Multimodal Prediction of User's Performance in High-Stress Dialogue Interactions

Setareh Nasihati Gilani
sngilani@ict.usc.edu
USC ICT

Kimberly Pollard
kimberly.a.pollard.civ@army.mil
Army Research Laboratory

David Traum
traum@ict.usc.edu
USC ICT

ABSTRACT

High-Stress interactions include cases in which decisions must be made, communicated, and agreed upon in a short amount of time to avoid dire consequences. Such interactions can be a source of different multimodal signals indicating participant cognitive and emotional states, which can vary with factors such as the difficulty of the interaction. By utilizing behavioral cues, a multimodal deep neural network (with audio, video, and text modalities) was developed to predict the performance of users in these interactions. An ablation study was conducted to compare impact of different modalities. Our best model can predict the user performance with 73% accuracy in a 3-class classification task.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *Empirical studies in HCI*; Laboratory experiments; • **Computing methodologies** → *Neural networks.*

KEYWORDS

Human-Agent Interaction; Affective Computing; emotional expressions; behavior analysis; multimodal neural networks

ACM Reference Format:
Setareh Nasihati Gilani, Kimberly Pollard, and David Traum. 2023. Multimodal Prediction of User's Performance in High-Stress Dialogue Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610661.3617166

1 INTRODUCTION

The personality of interlocutors can play a profound role in shaping the structure and dynamics of the conversation. Numerous studies have delved into the intricate relationship between personality and character traits in dialogue systems, and how tailoring the behavior of one interlocutor based on the personality profile of the conversant can result in superior outcomes [21, 27, 30, 38].

However, the majority of interactions examined in prior research have been conducted in low-stress contexts, such as during a casual conversation or while engaging in specific tasks without the imposition of time constraints. To build upon this existing knowledge and explore the impact of these emotional connections in more demanding scenarios, this paper investigates the dynamics of human-virtual character interactions within a high-stress environment. Specifically, the study focuses on the ability of participants to perform a critical task, namely rescuing residents from a fire-rescue scenario, while simultaneously developing and maintaining rapport with a virtual character.

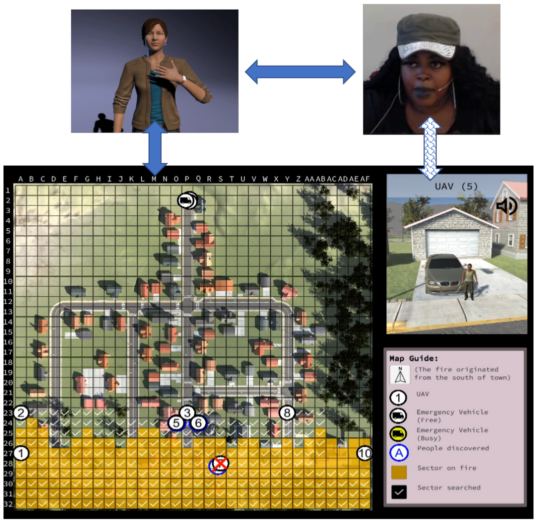

Figure 1: Overview of the simulation environment, the spokesperson, and the operator (aka the user)

Our objective is to develop a predictive model for the outcome (aka user's performance) in time-sensitive scenarios, utilizing multimodal signals (audio, video, and text) acquired during interaction windows. We analyzed data from Chaffey et al. [4], in which human participants in a disaster relief scenario act as *Operator* of a robot swarm and seek out and engage with diverse synthetic individuals (*residents*), to convince them to evacuate, in an attempt to save them from an approaching fire (Figure 1). This model will serve as a valuable tool for evaluating a user's ability to act efficiently and influence others under stressful conditions and time constraints. The results will provide insights into the complex interplay between multimodal signals and performance in time-sensitive scenarios.

2 RELATED WORK

Multimodal behavioral analysis has been used in multiple domains. In the cognitive science domain, research has shown that sequence learning and multimodal behavioral analysis can be used as a tool to assess human behavior towards learning ability, temporary memory and attention [1, 15, 25]. For an individual to perform well in a cognitive task, paying attention and being engaged in the task

Publication rights licensed to ACM. ACM acknowledges that this contribution was authored or co-authored by an employee, contractor or affiliate of the United States government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.
ICMI '23 Companion, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3617166

Table 1: An interaction example drawn from the data: Interaction between a stubborn person with the operator.

Speaker	Dialogue Utterance
Resident	Hello?
Operator	Hey, what's going on? Are you okay? I need you to evacuate right now. Immediately. You're in danger.
Resident	I'm not leaving my home. I have too much work here to just leave it all behind.
Operator	No, no, no. You're going to leave right now. There's a fire. Do you understand? You need to leave right now.
Resident	I've spent years on my collection, and there isn't time to take it all.
Operator	I understand that. Listen! You have to go, sir. You have to. Please listen. Your life is in danger.
Resident	You really think it's that bad?
Operator	It's very bad. It's spreading really fast.
Resident	Okay, I'm not stupid, just let me grab my bag and I'll head out.
Operator	Okay, thank you, sir

are crucial. Previous research has investigated the relationship between attention, task engagement, and human emotions expressed through bodily and facial expressions [3, 11, 36]. In [1], the authors propose a multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signals.

Many applications include tasks of sentiment analysis and sentiment prediction of a user. These include studies that have combined visual and audio features [7, 35], speech content [22, 32], and even physiological signals to recognize emotions [26]. In [16], the authors have focused on "Comfortability", an "internal state that focuses on the person's desire to maintain or withdraw from an interaction". They have proposed several multimodal classifiers (with various facial and upper-body movements as input) to recognize the comfortability of humans in a Human-Robot interaction. In [32], a multimodal deep neural classifier was proposed to predict the best times for an agent's empathetic response in a human-agent interaction. They argue that emotional tone in language in addition to facial expressions are strong indicators of dramatic sentiment in conversation that warrant an empathetic response, and therefore they are using visual, audio and language modalities in their prediction model.

3 DATA

The data used is from an experiment first designed and introduced by Chaffey et al. [5]. The study included 31 participants recruited through Craigslist (age range 22-49 with an average age of 29, 19 male and 12 female, from a range of ethnicities.) We briefly describe the overall scenario, the recorded data and the performance metric.

Scenario. The simulation presents a dynamic scenario where human participants (playing the role of *operator*) are tasked with rescuing *residents* from a small town threatened by an imminent wildfire. During the simulation, the *operator* is under severe time constraints to evacuate all residents before the fire engulfs their location. To evacuate each *resident*, the *operator* must first locate them and then convince them to either follow a drone to safety or use an evacuation vehicle for those in need physically. The *operator* controls a swarm of drones that can search for the 5 residents who are located randomly in the simulation map. To facilitate controlling the swarm and reduce the cognitive load associated with managing them, the *operator* also has access to a virtual assistant (*spokesperson*). The *spokesperson* can translate the high-level instructions from the *operator* (in natural language) into step-by-step commands to the drones and the evacuation vehicles. Finding the right tasks to delegate to the spokesperson is a crucial part of a successful evacuation, as an operator is unlikely to have the time and cognitive resources needed to do everything themselves. The simulation environment provides a real-time map of the town that shows the location of the *drones* and the evacuation vehicle, the areas that have already been searched, the fire's location, and the whereabouts of any rescued residents. Figure 1 illustrates an overview of the simulation environment, the *spokesperson*, and the *operator*.

Resident Interactions. The recorded information from the interactions contains a complete log of simulation events, residents' dialogues, the Operator's performance in terms of the number of rescued residents, their frontal video of the operator, and the screen recording of the simulation environment. We extracted sections of the data where the operator is conversing with one of the residents directly. Table 1 shows an example of a resident interacting with the operator. There were 104 such resident interactions (avr length 34.68 seconds) from all the subjects. The convincing rate among these interactions was 85%. We divide the interaction period into several segments based on the speaker tag. 2141 segments were extracted from the interactions. Segments are either utterances or silent periods. The two main speakers in the interaction windows are the operator (34% of segments) and the resident (27% of segments), but occasionally we also have the spokesperson jump into the conversation (13% of segments). The rest (26% of segments) are silent segments.

Performance Metric. We use **interaction length** as a performance metric since our experiment is designed to assess the operator's ability to efficiently and quickly convince the residents. This metric is defined for each interaction, measuring the operator's performance in that specific interaction. The length of the successful interactions (resident convinced) represented a double Gaussian distribution, separating the data into two major parts, therefore, we chose to have 2 classes of successful interactions and one class for unsuccessful interactions (resident not convinced) which represented close to a normal distribution; therefore categorizing our data into 3 classes: "Successful-short" and "successful-long" each makeup for 42.5%, and "unsuccessful" class for 15% of the data.

4 METHOD

4.1 Multimodal Feature Extraction

4.1.1 Visual Features. We use the OpenFace [2] toolkit to extract raw features per frame from the operator's video. The extracted 32-dimensional feature vectors including the estimated eye gaze direction vector in 3D, head pose, and 17 Facial Action Units (AUs) intensity [9] indicating the facial muscle movements. These visual descriptors have been shown to be strong indicators of human emotions and sentiments [31]. Therefore, we also extracted values for six universal emotions [8] {anger, happiness, sadness, fear, disgust,

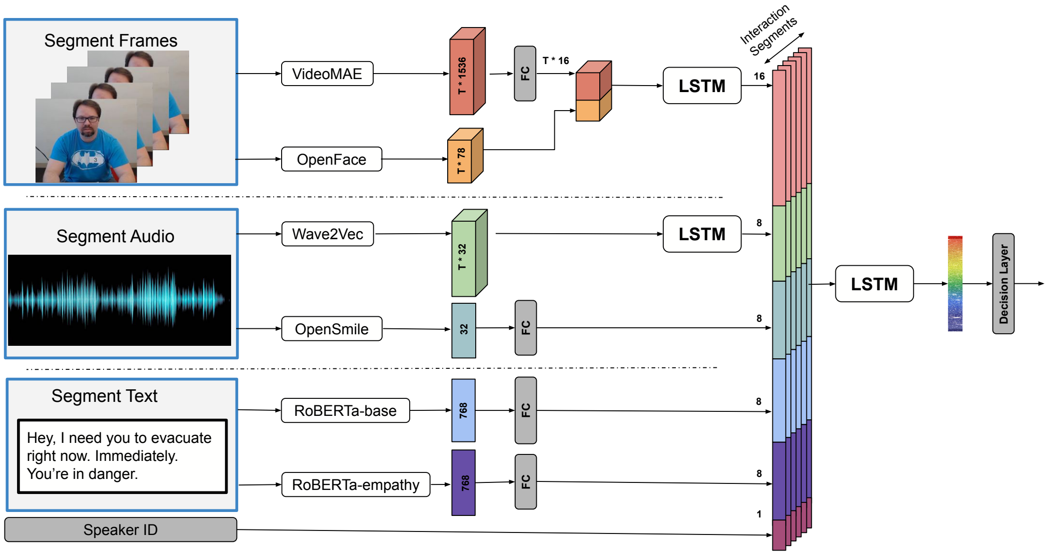

Figure 2: Overall Architecture of Our Multi-Modal Information Infusion.

surprise} based on [23]. We also used VideoMAE, a model that has been recently proposed by Tong et al. [33] as a data-efficient learner for self-supervised video pre-training (SSVP). It is inspired by the recent success of ImageMAE[14] and extended the masked autoencoders (MAE) to video, achieving state-of-the-art performance on several video classification benchmarks. We use the pre-trained VideoMAE model from huggingface to extract dense representation of 1536-dimensional vectors for each frame of the operator's videos.

4.1.2 Linguistic features. Audio recordings of the interactions were transcribed using the Whisper model [24], and manually verified. Two linguistic models, RoBERTa [19] and its fine-tuned version RoBERTa-empathy [12], processed the text from segments, with silent segments using an empty string as input. RoBERTa is a state-of-the-art English language model based on transformers, while RoBERTa-empathy is fine-tuned checkpoint of the *DistilRoBERTA-base* model [28] trained on empathy and distress datasets, specialized for emotional content, rather than semantic information. The models' pre-trained weights are from huggingface [37], each producing 768-dimensional contextualized embeddings per word. We use mean-pooling to generate sentence-level representations.

4.1.3 Auditory Features. We extracted auditory features using two state-of-the-art models. wave2vec [29] model explores unsupervised pre-training for speech recognition, using a multi-layer convolutional neural network that is optimized via a noise contrastive binary classification task. We use the pre-trained model weights from huggingface to obtain the frame-by-frame auditory embeddings. OpenSMILE [10] is an open-source software toolkit that enables the extraction of features from audio signals. It is commonly used in the classification of speech and music signals. OpenSMILE can recognize the characteristics of a given speech or music segment, such as a speaker's emotion, age, gender, personality, depression, intoxication, or vocal pathological disorders. We use OpenSMILE to generate an auditory representation of the conversation between *operator* and *residents* with a focus on affective-driven features.

4.2 Proposed Architecture

We developed a novel multimodal deep neural network architecture (depicted in Figure 2) that integrates linguistic, auditory, and visual features. The model utilizes pre-trained models fine-tuned on our dataset for optimal performance. Inspired by conventional architectures like Deng et al. [7], our model processes each modality as distinct channels, then integrates all the information into a unified space before forwarding it to a classification layer. It incorporates temporal information at two levels: segment and interaction, achieved through two layers of Long Short-Term Memory (LSTM) networks applied to visual and auditory modalities. Linguistic modality doesn't require this temporal modeling due to its less prominent temporal nature in this context.

Visual frames from each individual segment undergo processing by both the VideoMAE and OpenFace models, resulting in feature vectors of dimensions 1536 and 32 respectively. To compress the dimensionality, a fully connected layer transforms VideoMAE's output into a 16-dimensional representation. The visual frame embeddings from both sources are then concatenated and fed into an LSTM layer with a hidden layer size of 16, enabling the derivation of visual segment-level embeddings. Sequences are then streamlined by padding or truncating them to consist of 30 segments per video, with each segment containing 64 frames (equal to 32 seconds).

In the auditory modality, we adopt a two-step approach to encode information into dense representations. Pre-trained models are used to obtain embeddings, and for silent segments, an empty voice is fed into the models. For Wave2vec input, audio files were resampled to 16K FPS, with each frame resulting in a 32-dimensional vector. Sequence preparation involves padding or truncation to ensure 30 segments per video, each lasting 32 seconds. An LSTM layer with a hidden size of 8 is used to generate a dense segment representation, similar to the video component. On the other hand, OpenSmile generates a single 6373-dimensional vector per input file, eliminating the need for pooling for segment-level representations. A fully connected layer compresses the information to an 8-dimensional representation.

To process the linguistic features, we also use two pre-trained models to encode the information into dense representations. First each transcribed sentence from each segment is encoded using RoBERTa-base and RoBERTa-empathy models. For silent segments, an empty string was given to the models as input. Then we use mean pooling (instead of an LSTM) to amalgamate sentence-level information into segment-level representations. We use mean pooling for two reasons: first, it reduces the overall count of trainable parameters and second, it is a widely adopted technique in NLP[17, 34]. For each set of encodings, we use a fully connected layer to compress and reduce the dimensionality of the information into an 8-dimensional representation.

After obtaining the multi-modal segment level representations, in the last part of the model, we first concatenate information from all modalities into one vector and add the speaker ID to respective segments. Then we feed the resulting vector into an LSTM layer to pool the segment-level information into a 16-dimensional representation of the whole interaction. Finally, we feed this interaction representation into a fully connected decision layer to predict the user's performance in the interaction.

5 RESULTS

We implemented the model explained in section 4.2 using the Keras library[6]. To train our model, we utilized the interaction length performance metric as our data label. As a result, the problem was transformed into a 3-class classification task.

"Sparse categorical cross-entropy" was chosen as the loss function, while the Adam optimizer was employed to update the weights during training iteratively. The model underwent 100 epochs of training, and to prevent overfitting, an early stopping mechanism was incorporated with a patience of 10 interactions. Early stopping was triggered if the validation loss failed to improve beyond the threshold of 1e-4. To assess the model's performance, 4-fold cross-validation with random shuffling with a "per-participant" basis was performed, so no participants' interaction data were included in both train and test sets for the same experiment.

Within each fold, the validation accuracy and validation f1-macro score were calculated, and the average values were reported. We also conduct an ablation study to evaluate the effectiveness of each of our modalities in our proposed prediction model where we separately train the model using only one of the embedding sources. We also tested two variations of our model: 1) affective: which only uses the affective embedding sources (OpenFase, RoBERTa-empathy, and OpenSMILE), and 2) generals: which only uses general embedding sources (VideoMAE, RoBERTa, wave2vec).

Table 2 summarizes our proposed model's accuracy and f1 score results and its different variations. Our best model achieved an accuracy of 73.08% and f1-macro score of 63.96% on the classification task. Our findings demonstrate that the linguistic modality contributes the most to the model's performance, surpassing the auditory and visual modalities. This observation is consistent with the results reported in other studies[13, 18, 20, 32]. Furthermore, our results indicate that the combination of general models outperforms the combination of more specialized models trained on affective datasets. This finding suggests that, despite not being specifically fine-tuned for this particular task, the large-scale training data of general models provides them with enough implicit knowledge to perform adequately on a behavior analysis-based task. These results offer possibilities for future research into the optimal combination of modalities for behavior analysis and suggest that general models may provide an effective starting point for this investigation.

Table 2: Mean cross-validation accuracy and F1-Macro metrics for different modalities on the classification task. Models with the (†) sign are affective models and others are general.

Modality	Model	Accuracy (%)	F1-Macro (%)
Visual	VideoMAE	45.19	41.17
	OpenFace†	39.48	27.32
Linguistic	RoBERTa	56.73	53.86
	RoBERTa-empathy†	50.00	41.37
Auditory	wave2vec	40.38	29.92
	OpenSMILE†	33.65	16.77
Multimodal	generals	57.69	46.46
	affectives†	55.56	50.58
	Full	**73.08**	**63.96**
Random		33.65	16.78

6 CONCLUSION & FUTURE DIRECTIONS

We explore the user's interaction with a simulation environment in a high-stress context wherein users (operators) must rescue virtual characters (residents) from a simulated wildfire that is rapidly approaching a simulated town. Operators must engage in dialogue with the residents to convince them to flee and to arrange their escape methods. We developed a deep neural network model to predict the user's performance using the multimodal information extracted from the recordings. We categorize the performance as one of the three classes of "successful-short", "successful-long" and "unsuccessful". The model processes the interaction in two levels each with an LSTM layer. The first level encodes the temporal dimensions of each segment within the interaction and the second level incorporates the interaction dynamics across multiple segments. Our best model achieved a 73.08% accuracy on our dataset.

Future directions involve exploring the relationship between different personality types of characters and user's performance/user's emotional expressions, exploring the effects of user's personality on task performance, and looking at different persuasion strategies used by the operator. Other directions would be exploring different architectures such as incorporating late-fusion methods to compare them with our early-fusion method.

7 ACKNOWLEDGMENT

The work depicted here was sponsored by the U.S. Army Research Laboratory (ARL). Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] Ashwin Ramesh Babu, Akilesh Rajavenkatanarayanan, James Robert Brady, and Fillia Makedon. 2018. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signal. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. 1–7.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[3] Karina S Blair, Bruce W Smith, Derek GV Mitchell, John Morton, Meena Vythilingam, Luiz Pessoa, Daniel Fridberg, Alan Zametkin, Eric E Nelson, Wayne C Drevets, et al. 2007. Modulation of emotion by cognition and cognition by emotion. *Neuroimage* 35, 1 (2007), 430–440.

[4] Patricia Chaffey, Ron Artstein, Kallirroi Georgila, Kimberly A Pollard, Setareh Nasihati Gilani, David M Krum, David Nelson, Kevin Huynh, Alesia Gainer, Seyed Hossein Alavi, et al. 2020. Human swarm interaction using plays, audibles, and a virtual spokesperson. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*, Vol. 11413. SPIE, 272–285.

[5] Patricia Chaffey, Ron Artstein, Kallirroi Georgila, Kimberly A Pollard, Setareh Nasihati Gilani, David M Krum, David Nelson, Kevin Huynh, Alesia Gainer, Seyed Hossein Alavi, Rhys Yahata, and David Traum. 2019. Developing a virtual reality wildfire simulation to analyze human communication and interaction with a robotic swarm during emergencies. In *Workshop on Human Language Technologies in Crisis and Emergency Management*.

[6] François Chollet et al. 2015. Keras. https://keras.io.

[7] Didan Deng, Yuqian Zhou, Jimin Pi, and Bertram E Shi. 2018. Multimodal utterance-level affect analysis using visual, audio and text features. *arXiv preprint arXiv:1805.00625* (2018).

[8] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[9] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[11] Kristin Fraser, Irene Ma, Elise Teteris, Heather Baxter, Bruce Wright, and Kevin McLaughlin. 2012. Emotion, cognitive load and learning outcomes during simulation training. *Medical education* 46, 11 (2012), 1055–1062.

[12] Jochen Hartmann. 2022. Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[15] Larry E Humes and Shari S Floyd. 2005. Measures of working memory, sequence learning, and speech recognition in the elderly. (2005).

[16] Maria Elena Lechuga Redondo, Radoslaw Niewiadomski, Rea Francesco, and Alessandra Sciutti. 2022. Comfortability Recognition from Visual Non-verbal Cues. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 207–216.

[17] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045* (2017).

[18] Xia Li and Minping Chen. 2020. Multimodal sentiment analysis with multi-perspective fusion network focusing on sense attentive language. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*. Springer, 359–373.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[20] Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *arXiv preprint arXiv:2112.01368* (2021).

[21] Eric J Moody, Daniel N McIntosh, Laura J Mann, and Kimberly R Weisser. 2007. More than mere mimicry? The influence of emotion on rapid facial reactions to faces. *Emotion* 7, 2 (2007), 447.

[22] Yukiko I Nakano, Eri Hirose, Tatsuya Sakato, Shogo Okada, and Jean-Claude Martin. 2022. Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 5–14.

[23] Maja Pantic and Leon J. M. Rothkrantz. 2000. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1424–1445.

[24] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[25] Akilesh Rajavenkatanarayanan, Ashwin Ramesh Babu, Konstantinos Tsiakas, and Fillia Makedon. 2018. Monitoring task engagement using facial expressions and body postures. In *Proceedings of the 3rd International Workshop on Interactive and Spatial Computing*. 103–108.

[26] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–9.

[27] Krystyna Rymarczyk, Cezary Biele, Anna Grabowska, and Henryk Majczynski. 2011. EMG activity in response to static and dynamic facial expressions. *International Journal of Psychophysiology* 79, 2 (2011), 330–333.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[29] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862* (2019).

[30] Beate Seibt, Andreas Mühlberger, Katja U Likowski, and Peter Weyers. 2015. Facial mimicry in its social setting. *Frontiers in psychology* 6 (2015), 1122.

[31] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2011. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing* 3, 2 (2011), 211–223.

[32] Leili Tavabi, Kalin Stefanov, Setareh Nasihati Gilani, David Traum, and Mohammad Soleymani. 2019. Multimodal learning for identifying opportunities for empathetic responses. In *2019 International Conference on Multimodal Interaction*. 95–104.

[33] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602* (2022).

[34] Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A Cross-Task Analysis of Text Span Representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Online, 166–176. https://doi.org/10.18653/v1/2020.repl4nlp-1.20

[35] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing* 11, 8 (2017), 1301–1309.

[36] Lotte F Van Dillen, Dirk J Heslenfeld, and Sander L Koole. 2009. Tuning down the emotional brain: an fMRI study of the effects of cognitive load on the processing of affective images. *Neuroimage* 45, 4 (2009), 1212–1219.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP 2020: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[38] Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2020. Improving dialog systems for negotiation with personality modeling. *arXiv preprint arXiv:2010.09954* (2020).