INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A Multimodal Investigation of Speech, Text, Cognitive and Facial Video Features for Characterizing Depression With and Without Medication

Michael Neumann, Hardik Kothare, Doug Habberstad and Vikram Ramanarayanan

Modality.ai, Inc., San Francisco, CA

v@modality.ai

Abstract

Clinical depression is one of the most common mental disorders and technology for remote assessment of depression, including monitoring of treatment responses, is gaining more and more importance. Using a cloud-based multimodal dialog platform, we conducted a crowdsourced study to investigate the effect of depression severity and antidepressant use on various acoustic, linguistic, cognitive, and orofacial features. Our findings show that multiple features from all tested modalities show statistically significant differences between subjects with no or minimal depression and subjects with more severe depression symptoms. Moreover, certain acoustic and visual features show significant differences between subjects with moderately severe or severe symptoms who take antidepressants and those who do not take any. Machine learning experiments show that subjects with and without medication can be better discriminated from each other at higher severity levels.

Index Terms: multimodal dialog system, remote patient monitoring, depression, anti-depressants

1. Introduction

Clinical depression is a leading cause of disability worldwide. In 2020, an estimated 8.4% of U.S. adults had at least one episode of major depression [1], and the prevalence of depression symptoms increased notably during the COVID-19 pandemic [2]. Diagnosis, detection and monitoring of mental and neurological conditions, including major depressive disorder (MDD), remain a critical need today. This necessitates the development and validation of scalable, multimodal, and cost-effective technology for automatic assessment of individuals' health and well-being in the user's natural information technology environment [3]. Because depression affects speech features [4, 5, 6], facial activity and expressiveness [7, 8], these signals have the potential to serve as objective markers that can be analyzed automatically by means of speech processing and computer vision. An extensive review on speech analysis in depression assessment can be found in [9]. In this contribution, we aim at identifying measures that can be used to monitor treatment response with respect to antidepressants and anti-anxiety medication.

While earlier efforts for remote patient monitoring (RPM) usually required the use of dedicated hardware [10], recent work in the field focuses on using available consumer devices such as smartphones and wearables for various types of monitoring [11]. Regarding the mode of interaction with patients, multimodal conversational agents offer numerous advantages in RPM, including cost-effectiveness and scalability compared to human-to-human interviews, higher engagement levels and user acceptance compared to text-only interactions [12], and the ability to automatically analyze features from different modalities.

For the present study the Modality service was used, which is a cloud-based multimodal dialog system [13, 14] that can be used to elicit evidence required for detection or progress monitoring of neurological or mental health conditions through automated screening interviews. Users participate in a structured conversation with Tina, a virtual dialog agent. Depending on the health condition to be monitored, different dialog protocols can be employed. For this study, a comprehensive set of tasks to probe motor speech and cognitive functions was used.

In this paper, we present a study on crowdsourced multimodal speech and video data from people with depression at various symptom severity levels. We investigate how well speech, text, cognitive and orofacial features distinguish subjects with differing depression severity and antidepressant use. We present several extensions to the previously presented dialog system [14], including a set of tasks to assess cognitive functions, and the use of automatic speech recognition (ASR) to extract linguistic features in addition to acoustic and visual metrics.

The main contributions of this paper are a comprehensive description of a crowdsourced data collection using a multimodal, web-based dialog system, and an exploratory investigation of the effects of antidepressants on various speech, text, and video features in subjects with different depression severity.

2. Data collection

The data collection platform Prolific[1] was used to recruit participants. The pre-screening question for participant selection was, *Are you currently taking any medication to treat symptoms of depression, anxiety or low-mood (e.g. SSRIs)?* The set of available answer options was: *No; Prefer not to say; Yes, I'm taking anti-depressants; Yes, I'm taking anti-anxiety medication; Yes, I'm taking anti-psychotics; Yes, I'm taking a combination of these.*

After informed consent, a virtual agent guided participants through a set of structured speaking exercises as well as open-ended questions to elicit different types of speech, including read speech, automatic speech (counting up from one), and spontaneous speech (picture description task and one open ended question). The study protocol was granted exempt status by an external Institutional Review Board. This conversational protocol is based on previous work in remote depression monitoring [15]. In addition, the protocol contains a set of tasks designed to assess cognitive abilities, including immediate and delayed word recall, forward and backward digit spans, a category fluency task (naming as many animals as one can think

[1] https://www.prolific.co/

10.21437/Interspeech.2023-2194

Table 1: *Participant statistics. Age and PHQ-8 score are shown as mean (standard deviation).*

Cohort	# subjects	Age	PHQ-8
NO-MED	118 (76 female)	40.9 (15.5)	8.6 (6.6)
MED	151 (99 female)	39.3 (13.0)	10.9 (6.4)

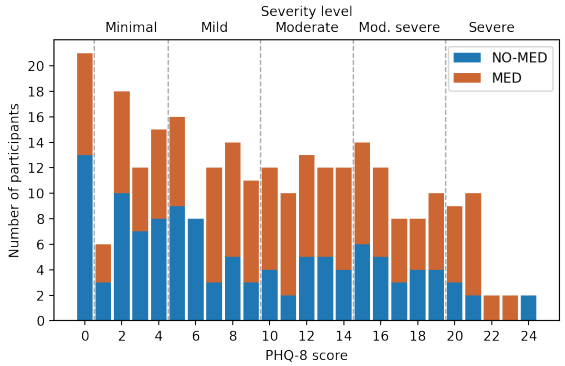


Figure 1: *Distribution of PHQ-8 scores in the MED and NO-MED cohorts.*

of), and a three-step task where participants are asked to touch their chin, nose and ear in a specific order [16]. These tasks are based on common assessments of cognitive domains, such as executive functioning, working memory, encoding memory, and retrieval memory [17, 18, 19].

After the conversation with Tina, participants were asked to fill out the following questionnaires: a usability survey, the Communicative Participation Item Bank - short form (CPIB-S) [20], and the Patient Health Questionnaire eight-item depression scale (PHQ-8), a standard instrument for depression assessment [21].

A total of 269 participants completed the assessment in the time from 2022-02-18 until 2022-08-10. 118 participants reported that they do not take medication to treat depression or anxiety (labeled as NO-MED) and 151 participants reported taking medication (MED). Table 1 shows participant statistics. Based on the severity levels described in [22, 23], the cutpoints 5, 10, 15, and 20 on the PHQ-8 score were used to study different sub cohorts based on depression severity. Fig. 1 shows the distribution of PHQ-8 scores. Audio was recorded at 44.1 kHz sampling rate.

3. Feature Extraction and Data Preprocessing

The Modality service has built-in analytics modules that extract a variety of acoustic and visual speech metrics from segmented user utterances in real-time. Acoustic metrics are extracted using Praat [24] and Kaldi [25]. Visual metrics are computed based on facial landmarks, which are extracted using the face detector in the *dnn* module of *OpenCV* (`https://opencv.org/`), and the Dlib [26] facial landmark detector. Details about the analytics modules and the extracted metrics have been previously described [27].

In addition, linguistic metrics were computed based on automatic transcriptions obtained through the Amazon Transcribe service.[2] We used the python package spaCy[3] to compute

[2] `https://aws.amazon.com/transcribe/`
[3] `https://spacy.io/`

Table 2: *Overview of extracted metrics. For visual metrics, functionals (minimum, maximum, average) are applied to produce one value across all video frames of an utterance. Visual distance metrics are measured in pixels and are normalized by dividing them by the intercanthal distance (distance between inner corners of the eyes) for each subject.*

	Domain	Metrics
Audio	Energy	shimmer (%), intensity (dB), signal-to-noise ratio (dB)
	Timing	speaking and articulation duration (sec.), articulation and speaking rate (WPM), percent pause time (PPT, %), canonical timing agreement (CTA, %)
	Voice quality	cepstral peak prominence (CPP, dB), harmonics-to-noise ratio (HNR, dB)
	Frequency	mean, max., min. fundamental frequency F0 (Hz), first three formants F1, F2, F3 (Hz), slope of 2nd formant (Hz/sec.), jitter (%)
Video	Mouth measurements	lip aperture/opening, lip width, mouth surface area, mean symmetry ratio between left and right half of the mouth
	Movement	velocity, acceleration, jerk, and speed of lower lip and jaw center
	Eyes	number of eye blinks per sec., eye opening, vertical displacement of eyebrows
Text	Lexico-semantic	word count, percentage of content words, noun rate, verb rate, pronoun rate, noun-to-verb ratio, noun-to-pronoun ratio, closed class word ratio, idea density
	Cognitive scores	percentage of correct words (immediate and delayed word recall), digit span forward/backward score (ranges from 0 to 2)

lexico-semantic features for the spontaneous speech parts of the conversation. These features are inspired by [28].

The cognitive tasks on word and digit recall were scored automatically based on the ASR output. For word recall, the score is expressed as the percentage of correct words (regardless of the order). For the digit span tasks, we assigned a score of 2 if all digits were correctly repeated in the correct order, a score of 1 if the correct digits were present, but in a different order, and a score of 0 otherwise. Table 2 provides an overview of all extracted metrics. For the analysis, two types of features were constructed based on these metrics: (1) task-specific features, i.e extracted metrics *per* speech task (Sec. 2) are considered as features[4] (e.g. *percent pause time for picture description* would be one feature), and (2) aggregate features, which are computed as the mean of a metric across all user turns.

To remove outliers from acoustic and visual metrics, we employed a distribution based outlier detection algorithm. First, all metric values that are more than five standard deviations away from the population mean are removed. These are considered extreme outliers, which potentially skew the mean. Then, the mean is re-computed and values outside ± 3 standard deviations are flagged as outliers and removed from any further analysis.

Missing data is a common problem in such remote, unsupervised data collections, which affects most machine learning algorithms because they cannot handle missing feature values. Reasons for missing data include incomplete dialogs, removed outliers as described above, or data transmission errors, which result in failures in the analytics modules. Typically, missing data can either be removed by discarding affected samples altogether or can be filled through interpolation methods. We re-

[4] Note, *acoustic* metrics were not extracted from all utterances, but only for those tasks where they are appropriate, whereas *visual* metrics were computed for each utterance.

frain from interpolating data because this can potentially distort the results, particularly if a large number of gaps is filled for a given feature. We use a threshold to determine a trade-off between removing features with many missing values and removing samples (participants). Specifically, features with more than 5% missing values in the data set are taken out completely, before remaining missing data is removed by taking out affected participant sessions.

4. Analyses and Observations

The central research question we aim to answer is the following: Based on remotely recorded data with users' end devices, which speech, orofacial, cognitive, and linguistic features show a statistically significant difference between subjects who take medication to treat depression symptoms and those who do not take such medication?

On the way to answering this, a first step in our analysis was to compare features from different depression severity sub cohorts to each other – regardless of medication use. This was done to investigate the efficacy of remotely collected speech data to distinguish, for example, minimal from severe depression. Specifically, we compared subjects with no or minimal depression (PHQ-8 score below 5) to different groups of patients reporting symptoms of depression at different cutoff points as described by [22]. We compared subjects with no or minimal depression to all patients with mild to severe symptoms of depression (PHQ-8 score $\geq$5), patients with moderate to severe symptoms (PHQ-8 $\geq$10), and patients with moderately severe to severe symptoms (PHQ-8 $\geq$15). Because of the small sample size of patients who reported severe symptoms (PHQ-8 $\geq$20, n=25) this group was not included as an individual cohort. In the following presentation of findings, we focus on aggregate features (averaged metrics across all user turns in a session, cf. Sec. 3).

Non-parametric Kruskal-Wallis tests were conducted for each individual feature to identify features that exhibit a statistically significant difference ($p = 0.05$) between two groups and effect sizes in terms of Glass' Delta were computed for these features. We used a Benjamini-Hochberg correction to control for false discovery rate [29]. Features from all tested modalities were present. Most prominently, we observed differences in kinematic features of the lip and jaw, such as average and maximum speed, which indicate slower movement in more severely depressed groups. Among the acoustic features, articulation rate, number of syllables and speaking duration for read sentences show signal. Cognitive measures for the digit span and immediate word recall tasks indicate differences between cohorts. Linguistic features related to noun rate, pronoun rate, content words and noun:pronoun and noun:verb ratios show signal for some comparisons.

While looking into effects of medication use, we focused on three sub cohorts PHQ-8 $\geq$\{5,10,15\}.[5] Each sub cohort was divided into MED and NO-MED subjects and Kruskal-Wallis tests were conducted to identify features that are different between these groups. With this setup, we wanted to find out whether observed effects between MED and NO-MED groups are stronger in more severely depressed participants. To investigate how well the extracted features can discriminate the groups, binary classification experiments with a random forest classifier were done using stratified 5-fold cross validation. Implementation was done with scikit-learn[6], and we used the default parameters for the random forest classifier. The following feature sets were used as input to the classifier: (1) acoustic only, (2) visual only, (3) cognitive only, (4) linguistic only, (5) the combination of all these features, and (6) the subset of features that showed statistically significant difference between any of the tested cohort pairs (Fig. 3a). We found that overall the best results were achieved with feature sets (2) and (6) across the different severity levels. Acoustic speech features yielded better results for the PHQ-8>15 group compared to the other cohorts.

Fig. 3 shows the effect sizes for statistically significant differences (Fig. 3a) and receiver operating characteristic (ROC) curves for the classification experiments, using feature set (6)

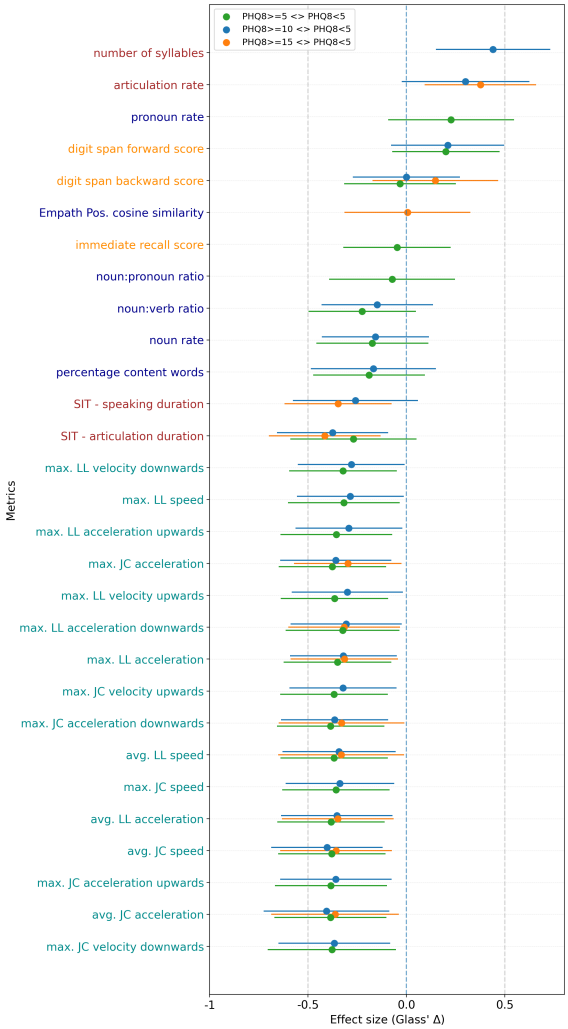
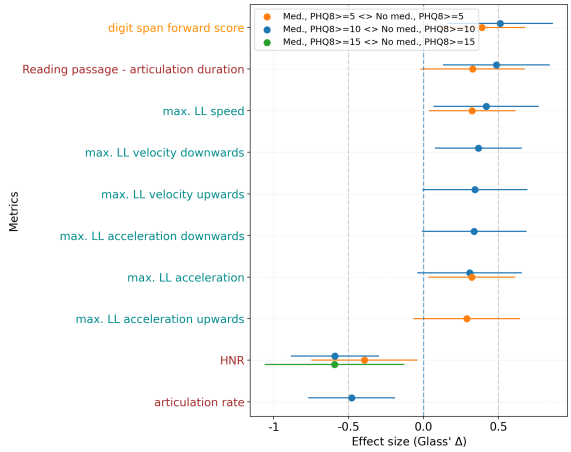


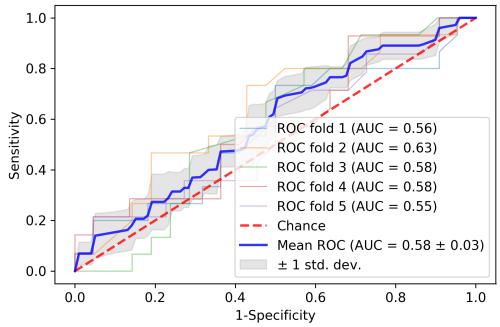
Figure 2: *Effect sizes of **acoustic**, **cognitive**, **linguistic**, and **visual** metrics that show statistically significant differences at $p < 0.05$, shown with 95% confidence interval. Negative effects indicate higher values in the **PHQ-8 <5** group. LL: lower lip, JC: jaw center, SNR: signal-to-noise ratio, SIT: sentence intelligibility test.*

[5]The PHQ<5 group was left out of this analysis because we were mainly interested in the effect of medication in subjects with at least mild depression. In the present setup, more severely depressed subjects were considered as a subset of groups at lower cutpoints, e.g. participants in the PHQ$\geq$15 group are also included in the PHQ$\geq$\{5,10\} groups.

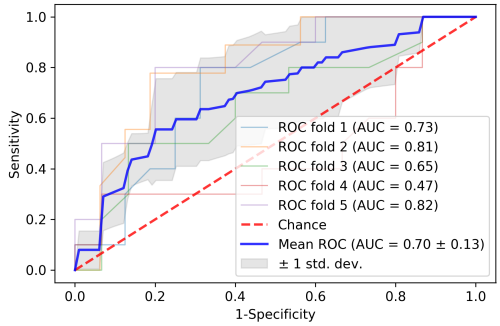
[6]https://scikit-learn.org/



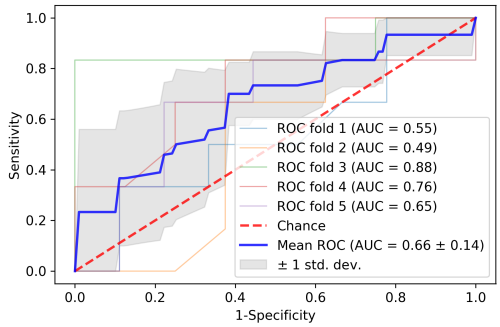
(a) *Effect sizes of **acoustic**, **visual**, and **cognitive** metrics that show statistically significant differences between MED and NO-MED cohorts at $p < 0.05$, shown with 95% confidence interval. HNR: harmonics-to-noise ratio.*



(b) *PHQ-8 $\geq$ 5 (# samples: 108 MED, 73 NO-MED)*



(c) *PHQ-8 $\geq$ 10 (# samples: 77 MED, 49 NO-MED)*



(d) *PHQ-8 $\geq$ 15 (# samples: 42 MED, 30 NO-MED)*

Figure 3: *Effect sizes and ROC curves for telling apart MED from NO-MED subjects at different severity levels. Features from (a) were used as input to a random forest classifier.*

– features from the effect size plot (Fig. 3b-3d). We observed moderate effect sizes (absolute effects around or less than 0.5) for a number of features, including harmonics-to-noise ratio (HNR), articulation rate, reading passage articulation duration, velocity and acceleration measures of the lower lip, and the digit span forward score. The ROC curves illustrate that the classification into MED and NO-MED cohorts is most difficult for the PHQ$\geq$5 group, whereas the area under curve (AUC) score for the PHQ$\geq$10 and PHQ$\geq$15 groups is well above chance level. Generally, these results need to be interpreted with caution because of the small samples.

5. Discussion

For pharmaceutical trials and clinical monitoring purposes, it is important to identify biomarkers that can help track treatment response to depression medication. This study identified key acoustic, cognitive, linguistic and orofacial features that show potential for characterizing depression, both in terms of severity of symptoms and antidepressant use. Our observation of moderate effect sizes for rate and timing measures of speech and pausing (articulation rate and duration, as well as lower lip velocity and acceleration) in distinguishing participants with and without medication are consistent with Mundt et al. (2012), who found that these measures are consistent in classifying patients with major depressive disorder (MDD) enrolled in a four-week, randomized, double-blind, placebo-controlled study as treatment Responders or Nonresponders, based on a 50% or greater improvement from baseline [30]. Liu et al. (2017) also found that speech pause time captures clinical treatment with antidepressants [31]. Our findings are also consistent with Abbas et al. (2021), who found that MDD patients' responses to antidepressant treatment (ADT) demonstrated significant increases in multiple digital markers including facial expressivity and amount of speech [32].

That being said, it is important to temper the promise of our findings with several important caveats. Our sample size is not large enough to truly claim generalizability of findings. The smaller the sample, the larger the risk of having model "blind spots" that in turn lead to variable estimates of true model performance on unseen real world data, giving algorithm designers an inaccurate sense of how well a model is performing during development [33]. This is also one reason why we do not consider deep neural network models, which require a large number of training data samples, for classification in this study (the other being the relative difficulty in interpreting such models). Furthermore, we have considered a wide range of antidepressants as well as anti-psychotics in the study, which contributes to the variability (the two types of medication have different treatment and side effects; however, we included both for the sake of a larger sample). Finally, findings regarding the cognitive scores need to be taken with a grain of salt, because the scores rely on ASR output, and errors can be propagated.

In sum, we have identified acoustic, cognitive, linguistic and orofacial features that show potential for characterizing depression, both in terms of severity of symptoms and antidepressant use. Future work will focus on confirming the robustness and generalizability of these findings on a larger and more balanced sample, especially for people with severe symptoms, and on investigating effects with respect to different drug classes, with a more in-depth analysis of the specific causes for the observed differences in features with medication use.

6. References

[1] S. Abuse and M. H. S. Administration, "2020 national survey of drug use and health (nsduh)," https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases, 2020.

[2] C. K. Ettman, S. M. Abdalla *et al.*, "Prevalence of depression symptoms in us adults before and during the covid-19 pandemic," *JAMA network open*, vol. 3, no. 9, pp. e2 019 686–e2 019 686, 2020.

[3] S. Kumar, W. Nilsen *et al.*, "Mobile health: Revolutionizing healthcare through transdisciplinary research," *Computer*, vol. 46, no. 1, pp. 28–35, 2012.

[4] D. J. France, R. G. Shiavi *et al.*, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[5] V. Jain, J. L. Crowley *et al.*, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 87–91.

[6] H. Kaya, F. Eyben *et al.*, "Cca based feature selection with application to continuous depression recognition from acoustic speech features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3729–3733.

[7] H. Meng, D. Huang *et al.*, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 21–30.

[8] M. Nasir, A. Jati *et al.*, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 43–50.

[9] N. Cummins, S. Scherer *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015.

[10] R. F. Dickerson, E. I. Gorlin, and J. A. Stankovic, "Empath: a continuous remote emotional health monitoring system for depressive illness," in *Proceedings of the 2nd Conference on Wireless Health*, 2011, pp. 1–10.

[11] Y. Ranjan, Z. Rashid *et al.*, "Radar-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices," *JMIR mHealth and uHealth*, vol. 7, no. 8, p. e11734, 2019.

[12] J. O. Egede, D. Price *et al.*, "Design and evaluation of virtual human mediated tasks for assessment of depression and anxiety," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021, pp. 52–59.

[13] D. Suendermann-Oeft, A. Robinson *et al.*, "Nemsi: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.

[14] M. Neumann, O. Roessler *et al.*, "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, 2020, pp. 47–52.

[15] J. C. Mundt, P. J. Snyder *et al.*, "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.

[16] W. Burke, D. Pautler *et al.*, "On the feasibility of remotely administered and self driven cognitive assessments using a multimodal dialog agent," in *Annual Meeting of the Cognitive Neuroscience Society (CNS)*, San Francisco, CA, USA, April 2022.

[17] Z. Shao, E. Janse *et al.*, "What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults," *Frontiers in psychology*, vol. 5, p. 772, 2014.

[18] D. L. Woods, M. M. Kishiyama *et al.*, "Improving digit span assessment of short-term verbal memory," *Journal of clinical and experimental neuropsychology*, vol. 33, no. 1, pp. 101–111, 2011.

[19] C. Cerami, B. Dubois *et al.*, "Clinical validity of delayed recall tests as a gateway biomarker for alzheimer's disease in the context of a structured 5-phase development framework," *Neurobiology of aging*, vol. 52, pp. 153–166, 2017.

[20] C. Baylor, K. Yorkston *et al.*, "The communicative participation item bank (cpib): Item bank calibration and development of a disorder-generic short form," 2013.

[21] K. Kroenke, T. W. Strine *et al.*, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[22] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.

[23] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[25] D. Povey, A. Ghoshal *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[26] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[27] M. Neumann, O. Roesler *et al.*, "Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 2021*. ISCA, 2021, pp. 4783–4787. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1801

[28] V. Boschi, E. Catricala *et al.*, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in psychology*, vol. 8, p. 269, 2017.

[29] 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.

[30] J. C. Mundt, A. P. Vogel *et al.*, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[31] Z. Liu, H. Kang *et al.*, "Speech pause time: A potential biomarker for depression detection," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2017, pp. 2020–2025.

[32] A. Abbas, C. Sauder *et al.*, "Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study," *Frontiers in digital health*, vol. 3, p. 610006, 2021.

[33] V. Berisha, C. Krantsevich *et al.*, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, vol. 4, no. 1, p. 153, 2021.