies for audio feature engineering on the performance of the model. In particular, we evaluated four different approaches for audio feature engineering: using raw data Polynomial Regression (PR) coefficients, variation data PR coefficients, per-frame variation data summary, and raw data summary. The performance for each approach was evaluated in terms of the Concordance Correlation Coefficient (CCC) on the validation set, as shown in Table 2.

The results showed that using raw data PR coefficients provided the highest validation CCC of 0.623, significantly outperforming the other approaches. This finding suggests that the raw data PR coefficients more effectively capture the essential characteristics of the audio signals than the other methods. This approach was therefore chosen for audio feature engineering in our model.

3.1.2 Choice of video feature engineering. In the second part of the ablation study, we evaluated different strategies for video feature engineering. Again, we tested four different approaches: raw data PR coefficients, variation data PR coefficients, per-frame variation data summary, and raw data summary. The performance for each

approach was evaluated in terms of the CCC on the validation set, as shown in Table 3.

The results indicated that using the per-frame variation data summary provided the highest validation CCC of 0.455, outperforming the other methods. This finding suggests that the per-frame variation data summary more effectively captures the essential characteristics of the video signals, specifically the movements of the participants, than the other approaches. Therefore, this approach was chosen for video feature engineering in our model.

Used Feature	Feature Engineering	Val CCC
Video only	raw data PR coef	0.184
Video only	variation data PR coef	0.413
Video only	**per-frame var data summary**	**0.455**
Video only	raw data summary	0.223

Table 3: Validation performance of different video feature engineering, where PR coef is polynomial regression coefficient and Val CCC is the validation concordance correlation coefficient

3.1.3 Impact of multiple modalities. We then examined the impact of using multiple modalities on the performance of our model. Three configurations were tested: using audio features only, using video features only, and using both audio and video features. The performance for each configuration was evaluated in terms of the CCC on the validation and test sets, as shown in Table 4.

The results indicated that using both audio and video features provided the highest validation CCC of 0.732 and test CCC of 0.68. In comparison, using only audio features yielded a validation CCC of 0.623, and using only video features resulted in a validation CCC of 0.548.

These results suggest that incorporating both audio and video modalities significantly improved the performance of the model. The combined use of audio and video features provided a more comprehensive representation of the interaction sessions, capturing both the speech characteristics and body movements of the participants. This holistic view of the interaction sessions was instrumental in improving the model's ability to predict participant engagement. Therefore, we chose to incorporate both audio and video features in our model.

This analysis underscores the importance of utilizing multiple modalities in engagement estimation tasks, as it allows the model to capture a broader range of features and patterns in the data, thereby improving prediction accuracy.

Used Feature	Time Windows Size	Val CCC	Test CCC
Audio	5s	0.623	-
Video	5s	0.548	-
Audio + Video	**5s**	**0.732**	**0.68**

Table 4: Impact of multiple modalities, where CCC is the concordance correlation coefficient

3.2 Model Fine-tuning

Several parameters are found to be determinant to the performance of our model.

3.2.1 Size of Time Window. Increasing the time window size from 5 seconds to 10 seconds resulted in improved system performance. This enhancement was evident in both the Validation CCC and Test CCC metrics as shown in Table 5. Utilizing longer time windows allowed for a more comprehensive representation of the data, capturing relevant information and underlying patterns over extended periods. The broader temporal range also helped mitigate the impact of short-term fluctuations and noise, resulting in a more stable feature set. Additionally, longer time windows facilitated the identification of long-term dependencies and trends within the data. Overall, the findings support the effectiveness of longer time windows in improving system performance.

Used Feature	Time Window Size	Val CCC	Test CCC
Audio + Video	5s	0.732	0.664
Audio + Video	**10s**	**0.741**	**0.68**

Table 5: performance of different length of the time window, where CCC is the concordance correlation coefficient

3.2.2 Target and opponent feature dimension. To determine the importance of features in predicting the engagement level of the target subject, we limited the shape of the final feature vector of the target subject and the opponent. Our findings presented in Table 6 indicate that the model performs better when relying more on the opponent's constructed features. This suggests that the engagement level of the target subject is significantly influenced by the reactions and behavior of the conversational partner.

Target FD	Opponent FD	Val CCC	Test CCC
512	512	0.741	0.68
512	216	0.735	0.672
216	512	0.745	0.695

Table 6: FD stands for the dimensions of the feature vector, the best-practice feature combination and local window size highlighted in the previous tables are used here, where CCC is the concordance correlation coefficient

4 CONCLUSION

In this study, we have developed a lightweight machine learning model that can assess the level of engagement between two individuals in a real-time conversation. Our model incorporates various techniques for feature engineering, allowing us to accurately capture the subjects' body motion patterns and relevant audio features from a recent time period. Through an ablation study, we have confirmed the importance of using multiple modes of input and our chosen feature engineering methodology. Additionally, our model has achieved state-of-the-art performance in the MultiMediate'23 engagement estimation challenge, further validating the efficacy of our approach. Moving forward, our research will focus on training with more advanced neural networks like transformers [19], with the aim of fully leveraging the valuable multi-modal information we have identified.

REFERENCES

[1] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. 2006. Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of school psychology* 44, 5 (2006), 427–445.

[2] S. Bultmann and Sven Behnke. 2021. Real-Time Multi-View 3D Human Pose Estimation using Semantic Feedback to Smart Edge Sensors. *Robotics: Science and Systems* (2021). https://doi.org/10.15607/RSS.2021.XVII.040

[3] Meredith Carroll, Mitchell Ruble, Mark Dranias, Summer Rebensky, Maria Chaparro, Joanna Chiang, and Brent Winslow. 2020. Automatic detection of learner engagement using machine learning and wearable sensors. *Journal of Behavioral and Brain Science* 10, 3 (2020), 165–178.

[4] A. Davoudi, Ruba Sajdeya, Ron Ison, Jennifer Hagen, P. Rashidi, C. Price, and P. Tighe. 2023. Fairness in the prediction of acute postoperative pain using machine learning models. *Frontiers in Digital Health* (2023). https://doi.org/10.3389/fdgth.2022.970281

[5] M Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1–20.

[6] Robert Dowd, Lauren H Jepson, Courtney R Green, Gregory J Norman, Roy Thomas, and Keri Leone. 2023. Glycemic Outcomes and Feature Set Engagement Among Real-Time Continuous Glucose Monitoring Users With Type 1 or Non–Insulin-Treated Type 2 Diabetes: Retrospective Analysis of Real-World Data. *JMIR diabetes* 8 (2023), e43991.

[7] Frankie Fan and Yun Shi. 2022. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Preprints* (2022). https://doi.org/10.20944/preprints202201.0365.v1

[8] Nan Gao, Max Marschall, Jane Burry, Simon Watkins, and Flora D Salim. 2022. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data* 9, 1 (2022), 261.

[9] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).

[10] Taiga Haruta, Mariko Oda, and Kohei Arai. 2022. Emotion Estimation Method with Mel-frequency Spectrum, Voice Power Level and Pitch Frequency of Human Voices through CNN Learning Processes. *International Journal of Advanced Computer Science and Applications* 13, 11 (2022).

[11] Rui Meng, Zhen Yue, and A. Glass. 2020. Predicting User Engagement Status for Online Evaluation of Intelligent Assistants. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.1007/978-3-030-72113-8_29

[12] Microsoft. [n. d.]. Azure Kinect Specification. https://learn.microsoft.com/en-us/azure/kinect-dk/windows-comparison

[13] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. 2023. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*.

[14] Isabella Poggi. 2007. Mind, hands, face and body. *A goal and belief view of multimodal communication. Weidler, Berlin* (2007).

[15] Joseph Rovetti, Huiwen Goy, Michael Zara, and Frank A Russo. 2022. Reduced semantic context and signal-to-noise ratio increase listening effort as measured using functional near-infrared spectroscopy. *Ear and Hearing* 43, 3 (2022), 836–848.

[16] George K Sidiropoulos, George A Papakostas, Chris Lytridis, Christos Bazinas, Vassilis G Kaburlasos, Efi Kourampa, and Elpida Karageorgiou. 2020. Measuring engagement level in child-robot interaction using machine learning based data analysis. In *2020 international conference on data analytics for business and industry: way towards a sustainable economy (ICDABI)*. IEEE, 1–5.

[17] Lynda Tait, Max Birchwood, and Peter Trower. 2002. A new scale (SES) to measure engagement with community mental health services. *Journal of mental health* 11, 2 (2002), 191–198.

[18] Mustafa Uğur Uçar and Ersin Özdemir. 2022. Recognizing Students and Detecting Student Engagement with Real-Time Image Processing. *Electronics* 11, 9 (2022), 1500.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).