On the Utility of Audiovisual Dialog Technologies and Signal Analytics for Real-time Remote Monitoring of Depression Biomarkers

Michael Neumann[*], **Oliver Roesler**[*], **David Suendermann-Oeft**[*] **and Vikram Ramanarayanan**[*†]
[*] Modality.ai, Inc.
[†] University of California, San Francisco
vikram.ramanarayanan@modality.ai

Abstract

We investigate the utility of audiovisual dialog systems combined with speech and video analytics for real-time remote monitoring of depression at scale in uncontrolled environment settings. We collected audiovisual conversational data from participants who interacted with a cloud-based multimodal dialog system, and automatically extracted a large set of speech and vision metrics based on the rich existing literature of laboratory studies. We report on the efficacy of various audio and video metrics in differentiating people with mild, moderate and severe depression, and discuss the implications of these results for the deployment of such technologies in real-world neurological diagnosis and monitoring applications.

1 Introduction

Diagnosis, detection and monitoring of neurological and mental health in patients remain a critical need today. This necessitates the development of technologies that improve individuals' health and well-being by continuously monitoring their status, rapidly diagnosing medical conditions, recognizing pathological behaviors, and delivering just-in-time interventions, all in the user's natural information technology environment (Kumar et al., 2012). However, early detection or progress monitoring of neurological or mental health conditions, such as clinical depression, Amyotrophic Lateral Sclerosis (ALS), Alzheimer's disease, dementia, etc., is often challenging for patients due to multiple reasons, including, but not limited to: (i) lack of access to neurologists or psychiatrists; (ii) lack of awareness of a given condition and the need to see a specialist; (iii) lack of an effective standardized diagnostic or endpoint for many of these health conditions; (iv) substantial transportation and cost involved in conventional or traditional solutions; and in some cases, (v) shortage of medical specialists in these fields to begin with (Steven and Steinhubl, 2013).

We developed NEMSI (Suendermann-Oeft et al., 2019), or the NEurological and Mental health Screening Instrument, to bridge this gap. NEMSI is a cloud-based multimodal dialog system that conducts automated screening interviews over the phone or web browser to elicit evidence required for detection or progress monitoring of the aforementioned conditions, among others. While intelligent virtual agents have been proposed in earlier work for such diagnosis and monitoring purposes, NEMSI makes novel contributions along three significant directions: First, NEMSI makes use of devices available to everyone everywhere (web browser, mobile app, or regular phone), as opposed to dedicated, locally administered hardware, like cameras, servers, audio devices, etc. Second, NEMSI's backend is deployed in an automatically scalable cloud environment allowing it to serve an arbitrary number of end users at a small cost per interaction. Thirdly, the NEMSI system is natively equipped with real-time speech and video analytics modules that extract a variety of features of direct relevance to clinicians in the neurological and mental spaces.

A number of recent papers have investigated automated speech and machine vision features for predicting severity of depression (see for example France et al., 2000; Joshi et al., 2013; Meng et al., 2013; Jain et al., 2014; Kaya et al., 2014; Nasir et al., 2016; Pampouchidou et al., 2016; Yang et al., 2017). These include speaking rate, duration, amplitude, and voice source/spectral features (fundamental frequency (F0), amplitude modulation, formants, and energy/power spectrum, among others) computed from the speech signal, and facial dynamics (for instance, landmark/facial action unit motions, global head motion, and eye blinks) and statistically derived features from emotions, action units, gaze, and pose derived from the video signal. We use these studies to inform our choices of speech and video metrics computed in real time, allowing clinicians to obtain useful analytics for their patients moments after they have interacted with the NEMSI dialog system.

We need to factor in additional considerations while deploying analytics modules as part of scalable real-time cloud-based systems in practice. Many of the studies above analyzed data recorded either offline or in laboratory conditions, implicitly assuming signal conditions which may hold differently or not at all during real world use. These considerations include, but are not limited to: (i) wide range of acoustic environments and lighting conditions resulting in variable background noise and choppy/blocky video at the user's end[1], (ii) limitations on a given user's network connection bandwidth and speed; (iii) the quantum of server traffic (or the number of patients/users trying to access the system simultaneously); and (iv) device calibration issues, given the wide range of user devices. This paper investigates the utility of a subset of audio and video biomarkers for depression collected using the NEMSI dialog system in such real-world conditions.

The rest of this paper is organized as follows: Sections 2 and 3 first present the NEMSI dialog system and the data collected and analyzed. Section 4 then details the speech and video feature extraction process. Section 5 presents statistical analyses of different groups of depression cohorts as determined by the reported PHQ-8 score, before Section 6 rounds out the paper, discussing the implications of our observations for real-world mental health monitoring systems.

2 System

2.1 NEMSI dialog ecosystem

NEMSI (NEurological and Mental health Screening Instrument) is a cloud-based multimodal dialog system. Refer to Suendermann-Oeft et al. (2019) for details regarding the system architecture and various software modules.

NEMSI end users are provided with a website link to the secure screening portal as well as login credentials by their caregiver or study liaison (physician or clinic). Once appropriate microphone and camera checks that the captured audio and video are of sufficient quality are complete, users hear the dialog agent's voice and are prompted to start a conversation with the agent, whose virtual image also appears in a web window. Users are also able to see their own video, if so needed, in a small window in the upper right corner of the screen. The virtual agent then engages with users in a conversation using a mixture of structured speaking exercises and open-ended questions to elicit speech and facial behaviors relevant for the type of condition being screened for.

Analytics modules extract multiple speech (for instance, speaking rate, duration measures, F0, etc.) and video features (such as range and speed of movement of various facial landmarks) and store them in a database, along with information about the interaction itself such as the captured user responses, call duration, completion status, etc. All this information can be accessed by the clinicians after the interaction is completed through an easy-to-use dashboard which provides a high-level overview of the various aspects of the interaction (including the video thereof and analytic measures computed), as well as a detailed breakdown of the individual sessions and the underlying interaction turns.

3 Data

Depending on the health condition to be monitored and on the clinician's needs, different protocols can easily be employed in the NEMSI system. For the present study, we designed a protocol targeting the assessment of depression severity, based on (Mundt et al., 2007). The protocol elicits five different types of speech samples from participants that are consistently highlighted in the literature: (a) free speech (open-ended questions about subjects' emotional and physical state), (b) automated speech (counting up from 1), (c) read speech, (d) sustained vowels, and (e) measure of diadochokinetic rate (rapidly repeating the syllables /pa ta ka/).

After dialog completion, participants are asked to answer the Patient Health Questionnaire eight-item depression scale (PHQ-8), a standard scoring system for depression assessment (Kroenke et al., 2009). The self-reported PHQ-8 score serves as a reference point for our analysis. Further, we ask for information about age, sex, primary language and residence.

In total, we collected data from 307 interactions. After automatic data cleaning[2], 208 sessions re-

[1] Such conditions often arise despite explicit instructions to the contrary.

[2] We removed interactions for which PHQ-8 answers or relevant speech metrics were missing and sessions for which no face was detected in the video

mained for analysis. From those 208 participants, 98 were females, 97 were males and 13 did not specify. Mean participant age is 36.5 (SD = 12.1). 184 participants specified English as their primary language, 9 other languages and 15 did not specify. 176 participants were located in the US, 8 in the UK, 5 in Canada, 4 in other countries and 15 did not specify. Figure 1 shows the distribution of PHQ-8 scores among women and men.

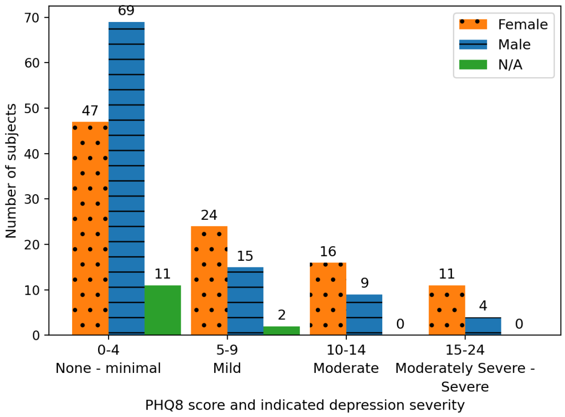

Figure 1: Distribution of PHQ-8 scores by gender.

4 Signal Processing and Metrics Extraction

4.1 Speech Metrics

For the speech analysis, we focus on *timing measures*, such as speaking rate and percentage of pause duration, *frequency domain measures*, such as fundamental frequency (F0) and jitter, and *energy-related measures*, such as volume and shimmer. We have selected commonly established speech metrics for clinical voice analysis (France et al., 2000; Mundt et al., 2012, 2007).

As described in Section 3, there are different types of speech samples, e.g. free speech and sustained vowels. Not all acoustic measures are meaningful for each type of stimuli. Table 2 presents all extracted metrics for the particular speech sample types.

All metrics are extracted with Praat (Boersma and Van Heuven, 2001). For the following measures, heuristics have been used to ignore obvious outliers in the analysis: articulation rate (excluded >350 words/min), speaking rate (excluded >250 words/min), percent pause time (excluded >80%).

Cutpoint (group sizes)	Free speech	Held Vowels
5 (127/81)	Percent pause time (a,f)	Volume (a,f), HNR (m), Mean F0 (m)
10 (168/40)	-	Jitter (f)
15 (193/15)	Volume (a,f,m)	Mean F0 (a), Volume (f)

Table 1: Speech metrics for which a statistically significant ($p < 0.05$) difference between sample populations is observed. In parentheses: f - females, m - males, a - all.

	Free speech	Read speech	Automated	Held vowels	DDK
SpRate		✓			
ArtRate		✓			
SylRate					✓
PPT	✓	✓	✓		
Mean F0				✓	
Jitter				✓	
HNR				✓	
Volume	✓	✓	✓	✓	✓
Shimmer				✓	

Table 2: Speech metrics for each type of speech sample. SpRate = speaking rate, ArtRate = articulation rate, SylRate = syllable rate, PPT = percent pause time, DDK = dysdiadochokinesia.

4.2 Visual Metrics

For each utterance, 14 facial metrics were calculated in three steps: (i) face detection, (ii) facial landmark extraction, and (iii) facial metrics calculation. For face detection, the Dlib[3] face detector was employed, which uses 5 histograms of oriented gradients to determine the (x, y)-coordinates of one or more faces for every input frame (Dalal and Triggs, 2005). For facial landmark detection the Dlib facial landmark detector was employed, which uses an ensemble of regression trees proposed by Kazemi and Sullivan (2014), to extract 68 facial landmarks according to MultiPIE (Gross et al., 2010). Figure 2 illustrates the 14 facial landmarks: RB (right eyebrow), URER (right eye, upper right), RERC (right eye, right corner), LRER (right eye, lower right), LB (left eyebrow), ULEL (left eye, upper left), LELC (left eye, left corner), LLEL (left eye, lower left), NT (nose tip), UL (upper lip center), RC (right corner of mouth), LC (left corner of mouth), LL (lower lip center), and JC (jaw center). These are then used to calculate the following facial metrics:

[3] http://dlib.net/

Cutpoint	Gender	Free speech	Read speech
5	All	width, vJC, S_R, S_ratio, utter_dur	S_R
	Female	S_ratio, utter_dur	eye_blinks
	Male	S_ratio, eyebrow_vpos, eye_open, eye_blinks	
10	All	S_ratio, utter_dur	S_R
	Female	open, width, LL_path, JC_path, S, S_R, S_L, eyebrow_vpos, eye_open	
	Male	open, width, LL_path, JC_path, S, S_R, S_L, S_ratio, eyebrow_vpos, eye_open, eye_blinks	open, width, S, S_R, S_L, eyebrow_vpos
15	All	vLL, vJC	vLL, S_ratio
	Female	width, vLL, vJC, S_ratio	vLL, S_ratio
	Male	width, S, S_L, eyebrow_vpos, eye_open, eye_blinks	eyebrow_vpos

Table 3: Facial metrics for which a statistically significant ($p < 0.05$) difference between sample populations is observed. For gender *All* not only female and male samples, but also samples for which no gender was reported are used.

- **Movement measures**: Average lips opening and width (open, width) were calculated as the Euclidean distances between UL and LL, and RC and LC, respectively. Average displacement of LL and JC (LL_path, JC_path) were calculated as the module of the vector between the origin and LL and JC. Average eye opening (eye_open) was calculated as the Euclidean distances between URER and LRER, and ULEL and LLEL. Average vertical eyebrow displacement (eyebrow_vpos) was calculated as the difference between the vertical positions of RB and NT, and LB and NT. All measures were computed in millimeters.

- **Velocity measures**: The average velocity of LL and JC (vLL, vJC) in mm/s was calculated as the first derivative of LL_path and JC_path with time.

- **Surface measures**: The average total mouth surface (S) in mm^2 was calculated as the sum of the surfaces of the two triangles with vertices RC, UL, LL (S_R) and LC, UL, LL (S_L). Additionally, the mean symmetry ratio (S_ratio_avg) between S_R and S_L was determined.

- **Duration measures**: Utterance duration (utter_dur) in seconds.

- **Eye blink measures**: The number of eye blinks (eye_blinks) in blinks per second calculated using the eye aspect ratio as proposed by Soukupová and Čech (2016).

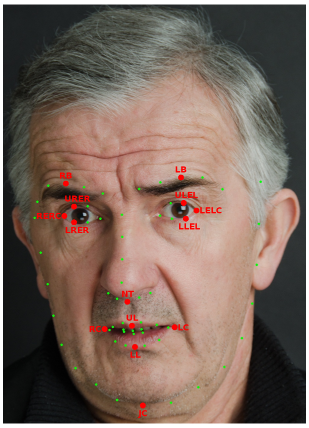

Figure 2: Illustration of the 68 obtained and 14 used facial landmarks.

5 Analyses and Observations

The central research question of this study is the following: for a given metric, is there a statistically significant difference between participant cohorts with and without depression of a given severity (i.e., above and below a certain cut-point PHQ-8 score)? The PHQ-8 has established cutpoints above which the probability of a major depression increases substantially (Kroenke and Spitzer, 2002). Ten is commonly recommended as cutpoint for defining current depression (see (Kroenke et al., 2009) for a comprehensive overview). For our analysis, we use the cutpoints 5, 10, and 15 which align with the PHQ-8 score intervals of mild, moderate and moderately severe depression. Concretely, for each metric and cutpoint, we divide the data into two sample populations: (a) PHQ-8 score below and (b) PHQ-8 equal or above the cutpoint. We conducted a non-parametric Kruskal-Wallis test for every combination to find out whether certain obtained metrics show a statistically significant difference between cohorts.[4]

[4] We decided to exclude the /pa ta ka/ exercise (measure of diadochokineic rate) from the analysis, because we observed that many participants did not execute it correctly (e.g. making pauses between repetitions).

5.1 Analysis of Speech Metrics

Table 1 presents the acoustic measures and speech sample types, for which a significant difference between sample populations was observed ($p < 0.05$). For read speech, there is no significant difference for any of the metrics. For free speech, percentage of pause time and volume are indicators to distinguish groups. For sustained vowels, we observe significant differences for volume, mean fundamental frequency, harmonics-to-noise ratio and jitter. There are differences between females and males, as indicated in the table.

5.2 Analysis of Visual Metrics

Table 3 shows the visual metrics for which a significant difference between sample populations was observed ($p < 0.05$) for free and read speech. Visual metrics are only analyzed for free speech and read speech because only limited movement of facial muscles can be observed for automated speech and sustained vowels. For read speech only, a few metrics show significant differences independent of the cutpoint and gender, while the number of metrics for free speech depends on both cutpoint and gender. For males, the measures that involve the eyes, i.e. eye_open, eyebrow_vpos and eye_blinks, show significant differences independent of the employed cutpoint. In contrast, when considering all samples, independent of the reported gender, and females, the metrics for which significant differences are observed depend on the cutpoint and speech sample. Cutpoint 5 mostly includes eye, surface and duration measures, while cutpoint 10 also includes movement measures. For cutpoint 15, significant differences can be observed for the velocity of the lower lip and jaw center for both free and read speech, when considering all samples or females.

6 Conclusion and Outlook

We investigated whether various audio and video metrics extracted from audiovisual conversational data obtained through a cloud-based multimodal dialog system exhibit statistically significant differences between depressed and non-depressed populations. For several of the investigated metrics such differences were observed indicating that the employed audiovisual dialog system has a potential to be used for remote monitoring of depression. However, more detailed investigations on the nature of value distributions of metrics, their dependency on subject age or native language, the quality of input signals or used devices, among other studies, are necessary to to see to which degree the results are generalizable. Additionally, the used PHQ-scores were self-reported and might therefore be less accurate than scores obtained under the supervision of a clinician. In future work, we will also collect additional interactions from larger and more diverse populations. Furthermore, we will perform additional analysis on the obtained data, such as regression analysis. Finally, we will extend the set of investigated metrics and investigate their efficacy for other neurological or mental health conditions.

References

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, USA.

Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837.

R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. 2010. Multi-pie. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, volume 28, pages 807–813.

Varun Jain, James L Crowley, Anind K Dey, and Augustin Lux. 2014. Depression estimation using audiovisual features and fisher vector encoding. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 87–91.

Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces*, 7(3):217–228.

Heysem Kaya, Florian Eyben, Albert Ali Salah, and Björn Schuller. 2014. 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)*, pages 3729–3733. IEEE.

V. Kazemi and J. Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA.

Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Santosh Kumar, Wendy Nilsen, Misha Pavel, and Mani Srivastava. 2012. Mobile health: Revolutionizing healthcare through transdisciplinary research. *Computer*, 46(1):28–35.

Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. 2013. 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*, pages 21–30.

James C Mundt, Peter J Snyder, Michael S Cannizzaro, Kara Chappie, and Dayna S Geralts. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 20(1):50–64.

James C Mundt, Adam P Vogel, Douglas E Feltner, and William R Lenderking. 2012. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587.

Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 43–50.

Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, et al. 2016. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 27–34.

Tereza Soukupová and J. Čech. 2016. Real-time eye blink detection using facial landmarks. In *21stComputer Vision Winter Workshop*, Rimske Toplice, Slovenia.

R Steven and M Steinhubl. 2013. Can mobile health technologies transform health care. *JAMA*, 92037(1):1–2.

David Suendermann-Oeft, Amanda Robinson, Andrew Cornish, Doug Habberstad, David Pautler, Dirk Schnelle-Walka, Franziska Haller, Jackson Liscombe, Michael Neumann, Mike Merrill, et al. 2019. Nemsi: A multimodal dialog system for screening of neurological or mental conditions. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 245–247.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59.