Towards Multimodal Dialog-Based Speech & Facial Biomarkers of Schizophrenia

VANESSA RICHTER, Modality.AI, USA and University of Stuttgart, Germany

MICHAEL NEUMANN, HARDIK KOTHARE, OLIVER ROESLER, JACKSON LISCOMBE, and DAVID SUENDERMANN-OEFT, Modality.AI, USA

SEBASTIAN PROKOP, Nathan Kline Institute for Psychiatric Research & Manhattan Psychiatric Center, USA and St. George's University, School of Medicine, Canada

ANZALEE KHAN, Nathan Kline Institute for Psychiatric Research & Manhattan Psychiatric Center, USA and Valis Biosciences, USA

CHRISTIAN YAVORSKY, Valis Biosciences, USA

JEAN-PIERRE LINDENMAYER, Nathan Kline Institute for Psychiatric Research & Manhattan Psychiatric Center, USA

VIKRAM RAMANARAYANAN, Modality.AI, USA and University of California, USA

We present a scalable multimodal dialog platform for the remote digital assessment and monitoring of schizophrenia. Patients diagnosed with schizophrenia and healthy controls interacted with Tina, a virtual conversational agent, as she guided them through a brief set of structured tasks, while their speech and facial video was streamed in real-time to a back-end analytics module. Patients were concurrently assessed by trained raters on validated clinical scales. We find that multiple speech and facial biomarkers extracted from these data streams show significant differences (as measured by effect sizes) between patients and controls, and furthermore, machine learning models built on such features can classify patients and controls with high sensitivity and specificity. We further investigate, using correlation analysis between the extracted metrics and standardized clinical scales for the assessment of schizophrenia symptoms, how such speech and facial biomarkers can provide further insight into schizophrenia symptomatology.

CCS Concepts: • **Applied computing → Health care information systems**.

Additional Key Words and Phrases: multimodal dialog system, schizophrenia, biomarkers, computer vision, facial landmarks, facial metrics, mediapipe, speech metrics

ACM Reference Format:
Vanessa Richter, Michael Neumann, Hardik Kothare, Oliver Roesler, Jackson Liscombe, David Suendermann-Oeft, Sebastian Prokop, Anzalee Khan, Christian Yavorsky, Jean-Pierre Lindenmayer, and Vikram Ramanarayanan. 2022. Towards Multimodal Dialog-Based Speech & Facial Biomarkers of Schizophrenia. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536220.3558075

The authors would like to thank Theresa Abad, Shuaib Ahmad, Saqib Bashir, Brianna Fitapelli, Beverly Insel, Mohan Parak and Benedicto Parker for help in collecting and processing the data.

1 INTRODUCTION

Schizophrenia is a chronic mental disorder with heterogeneous presentations that affects around 60 million (1%) of the world's adult population. Per the American Psychiatric Association (APA), when schizophrenia is active, symptoms can include delusions, hallucinations, disorganized speech, trouble with thinking and lack of motivation[1]. These symptoms are broadly categorized as either positive, which are pathological functions not present in healthy individuals (e.g., hallucinations and delusions), negative, which involve the loss of functions or abilities (e.g., apathy, lack of pleasure, blunted affect and poor thinking); or cognitive/disorganized (deficits in attention, memory and executive functioning, confused and disordered thinking and speech, trouble with logical thinking and sometimes bizarre behavior or abnormal movements) [3]. Crucially, the causes of schizophrenia are complex and are not fully understood, per the National Institute of Mental Health (NIMH), so current treatments typically focus on managing symptoms and solving problems related to day to day functioning[2]. The development of novel biomarkers for assessment and monitoring of schizophrenia, particularly if available remotely and in a scalable manner, can greatly help treatment. Indeed, demonstrating utility and validity of remote and automated assessments conducted outside of controlled experimental or clinical settings can facilitate scaling such measurement tools to aid in risk assessment and tracking of treatment response [1].

Previous work has shown that speech acoustic, language and facial movements can serve as useful biomarkers of schizophrenia [25]. Indeed, multiple research papers have put forth acoustic and temporal analysis of speech as a potential tool for the objective analysis in schizophrenia [7, 9, 16, 23, 27, 30, 32], see also [17] for a systematic review. Several papers have demonstrated the utility of linguistic features – for example, sociolinguistic features, discourse coherence, syntactic complexity, poverty of content, referential coherence, and metaphorical language – as being characteristic of schizophrenia [8, 21]. Yet others have confirmed the utility of facial expressions and movement biomarkers for this purpose [1, 18, 22]. In this paper, we demonstrate that such speech, facial movement, and text features can be extracted from conversational interviews conducted with a scalable multimodal dialog platform, using participants' available end devices. We show that these features (a) are reliable and robust across sessions, (b) show statistically significant differences between patients and controls, and (c) can be used to classify patients with schizophrenia from healthy controls with high sensitivity and specificity. We further provide insights one can gain into schizophrenia symptomatology using correlation analyses between the aforementioned features and validated clinical scales.

2 SYSTEM & DATA

We use NEMSI (Neurological and Mental health Screening Instrument [31]), a cloud-based multimodal dialog system, to conduct automated structured interactions with study participants, who can start a conversation with a virtual agent, Tina, through a personalised web URL. At the beginning of the call, tests of the speaker, microphone, and camera need to be passed to ensure that the participants' devices are correctly configured so that the collected data has sufficient quality. Once all device tests pass, Tina guides participants through an interactive interview comprising several structured exercises that are in turn designed to elicit specific speech, facial, or motor behaviours [20, 26].

We assessed 24 patients (mean age 41, range 25 – 58) with a diagnosis of schizophrenia at a state psychiatric facility in New York, NY. We administered several clinical instruments for schizophrenia, including the standard positive and negative syndrome scale (PANSS) [14], the more recent brief negative symptom scale (BNSS) [15], Calgary depression scale for schizophrenia (CDSS) [2] and the clinical global impression severity scale (CGI-S) [13]. We also administered the Simpson Angus Scale (SAS) [29], Abnormal Involuntary Movement Scale (AIMS) [13], and Barnes Akathisia Rating

[1] https://www.psychiatry.org/patients-families/schizophrenia/what-is-schizophrenia
[2] https://www.nimh.nih.gov/health/topics/schizophrenia

Scale (BARS) [4] to assess and monitor abnormal involuntary movement. To assess reliability, the second visit occurred within one week and was done by the same clinician. We also collected data from 20 healthy controls (mean age 42, range 23 – 56). We obtained written informed consent from all participants at the time of screening after explaining details of the study. Both patients and controls had their assessment overseen by a psychiatrist. The study was approved by the Nathan S. Kline Institute for Psychiatric Research. The conversational flow elicited speech samples of the following types from participants: (a) sustained vowel phonation, (b) alternating motion rate diadochokinesis (DDK), (c) read speech, (d) a picture description task, and (e) spontaneous speech. For read speech, participants were asked to read speech intelligibility test (SIT) sentences and a reading passage (Bamboo Passage). For (e), participants were asked to speak about any topic of their choice with a few topic suggestions listed on the screen. Even though these tasks do not require a true dialog, having a virtual agent to elicit participants' behavior allows for scalability while providing a natural but objective interview environment and immersive user experience. In addition, the use of a dialog setting aims to increase user engagement in the tasks.

Table 1. Overview of speech, text and facial metrics. Please note that facial metrics are only roughly and exemplary described here. They are explained in Table 2 and Table 3 in more detail.

Domain	Metrics
Energy	shimmer (%), intensity (dB)
Timing	speaking and articulation duration (sec.), speaking and articulation rate (WPM), percent pause time (PPT, %)
Specific to DDK	cycle-to-cycle temporal variability (cTV, sec.), syllable rate (syl./sec.), number of syllables
Voice quality	cepstral peak prominence (CPP, dB), harmonics-to-noise ratio (HNR, dB)
Frequency	mean fundamental frequency (F0, Hz), first three formants (Hz), slope of 2nd formant (Hz/sec.), jitter (%)
Text	word count, percentage content words
Lower face	facial dynamics of jaw, lip and mouth (such as velocity of mouth opening/ closing)
Middle face	facial dynamics of nose and cheek (such as acceleration of cheek raising)
Upper face	facial dynamics of eye and eyebrow (such as local maxima of eyebrow height)

3 METHODS

3.1 Speech and text metrics

We extracted speech metrics using Praat [5] – see Table 1 for a complete list. These include timing measures such as percent pause time (PPT), articulation duration and speaking duration including pauses; frequency domain measures such as fundamental frequency (F0), jitter and formant frequencies (F1, F2, F3); energy-related measures like articulation intensity and shimmer; voice quality measures such as Harmonics-to-Noise Ratio (HNR) and Cepstral Peak Prominence (CPP). Speaking and articulation rates were computed by dividing the expected number of words for read sentences by the total time in seconds. We further computed simple text-based lexico-semantic features based on *automatic speech recognizer* (ASR) transcriptions of spontaneous speech utterances obtained using AWS Transcribe[3]. These features included simple word count and the percentage of content words, along with other lexico-semantic features such as noun rate, verb rate, and idea density, as described in [6][4].

[3] https://aws.amazon.com/transcribe/

[4] Note that the lexico-semantic features did not show significant signal in this study, possibly due to short duration of spontaneous speech samples.

3.2 Facial metrics

We evaluate a set of facial metrics based on landmarks generated in real-time using the mediapipe FaceMesh algorithm[5], allowing to obtain a fine-grained face mesh consisting of 468 3D facial landmarks. The obtained mesh is normalized in two steps. First, the landmarks are transformed to a common coordinate system by subtracting the center of mass of the face mesh of each of them. Second, the scale of the facial mesh is normalized by dividing each landmark by the intercanthal distance, which is the distance between the inner corners of the eyes. Out of these 468 normalized facial landmarks, two subsets of key landmarks are chosen for the calculation of dynamic features based on a) distances and b) angles. These dynamic features comprise motion measurements of different facial parts and are related to the facial action units coding system defined by Ekman [10]. While the features based on distances represent general facial movements, the ones based on angles are thought to be representative of emotional expressions [28]. The dynamics of these measures are calculated by capturing the frame-by-frame differences in the determined angular or distance values. The selected features that capture the movement dynamics are displayed in Table 2.

Table 2. Dynamic feature set extracted for both distance and angle-based facial actions. Adapted from Gomez et al. [12].

F_{1-5}	velocity: average absolute, standard deviation; ratios: RMS/ maximum abs., average abs./ max. (abs.)
F_{6-8}	acceleration: average absolute, standard deviation; ratio: RMS/ maximum
F_{9-13}	jerk: average (absolute), maximum (absolute), RMS
F_{14}	ratio: maximum/ minimum angle or distance measures (respectively)
F_{15}	ratio: sum of positive velocities/ sum of negative velocities
F_{16}	number of local maxima (i.e. number of distance/ angle peaks)

3.2.1 Distance-based metrics. Eight distances representing different facial actions are chosen, as shown in Table 3. The selection of landmarks to compute the distances is based on [12]. Distances are calculated for each normalized face mesh. The obtained distance values are assessed frame-by-frame by taking the average of a set of pairwise distances between the selected landmarks (for more detail and visualization, see Table 2 and Figure 3 in [12]). Combining the distance-based facial actions with the dynamic feature set results in 128 (8x16) distance-based metrics.

3.2.2 Geometric-based metrics. We calculate the geometric features defined for emotion detection on images by Siam et al. [28]. To fully capture the information contained in the video data, we enrich their features with the above-defined motion cues. The geometric features are based on 27 landmarks chosen as key vertices of the emotion face mesh. Locations are selected based on the probability that a landmark is affected by a specific emotion-related action unit. Key vertices are combined to 38 edges. Facial motions are captured by the dynamics of angles between edges. In sum, 10 geometric facial actions are evaluated in terms of the dynamics shown in Table 2 which results in 160 (10x16) metrics for the dynamic, geometry-based features.

4 ANALYSES & OBSERVATIONS

Metrics should be robust and as independent of daily performance as possible. We consider a high correlation between sessions within a study participant as an important indication of reliability. Please note that it is a requirement here that sessions are conducted at close time intervals, as symptoms may change over time. In contrast to rather subjective and time-consuming assessments like BNSS or PANSS, our approach allows for an objective and time-efficient evaluation of

[5] https://google.github.io/mediapipe/

Table 3. Facial parts taken into account for geometric- and distance-based features.

Part of face	Distance-based [12]	Geometric/ angles between key vertices of... [28]
Upper face	Opening of the left eye Opening of the right eye Right eyebrow height Left eyebrow height	...outer/ middle/ inner eyebrow (right) ...upper eyelid/ lower eyelid/ outer eye (right) ...nose end/ nose bridge/ inner eyebrow (right)
Middle face		...nose end (right)/ nose tip/ nose end (left) ...upper jaw (right)/ nose tip/ upper jaw (left) ...mouth end/ cheek/ nose end (left) ...mouth end/ upper jaw/ nose end (left) ...mouth end (right)/ nose tip/ mouth end (left)
Lower face	Mouth width Mouth opening Lip stretch Jaw opening	...upper lip (middle)/ mouth end (right)/ lower lip (middle) ...upper lip (middle)/ mouth end (right)/ upper lip (middle)/ mouth end (left)

a person's behavior. While the inter-rater agreement for the BNSS scores is very high (0.94), it is considerably lower for the PANSS scores (0.73). We are able to achieve higher test-retest-reliability (measured as Spearman correlation within participants between sessions) for a significant subset of the tested metrics relative to the inter-rater agreement of the PANSS scores. Test-retest-reliabilities are shown in brackets behind the evaluated metrics in Figure 1.

Next, to test the hypothesis that the medians of our extracted features differed significantly between schizophrenia patients and healthy controls at the $\alpha = 0.05$ level, we conducted non-parametric Kruskal-Wallis tests [19]. Figure 1 shows the effect sizes of metrics found to be significant. We find that patients show reduced speaking rates (and therefore greater durations) and different voice quality (higher HNR and CPP), as compared to healthy controls. In terms of facial dynamics, patients show lower average velocity, acceleration and jerk for various action units across all facial parts and also less variety of these measurements shown by lower standard deviations. In addition, we observe a higher number of local maxima in patients' movements for both distance- and geometric-based metrics, meaning that patients demonstrate flatter and less intense facial movements. We observe that features obtained from the picture description task, the reading passage, and spontaneous speech predominantly show good signal. All linguistic features had a test-retest reliability below 0.6 and were thus not included in the analysis. A potential reason for this might be errors in the ASR output that affect the quality of the features.

To further test the efficacy and generalizability of speech and facial features in distinguishing patients and controls, we ran 5-fold classification experiments using a random forest classifier. For this experiment, features were mean-aggregated across user turns to reduce the number of features and smoothen out some variability – we found that this yields better and more stable results. We ran each classification experiment ten times with different random seeds to be able to report variations in the results. Our results show that each modality performs well by itself. Speech metrics (mean AUC: 0.84 ± 0.02) perform better than facial metrics (mean AUC: 0.75 ± 0.04) in classifying patients versus controls. The performance was also evaluated by taking all metrics independent of statistical significance (mean AUC: 0.76 ± 0.02) into account compared to choosing only those metrics that showed a significant effect in the effect size plot (mean AUC: 0.84 ± 0.02). To further ensure validity and robustness, we run classification with only those metrics (speech + facial) that demonstrated a moderate to high (> 0.6) test-retest-reliability across sessions in addition to statistically significance. We achieve best results (mean AUC: 0.86 ± 0.02) with these specifications. Classification results for the

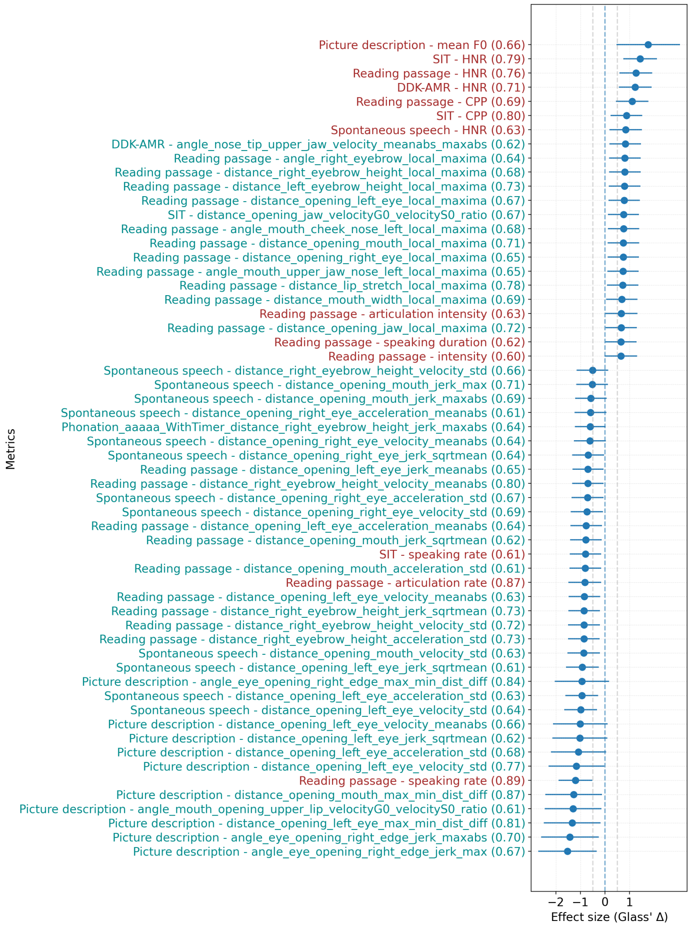

Fig. 1. Effect sizes of **speech** and **facial** metrics with a moderate to high test-retest-reliability (> 0.6) that show statistically significant differences between controls and patients at $\alpha = 0.05$. Error bars show the 95% confidence interval. Positive values indicate features where patients had higher values than controls.

best run are shown in Figure 2. The comparison shows that a combination of all modalities and taking only statistically significant metrics with high test-retest-reliability improves the performance further compared to evaluating each modality alone. To dig deeper into the extent to which speech and facial metrics capture characteristics of schizophrenia symptomatology, we evaluate if and how our features correlate with patients' negative symptoms. For the analysis, we chose only metrics that show a moderate to high test-retest-reliability (> 0.6) and evaluated their correlation with a subset of the BNSS scores representing lack of normal distress, anhedonia (inability to feel pleasure), asociality (lack of motivation or ability in social interaction), avolition (loss of ability to do or experience things), blunted affect, and alogia (poverty of speech). For facial metrics, we find high (> 0.8) negative correlations between a large number of negative symptoms and facial actions capturing motion of the angles involving the nose for the picture description task (e.g. −0.92 between avolition and the standard deviation of acceleration for the angle between the nose tip and the lips). Correlations are weaker between BNSS scores and tasks such as DDK. This might be due to the nature of standardized

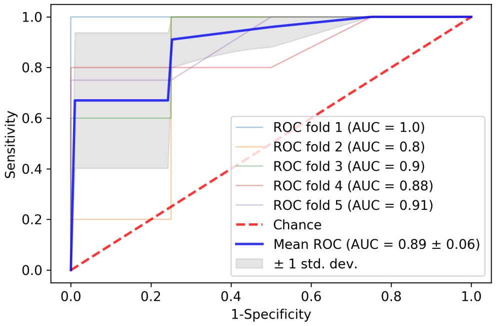

Fig. 2. ROC curve of combined statistically significant speech and facial metrics with moderate to high test-retest-reliability in classifying patients vs. controls.

tasks versus those that allow for spontaneous facial expressions and therefore may better reveal motivational and emotional behavior. Regarding speech metrics, we find, among others, moderate correlations of blunted affect in vocal expression (−0.5), lack of normal distress (−0.5) and total scores (−0.54) with speaking rate in the reading task.

5 DISCUSSION

Our results show that multiple speech and facial biomarkers extracted during the course of a multimodal dialog interaction can capture significant speech and oromotor differences between schizophrenia patients and healthy controls with high sensitivity and specificity. While each modality alone is able to achieve decent results in classifying patients and controls, combining the statistically significant metrics across modalities further enhances performance and showcases the benefits of our multimodal approach. For both speech and facial modalities, our findings are supported by clinical studies [11, 24]. Pueschel et al. found that speech behavior and sound characteristics of patients and controls differed significantly as patients tended to make longer pauses, speak more monotonously in a low voice, and shift their vocal pitch toward higher frequencies. It was also found that speech disorders persisted at hospital release, even though "acute psychopathology had significantly improved", which supports the utility of scalable health assessments at home. The clinical FACS-based assessments of [11] show, in line with our findings, reduced facial expressivity in terms of both quantitative and qualitative aspects in patients. It is also worth noting that while the standard clinician rating scales for schizophrenia take around 1.5 hours to administer (both PANSS and BNSS), the interaction with Tina is much shorter, ∼10 minutes, which is a significant advantage for scalability and adoptability. In addition, the evaluation of the metrics' test-retest-reliability shows that our multimodal system is able to deliver objective and robust metrics, which outperform the inter-rater correlation of PANSS assessments. While this implies several potential benefits for automated monitoring and assessment of schizophrenia symptoms, there are several caveats and areas for improvement to address in future research. For instance, our results are based on a limited sample size, and measurements of participants at a single point in time, which can in turn affect the robustness of statistics. Additionally, while correlation analysis revealed interesting patterns between select biomarkers and symptoms, correlation does not imply causation, and so we need to unpack and understand these findings in more detail, and explore their robustness, explainability, and generalizability. Longitudinal studies that track participant biomarkers over time will allow us to better understand and model time- and symptom-specific variabilities across patients.

REFERENCES

[1] Anzar Abbas, Bryan J Hansen, Vidya Koesmahargyo, Vijay Yadav, Paul J Rosenfield, Omkar Patil, Marissa F Dockendorf, Matthew Moyer, Lisa A Shipley, M Mercedez Perez-Rodriguez, et al. 2022. Facial and Vocal Markers of Schizophrenia Measured Using Remote Smartphone Assessments: Observational Study. *JMIR Formative Research* 6, 1 (2022), e26276.

[2] Donald Addington, Jean Addington, and B Schissel. 2000. Calgary Depression Scale for Schizophrenia (CDSS). *American Psychiatric Association. Task Force for the Handbook of Psychiatric Measures. American Psychiatric Association. Washington DC* (2000), 504–507.

[3] Nancy C Andreasen and Scott Olsen. 1982. Negative v positive schizophrenia: Definition and validation. *Archives of general psychiatry* 39, 7 (1982), 789–794.

[4] Thomas RE Barnes. 1989. A rating scale for drug-induced akathisia. *The British Journal of Psychiatry* 154, 5 (1989), 672–676.

[5] Paul Boersma and Vincent Van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International* 5, 9/10 (2001), 341–347.

[6] Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology* 8 (2017), 269.

[7] Debsubhra Chakraborty, Zixu Yang, Yasir Tahir, Tomasz Maszczyk, Justin Dauwels, Nadia Thalmann, Jianmin Zheng, Yogeswary Maniam, Nur Amirah, Bhing Leet Tan, et al. 2018. Prediction of negative symptoms of schizophrenia from emotion related low-level speech signals. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6024–6028.

[8] Cheryl M Corcoran, Vijay A Mittal, Carrie E Bearden, Raquel E Gur, Kasia Hitczenko, Zarina Bilgrami, Aleksandar Savic, Guillermo A Cecchi, and Phillip Wolff. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia research* 226 (2020), 158–166.

[9] Michael A Covington, SL Anya Lunden, Sarah L Cristofaro, Claire Ramsay Wan, C Thomas Bailey, Beth Broussard, Robert Fogarty, Stephanie Johnson, Shayi Zhang, and Michael T Compton. 2012. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia research* 142, 1-3 (2012), 93–95.

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

[11] Wolfgang Gaebel and Wolfgang Woelwer. 2004. Facial expression in the course of schizophrenia and depression. *European archives of psychiatry and clinical neuroscience* 254 (11 2004), 335–42. https://doi.org/10.1007/s00406-004-0510-5

[12] Luis F. Gomez, Aythami Morales, Juan R. Orozco-Arroyave, Roberto Daza, and Julian Fierrez. 2021. Improving Parkinson Detection using Dynamic Features from Evoked Expressions in Video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1562–1570. https://doi.org/10.1109/CVPRW53098.2021.00172

[13] William Guy. 1976. *ECDEU assessment manual for psychopharmacology*. US Department of Health, Education, and Welfare, Public Health Service ....

[14] Stanley R Kay, Abraham Fiszbein, and Lewis A Opler. 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia bulletin* 13, 2 (1987), 261–276.

[15] Brian Kirkpatrick, Gregory P Strauss, Linh Nguyen, Bernard A Fischer, David G Daniel, Angel Cienfuegos, and Stephen R Marder. 2011. The brief negative symptom scale: psychometric properties. *Schizophrenia bulletin* 37, 2 (2011), 300–305.

[16] Rony Krell, Wenqing Tang, Katrin Hänsel, Michael Sobolev, Sunghye Cho, Sarah Berretta, and Sunny X Tang. 2021. Lexical and acoustic correlates of clinical speech disturbance in schizophrenia. In *International Workshop on Health Intelligence*. Springer, 27–35.

[17] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology* 5, 1 (2020), 96–116.

[18] Manas K Mandal, Rakesh Pandey, and Akhouri B Prasad. 1998. Facial expressions of emotions and schizophrenia: a review. *Schizophrenia bulletin* 24, 3 (1998), 399–412.

[19] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.

[20] Michael Neumann, Oliver Roesler, Jackson Liscombe, Hardik Kothare, David Suendermann-Oeft, David Pautler, Indu Navar, Aria Anvar, Jochen Kumm, Raquel Norel, Ernest Fraenkel, Alex Sherman, James Berry, Gary Pattee, Jun Wang, Jordan Green, and Vikram Ramanarayanan. 2021. Investigating the Utility of Multimodal Conversational Technology and Audiovisual Analytic Measures for the Assessment and Monitoring of Amyotrophic Lateral Sclerosis at Scale. Brno, Czech Republic, 4783–4787. https://doi.org/10.21437/Interspeech.2021-1801

[21] Lena Palaniyappan. 2021. More than a biomarker: could language be a biosocial marker of psychosis? *npj Schizophrenia* 7, 1 (2021), 1–5.

[22] Alberto Parola, Ilaria Gabbatore, Laura Berardinelli, Rogerio Salvini, and Francesca M Bosco. 2021. Multimodal assessment of communicative-pragmatic features in schizophrenia: a machine learning approach. *NPJ schizophrenia* 7, 1 (2021), 1–9.

[23] Alberto Parola, Arndis Simonsen, Vibeke Bliksted, and Riccardo Fusaroli. 2020. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizophrenia research* 216 (2020), 24–40.

[24] Joerg Pueschel, Hans Stassen, G. Bomben, Christian Scharfetter, and Daniel Hell. 1998. Speaking behavior and speech sound characteristics in acute schizophrenia. *Journal of psychiatric research* 32 (03 1998), 89–97. https://doi.org/10.1016/S0022-3956(98)00046-6

[25] Vikram Ramanarayanan, Adam C Lammert, Hannah P Rowe, Thomas F Quatieri, and Jordan R Green. 2022. Speech as a Biomarker: Opportunities, Interpretability, and Challenges. *Perspectives of the ASHA Special Interest Groups* (2022), 1–8.

[26] Vikram Ramanarayanan, Oliver Roesler, Michael Neumann, David Pautler, Doug Habberstad, Andrew Cornish, Hardik Kothare, Vignesh Murali, Jackson Liscombe, Dirk Schnelle-Walka, et al. 2020. Toward Remote Patient Monitoring of Speech, Video, Cognitive and Respiratory Biomarkers Using Multimodal Dialog Technology.. In *INTERSPEECH*. 492–493.

[27] Viliam Rapcan, Shona D'Arcy, Sherlyn Yeap, Natasha Afzal, Jogin Thakore, and Richard B Reilly. 2010. Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical engineering & physics* 32, 9 (2010), 1074–1079.

[28] Ali Siam, Naglaa Soliman, Abeer Algarni, Fathi Abd El-Samie, and Ahmed Sedik. 2022. Deploying Machine Learning Techniques for Human Emotion Detection. *Computational Intelligence and Neuroscience* 2022 (02 2022). https://doi.org/10.1155/2022/8032673

[29] GM Simpson and JWS Angus. 1970. A rating scale for extrapyramidal side effects. *Acta Psychiatrica Scandinavica* 45, S212 (1970), 11–19.

[30] Yashish M Siriwardena, Carol Espy-Wilson, Chris Kitchen, and Deanna L Kelly. 2021. Multimodal Approach for Assessing Neuromotor Coordination in Schizophrenia Using Convolutional Neural Networks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 768–772.

[31] David Suendermann-Oeft, Amanda Robinson, Andrew Cornish, Doug Habberstad, David Pautler, Dirk Schnelle-Walka, Franziska Haller, Jackson Liscombe, Michael Neumann, Mike Merrill, Oliver Roesler, and Renko Geffarth. 2019. NEMSI: A Multimodal Dialog System for Screening of Neurological or Mental Conditions. In *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA)*. Paris, France.

[32] Eric J Tan, Denny Meyer, Erica Neill, and Susan L Rossell. 2021. Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations. *Schizophrenia research* 238 (2021), 91–98.