Emotion Expression Estimates to Measure and Improve Multimodal Social-Affective Interactions

Jeffrey A. Brooks
Hume AI
United States

Vineet Tiruvadi
Hume AI
United States
vineet@hume.ai

Alice Baird
Hume AI
United States

Panagiotis Tzirakis
Hume AI
United States

Haoqi Li
Hume AI
United States

Chris Gagne
Hume AI
United States

Moses Oh
Hume AI
United States

Alan Cowen
Hume AI
United States

ABSTRACT

Large language models (LLMs) are being adopted in a wide range of applications, but an understanding of other social-affective signals is needed to support effective human-computer-interaction (HCI) in multimodal interfaces. In particular, robust, accurate measurements of human emotional expression can be used to tailor responses to human values and preferences. In this paper, we present two models available from an API-based suite of emotional expression models that measure nuanced facial and vocal signals, providing rich, high-dimensional emotional expression estimates (EEEs). We demonstrate the ability of EEEs to provide insight into two established datasets and present methods for integrating EEEs into large language model (LLM) applications. We discuss how this approach is a step towards more reliable tools for clinical screening and scientific study, as well as empathic digital assistants that can be used in therapeutic settings.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Computer vision*; *Natural language processing*.

KEYWORDS

Multimodal Sentiment Analysis, Emotion Recognition, Affective Computing, Emotion Science, Mental Health

ACM Reference Format:
Jeffrey A. Brooks, Vineet Tiruvadi, Alice Baird, Panagiotis Tzirakis, Haoqi Li, Chris Gagne, Moses Oh, and Alan Cowen. 2023. Emotion Expression Estimates to Measure and Improve Multimodal Social-Affective Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3610661.3616129

ICMI '23 Companion, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3616129

1 INTRODUCTION

Human communication is rich with non-verbal expressions: we smile, sigh, laugh, and nod; frown, shout, slump, and cast sidelong glances. Expressions are thought to convey distinct emotion-related content that forms a critical components of human communication [18]. Rigorous testing of hypotheses following from this assumption requires reliable, validated measurements of multiple modalities of expression. While artificial intelligence (AI) models for understanding language have advanced, standard approaches overlook cues of social-affective context embedded in the face, body, and voice necessary to fully understand human communicative behavior. Further, accurate assessment of emotional and mood-related signals is particularly crucial for healthcare applications, in neuropsychiatric domains [17, 22, 28] and beyond [14, 26].

Computational approaches to the measurement of emotional and expressive states (driven by the field of affective computing [18]) have been constrained by theories and methods from emotion science, which has historically relied on small and homogenous samples and time-consuming approaches to behavioral measurement (e.g., manual coding of facial action units, or AUs; [10]). Machine learning (ML) advances [6] are paving the way for models that better understand and respond to human expressions [13]. Such developments leverage data-driven findings from large-scale audiovisual datasets, replacing the low-dimensional, theory-driven methods of emotion science with high-dimensional data-driven taxonomies of emotional expression.

These multidimensional insights have been translated into ML models able to measure expressive behaviors across up-to 40 dimensions of emotional expression, defined here as emotional expression estimates (EEEs) [6, 13]. EEEs have in recent years begun to be adopted by the machine-learning driven affective computing [2], and multimodal sentiment [8] communities, being applied to a variety of behavioral cues including vocal bursts [3, 25]. ML-based approaches may offer valuable insights into disease states, general health, and social-affective functioning, and provide reliable measures for clinical monitoring and interactive rapport [1, 9, 11, 12, 15, 21, 24].

In this study, we demonstrate the efficacy of analyzing and integrating EEEs into large-scale data analysis and with language-based interactive technology. We characterize the distribution of expressed emotions in two open datasets, RAVDESS [16] and CANDOR [19], and propose a method to integrate expression estimates

into LLM-based applications, which we introduce as Language-based Emotional Expression Description (LEED).

2 METHODS

2.1 Datasets

CANDOR To demonstrate the ability of EEEs to measure and understand complex emotional behaviors in real-world contexts, we used the recently released *CANDOR* dataset ("Conversation: A Naturalistic Dataset of Online Recordings") [19]. *CANDOR* includes 1,656 recorded conversations conducted over video chat in English, averaging approximately 30 minutes in length. For *CANDOR*, we extracted EEEs for facial expression and prosody at the level of each spoken turn.

RAVDESS We also made use of a subset of the Ryerson Audio-Visual Database of Emotional Speech and Song (*RAVDESS*) [16]. We focused on the emotional speech component of RAVDESS, which includes 1,400 utterances from 24 professional actors (12 female, 12 male). Each utterance is composed of two sentences from actors recorded with different classes of emotion (calmness, happiness, sadness, anger, fear, surprise, disgust, and neutral), spoken with two intensities (normal and strong). Thus, compared to *CANDOR*, *RAVDESS* is a comparatively more balanced and controlled dataset, at the cost of naturalness. For *RAVDESS*, we extracted EEEs for facial and prosodic speech expression for each utterance spoken.

2.2 EEE Models and Featurization

To extract EEEs, we utilized a suite of unimodal deep learning-based models (accessed via the Hume AI API[1]) to measure the emotion expressed in two key modalities of expressive behavior: Facial Expression, and Prosodic Speech Expression.

The facial expression model is a fine-tuned adaptation of the FaceNet Inception Resnet v1 [20] pre-trained on the VGGFace2 dataset [7]. The model is fine-tuned on the pixel intensities from hundreds of thousands of cropped faces of the subjects mimicking an array of expressive behaviors. We used the Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [29] model to detect the face in the images.

The prosody expression model is a fine-tuned adaption of OpenAI's Whisper-Small model [2]. The model was pre-trained on 680 k hours of audio in a multi-task setting. We fine-tuned the model on thousands of hours of audio data from individuals mimicking an array of expressions with their voice. Timestamps extracted from the transcription of each audio file (utilising the Deepgram[3] - AI speech transcription models) was used to trim the audio files based on spoken language. The outputs of both models was 48 emotional expression dimensions.

2.3 Language-based Emotional Expression Description (LEED)

Interactive Agents such at ChatGPT function by the user prompting the model with natural language. In order to embed estimates of expression in real-time from the user interacting with the agent we proposed a method for Language-based Emotional Expression Description (LEED). To perform LEED, we took the rating of each EEE and converted it to an adverb in a specified range We then concatenated each summary with a prefix of the form "They sounded ..." (for prosodic speech expression) or "They looked .." (for facial expression) and then lastly prefix the transcribed text spoken, e.g.:

- Facial: {"Joy": 0.84} –> "They look extremely joyful."
- Prosodic: {"Sad": 0.34} –>"They said "" and sound quite sad."

For the interested reader we provide a GitHub repository with further examples on how this can be performed [4].

2.3.1 LEEDs Proof-of-Concept. To understand what joint sequences of language and expressions reveal, given their potential to predict real-world outcomes better than language alone, we developed a transfer learning pipeline. To do so, we utilised the RAVDESS datasets. First, we calculated the LEED of each modalities EEEs (+ transcription for prosodic speech expression). For facial expression, the LEED is taken either from the mean (μ) over the 3 frame per second representation, or extracted at each time-step. We then fuse the modalities with two fusion strategies, for this purpose what we call 'early-fusion' is concatenation of the embeddings from each modalities LEED, and 'late-fusion' refers to concatenation of each modality's LEEDs before extracting embeddings.

We then continue by feeding the concatenated language-based summary to OpenAI Ada V2 model ("text-embedding-ada-002"), and extracting embeddings which serve as the 1 535 dimensional feature input to the classifier. The Ada V2 model belongs to the suite of GPT (Generative Pre-trained Transformer) models and is particularly advantageous for capturing contextual information.

For the subsequent classification, we split the data speaker-independently into a training set (20 speakers) and test set (4 speakers). We employed a support vector machine (SVM) optimising the complexity parameters ($C \in 10^{-4}, 10^{-3}, ..., 1$), reporting the result for the best value of C on the test set. To evaluate the performance of the model we utilised unweighted average recall (UAR).

3 RESULTS AND DISCUSSION

We first demonstrated the utility of using EEEs to characterize large-scale datasets using the estimates extracted for *CANDOR* and *RAVDESS*. Then, we proposed a method to summarise EEEs for LLM-based interactive agents before validating this via transfer learning experiments.

3.1 Distribution and representation of expressed emotions

3.1.1 EEEs reflect ground-truth ratings. We first focused on *RAVDESS* given that it reports "ground truth" emotion ratings to which we compared the estimates of our models.

Figure 2 shows a t-distributed stochastic neighbor embedding (*t-SNE*) representation of the EEEs from the face and speech prosody model embeddings, labelled by the 8-class ground-truth for emotion given by the dataset authors [16]. This method projects high-dimensional data onto two nonlinear axes, such that the local distances between data points are accurately preserved while more distinct data points are separated by longer, more approximate,

[1] https://beta.hume.ai
[2] https://openai.com/blog/introducing-chatgpt-and-whisper-apis
[3] https://developers.deepgram.com/docs/model
[4] https://github.com/HumeAI/expressive-prompt-engineering

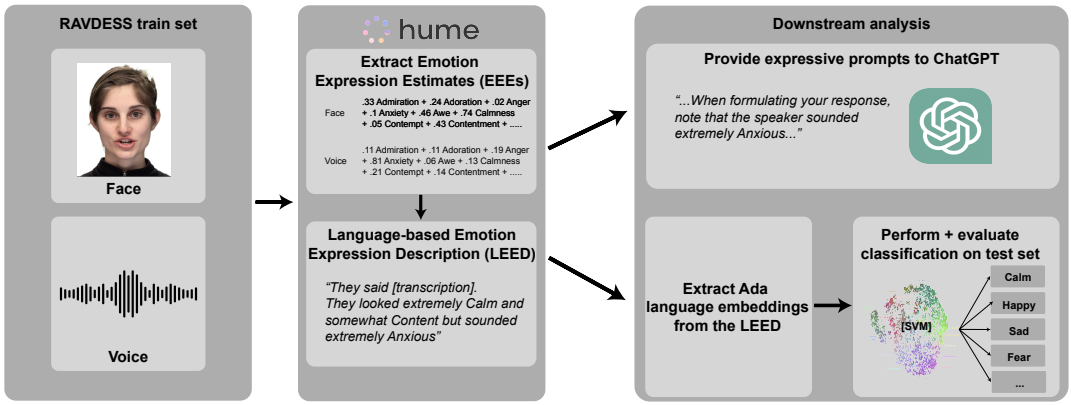

Figure 1: Schematic of our experimental and analytic approach.

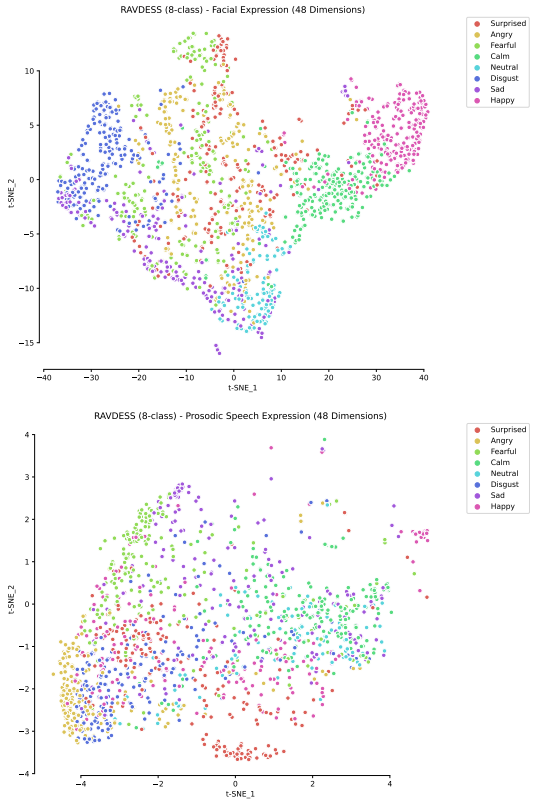

Figure 2: t-SNE representation of the EEEs in the RAVDESS dataset for Facial (above) and Prosodic Expression (below).

distances. It can be seen that there is distinct clustering of the EEEs based on the ground-truth labels. Emotions with expected similar arousal levels show a slight overlap (e.g. 'calmness' and'neutral'), and emotions expressed through similar facial landmarks also have overlapping clusters. For example, 'surprised' and 'fearful' share similarities in facial expression, (e.g, raised eyebrows, widened eyes) which is reflected by the outputs of our models.

EEEs are also sensitive to nuances in expressive state that can provide insights differing from traditional assumptions about the distribution and conceptualization of emotion. Despite the overall clustering by ground-truth emotion labels, visual inspection of the t-SNE plots reveals that emotion classes do not strictly organize into discrete clusters, and instead show continuity: individual points show smooth gradients of meaning between emotion classes, as well as a lack of homogeneity within classes. That is, individual expressive states can lie somewhere between, for example, "calmness" and "happiness", and these blended states are readily measured by our high-dimensional output space.

3.1.2 Distribution of EEEs. EEEs can also provide insight into the convergence and divergence of different modalities and datasets within the emotion space. Figure 3 depicts the distribution of expression intensities (measured using EEEs) for both facial expression and prosodic speech expression. We can see that facial expression and prosody are not strictly co-dependent, and further analysis of these divergence in estimates between modalities could provide insights relevant to understanding socio-affective context in and out of healthcare applications.

3.1.3 Correlation of expression across modalities. Next, we analysed the the EEEs from both datasets using two modalities: facial expression and prosodic speech expression. Correlations between the dimensions of the EEEs demonstrate covariance structures that differ between the two datasets (Figure S4 top vs bottom), indicating that the outputs of our EEE models cannot be reduced to a small number of classes that show universal covariance structure (as has been traditionally assumed). Additionally, the distinct modalities of face and prosody exhibit distinct covariance between the EEE dimensions (Figure S4 left vs right). Ultimately, we found that the face and voice outputs were partially, but significantly, correlated within both *CANDOR* (Spearman's $rho = 0.3109$, $p < .001$) and *RAVDESS* (Spearman's $rho = 0.3515$, $p < .001$). With further validation, these multdimensional analyses of EEEs may yield signals that better discern healthy emotional expression from those found in psychiatric illnesses such as depression, and schizophrenia, needs that are discussed in [4, 5, 23, 27].

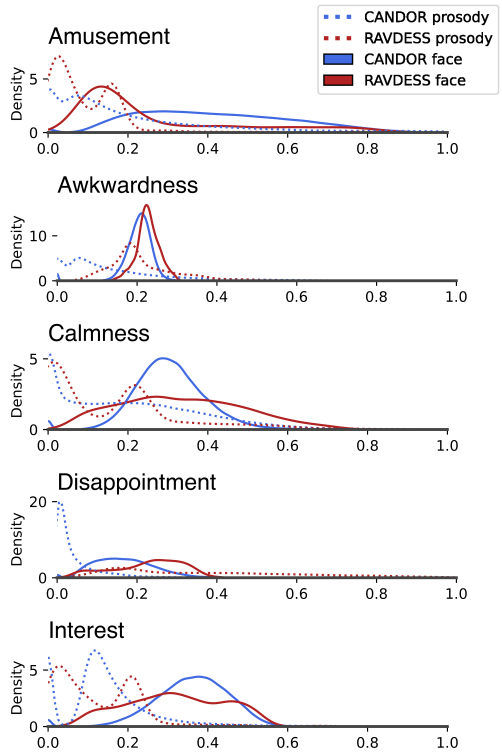

Figure 3: Distribution of expression intensity across modalities and datasets. Distributions are shown for the top 5 emotions (assessed by mean rank ordering of predictions) across the CANDOR and RAVDESS datasets in each modality (facial expression, prosodic speech expression).

3.2 Integrating EEE into interactive technology

Language-based summary of expression can then be utilised two-fold: (1) to provide expressive information to a LLM based chatbot, (2) to enhance LLM embeddings for a more meaningful understanding of affective targets. Towards this end, we demonstrated the utility of integrating EEE into language models, particularly LLMs as a proof-of-concept for integration with interactive agents (Table 1). Our approach, which we introduce as LEEDs, has at its core a description of the expression estimates in natural language.

From these results we see that the facial expression EEEs have a strong above chance (chance level 0.125 UAR) baseline performance of 0.69 UAR. At first taking the mean over the time-domain and performing LEED we find does not lead to any improvements, it was from this that we found more meaningful was to harness the sequential nature of the facial expression model, and perform lead at each time-step in the EEE, before extracting the language embedding. This method led to an improvement over the baseline EEE (up to 0.72 UAR), driven by the higher-resolution which potentially allowed for more contextual information to be incorporated.

We then explore fusing prosodic EEEs with face to explore this benefit, as can be seen in Table 1, the 'early-fusion' (described in Section 2.3.1, performs better than fusing before the language

Table 1: Transfer learning results with RAVDESS dataset integrating language-based emotional expression description (LEED), of both facial expression (face), mean (μ) and frame-wise (3fps), and prosodic speech expression (prosody) EEEs, as (Ada) language-embeddings. Reporting, feature set dimensions (Dim.) Unweighted Average Recall (UAR) for 8 classes.

Modality	Method	Dim.	UAR
Face	EEEs (Raw)	48	0.69
Face (μ)	LEED (Ada)	1536	0.66
Face (3fps)	LEED (Ada)	1536	0.72
Prosody + Face (3fps)	LEED *early* (Ada)	1536×2	**0.79**
Prosody + Face (3fps)	LEED *late* (Ada)	1536	0.78

embedding is extracted. This result would benefit from a deeper and more extended analysis, to explore the reasoning behind the fusion stratergies, however it does potentially indicate that there is benfit to the both modaitilyy that is lost through an early-fusion prior to embedding extraction.

Overall through the integration of LEEDs we see that language embeddings can target emotion spaces, even in datasets with limited language diversity such as RAVDESS. We are encouraged by these results and aim to validate them more deeply in future studies.

3.3 Limitations

This work has several limitations. First, we analysed standardized datasets without representation of disease or disorder, and further analysis directly in clinical populations is required. The datasets we used also pose limitations for generalizability: CANDOR only includes human-human interactions and RAVDESS is composed of acted expressions, which means that they could lack some of the expressive features of human-computer interactions and spontaneous naturalistic expression, respectively. Second, individual modalities are not integrated into a unified multimodal assessment, which could be an important direction for future research. Third, our approach is a proof-of-principle, requiring more rigorous assessment for healthcare inferences.

4 CONCLUSIONS

Here, we characterized emotional expression estimates (EEE) as informative measures for emotional expression analysis and demonstrated a link to interactive agent integration through LLMs. Further work is needed to improve the interpretability of EEEs, including characterizing the relationship between EEEs and the models' earlier embedding layers with low-level stimulus features.

The next phase for the study of our high-dimensional EEEs involves extending our models into healthcare-specific domains through domain-adaptation and transfer learning approaches, as well as extending alignment approaches such as reinforcement learning with human feedback (RLHF) to include EEEs. Our platform aims to facilitate the ethical adoption and integration of EEEs into socially intelligent interactive technologies such as virtual healthcare agents through cloud-based platforms for EEE models.

REFERENCES

[1] R Michael Bagby, Andrew G Ryder, Deborah R Schuller, and Margarita B Marshall. 2004. The Hamilton Depression Rating Scale: has the gold standard become a

lead weight? *American Journal of Psychiatry* 161, 12 (2004), 2163–2177.
[2] Alice Baird, Panagiotis Tzirakis, Jeffrey A Brooks, Chris B Gregory, Björn Schuller, Anton Batliner, Dacher Keltner, and Alan Cowen. 2022. The ACII 2022 Affective Vocal Bursts Workshop & Competition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–5.
[3] Alice Baird, Panagiotis Tzirakis, Jeffrey A Brooks, Lauren Kim, Michael Opara, Christopher B Gregory, Jacob Metrick, Garrett Boseck, Dacher Keltner, and Alan S Cowen. 2022. State & Trait Measurement from Nonverbal Vocalizations: A Multi-Task Joint Learning Approach. *Proc. Interspeech 2022* (2022), 2028–2032.
[4] Maryam Bijanzadeh, Ankit N Khambhati, Maansi Desai, Deanna L Wallace, Alia Shafi, Heather E Dawes, Virginia E Sturm, and Edward F Chang. 2022. Decoding naturalistic affective behaviour from spectro-spatial features in multiday human iEEG. *Nature Human Behaviour* 6, 6 (2022), 823–836.
[5] Michael L Birnbaum, Avner Abrami, Stephen Heisig, Asra Ali, Elizabeth Arenare, Carla Agurto, Nathaniel Lu, John M Kane, and Guillermo Cecchi. 2022. Acoustic and facial features from clinical interviews for machine learning–based psychiatric diagnosis: Algorithm development. *JMIR Mental Health* 9, 1 (2022), e24699.
[6] Jeffrey Brooks, Panagiotis Tzirakis, Alice Baird, Lauren Kim, Michael Opara, Xia Fang, Dacher Keltner, Maria Monroy, Rebecca Corona, Jacob Metrick, and Alan S. Cowen. 2023. Deep learning reveals what facial expressions mean to people in different cultures. *Nature Human Behaviour* 7 (2023), 240–250.
[7] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman. 2017. VGGFace2: A dataset for recognising faces across pose and age. *arXiv* (2017).
[8] Lukas Christ, Shahin Amiriparian, Alice Baird, Panagiotis Tzirakis, Alexander Kathan, Niklas Müller, Lukas Stappen, Eva-Maria Meßner, Andreas König, Alan Cowen, et al. 2022. The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 5–14.
[9] Andrea L Crowell, Steven J Garlow, Patricio Riva-Posse, and Helen S Mayberg. 2015. Characterizing the therapeutic response to deep brain stimulation for treatment-resistant depression: a single center long-term perspective. *Frontiers in integrative neuroscience* 9 (2015), 41.
[10] Paul Ekman and W.V. Friesen. 1978. Facial Action Coding System. *Consulting Psychologist Press* (1978).
[11] María F Jiménez-Herrera, Mireia Llauradó-Serra, Sagrario Acebedo-Urdiales, Leticia Bazo-Hernández, Isabel Font-Jiménez, and Christer Axelsson. 2020. Emotions and feelings in critical and emergency caring situations: A qualitative study. *BMC nursing* 19 (2020), 1–10.
[12] Alane E Kazdin and Theodore A Petti. 1982. Self-report and interview measures of childhood and adolescent depression. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 23, 4 (1982), 437–457.
[13] Dacher Keltner, Jeffrey Brooks, and Alan S. Cowen. 2022. Semantic space theory: Data-driven insights into basic emotions. *Current Directions in Psychological Science* 32, 3 (2022).
[14] Katarina Krkovic, Annika Clamor, and Tania M Lincoln. 2018. Emotion regulation as a predictor of the endocrine, autonomic, affective, and symptomatic stress response and recovery. *Psychoneuroendocrinology* 94 (2018), 112–120.
[15] G.-A. Levow and S. Duncan. 2012. Contrasting cues to verbal and non-verbal backchannels in multi-lingual dyadic rapport. *Thirteenth Annual Conference of the International Speech Communication Association* (2012).
[16] S.R. Livingstone and F.A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* (2018), e0196391.
[17] Iris B Mauss and Michael D Robinson. 2009. Measures of emotion: A review. *Cognition and emotion* 23, 2 (2009), 209–237.
[18] Rosalind W Picard. 2000. *Affective computing*. MIT press.
[19] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2022. Advancing an Interdisciplinary Science of Conversation: Insights from a Large Multimodal Corpus of Human Speech. *ArXiv* (2022).
[20] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *arXiv* (2015).
[21] Jesus Serrano-Guerrero, Mohammad Bani-Doumi, Francisco P Romero, and Jose A Olivas. 2022. Understanding what patients think about hospitals: A deep learning approach for detecting emotions in patient opinions. *Artificial Intelligence in Medicine* 128 (2022), 102298.
[22] Maryam M Shanechi. 2019. Brain–machine interfaces from motor to mood. *Nature neuroscience* 22, 10 (2019), 1554–1564.
[23] Cara Tannenbaum, Joel Lexchin, Robyn Tamblyn, and Sarah Romans. 2009. Indicators for measuring mental health: towards better surveillance. *Healthcare Policy* 5, 2 (2009), e177.
[24] Andreas Triantafyllopoulos, Alexander Kathan, Alice Baird, et al. 2023. HEAR4HEALTH: A blueprint for making computer audition a staple of modern healthcare. *arXiv* (2023).
[25] Panagiotis Tzirakis, Alice Baird, Jeffrey Brooks, Christopher Gagne, Lauren Kim, Michael Opara, Christopher Gregory, Jacob Metrick, Garrett Boseck, Vineet Tiruvadi, et al. 2023. Large-Scale Nonverbal Vocalization Detection Using Transformers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[26] Chrisanthy Vlachakis, Konstantina Dragoumani, Sofia Raftopoulou, Meropi Mantaiou, Louis Papageorgiou, Spyridon Champeris Tsaniras, Vasileios Megalooikonomou, and Dimitrios Vlachakis. 2018. Human emotions on the onset of cardiovascular and small vessel related diseases. *in vivo* 32, 4 (2018), 859–870.
[27] A.E. Whitton, M.T. Treadway, and D.A. Pizzagalli. 2015. Reward processing dysfunction in major depression, bipolar disorder, and schizophrenia. *Current Opinion in Psychiatry* 28, 1 (2015), 7–12.
[28] Alik S Widge, Donald A Malone Jr, and Darin D Dougherty. 2018. Closing the loop on deep brain stimulation for treatment-resistant depression. *Frontiers in neuroscience* 12 (2018), 175.
[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23 (2016), 1499–1503.

5 ONLINE RESOURCES

For more information, visit https://hume.ai

6 SUPPLEMENTARY

6.1 EEE Correlation Matrices

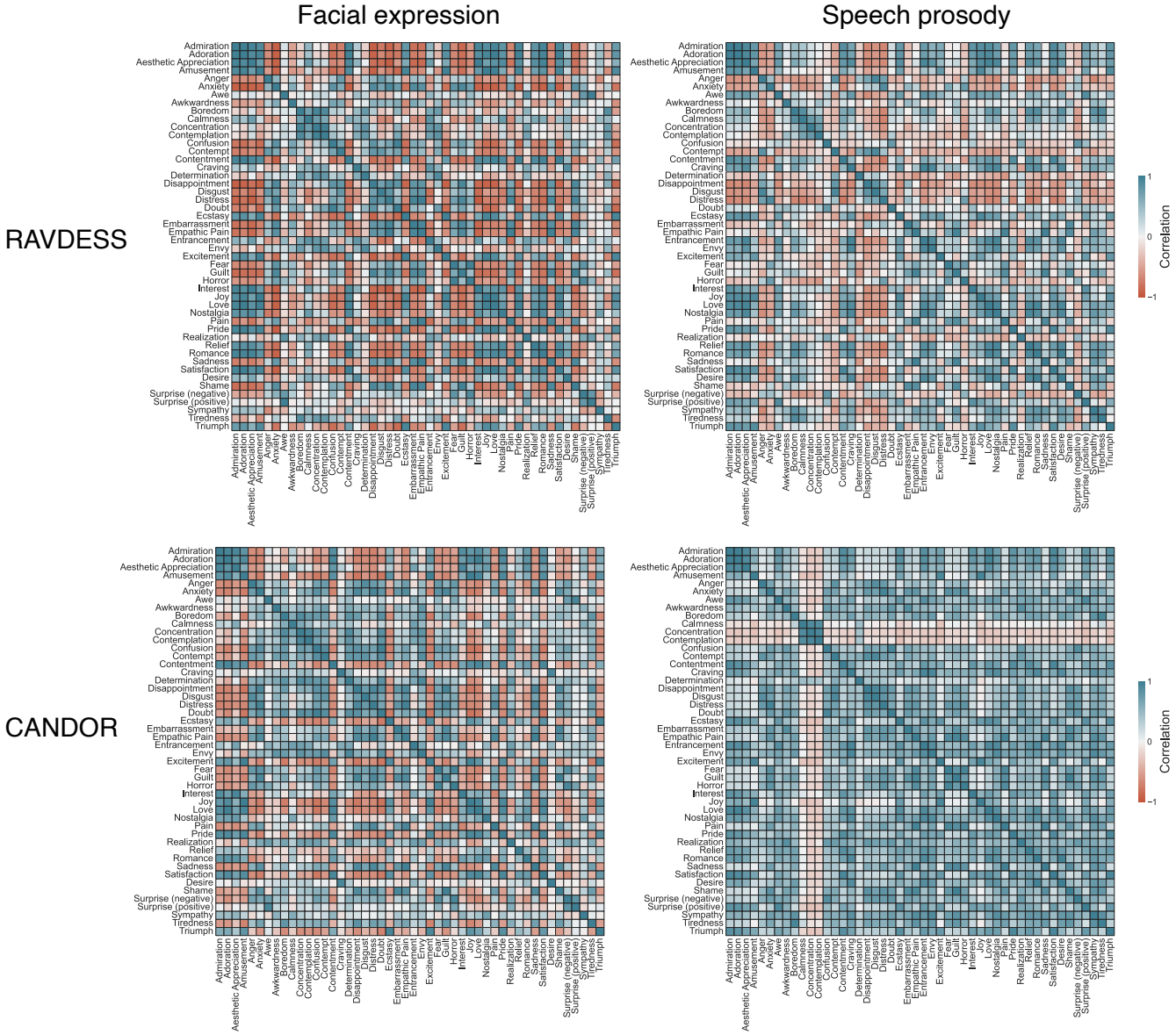

Figure 4: Correlation matrix of the emotional expression estimates. (a) Facial Expression and (b) Prosodic Speech Expression for RAVDESS. Contrast with CANDOR (c) Facial Expression and (d) Speech Prosody.