Multimodal Speech, Language and Orofacial Analysis for Remote Assessment of Positive, Negative and Cognitive Symptoms in Schizophrenia

Michael Neumann[1], *Hardik Kothare*[1], *Beverly Insel*[2], *Anzalee Khan*[2], *Danyah Nadim*[2], *Jean-Pierre Lindenmayer*[2], *Vikram Ramanarayanan*[1,3]

[1]Modality.AI, Inc., San Francisco, CA, USA; [2]Nathan Kline Institute for Psychiatric Research at Manhattan Psychiatric Center, New York, NY, USA; [3]University of California, San Francisco, San Francisco, CA, USA

michael.neumann@modality.ai

Abstract

This study provides a comprehensive analysis of digital speech, orofacial and linguistic features for the assessment of schizophrenia. We recorded audio and video from 94 people with schizophrenia (pSCZ) and 100 healthy controls (HC) and extracted features automatically. Clinical rating scales were administered to assess positive, negative, and cognitive symptoms. We show that pSCZ exhibit significant alterations in speech timing, orofacial dynamics, and lexical richness, as compared to HC. A multimodal classification approach achieved high accuracy (96% AUC, 87% UAR), with speech features contributing most to discrimination. Correlation analysis revealed that speech timing and lip velocity measures are correlated with blunted affect and alogia. Linguistic features correlate well with positive symptoms, particularly conceptual disorganization and excitement. Cognitive abilities are most strongly associated with speech timing and specific linguistic features.

Index Terms: speech biomarkers, multi-modal dialog, remote monitoring, clinical trials, schizophrenia

1. Introduction

Speech and language characteristics are well-known to be informative markers for mental disorders such as depression or schizophrenia spectrum disorder (SSD) [1, 2, 3]. People with schizophrenia (pSCZ) commonly exhibit reduced word and sentence production, slower articulation rate, impaired processing of complex syntax, and deficits with semantic verbal fluency [4, 5]. In addition to speech and language deficits, changes in orofacial characteristics can be observed, such as reduced expressiveness (blunted affect) and aberrant movements [6, 7].

Beyond studying the characteristics that differ between pSCZ and healthy controls (HC), there is a high need to identify markers relevant for specific symptoms due to the highly heterogeneous presentation of symptoms in SSD, which are commonly categorized into positive, negative, and cognitive symptoms [8, 9]. The presence of these symptoms and the interaction between them varies across individuals, and the lack of objective, reliable biomarkers makes it challenging to assess and monitor the course of the disease.

Assessments of speech characteristics and facial expression are a promising direction towards objective measurements, which can complement established clinical rating scales. One challenge for the wide-spread adoption and clinical employment of such evaluations is the scalability and automation using digital technologies. Therefore, we aim to provide supportive evidence for the feasibility and clinical utility of fully automated speech assessments using a web-based multimodal dialog system.

Several studies have investigated speech and language assessments in SSD, often focusing on a single modality, such as linguistic characteristics or acoustic speech markers. Acoustic analyses have revealed significant differences between pSCZ and HC, with features like pitch, jitter, shimmer, and voice breaks showing statistical differences [10]. Machine learning approaches have been applied to classify SSD based on acoustic features, with studies reporting classification accuracies ranging from 70-80% using temporal speech fluency features [11] to 86.2% using openSMILE's eGeMAPS feature set [12]. Some studies have further linked acoustic features to symptom severity, showing weak to moderate correlations with negative [13, 12] and positive [14] symptoms of SSD. Linguistic features have also been explored. Recent work has employed deep learning approaches to integrate linguistic and acoustic information [15], though these methods suffer from limited interpretability, which is crucial in clinical applications. Instead of focusing mainly on diagnosis, Martin et al. have recently proposed a paradigm shift towards estimating clinical signs and symptom severity [16].

A common weakness of many previous studies on SSD is the small sample size of sometimes only 10 to 30 subjects [11, 17, 15]. Furthermore, most studies focused on a single modality and often on rather constrained, very specific feature sets (e.g., MFCC features in [13] or voice quality features in [10]). *In contrast, this contribution aims to provide a comprehensive, multimodal analysis of a broad set of speech, language, and orofacial characteristics, computed based on audiovisual recordings from 94 pSCZ and 100 HC.* We investigate the relationship of these features with all three symptom domains, positive, negative and cognitive symptoms. Specifically, this study addresses the following research question: *what is the relative clinical utility of different modalities for the assessment of schizophrenia, both with respect to discriminability between pSCZ and HC, and in characterizing positive, negative and cognitive symptoms associated with the disorder?*

2. Data Collection and Dataset

Audiovisual data was collected using the Modality platform [18, 19], a web-based multimodal dialog system for self-guided clinical assessments. A virtual guide led study participants through a set of tasks designed to elicit certain speech and facial behaviors, including read speech (Bamboo reading passage, 99 words), sentence intelligibility test (SIT), sustained vowel phonation, diadochokinesis (DDK), and an open question to elicit spontaneous speech. The language of this data collection is English. Acoustic speech and orofacial video features were extracted for all assessment tasks. For linguistic features, we focused the spontaneous speech task where partici-

Table 1: *Participant statistics. Age and clinical ratings are reported as* mean (standard deviation). *BNSS Total ranges from 0 to 78, PANSS Negative and Positive Total range from 7 to 49 each, and PANSS Total ranges from 30 to 210. f: female, m: male.*

	Number of participants	# sessions	Age	BNSS Total	PANSS Neg.	PANSS Pos.	PANSS Total
pSCZ	94 (22 f, 71 m, 1 unspecified)	180	42.6 (11.6)	37.8 (10.2)	23.1 (3.8)	15.6 (4.7)	73.2 (13.3)
HC	100 (49 f, 51 m)	195	36.4 (11.4)	-	-	-	-

pants were asked to talk about a topic of their choice. Recordings from 94 pSCZ and 100 HC were collected in cooperation with the Nathan Kline Institute, a total of 375 sessions (see Table 1). Patients were inpatients with a DSM 5 diagnosis of schizophrenia and the digital assessments were done in clinic under supervision. Clinician ratings were available for the brief negative symptom scale (BNSS, [20]), the positive and negative syndrome scale (PANSS, [21]), and the brief assessment of cognition in Schizophrenia (BACS, [22]).[1] The BNSS and PANSS are measures of symptom severity, with higher scores indicating more severe symptoms. The BACS item scores were converted into z-scores based on healthy norms published in [23], and a composite BACS score was derived by calculating the mean of the six items' z-scores.

3. Methods

3.1. Multimodal Feature Extraction

We extracted acoustic, language and facial video features fully automatically using digital signal processing, automatic speech recognition, and computer vision techniques. The speech and facial video features have been described in detail in previous work [24], whereas the linguistic and graph-based language features are a novel addition to our multimodal dialog system and will be described in more detail in the following. Table 2 provides an overview of all features. We employed a distribution-based algorithm to remove speech and facial feature outliers, which may result from background noise, poor lighting, or task errors. First, extreme outliers beyond five standard deviations from the mean were removed, as they could skew the distribution. This threshold was chosen empirically after analyzing data distributions. The mean was then recomputed, and values outside ± 3 standard deviations were flagged and excluded from further analysis. For the spontaneous speech task, samples with less than 10 words in the automatic transcription were removed. This threshold was set upon inspection of the data; such samples were identified as incorrect task performance.

3.1.1. Acoustic Speech Features

We used Praat (v6.2.17) [25] to extract speech features, including timing measures, such as percent pause time (PPT), speaking duration (including pauses), and articulation duration (excluding pauses), frequency-related measures, such as fundamental frequency (F0), energy-related measures, such as shimmer, and voice quality measures, such as cepstral peak prominence (CPP). We also computed Canonical Timing Alignment (CTA), a measure of the alignment of word and silence boundaries between the participant's speech and a canonical speech production of the same text [26].

3.1.2. Orofacial Video Features

These features are based on facial landmarks generated with MediaPipe Face Mesh [27]. First, MediaPipe Face Detection, based on BlazeFace [28], is used to determine the (x, y)-coordinates of the face for every video frame. Then, facial landmarks are extracted using MediaPipe Face Mesh. We used 14 key landmarks to compute features like kinematics of the articulators (jaw, lower lip), surface area of the mouth, and eyebrow raises. These landmarks include center and corners of the lips, jaw center, nose tip, center and corners of the eyes, and the center of the eyebrows. Lastly, the features were normalized by dividing them by the inter-caruncular or inter-canthal distance, to handle variability across participant sessions due to position and movement relative to the camera [29].

3.1.3. Linguistic and Graph-Based Language Features

The choice of linguistic features was motivated by the review article by Boschi et al. [30]. Speech samples were automatically transcribed using AWS Transcribe.[3] We computed linguistic features from the transcriptions using the Python package spaCy, version 3.5.3. In addition to more traditional linguistic features, we explored graph-based language features, which offer a different means of analyzing the structure of text. These features are based on graph-theoretical methods and have been shown to capture aspects of normal and dysfunctional flow of thought [31, 32].[4] Transcriptions were transformed into two kinds of graphs: a naive graph, where every distinct word (type) in the text is a node, and a part-of-speech (POS) graph, where every distinct POS is a node. Nodes are connected by an edge if they correspond to consecutive words. We followed [32] and computed ten features for each graph, see Table 2.

3.2. Clinical Validation

To identify features that show statistically significant differences between pSCZ and HC, we applied non-parametric Kruskal-Wallis tests, using data from every participant's first recording session only (to avoid effects of repeated measurements). We report effect sizes in terms of Glass' delta, which is based on the control group's standard deviation as opposed to a pooled standard deviation. Additionally, we evaluated a logistic regression model with L1 regularization in a nested 5-fold cross-validation (CV) setup to assess classification performance and generalizability to unseen data. The regularization parameter C was tuned on each training fold using 5-fold inner CV.

Correlations between all extracted features and clinical scales were assessed with Spearman's rank correlation coefficients, also based on participants' baseline sessions. For this analysis, we focused on specific symptoms, for which changes in speech, language usage, and/or orofacial expressiveness can be expected, namely BNSS Alogia, Blunted Affect, Distress, PANSS N1 (Blunted Affect), N6 (Lack of spontaneity and flow of conversation), P2 (Conceptual disorganization), P4 (Excitement), G7 (Motor Retardation), and G9 (Unusual thought content). For cognitive ratings, correlations were assessed for the composite BACS score and for the six individual z-scores.

[1] BACS scores were only available for a subset of 40 pSCZ.

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

[4] https://github.com/guillermodoghel/speechgraph

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

	Domain	Extracted Metrics
Audio	Energy	shimmer (%), signal-to-noise ratio (SNR, dB)
	Timing	speaking and articulation duration (sec.), percent pause time (PPT, %), canonical timing alignment (CTA, %), 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 and standard deviation of the fundamental frequency F0 (Hz), first three formants F1, F2, F3 (Hz), jitter (%)
Text	Lexico-semantic	noun rate, verb rate, demonstrative rate, pronoun rate, adjective rate, adverb rate, conjunction rate, possessive rate, noun-pronoun ratio, noun-verb ratio, closed-class word rate, open-class word rate, percentage content words, light verb rate, idea density, number of repetitions, Honore's statistic, Brunet's index, type-token ratio, average word length
	Morphosyntactic	inflected verb rate, gerund rate
	Discourse-Pragmatic	word count, number of subjects, number of objects, number of places, number of actions
	Syntactic	average dependency tree height
	Sentiment	Empath[2] positive and negative cosine similarity
	Graph based features	No. of nodes, No. of edges, No. of parallel edges (PE), average degree of the graph, standard deviation of the average degree, No. of nodes in the largest connected component (LCC) and the largest strongly connected component (LSC), No. of self-loops (L1), loops with two nodes (L2), and loops with three nodes (L3)
Video	Mouth (distances)	lip aperture/opening, lip width, mouth surface area,
		mean symmetry ratio between left and right half of the mouth
	Lip/Jaw Movement	speed of the lower lip and jaw center
	Eyes	eye opening, vertical displacement of the eyebrows

For all aforementioned analyses, features were standardized (z-scored) for females and males separately, to account for sex-specific differences in certain feature domains (e.g., frequency-related and voice quality).

4. Results

4.1. Clinical Utility: Statistical Tests and Classification

For the Kruskal-Wallis tests, we report findings only for features that are significantly different between pSCZ and HC ($p < 0.05$) and show at least a moderate effect size of 0.5 (in absolute terms). We observed large effect sizes for timing related speech features. DDK syllable rate, speaking duration for spontaneous speech, and CTA for the reading passage are reduced in pSCZ, whereas percent pause time and speaking duration for the reading task are increased. SNR and CPP are on average higher in the pSCZ group for multiple tasks. Among the orofacial measures, two features show significant differences: average eye opening (greater in pSCZ) and average lip width (smaller in pSCZ). Linguistic features, which are all computed from the spontaneous speech task, indicate overall a smaller amount of speech for pSCZ (reduced word count, lower number of nodes and edges in speech graphs). Furthermore, the number of repetitions is increased, and a decreased Brunet's index indicates less lexical richness in the speech samples of pSCZ.

The results of the cross-validation classification experiment are shown in Table 3, in terms of area under the ROC curve (ROC-AUC) and unweighted average recall (UAR). While each modality alone yields high accuracy, the multimodal combination of all features performs best with 96% AUC and 87% UAR. Speech acoustic features appear to contribute the most information, achieving results close to the best model. To gain insights about the most informative features for this task, we ranked features by their mean coefficient across all validation folds. Figure 1 shows the top ten features' coefficients and standard deviation across folds. The most relevant feature is CTA for the reading passage, a measure of timing alignment and intelligibility, followed by SNR for spontaneous speech. Orofacial and linguistic features are also among the top ten, underscoring the beneficial effect of a multimodal assessment approach.

Table 3: *Results for binary classification of pSCZ and HC.*

Modality	ROC-AUC	UAR
Audio	0.94	**0.87**
Video	0.84	0.75
Text	0.79	0.74
Multimodal	**0.96**	**0.87**

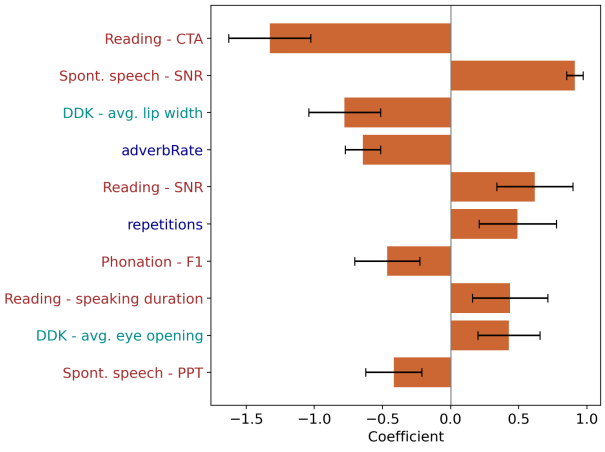

Figure 1: *The ten most useful features ranked by their absolute coefficient value, including the mean and standard deviation (error bars) of coefficients over all cross-validation folds.*

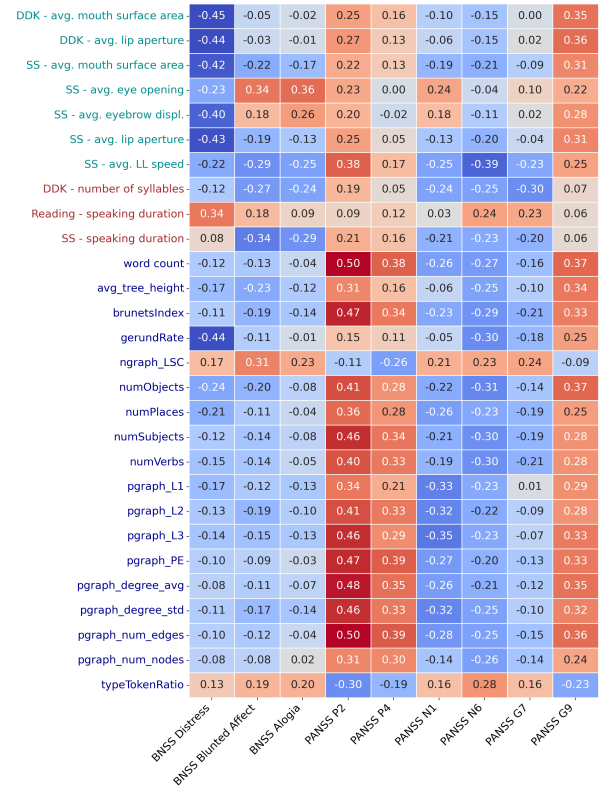

Figure 2: *Correlations between orofacial, speech and linguistic features and clinical ratings of specific symptoms.*

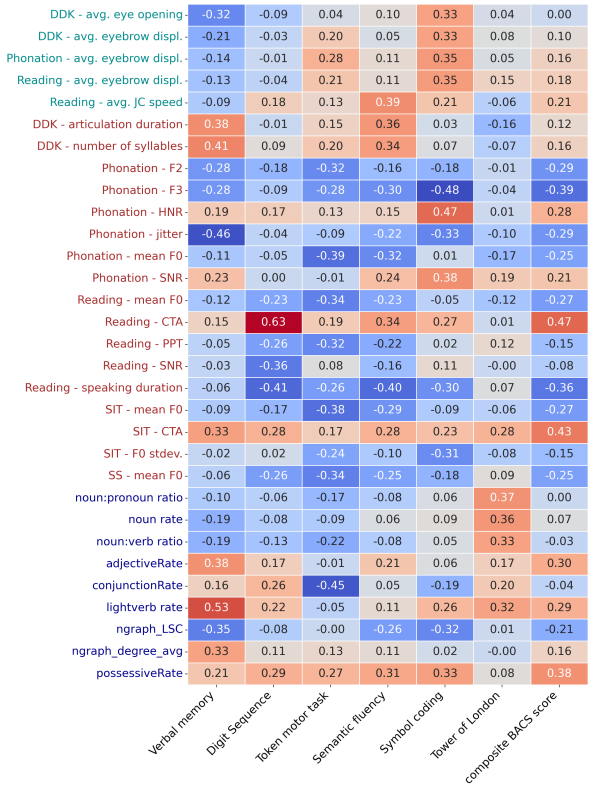

Figure 3: *Correlations between orofacial, speech and linguistic features and the BACS cognitive ratings.*

4.2. Correlation Analysis

Figure 2 shows correlations between features and selected symptom scores of the BNSS and PANSS, and Figure 3 shows correlations for cognitive scores of the BACS assessment. Only features are shown for which at least one correlation with a clinical rating scale item is significant ($p < 0.05$) and has an absolute correlation coefficient of at least 0.3. For blunted affect (BNSS Blunted Affect and PANSS N1) and alogia we observe moderate negative correlations with timing and velocity related measures (number of DDK syllables, speaking duration, lower lip speed) and positive correlations with average eye opening and the largest strongly connected component in the word graph (LSC). Lack of normal distress (BNSS Distress) is moderately correlated with many facial features, potentially indicating reduced facial expressiveness overall. The PANSS N6 score about flow of conversation shows weak to moderate correlations with speech and linguistic features. We observe that many linguistic features are well correlated with the positive symptoms PANSS P2 (Conceptual disorganization) and P4 (Excitement). For the PANSS G7 score on motor retardation, we observe weak to moderate negative correlations with average lower lip speed and the number of DDK syllables, and the G9 score (Unusual thought content) shows moderate positive correlations with most linguistic and orofacial features.

For the BACS composite score we observe the highest correlations with CTA, possessive rate, and certain voice quality features for the phonation task. Looking at individual BACS item scores, two features stand out with correlation coefficients greater than 0.5, namely CTA (reading passage), which is strongly correlated with the digit sequence score, and light verb rate, which is correlated with the verbal memory score.

5. Discussion & Conclusions

We presented a comprehensive, multimodal examination of a broad set of interpretable speech, language, and orofacial measures computed from remote audiovisual recordings from 94 pSCZ and 100 HC, and demonstrated their complementary clinical utility at characterizing various aspects of SSD, including positive, negative and cognitive symptoms observed therein. Such an analysis has important implications supporting the use of such multimodal clinical analytics for remote assessment and monitoring in clinical trials and care management.

While several interpretable features intuitively demonstrate promise in characterizing positive (language graph features), negative (facial features) and cognitive (CTA, lexico-semantic features) symptoms, several others are counterintuitive discoveries. For instance, we observe that SNR and CPP are on average higher in the pSCZ group for multiple tasks, indicating clearer, possibly louder speech with better voice quality (less aperiodicity) than in the HC group, which is unexpected. It is noteworthy that the PANSS N1 score shows higher correlations with pgraph features than the BNSS blunted affect score, although both scales measure blunted affect. In addition to negative and positive symptoms, we selected the PANSS G7 item on motor retardation for this analysis because we hypothesized that orofacial features provide potential markers to assess slowed movement and reduced activity levels. A weak, but significant correlation with average lower lip speed suggests such potential. Future work is required to investigate whether these findings are indeed robust. Additionally, these findings lay the groundwork toward the development of more sensitive composite multimodal index scores that can predict and assess individual symptoms with applications towards schizophrenia clinical trials and patient assessment.

6. References

[1] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a Biomarker: Opportunities, Interpretability, and Challenges," *Perspectives of the ASHA Special Interest Groups*, vol. 7, no. 1, pp. 276–283, 2022.

[2] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, "Quantifying Articulatory Impairments in Neurodegenerative Motor Diseases: A Scoping Review and Meta-Analysis of Interpretable Acoustic Features," *International Journal of Speech-Language Pathology*, pp. 1–14, 2022.

[3] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105020, 2023.

[4] F. Ehlen, C. Montag, K. Leopold, and A. Heinz, "Linguistic findings in persons with schizophrenia—a review of the current literature," *Frontiers in Psychology*, vol. 14, p. 1287706, 2023.

[5] J. De Boer, M. Van Hoogdalem, R. Mandl, J. Brummelman, A. Voppel, M. Begemann, E. Van Dellen, F. Wijnen, and I. Sommer, "Language in schizophrenia: relation with diagnosis, symptomatology and white matter tracts," *npj Schizophrenia*, vol. 6, no. 1, p. 10, 2020.

[6] F. Trémeau, D. Malaspina, F. Duval, H. Corrêa, M. Hager-Budny, L. Coin-Bariou, J.-P. Macher, and J. M. Gorman, "Facial expressiveness in patients with schizophrenia compared to depressed patients and nonpatient comparison subjects," *American Journal of Psychiatry*, vol. 162, no. 1, pp. 92–101, 2005.

[7] S.-M. Wang, W.-C. Ouyang, H.-M. Hsu, and L.-T. Hsu, "An instrumental measure of hand and facial movement abnormalities in patients with schizophrenia," *Frontiers in psychiatry*, vol. 13, p. 803661, 2022.

[8] N. C. Andreasen and W. M. Grove, "Evaluation of positive and negative symptoms in schizophrenia," *Psychiatry and Psychobiology*, vol. 1, no. 2, pp. 108–122, 1986.

[9] P. D. Harvey, D. Koren, A. Reichenberg, and C. R. Bowie, "Negative symptoms and cognitive deficits: what is the nature of their relationship?" *Schizophrenia Bulletin 32(2)*, pp. 250–258, 2006.

[10] Q. Zhao, W.-Q. Wang, H.-Z. Fan, D. Li, Y.-J. Li, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, and S.-P. Tan, "Vocal acoustic features may be objective biomarkers of negative symptoms in schizophrenia: A cross-sectional study," *Schizophrenia Research*, vol. 250, pp. 180–185, 2022.

[11] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, and I. Hoffmann, "Identifying schizophrenia based on temporal parameters in spontaneous speech," in *Proc. Interspeech 2018*. International Speech Communication Association (ISCA), 2018.

[12] J. De Boer, A. Voppel, S. Brederoo, H. Schnack, K. Truong, F. Wijnen, and I. Sommer, "Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool," *Psychological medicine*, vol. 53, no. 4, pp. 1302–1312, 2023.

[13] J. Huang, Y. Zhao, Z. Tian, W. Qu, X. Du, J. Zhang, Y. Tan, Z. Wang, and S. Tan, "Evaluating the clinical utility of speech analysis and machine learning in schizophrenia: A pilot study," *Computers in Biology and Medicine*, vol. 164, p. 107359, 2023.

[14] M. Berardi, K. Brosch, J.-K. Pfarr, K. Schneider, A. Sültmann, F. Thomas-Odenthal, A. Wroblewski, P. Usemann, A. Philipsen, U. Dannlowski *et al.*, "Relative importance of speech and voice features in the classification of schizophrenia and depression," *Translational Psychiatry*, vol. 13, no. 1, p. 298, 2023.

[15] Y.-J. Huang, Y.-T. Lin, C.-C. Liu, L.-E. Lee, S.-H. Hung, J.-K. Lo, and L.-C. Fu, "Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 947–956, 2022.

[16] V. P. Martin and J.-L. Rouas, "Estimating symptoms and clinical signs instead of disorders: the path toward the clinical use of voice and speech biomarkers in psychiatry," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 606–10 610.

[17] V. Richter, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, D. Suendermann-Oeft, S. Prokop, A. Khan, C. Yavorsky, J.-P. Lindenmayer *et al.*, "Towards multimodal dialog-based speech & facial biomarkers of schizophrenia," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 171–176.

[18] V. Ramanarayanan, "Multimodal technologies for remote assessment of neurological and mental health," *Journal of Speech, Language, and Hearing Research*, pp. 1–8, 2024.

[19] V. Ramanarayanan, D. Pautler, L. Arbatti, A. Hosamath, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, A. Cornish, D. Habberstad *et al.*, "When words speak just as loudly as actions: Virtual agent based remote health assessment integrating what patients say with what they do," in *Proc. Interspeech 2023*. International Speech Communication Association (ISCA), 2023.

[20] B. Kirkpatrick, G. P. Strauss, L. Nguyen, B. A. Fischer, D. G. Daniel, A. Cienfuegos, and S. R. Marder, "The brief negative symptom scale: psychometric properties," *Schizophrenia bulletin*, vol. 37, no. 2, pp. 300–305, 2011.

[21] S. R. Kay, A. Fiszbein, and L. A. Opler, "The positive and negative syndrome scale (panss) for schizophrenia," *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261–276, 1987.

[22] R. S. Keefe, T. E. Goldberg, P. D. Harvey, J. M. Gold, M. P. Poe, and L. Coughenour, "The brief assessment of cognition in schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery," *Schizophrenia research*, vol. 68, no. 2-3, pp. 283–297, 2004.

[23] R. S. Keefe, P. D. Harvey, T. E. Goldberg, J. M. Gold, T. M. Walker, C. Kennel, and K. Hawkins, "Norms and standardization of the brief assessment of cognition in schizophrenia (bacs)," *Schizophrenia research*, vol. 102, no. 1-3, pp. 108–115, 2008.

[24] M. Neumann, H. Kothare, and V. Ramanarayanan, "Multimodal speech biomarkers for remote monitoring of als disease progression," *Computers in Biology and Medicine*, vol. 180, 2024.

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

[26] J. Liscombe, M. Neumann, H. Kothare, O. Roesler, D. Suendermann-Oeft, and V. Ramanarayanan, "On Timing and Pronunciation Metrics for Intelligibility Assessment in Pathological ALS Speech," in *Speech Motor Control Conference (SMC)*, 2022.

[27] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *CoRR*, vol. abs/1907.06724, 2019. [Online]. Available: http://arxiv.org/abs/1907.06724

[28] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *CoRR*, vol. abs/1907.05047, 2019. [Online]. Available: http://arxiv.org/abs/1907.05047

[29] O. Roesler, H. Kothare, W. Burke, M. Neumann, J. Liscombe, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, and V. Ramanarayanan, "Exploring facial metric normalization for within- and between-subject comparisons in a multimodal health monitoring agent," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22 Companion. New York, NY, USA: Association for Computing Machinery, 2022, p. 160–165.

[30] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa, "Connected speech in neurodegenerative language disorders: a review," *Frontiers in psychology*, vol. 8, p. 269, 2017.

[31] N. B. Mota, N. A. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech graphs provide a quantitative measure of thought disorder in psychosis," *PloS one*, vol. 7, no. 4, p. e34928, 2012.

[32] F. Carrillo, N. Mota, M. Copelli, S. Ribeiro, M. Sigman, G. Cecchi, and D. Fernandez Slezak, "Automated speech analysis for psychosis evaluation," in *Machine Learning and Interpretation in Neuroimaging: 4th International Workshop, MLINI 2014, Held at NIPS 2014, Montreal, QC, Canada, December 13, 2014, Revised Selected Papers 4*. Springer, 2016, pp. 31–39.