Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Identifying Vocal and Facial Biomarkers of Depression in Large-Scale Remote Recordings: A Multimodal Study Using Mixed-Effects Modeling

Nelson Hidalgo Julia[*1], *Robert Lewis*[*1], *Craig Ferguson*[1], *Simon Goldberg*[2], *Wendy Lau*[2], *Caroline Swords*[2], *Gabriela Valdivia*[2], *Christine Wilson-Mendenhall*[2], *Raquel Tartar*[3], *Rosalind Picard*[1], *Richard Davidson*[2]

[1]Media Lab, Massachusetts Institute of Technology, United States; [2]Healthy Minds Institute, University of Wisconsin-Madison, United States; [3]Healthy Minds Innovations, United States

nelsonh@mit.edu, roblewis@mit.edu

Abstract

We examine vocal and facial data from a new study with n=954 depressed participants, each characterized by six time points of the eight-item Patient Health Questionnaire survey (PHQ-8). Patients interacted with a smartphone app over four weeks, with a 3-month follow-up. The app's animated character asked participants to describe, for 90 seconds, an emotional experience from the past 24 hours. We obtained 4,875 audio-video recordings, and applied linear mixed-effects models to examine associations between depression severity and 30 acoustic, linguistic and facial action unit features. Significant associations were found with speech timing and prosody, voice quality, linguistic sentiment, the use of self-referential pronouns, and facial action units related to smiling. We also show that these features allow accurate estimation of depression severity in multimodal mixed-effects machine learning models.

Index Terms: multimodal, longitudinal depression assessment

1. Introduction

While conventional methods for mental health care are often effective for those who can access them, demand far exceeds supply, and thus there is an unmet need. Now that smartphones are ubiquitous, there has been a surge in interest in developing digital technologies to support monitoring and treatment, with the view to incorporate them into clinical workflows or direct-to-consumer products [1]. An important requirement for these systems is accurate assessment. When assessing a patient, a clinician will take into account both their description of their well-being and life experience, as well as details on how they present (e.g., tone, choice of language, and affective responses). They also consider *inter-individual* and *intra-individual* differences – i.e., how does this patient present relative to other patients and relative to their own baseline levels of relevant factors (e.g., expressivity). A digital system should take into account these same components. Given recordings of speech and facial video are readily obtainable from smartphones, identifying longitudinal mental health biomarkers from these modalities can enable scalable digital tools to support mental health.

In this paper we focus on depression – one of the most common mental health conditions – and we investigate its association with linguistic and acoustic features of recorded speech, as well as with facial features from simultaneously recorded video. Most relevant to our acoustic analysis, Cummins et al. study the relationship between acoustic features and depression in a longitudinal study with 585 participants and bi-weekly speech and PHQ8 measurements [2]. They identify several acoustic features that are negatively associated with depression severity: speaking rate, articulation rate, and how loudly a participant speaks. Other smaller scale longitudinal studies have been conducted on depression and acoustic features, confirming the importance of speech timing (or *fluency*) features [3], as well as identifying other potential markers of depression severity such as *harmonics-to-noise ratio* (HNR) and *shimmer* measured on vowels [4]. The findings of many cross-sectional and longitudinal studies are summarised in recent reviews [5, 6].

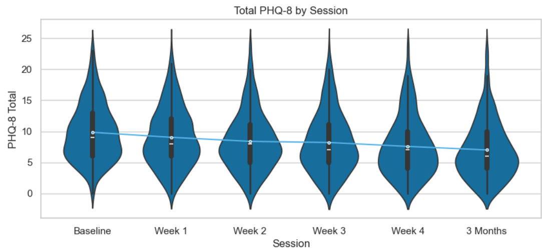

Figure 1: *Distribution of PHQ-8 scores over the study shows a wide range of scores and decreasing trend over time.*

Regarding language, much prior work has used social media data [7], though more recently participant diaries [8] and clinical interviews [9] have been studied. The Linguistic Inquiry and Word Count (LIWC) is used to compute word frequencies within psychological categories, and studies show that self-referential pronouns and negative words correlate with depression [10, 11]. Large language models (LLMs) offer deeper linguistic context understanding, with recent findings showing LLM-rated sentiment in open-ended symptom descriptions predicts depression changes [8]. Facial expressions have also been studied for depression detection, initially in controlled settings [12, 13], though mobile technology now enables remote assessments. Recent work found that facial landmarks near the lips are most predictive of depression [14].

Prior studies have also looked at multi-modal affective features. The *DAIC-WOZ* dataset [15] contains speech and facial video of participants in a lab interacting with either a human or virtual interviewer and was studied extensively in the AVEC challenges [6]. Furthermore, recent work has investigated depression estimation from YouTube Vlogs, showing that audiovisual features significantly improves accuracy (F1 of 63.9 visual and 67.8 audio to 73.5 audiovisual) for predicting vlogger's informally stated depression [16].

The novel dataset we present here (4,875 observations from 954 participants) is one of the largest longitudinal studies exploring facial-vocal biomarkers of evolving well-being within a fully remote mobile intervention. In our first analysis, we find significant vocal and facial action unit biomarkers associated with depression severity. Furthermore, we implement a multi-

[*]Both authors contributed equally to this work

10.21437/Interspeech.2025-2560

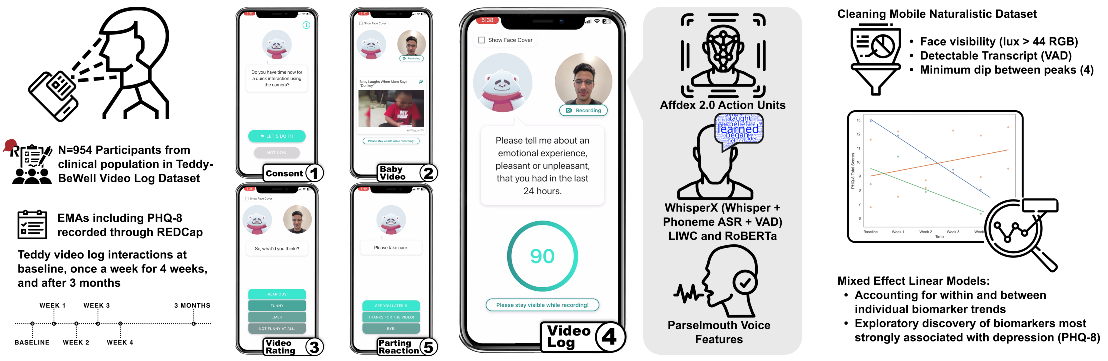

Figure 2: *Schematic of the BeWell study protocol. Participants describe an emotional experience they have had in the last 24-hours. We extract acoustic, linguistic and facial features from these recordings and analyze their association with depression.*

modal machine learning model to estimate depression that improves over baselines. We probe our models through an ablation study and SHAP analysis to offer insights about the importance of different modalities for estimating depression severity.

2. Methods

2.1. Study Protocol

This study uses data from the Behavior, Biology, and Well-Being (BeWell) randomized clinical trial (NCT05183867): an intervention study where participants are randomized to arms receiving meditation, well-being psychoeducation, or *usual care* treatment. The dataset comprises 954 participants interacting with a custom behavioral research platform, *Teddy* (Figure 2). Regarding demographics, the self-reported genders are 75.74% female, 21.03% male, 2.18% non-binary, and 1.05% other. Ages include young adults (18-30) 27.72%, adults (30-65) 71.58%, and seniors (65+) 0.69%. Self-reported races are 77.44% white, 8.97% black, 7.19% more than one, 5.51% Asian, and 0.89% indigenous or Pacific Islander.

The study follows a clinical population, where recruited participants scored over 5 on the nine-item Patient Health Questionnaire (PHQ-9) and many also underwent a Structured Clinical Interview for DSM-5 (SCID-5). As shown in Figure 2, the Teddy-BeWell longitudinal dataset was collected at baseline (before study start), once a week for 4 weeks, and then at a 3-month follow-up. At each of the 6 checkpoints, participants completed a PHQ-8 [17] to assess depressive symptom severity. Participants then interacted with the Teddy app that (1) requested the user's consent for the recording, (2) showed them a funny video, (3) asked the user to rate the video, (4) prompted the user: "Please tell me about an emotional experience, pleasant, or unpleasant that you had in the last 24 hours" with the camera and microphone turned on, and (5) asked the user to close the app with a farewell. Our analysis focuses on step (4), which creates a 90-second video log. The length of 90 seconds was decided upon after user testing, which indicated that the longer duration facilitated the sharing of more genuine experiences. The distribution of PHQ-8 scores is shown in Figure 1, displaying a diverse range of depression severity levels.

2.2. Audiovisual Feature Pre-processing

We compute a set of vocal and facial features. The language transcripts were obtained using WhisperX [20], a model that incorporates voice activity recognition (VAD) and a phoneme model into OpenAI's Whisper to produce more correct time-stamped transcripts. LIWC-22 [11] was used to extract word-frequency features posited to be associated with well-being based on psychological theories [10]. The *RoBERTa* language model fine-tuned on the GoEmotions dataset [21, 22] was used to extract fine-grained sentiment characteristics that account for the context of words. The positive and negative sentiment features were derived from the average intensity of all positive and negative emotions estimated by RoBERTa.

Acoustic features were extracted using *Parselmouth* [23]. These features, building from Cummins et al. [2, 24], measure properties of the speech production process that may become impaired during depression. These include *speech timing* (e.g., articulation rate and pauses), *prosody* (e.g., loudness and pitch variation), *voice quality* (i.e., related to the generation of the speech signal in the larynx), and *spectral properties* (i.e., related to the shaping of the vocal tract to produce phonemes). Facial expression action units (AUs) were extracted using Affdex 2.0 [25], a proprietary algorithm chosen because it is rated to perform fairly across diverse demographics based on both mobile and computer settings. The AU features are based on prior theories from controlled lab experiments, including smile-related expressions and head pose (looking up or down) [26].

We took several steps to ensure the quality of the data. For the vocal features, we excluded recordings that: (i) do not contain any interpretable speech in the transcripts, and (ii) where the participant speaks for less than 10% of the recording (using VAD). For the facial features, users were excluded if their face tracking confidence was below the fifth percentile, luminance darker than 40 RGB, and head scale larger than 3.3 interocular distance. The facial recording quality exclusion had a user retention of 99%, 98%, 98%, 100%, and 100% for white, black, more than one, asian, and indigenous racial groups respectively. We combine the resulting features to the PHQ-8 scores, and this results in a total of 5,096 sessions across 1,112 users that at minimum have successful audio recordings.

For the analysis in Section 2.3, we exclude users with less than 3 observations so that we can examine trajectories of individual change. Given this analysis considers each feature in isolation, we preserve all clean sessions by feature to study as many observations as possible. This results in 4,875 sessions from 954 users for acoustic and linguistic features, and 4,307 sessions from 905 users for facial features. For the multimodal ML analysis in Section 2.4, to ensure no missingness for a modality, we only use observations that have all modalities present – this results in a total of 4,055 sessions from 1,086

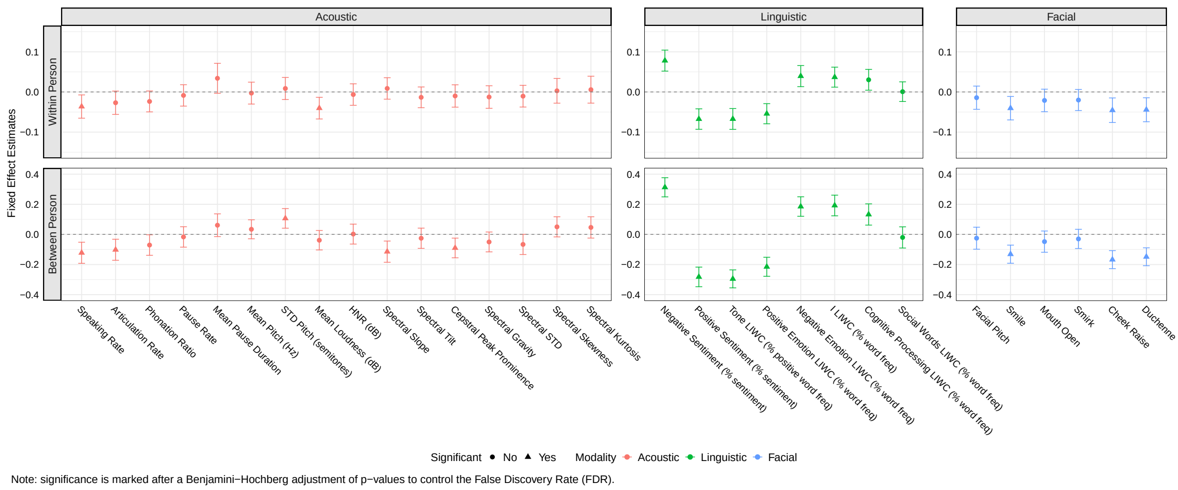

Figure 3: *Linear mixed-effects results of within- and between-individual associations between audiovisual features and PHQ-8 total score ($\alpha = 0.05$ using a Benjamini-Hochberg procedure [18]). We see (i) several acoustic speech timing and prosodic features are significant; (ii) linguistic sentiment and emotional tone – as well as use of self-referential and cognitive processing language – are significant with strong effects; and (iii) facial action units related to smiling are significant. Coefficients are pseudo-standardised [19].*

users with face and audio recordings.

2.3. Analysis 1. Linear Mixed-Effects Modeling (LME)

To identify within- and between-person features associated with depression scores, we use a linear mixed-effects model (LME). Let y_{ti} represent the PHQ-8 score at time t for participant i. Our LME model incorporates two key predictors: b_i, the participant-specific average of the audiovisual feature over time; and w_{ti}, the within-individual deviation of the audiovisual feature from this average at time t. This formulation disaggregates within-individual variation from between-individual variation through individual-mean centering [27]. The intercept β_{0i} is composed of a fixed effect γ_{00}, a between-individual fixed effect γ_{01}, and a random effect u_{0i}. The slope coefficient β_{1i} for w_{ti} is modeled as a fixed effect γ_{10}, while the time slope coefficient β_{2i} consists of a fixed effect γ_{20} and a random effect u_{2i}. Formally:

$$y_{ti} = \beta_{0i} + \beta_{1i} w_{ti} + \beta_{2i} t + r_{ti} \tag{1}$$
$$\beta_{0i} = \gamma_{00} + \gamma_{01} b_i + u_{0i} \tag{2}$$
$$\beta_{1i} = \gamma_{10} \tag{3}$$
$$\beta_{2i} = \gamma_{20} + u_{2i} \tag{4}$$

This estimates two key effects: γ_{10} is the within-person association between fluctuations in the audiovisual feature and fluctuations in depression severity, while γ_{01} is the between-person association, indicating how an individual's average feature level relates to depression severity. The fixed effect coefficients γ_{00} (the intercept) and γ_{20} (the association between depression severity and time) help control the model.

We adjust the model in two ways to take into account assumption violations. First, auto-regressive structure of type 1 is specified for the autocorrelation between residuals. Second, we also tested our models for heteroskedasticity and normality of residuals. While residuals were normally distributed, our models showed heteroskedasticity in various cases and required cluster-robust error correction of type *CR1* [28]. We used the R libraries *nlme* for LME with AR(1) correlation structure and *clubSandwich* for cluster-robust error correction. We report standardized coefficients using the *pseudo* approach [19].

2.4. Analysis 2. Multimodal Machine Learning

We subsequently analyze different machine learning (ML) models that use all of the features simultaneously. Our proposed model is a mixed-effects random forest model [29] which has previously been found to be useful when estimating depression severity from wearable data [30]. In this approach, the ML model is used as the estimator of the fixed effects, $f(\mathbf{Z}_i; \Theta)$ where $\mathbf{Z}_i$ is a matrix of features and Θ the hyper-parameters, while random intercepts u_{0i} are estimated per user. Formally:

$$y_{ti} = f(\mathbf{Z}_i; \Theta) + u_{0i} + r_{ti} \tag{5}$$

We perform 5-fold nested cross-validation. For each fold, 80% of the data is used for training, and the remaining 20% for testing. Hyperparameter tuning is performed on the independent subset of users with <3 sessions for each modality. We compare against a personalized baseline, which predicts that unseen PHQ-8 is equal to that participant's average PHQ-8 score (at training). Finally, we analyze feature importance using a feature ablation approach and SHapley Additive exPlanations (SHAP) for the predictions of the best model [31].

3. Results & Discussion

3.1. Analysis 1. Linear Mixed-Effects Results

Figure 3 displays the results of Analysis 1. Focusing first on the within-person associations between depression severity and acoustic features, we see that *Speaking Rate* and *Mean Loudness* are significantly negatively associated with depression – on days where participants are less depressed relative to their baseline, they speak faster and louder. *Speaking Rate* is also significantly associated with depression severity between individuals, along with one other *speech timing* feature: *Articulation Rate*. These associations corroborate findings from a recent longitudinal study [2] and may be explained by *psychomotor retardation*, which is a common symptom in depression. They might also suggest increased hesitation in speech planning and production in individuals with more severe depression, which could relate to cognitive impairment. Two features related to *voice quality* are also significantly negatively associated with depression

Table 1: *Performance metrics (MSE, MAE, and R^2) for different ML models. The best values for each metric are in* ***bold***.

Modality	Model	MSE$^{\downarrow}$ (SD)	MAE$^{\downarrow}$ (SD)	$R^{2\uparrow}$ (SD)
Multimodal Standard Approach				
All	Group Mean	22.57 (1.85)	3.79 (0.14)	0.000 (0.000)
All	Random Forest (RF)	20.96 (1.61)	3.63 (0.13)	0.070 (0.010)
Multimodal Mixed-Effects Approach				
All	Individual Mean	11.40 (0.99)	2.47 (0.11)	0.492 (0.061)
All	ME-LASSO	10.50 (0.92)	2.44 (0.09)	0.533 (0.043)
All	ME-RF	**10.40 (0.95)**	**2.43 (0.11)**	**0.538 (0.044)**
All	ME-SVR	10.61 (0.85)	2.44 (0.09)	0.528 (0.038)
Individual Modalities (Mixed RF)				
Acoustic	ME-RF	10.55 (0.90)	2.44 (0.10)	0.531 (0.041)
Linguistic	ME-RF	10.46 (0.89)	2.44 (0.09)	0.534 (0.043)
Facial	ME-RF	10.67 (0.86)	2.45 (0.09)	0.525 (0.042)

across individuals: the *Cepstral Peak Prominence (CPP)* and *Spectral Slope*. The CPP feature was designed to measure *dysphonia* (hoarseness of the voice) – with the negative association suggesting individuals with lower levels of depression had less dysphonic voices. However, the *Spectral Slope* – which quantifies the difference in energies between high and low frequency bands – has an opposite trend than expected. Another counterintuitive result is that variability in pitch is positively associated with higher depression severity. It is often posited that more depressed individuals speak more monotonously which is at odds with our reported association, though we note that our finding does align with a recent meta-analysis [6]. These inter-individual findings could also relate to the gender imbalance in our dataset: there are many more women and their average depression level is higher relative to other genders.

The linguistic features in Figure 3 show several strong and intuitive effects, highlighting the importance of also analyzing the content of speech when studying depression. Negative sentiment, as assessed using RoBERTa, shows the strongest effect of all features – both intra-individually and inter-individually. Its directionality is supported by the *Negative Emotion LIWC* feature, as well as by the other RoBERTa and LIWC features that are positively valenced. Furthermore, the use of more self-referential pronouns (*I LIWC*) is positively associated with depression severity, which aligns with prior findings [10, 32]. The positive inter-individual relation between cognitive processing and depression contradicts prior findings and may be a result of the specific task performed. The facial action unit features show that smiling – including the genuine *Duchenne* smile – and raising one's cheeks are significantly associated with lower self-reported depression severity.

It is interesting that significant associations of the linguistic and facial features are consistent within and between individuals, while this is not always the case for the acoustic features. The weaker intra-individual effects for acoustic features related to voice quality and tonality could potentially be explained by the challenges of ambulatory, smartphone-derived speech recordings: fine-grained changes in these properties are subtle, and may be overshadowed by background noise or device-specific audio filtering [33].

3.2. Analysis 2. Multimodal Machine Learning Results

Table 1 shows that the random intercept mixed-effects (ME) machine learning models substantially outperform standard machine learning approaches. The mixed-effects models also perform slightly better than the challenging *Individual Mean* baseline with the mixed-effects random forest (*ME-RF*) performing best compared to the support vector regressor (ME-SVR) and Least Absolute Shrinkage and Selection Operator (ME-LASSO). Finally, we show that using all the modalities with *ME-RF* results in a lower error versus using single modalities in isolation. To understand feature importance, we performed a SHAP analysis on the *ME-RF* model, computed across each outer testing cross-validation fold and aggregated in Figure 4. The 10 most important features in the model include language, acoustic, and facial features, which further highlights the utility of using multimodal data to assess depression.

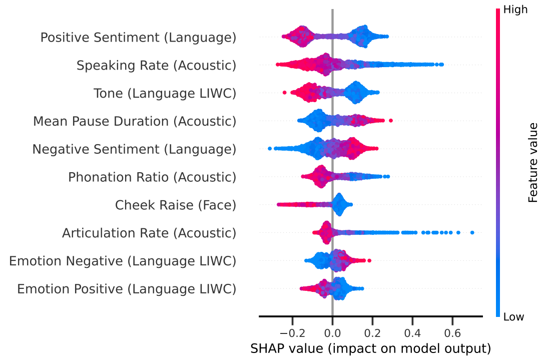

Figure 4: *SHAP plot from ME-RF shows that the most important features (top to bottom) include all three modalities.*

Our dataset exhibits demographic imbalances, with 75.74% female and 77.44% white participants. Model performance shows variation across groups, with higher R^2 for females (0.53, SD = 0.04) vs. non-females (0.52, SD = 0.09), and for non-white (0.58, SD = 0.06) vs. white participants (0.52, SD = 0.04). The lower explained variance for non-female participants highlights the need for future analysis of demographic-specific performance especially across genders.

4. Conclusion

This analysis identified significant acoustic, linguistic and facial biomarkers for depression from smartphone-derived audiovisual recordings. We also showed that these features estimate depression severity on held-out data using mixed-effects machine learning. Limitations stem from the mobile-based and ambulatory nature of the study which results in noisier recordings, potentially suppressing the detection of other relevant associations. Additionally, we only study associations relative to the total score of the PHQ-8 which does not account for the heterogeneity in depression symptom profiles. Future work should analyze how performance generalizes across demographics, how features associate with individual symptoms, and what interactions between features are significant.

5. Acknowledgments

We thank Nick Cummins for support on the acoustic feature processing and Jeffrey Girard for feedback on the statistical modeling framework. This work was supported by the National Center for Complementary & Integrative Health of the National Institutes of Health (# U24AT011289), NCCIH K23: K23AT010879, Hope for Depression Research Foundation Defeating Depression Award, and Joy Ventures. We also thank Affectiva Inc. for donating a free license to use Affdex 2.0

6. References

[1] H. Hsin, M. Fromer, B. Peterson, C. Walter, M. Fleck, A. Campbell, P. Varghese, and R. Califf, "Transforming Psychiatry into Data-Driven Medicine with Digital Measurement Tools," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–4, 2018.

[2] N. Cummins, J. Dineley, P. Conde, F. Matcham, S. Siddi, F. Lamers, E. Carr, G. Lavelle, D. Leightley, K. M. White *et al.*, "Multilingual markers of depression in remotely collected speech samples: A preliminary analysis," *Journal of affective disorders*, vol. 341, pp. 128–136, 2023.

[3] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[4] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Thirteenth annual conf. of the international speech communication assoc.*, 2012.

[5] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[6] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[7] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the international AAAI conference on web and social media*, vol. 7, no. 1, 2013, pp. 128–137.

[8] J. K. Hur, J. Heffner, G. W. Feng, J. Joormann, and R. B. Rutledge, "Language sentiment predicts changes in depressive symptoms," *Proceedings of the National Academy of Sciences*, vol. 121, no. 39, p. e2321321121, Sep. 2024, publisher: Proceedings of the National Academy of Sciences.

[9] L. Corbin, E. Griner, S. Seyedi, Z. Jiang, K. Roberts, M. Boazak, A. Bahrami Rad, G. D. Clifford, and R. O. Cotes, "A comparison of linguistic patterns between individuals with current major depressive disorder, past major depressive disorder, and controls in a virtual, psychiatric research interview," *Journal of Affective Disorders Reports*, vol. 14, p. 100645, 2023.

[10] Y. R. Tausczik and J. W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[11] R. Boyd, A. Ashokkumar, S. Seraj, and J. Pennebaker, "The development and psychometric properties of liwc-22," 02 2022.

[12] Q. Wang, H. Yang, and Y. Yu, "Facial expression video analysis for depression detection in Chinese patients," *Journal of Visual Communication and Image Representation*, vol. 57, pp. 228–233, Nov. 2018.

[13] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu, "Deep Neural Networks for Depression Recognition Based on 2D and 3D Facial Expressions Under Emotional Stimulus Tasks," *Frontiers in Neuroscience*, vol. 15, 2021.

[14] S. Nepal, A. Pillai, W. Wang, T. Griffin, A. C. Collins, M. Heinz, D. Lekkas, S. Mirjafari, M. Nemesure, G. Price, N. C. Jacobson, and A. T. Campbell, "MoodCapture: Depression Detection Using In-the-Wild Smartphone Images," Feb. 2024.

[15] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), May 2014.

[16] K. Min, J. Yoon, M. Kang, D. Lee, E. Park, and J. Han, "Detecting depression on video logs using audiovisual features," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–8, Nov. 2023, publisher: Palgrave.

[17] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1, pp. 163–173, 2009.

[18] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[19] L. Hoffman, *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge, 2015.

[20] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, arXiv:1907.11692.

[22] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," Jun. 2020, arXiv:2005.00547.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] N. Cummins, L. L. White, Z. Rahman, C. Lucas, T. Pan, E. Carr, F. Matcham, J. Downs, R. J. Dobson, and J. Dineley, "A methodological framework and exemplar protocol for the collection and analysis of repeated speech samples," *under review JMIR Research Protocols*, 2025.

[25] M. Bishay, K. Preston, M. Strafuss, G. Page, J. Turcot, and M. Mavadati, "AFFDEX 2.0: A Real-Time Facial Expression Analysis Toolkit," Nov. 2022, issue: arXiv:2202.12059 arXiv:2202.12059 [cs].

[26] L. I. Reed, M. A. Sayette, and J. Cohn, "Impact of depression on response to comedy: A dynamic facial coding analysis," pp. 804–809, 2007, num Pages: 804-809 Publisher: American Psychological Association.

[27] P. J. Curran and D. J. Bauer, "The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change," *Annual Review of Psychology*, vol. 62, no. Volume 62, 2011, pp. 583–619, Jan. 2011, publisher: Annual Reviews.

[28] F. L. Huang, W. Wiedermann, and B. Zhang, "Accounting for Heteroskedasticity Resulting from Between-Group Differences in Multilevel Models," *Multivariate Behavioral Research*, vol. 58, no. 3, pp. 637–657, May 2022.

[29] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," *Journal of Statistical Computation and Simulation*, vol. 84, no. 6, pp. 1313–1328, 2014.

[30] R. A. Lewis, A. Ghandeharioun, S. Fedor, P. Pedrelli, R. Picard, and D. Mischoulon, "Mixed effects random forests for personalised predictions of clinical depression severity," *ICML Computational Approaches to Mental Health Workshop*, 2021.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017.

[32] T. Brockmeyer, J. Zimmermann, D. Kulessa, M. Hautzinger, H. Bents, H.-C. Friederich, W. Herzog, and M. Backenstrass, "Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety," *Frontiers in Psychology*, vol. 6, Oct. 2015, publisher: Frontiers.

[33] J. Dineley, E. Carr, F. Matcham, J. Downs, R. J. B. Dobson, T. F. Quatieri, and N. Cummins, "Towards robust paralinguistic assessment for real-world mobile health (mHealth) monitoring: an initial study of reverberation effects on speech," in *Interspeech 2023*. Dublin, Ireland: ISCA, 2023, pp. 2373–2377.