Speech Prosody in Schizophrenia Spectrum Disorders: Perceptual Evaluation and Machine Classification

Anonymous submission to Speech Prosody 2026

Abstract

Abnormal prosody is a prominent component of the speech changes in schizophrenia spectrum disorders (SSD). We investigated whether prosodic information alone can distinguish SSD from healthy control (HC) speech through parallel human perception and machine learning experiments. Speech samples from 25 participants (15 SSD, 10 HC) underwent adaptive low-pass filtering to preserve prosodic contours while removing semantic content. Thirty-three raters with varying clinical expertise evaluated 50 filtered stimuli on a 4-point Likert scale. Aggregate ratings achieved 80.0% accuracy (AUC=0.820). Unexpectedly, clinical expertise showed no relationship with classification accuracy (r=-0.17, p=0.369). Machine learning classifiers trained on 108 acoustic features from 251 participants (162 HC, 89 SSD) achieved comparable performance, with Logistic Regression reaching 80.0% accuracy (AUC=0.805). Both approaches demonstrated that prosodic abnormalities in SSD are perceptually salient and computationally detectable independent of semantic content. These findings support prosody-based markers as potential language-independent biomarkers for screening applications, while highlighting the comparable performance of human perception and automated classification in utilizing suprasegmental speech information.

Index Terms: speech recognition, speech biomarker, clinical speech

1. Introduction

Speech abnormalities have long been recognized as clinically significant features of schizophrenia spectrum disorders (SSD), encompassing both content-level disorganization and suprasegmental characteristics [1, 2]. While much research has focused on linguistic and semantic aspects of speech in psychosis [3], prosodic features – including fundamental frequency (F0) patterns, rhythm, and intonation – represent a relatively understudied dimension that may carry diagnostic information.

Previous studies have documented that individuals with SSD often exhibit atypical prosodic patterns, commonly described as "flat affect" or monotonous speech [4, 5]. These observations raise a fundamental question: can prosodic information alone, isolated from semantic content, enable detection of psychosis? This question has both theoretical and practical implications. Theoretically, it addresses whether prosodic abnormalities in SSD are sufficiently distinctive to be perceptible independent of other speech characteristics. Practically, prosody-based markers could potentially serve as language-independent biomarkers for screening or monitoring purposes, with the possibility of expanding to other clinical conditions [6].

To investigate this question, we conducted two parallel experiments using speech samples from individuals with SSD and healthy controls. Following the methodology outlined in recent work on dialect classification [7], we applied adaptive low-pass filtering to isolate prosodic information while removing intelligible semantic content. This processing preserves pitch contours and rhythmic patterns while rendering the speech unintelligible – creating stimuli that contain prosodic information but lack lexical content. Our study addresses three specific research questions:

1. Can human raters distinguish SSD from health controls' (HC) speech based solely on prosodic information, and does clinical expertise influence this ability?
2. Can machine learning classifiers trained on acoustic features achieve accurate SSD/HC discrimination?
3. How do human perception and automated classification compare in utilizing prosodic information for psychosis detection?

By comparing human perception with machine learning approaches, we aim to understand both the perceptual salience of prosodic abnormalities in psychosis and the potential for developing automated assessment tools.

2. Previous Studies

Research on speech in schizophrenia has identified abnormalities across multiple dimensions. At the semantic and discourse level, studies have documented thought disorder, tangentiality, and reduced coherence [8, 9, 10, 11]. At the acoustic-prosodic level, individuals with SSD often exhibit reduced pitch variability, abnormal speech rate, and altered rhythm patterns [4, 5]. Compton et al., [5] demonstrated computationally-derived evidence of monotone speech, with reduced F0 variability correlating with clinical ratings of flat affect. Parola et al., [4] conducted a cross-linguistic meta-analysis showing that voice patterns can serve as markers of schizophrenia across diverse languages and populations. Recent work has also examined harmonic-to-noise ratio (HNR) and other voice quality measures as potential objective biomarkers of negative symptoms [12].

Low-pass filtering has been successfully employed to isolate prosodic information while obscuring articulatory detail [13, 14]. Parsons et al. [7] demonstrated that adaptive filtering methods, where cutoff frequencies are dynamically adjusted based on speaker-specific F0 characteristics, effectively preserve pitch contours while removing formant structure necessary for phoneme identification. While clinicians routinely observe and document prosodic changes in psychiatric assessment (e.g., contributing to "flat vs. labile vs. expansive affect", "monotonous vs. stilted speech"), systematic studies of human perception of these features and their relationship to diagnostic

impression remain limited [15]. Research on affective prosody has primarily focused on patients' ability to perceive emotional prosody [16, 17], rather than on how listeners perceive prosodic abnormalities in patients' speech. Studies examining clinical expertise suggest that trained raters [18] may be more sensitive to subtle speech abnormalities, though the extent to which clinical experience enhances detection of prosodic markers specifically remains unclear.

Automated speech analysis has increasingly been applied to psychiatric assessment [19], with recent studies demonstrating that machine learning classifiers can distinguish SSD from HC speech with substantial accuracy [20, 21]. However, most existing work analyzes complete speech samples containing both prosodic and semantic information. Studies specifically examining prosody-only classification in clinical populations are rare. Our previous work [22] highlighted the importance of robust feature extraction and the challenges of cross-toolkit consistency in clinical applications – a concern particularly relevant when developing automated assessment tools for clinical deployment.

3. Methods

3.1. Datasets

Our study combines data from two internal datasets with different collection protocols ("ACES" and "Remora"). All participants completed open-ended speech tasks designed to elicit naturalistic, spontaneous-style speech. General symptom severity was assessed with the Brief Psychiatry Rating Scale (BPRS) [23] and negative symptoms were assessed with the Scale for the Assessment of Negative Symptoms (SANS) [24]. From these combined datasets, 25 participants were randomly selected for the human perception experiment, while all available participants (N=251) were used for machine learning classification. For the human perception experiment, we stratified participants based on BPRS total scores into severity categories, with participant characteristics shown in Table 1: mild (18-31), moderate (20-37), and severe (33-67). All study procedures were approved by the Institutional Review Board, and all participants provided informed consent.

3.2. Stimulus preparation

To isolate prosodic information, we applied adaptive low-pass filtering following [7]. From each recording, we extracted the first and last 15 seconds of speech excluding silence. For each segment, F0 was estimated using Librosa [25] with a search range of 50-400 Hz. The cutoff frequency was computed using $\text{cutoff} = 420.2 \times (1 - e^{-0.0124 \times F_0})$, bounded between 200-500 Hz, ensuring the filter preserves F0 and lower harmonics while removing formant structure. A 5th-order Butterworth filter was applied, and the filtered audio was normalized to 80% of maximum amplitude.

3.3. Human Perception Experiment

Each of the 25 pairs of low-pass filtered samples were reviewed by 33 raters who are blinded to the diagnosis of the participant. Rater had varying clinical experience working with individuals with psychosis, categorized into five expertise levels from minimal (n=7) to extensively experienced (n=2, 10+ years). Raters also reported their experience with prosody and phonetics research (minimal: n=12, some: n=14, moderate: n=5, extensive: n=1). Raters were instructed that audio files had been processed to remove semantic content while preserving prosody. For each of the 25 participants, raters listened to two filtered audio segments (first and last 15 seconds) and provided a single rating on a 4-point Likert scale: 1 (Very Unlikely to have SSD), 2 (Somewhat Unlikely), 3 (Somewhat Likely), 4 (Very Likely to have SSD). Raters based judgments solely on prosodic features including rhythm, intonation, and speech patterns. Stimuli were randomized, and raters could replay segments as needed.

3.4. Machine Learning Classification

Acoustic features were extracted from both datasets using OpenSMILE's eGeMAPS configuration, which provides a standardized set of 88 acoustic parameters including F0 statistics (mean, range, percentiles), intensity measures, spectral features, voice quality metrics (HNR, jitter, shimmer), and mel-frequency cepstral coefficients (MFCCs). Timing features including pause statistics and speech rate measures were extracted separately, yielding an additional 107 temporal parameters. All features were extracted from the low-pass filtered audio at 16 kHz sampling rate with 60ms frame size and 10ms hop length to match the human perception stimuli processing. Features were aggregated at the participant level by averaging across recordings. We removed features with more than 50% zero or missing values, reducing the feature set from 194 to 108 features (20 timing features, 88 acoustic features), and imputed remaining missing values using median imputation.

To identify the most informative features, we compared eight feature reduction strategies: no reduction (baseline), variance threshold (removing features with variance < 0.01), correlation-based removal (eliminating features with > 0.95 correlation), univariate selection using F-statistic (top 50 features), mutual information-based selection (top 50), recursive feature elimination (RFE) with Random Forest (top 50), Random Forest importance ranking (top 50), and principal component analysis (PCA, 50 components). For each strategy, we evaluated four classifiers: Logistic Regression with L2 regularization (C=1.0, max 1000 iterations), Random Forest (100 estimators, max depth=None), Gradient Boosting (100 estimators, learning rate=0.1), and Support Vector Machine with RBF kernel (C=1.0, gamma='scale').

We employed participant-level group-based splitting using GroupShuffleSplit with a 70-30 train-test split, ensuring that all recordings across multiple tasks from a given participant appeared only in either the training or test set to prevent data leakage. Features were standardized using StandardScaler fit on training data and applied to test data. Model performance was evaluated using accuracy, F1-score, and area under the ROC curve (AUC). All experiments used a fixed random seed (42) for reproducibility.

4. Result

4.1. Human Perception Experiment Results

To evaluate raters' ability to distinguish SSD from HC based on prosodic features, we computed mean ratings across all 33 raters for each of the 25 sets of stimuli. Using an optimal threshold of 2.5 (determined by F1 score maximization), the aggregate ratings achieved 80.0% accuracy (20/25 correct classifications). Table 2 summarizes the classification performance and group comparison statistics. The classifier demonstrated high specificity (90.0%, 9/10 HC correctly identified) and moderate sensitivity (73.3%, 11/15 SSD correctly identified), with positive predictive value of 91.7% and negative predictive value

Table 1: *Demographic and Clinical Characteristics of Participants*

	HC (n=10)	**SSD-Mild (n=6)**	**SSD-Moderate (n=6)**	**SSD-Severe (n=3)**	***p*-value**	**All SSD (n=15)**
Age, y	30.3 (5.4)	23.9 (3.7)	25.8 (7.6)	26.9 (0.9)	0.150	25.2 (5.2)
Female, n (%)	7 (70%)	1 (17%)	2 (33%)	0 (0%)	–	3 (20%)
Race, n (%)						
White	4 (40%)	2 (33%)	1 (17%)	0 (0%)	–	3 (20%)
Black	3 (30%)	2 (33%)	3 (50%)	0 (0%)	–	5 (33%)
Asian	0 (0%)	1 (17%)	1 (17%)	0 (0%)	–	2 (13%)
Multiple	3 (30%)	1 (17%)	0 (0%)	1 (33%)	–	2 (13%)
Education, y	17.2 (2.4)	14.3 (1.6)	11.8 (2.2)	12.7 (2.1)	0.001	13.0 (2.2)
Clinical Characteristics						
BPRS Total	–	25.2 (5.1)	31.7 (6.2)	45.7 (18.6)	0.026	31.9 (11.5)
SANS Total	–	18.2 (11.1)	28.3 (13.4)	16.3 (8.1)	0.253	21.9 (12.2)

Results of ANOVA comparing groups are shown in the *p*-value column. BPRS = Brief Psychiatric Rating Scale; SANS = Scale for the Assessment of Negative Symptoms.

Table 2: *Human perception classification performance*

Metric	**Value**
Accuracy	80.0% (20/25)
Sensitivity	73.3% (11/15)
Specificity	90.0% (9/10)
Positive Predictive Value	91.7%
Negative Predictive Value	69.2%
AUC-ROC	0.820 (95% CI: 0.657–0.984)
Group Comparison	
SSD Mean Rating	2.79 (SD = 0.61)
HC Mean Rating	2.03 (SD = 0.56)
Group Difference	t(23) = 3.15, p = 0.0045
Cohen's d	1.31

of 69.2%. Receiver operating characteristic analysis yielded an AUC of 0.820 (95% CI: 0.657–0.984), indicating good discriminative ability. Mean ratings differed significantly between groups: SSD participants received higher ratings (M = 2.79, SD = 0.61) compared to HC participants (M = 2.03, SD = 0.56), t(23) = 3.15, p = 0.0045, Cohen's d = 1.31. This large effect size indicates that prosodic features provided substantial information for group discrimination. Individual rater accuracy ranged from 44.0% to 80.0% (M = 66.2%, Mdn = 68.0%, SD = 8.7%). Three raters achieved the maximum accuracy of 80.0%, correctly classifying 20 of 25 participants. Inter-rater agreement was moderate, with mean pairwise Spearman correlation of r = 0.39 (Mdn = 0.43, range: -0.43 to 0.80), suggesting that while raters generally agreed on which prosodic patterns indicated SSD, there was considerable individual variation in perceptual strategies.

To test whether clinical or research expertise influenced classification accuracy, we conducted one-way ANOVAs comparing mean accuracy across experience levels. For clinical experience, there was no significant difference in accuracy across the five expertise levels, F(4, 28) = 1.39, p = 0.263, η_p^2 = 0.17. Mean accuracy by clinical experience level ranged from 60.0% (extensive experience) to 71.4% (some experience), with no monotonic relationship between expertise and performance. Similarly, research experience in prosody and phonetics showed no significant effect on accuracy, F(3, 28) = 0.11, p = 0.957, η_p^2 = 0.01. Spearman correlations confirmed these null findings: clinical experience level showed a weak negative correlation with accuracy (r = -0.17, p = 0.369), while research experience showed essentially no relationship (r = 0.01, p = 0.973). Notably, the three raters achieving maximum accuracy (80%) had relatively low expertise levels: two had "some" clinical experience with "minimal" research experience, and one had "minimal" clinical experience with "some" research experience.

Examining individual samples revealed substantial variation in perceived prosodic abnormality. Two SSD participants received near-unanimous classification as SSD, with mean ratings of 3.84 and 3.74 respectively, and 100% of raters assigning them ratings of 3 or 4. Conversely, one HC participant was unanimously classified as HC, receiving a mean rating of 1.58 with zero ratings of 4 ("Very Likely SSD"). Four SSD participants were consistently misclassified as HC, with mean ratings below 2.5. Interestingly, three of these four were classified as moderate severity, suggesting that prosodic abnormalities may not directly track overall symptom severity. Indeed, when stratifying SSD participants by severity, we found no significant relationship between BPRS-based severity categories and mean prosodic ratings, F(2, 12) = 1.21, p = 0.33. Mild SSD cases received numerically higher ratings (M = 3.02, SD = 0.44) than moderate (M = 2.67, SD = 0.89) or severe cases (M = 2.51, SD = 0.42), though this trend did not reach significance in our sample. Both BPRS and SANS total scores showed no correlation with prosody ratings (BPRS: r = -0.027, p = 0.925; SANS: r = 0.239, p = 0.390). One HC participant was misclassified as SSD, receiving a mean rating of 3.23. This false positive case merits further investigation, as it suggests that prosodic patterns associated with SSD may occasionally occur in healthy individuals, or that other factors (e.g., speaking style, affective state during recording) can produce similar acoustic-prosodic profiles.

4.2. Machine Learning Classification Results

The machine learning classification analysis revealed that prosodic features extracted from low-pass filtered speech can distinguish SSD from HC participants with moderate to good accuracy. Table 3 presents the performance of selected model configurations.

The best performing model used Logistic Regression with correlation-based feature reduction, achieving 80.0% accuracy (F1=0.696, AUC=0.805) on the held-out test set. Using all 108

Table 3: *Performance comparison of selected machine learning models for SSD vs. HC classification using prosodic features. LR: Logistic Regression, RF: Random Forest, Acc.: Accuracy, N Feat.: Number of features, All: No feature reduction, Corr: Correlation-based feature removal, PCA-50: PCA with 50 components, Uni-50: Univariate selection with 50 features. Best performance shown in bold.*

Model	Acc.	F1	AUC	N Feat.
LR (All)	**80.00**	**69.57**	**80.53**	**108**
LR (PCA-50)	78.57	68.09	78.12	50
LR (Corr)	77.14	66.67	80.36	94
LR (Uni-50)	75.71	60.47	75.37	50
RF (PCA-50)	74.29	62.50	72.35	50
SVM (PCA-50)	74.29	62.50	74.29	50
SVM (All)	72.86	61.22	68.73	108
GB (PCA-50)	72.86	59.57	68.48	50

features without reduction outperformed both correlation-based feature removal (77.1% accuracy, 94 features) and PCA dimensionality reduction (78.6% accuracy, 50 components), suggesting that the feature set was well-suited for linear classification and that discriminative information was distributed across multiple acoustic dimensions.

Logistic Regression consistently outperformed other classifiers across feature reduction strategies, with mean accuracy of 75.7% compared to 72.9% for SVM, 70.7% for Random Forest, and 70.4% for Gradient Boosting. Among feature reduction methods, PCA (50 components) achieved the highest mean accuracy (75.0%) across all classifiers, followed by using all features (73.9%) and variance threshold (73.9%). The PCA model captured 99.1% of the cumulative variance with 50 components, demonstrating that the acoustic feature space could be effectively compressed while retaining discriminative information. Feature reduction methods that selected subsets based on univariate statistics or mutual information generally performed worse than methods that transformed the feature space (PCA) or removed redundancy (correlation-based removal). This suggests that discriminative information is distributed across multiple features rather than concentrated in a small subset, consistent with the multidimensional nature of prosodic abnormalities in schizophrenia.

5. Discussion

Our findings demonstrate that prosodic abnormalities in schizophrenia spectrum disorders are both perceptually salient and computationally detectable when isolated from semantic content. The comparable performance between human perception (80.0% accuracy, AUC=0.820) and machine learning (80.0% accuracy, AUC=0.805) suggests that prosodic features carry substantial diagnostic information across multiple assessment modalities.

Human raters reliably distinguished SSD from HC speech based solely on prosodic cues, with large between-group differences (Cohen's d=1.31). High specificity (90.0%) and moderate sensitivity (73.3%) suggest that prosodic abnormalities, when present, are highly distinctive, though not all individuals with SSD exhibit equally pronounced markers. Unexpectedly, clinical expertise showed no relationship with classification accuracy, suggesting that prosodic abnormalities may be sufficiently salient for untrained listeners to detect, or that clinical training emphasizes content-level rather than suprasegmental features. In other words, the perceptual distinction of speech from people with SSD based on prosodic cues relies less on specialized clinical or linguistic training, and more on general skills - perhaps social processing skills - that are accessible to untrained individuals. This would also provide one explanation for the observed relationships between speech and language impairment in SSD and poor functional outcomes [26] - because the impairments can readily be perceived by interlocutors in daily life. The moderate inter-rater agreement (mean r=0.39) indicates substantial individual variation in perceptual strategies.

The machine learning results demonstrate that automated classification can match human performance while offering scalability and consistency advantages. The success of Logistic Regression without feature reduction suggests that prosodic abnormalities manifest across multiple acoustic dimensions rather than being concentrated in a small subset of features. Explainable AI techniques, such as feature importance analysis and SHAP values, could identify which prosodic characteristics drive individual classifications, supporting clinical decision-making and trust in automated assessments.

Several limitations warrant consideration. Our human perception study used filtered speech samples, which may not fully capture prosodic variation in extended spontaneous conversation. The modest pool of raters limited power to detect relationships with symptom severity. Our filtering approach preserved pitch contours but removed other potentially diagnostic acoustic information.

Prosody-based assessment could provide an objective, language-independent screening tool for psychosis, valuable in multilingual settings or for monitoring disease progression. However, the moderate sensitivity indicates that prosodic markers alone are insufficient for diagnosis and should complement comprehensive clinical assessment. Future research should examine whether prosodic features track symptom changes over time, investigate which acoustic-prosodic parameters drive classification through explainable AI, establish cross-linguistic validity, and determine specificity to schizophrenia spectrum disorders versus other psychiatric conditions. We should also examine whether prosodic changes in SSD offer a viable avenue for intervention - whether pragmatic language training [27], for example, can normalize speech prosody and whether this then has a downsteam effect on social and occupational functioning.

In conclusion, our parallel experiments demonstrate that prosodic abnormalities in schizophrenia spectrum disorders are robustly detectable independent of semantic content, achieving approximately 80% classification accuracy through both human perception and machine learning. These findings support developing prosody-based assessment tools as potential language-independent biomarkers for psychosis screening and monitoring, with explainable AI offering pathways to enhance clinical interpretability.

6. References

[1] M. A. Covington, C. He, C. Brown, L. Naçi, J. T. McClain, B. S. Fjordbak, J. Semple, and J. Brown, "Schizophrenia and the structure of language: the linguist's view," *Schizophrenia research*, vol. 77, no. 1, pp. 85–98, 2005.

[2] X. Chang, W. Zhao, J. Kang, S. Xiang, C. Xie, H. Corona-Hernández, L. Palaniyappan, and J. Feng, "Language abnormalities in schizophrenia: binding core symptoms through contemporary empirical evidence," *Schizophrenia*, vol. 8, no. 1, p. 95, 2022.

[3] S. X. Tang, R. Kriz, S. Cho, S. J. Park, J. Harowitz, R. E. Gur,

M. T. Bhati, D. H. Wolf, J. Sedoc, and M. Y. Liberman, "Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders," *npj Schizophrenia*, vol. 7, no. 1, p. 25, 2021.

[4] A. Parola, A. Simonsen, J. M. Lin, Y. Zhou, H. Wang, S. Ubukata, K. Koelkebeck, V. Bliksted, and R. Fusaroli, "Voice patterns as markers of schizophrenia: building a cumulative generalizable approach via a cross-linguistic and meta-analysis based investigation," *Schizophrenia Bulletin*, vol. 49, no. Supplement_2, pp. S125–S141, 2023.

[5] M. T. Compton, A. Lunden, S. D. Cleary, L. Pauselli, Y. Alolayan, B. Halpern, B. Broussard, A. Crisafio, L. Capulong, P. M. Balducci *et al.*, "The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech," *Schizophrenia research*, vol. 197, pp. 392–399, 2018.

[6] A. S. G. Choi, J.-s. Kim, S.-h. Kim, M. S. Back, and S. Cho, "Crosslinguistic acoustic feature-based dementia classification using advanced learning architectures," in *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments@ LREC-COLING 2024*, 2024, pp. 95–100.

[7] P. Parsons, H. S. Bremnes, K. Kvale, T. Svendsen, and G. Salvi, "Effects of prosodic information on dialect classification using whisper features," in *Proc. Interspeech 2025*, 2025, pp. 2785–2789.

[8] M. Spitzer, "A cognitive neuroscience view of schizophrenic thought disorder," *Schizophrenia Bulletin*, vol. 23, no. 1, pp. 29–50, 1997.

[9] L. Palaniyappan, P. Homan, and M. F. Alonso-Sanchez, "Language network dysfunction and formal thought disorder in schizophrenia," *Schizophrenia bulletin*, vol. 49, no. 2, pp. 486–497, 2023.

[10] L. Pauselli, B. Halpern, S. D. Cleary, B. S. Ku, M. A. Covington, and M. T. Compton, "Computational linguistic analysis applied to a semantic fluency task to measure derailment and tangentiality in schizophrenia," *Psychiatry research*, vol. 263, pp. 74–79, 2018.

[11] L. Jeong, M. Lee, B. Eyre, A. Balagopalan, F. Rudzicz, and C. Gabilondo, "Exploring the use of natural language processing for objective assessment of disorganized speech in schizophrenia," *Psychiatric Research and Clinical Practice*, vol. 5, no. 3, pp. 84–92, 2023.

[12] 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.

[13] V. Raithel and M. Hielscher-Fastabend, "Emotional and linguistic perception of prosody: Reception of prosody," *Folia Phoniatrica et Logopaedica*, vol. 56, no. 1, pp. 7–13, 2004.

[14] M. A. Knoll, M. Uther, and A. Costall, "Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners," *Speech Communication*, vol. 51, no. 3, pp. 210–216, 2009.

[15] H. Ding and Y. Zhang, "Speech prosody in mental disorders," *Annual Review of Linguistics*, vol. 9, no. 1, pp. 335–355, 2023.

[16] V. Lucarini, M. Grice, F. Cangemi, J. T. Zimmermann, C. Marchesi, K. Vogeley, and M. Tonna, "Speech prosody as a bridge between psychopathology and linguistics: the case of the schizophrenia spectrum," *Frontiers in psychiatry*, vol. 11, p. 531863, 2020.

[17] V. P. Bozikas, M. H. Kosmidis, D. Anezoulaki, M. Giannakou, C. Andreou, and A. Karavatos, "Impaired perception of affective prosody in schizophrenia," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 18, no. 1, pp. 81–85, 2006.

[18] S. D. Targum, "Evaluating rater competency for cns clinical trials," *Journal of clinical psychopharmacology*, vol. 26, no. 3, pp. 308–310, 2006.

[19] G.-D. Liu, Y.-C. Li, W. Zhang, and L. Zhang, "A brief review of artificial intelligence applications and algorithms for psychiatric disorders," *Engineering*, vol. 6, no. 4, pp. 462–467, 2020.

[20] 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.

[21] R. Banks, C. Higgins, B. R. Greene, A. Jannati, J. Gomes-Osman, S. Tobyne, D. Bates, and A. Pascual-Leone, "Clinical classification of memory and cognitive impairment with multimodal digital biomarkers," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 16, no. 1, p. e12557, 2024.

[22] A. S. G. Choi, A. Richardson, R. Partlan, S. Tang, and S. Cho, "Comparative evaluation of acoustic feature extraction tools for clinical speech analysis," *arXiv preprint arXiv:2506.01129*, 2025.

[23] J. E. Overall and D. R. Gorham, "The brief psychiatric rating scale," *Psychological reports*, vol. 10, no. 3, pp. 799–812, 1962.

[24] N. C. Andreasen, "Scale for the assessment of positive symptoms," *Group*, vol. 17, no. 2, pp. 173–180, 1984.

[25] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[26] E. J. Tan, N. Thomas, and S. L. Rossell, "Speech disturbances and quality of life in schizophrenia: Differential impacts on functioning and life satisfaction," *Comprehensive Psychiatry*, vol. 55, no. 3, pp. 693–698, 2014.

[27] V. Bambini, G. Agostoni, M. Buonocore, E. Tonini, M. Bechi, I. Ferri, J. Sapienza, F. Martini, F. Cuoco, F. Cocchi *et al.*, "It is time to address language disorders in schizophrenia: A rct on the efficacy of a novel training targeting the pragmatics of communication (pragmacom)," *Journal of Communication Disorders*, vol. 97, p. 106196, 2022.