Speech Prosody in Schizophrenia Spectrum Disorders

Perceptual Evaluation and Machine Classification

Anna Seo Gyeong Choi, Ryan Partlan, Alexander Richardson, Sandy Yin, Nourhan Zalat, Katharina Brosch, Amir Nikzad, Simran Bhola, Khatiya Chelidze Moon, Sunghye Cho, Sunny X. Tang

Background

Schizophrenia Spectrum Disorders
Speech abnormalities are clinically recognized features of SSD, spanning:

- Semantic & discourse level
- Prosodic & suprasegmental features

"*Flat affect*," *monotonous speech*, *reduced F0 variability* are hallmark characteristics of SSD but understudied

Contribution

Key open question:
Can prosodic information alone, isolated from semantic content, enable detection of psychosis?

Theoretical:
Are prosodic abnormalities sufficiently distinctive on their own?

Practical:
Language-independent biomarkers for cross-lingual screening

Research Questions

RQ1

Can human raters distinguish SSD from HC speech based solely on prosodic information?

Does clinical expertise influence this ability?

Research Questions

RQ1

Can human raters distinguish SSD from HC speech based solely on prosodic information?

Does clinical expertise influence this ability?

RQ2

Can machine learning classifiers achieve accurate SSD / HC discrimination from prosodic features alone?

Which acoustic dimensions drive classification?

Research Questions

RQ1

Can human raters distinguish SSD from HC speech based solely on prosodic information?

Does clinical expertise influence this ability?

RQ2

Can machine learning classifiers achieve accurate SSD / HC discrimination from prosodic features alone?

Which acoustic dimensions drive classification?

RQ3

How do human perception and automated classification compare in utilizing prosodic information?

Are they comparable or complementary?

Speech Recordings

Low-Pass Filtering

Adaptive cutoff
200-500 Hz

Human Perception

Machine Learning Classification

251

Total participants

89

SSD participants

162

Healthy participants

33

Human raters

Low-Pass Filtering

Adaptive cutoff 200-500 Hz

Segment & Extract

First and last 15 seconds of speech extracted (silence excluded)
F0 estimated with Librosa (search range 50-400 Hz)

Adaptive Cutoff

cutoff = 420.2 × (1 − e^{−0.0124 × F0})
Bounded 200–500 Hz per speaker
Preserves pitch contours + lower harmonics

Filter & Normalize

5th-order Butterworth filter applied
Audio normalized to 80% max amplitude
Semantic content removed; prosody preserved

Low-Pass Filtering

Adaptive cutoff
200-500 Hz

Result:
intelligible speech
→ **unintelligible** stimuli that retain prosodic contours, rhythm, and intonation – but no lexical content

Human Perception Experiment

Rater Setup

33 raters, blinded to diagnosis

Varying clinical expertise (5 levels):
Minimal (n=7) → Extensive (n=2, 10+ yrs)

Prosody/phonetics research experience:
Minimal (n=12), Some (n=14), Moderate/Extensive (n=6)

Stimulus set:
25 participants x 2 audio segments each

Human Perception Experiment

Rater Setup

33 raters, blinded to diagnosis

Varying clinical expertise (5 levels):
Minimal (n=7) → Extensive (n=2, 10+ yrs)

Prosody/phonetics research experience:
Minimal (n=12), Some (n=14), Moderate/Extensive (n=6)

Stimulus set:
25 participants x 2 audio segments each

Rating Task

4-point Likert scale:

1. Very unlikely to have SSD
2. Somewhat unlikely
3. Somewhat likely
4. Very likely

Optimal threshold: 2.5
(determined by F1 score maximization)

Machine Learning Classification

Acoustic Features (108 total)

OpenSMILE eGeMAPS:
F0 stats, intensity, spectral features, HNR, jitter/shimmer, MFCCs

Timing features:
Pause statistics, speech rate (20 temporal features)

Extracted from low-pass filtered audio at 16 kHz

Machine Learning Classification

Acoustic Features (108 total)

OpenSMILE eGeMAPS:
F0 stats, intensity, spectral features, HNR, jitter/shimmer, MFCCs

Timing features:
Pause statistics, speech rate (20 temporal features)

Extracted from low-pass filtered audio at 16 kHz

Classifiers & Feature Reduction

4 Classifiers:
Logistic Regression, Random Forest, Gradient Boosting, SVM-RBF

8 Feature strategies:
No reduction, variance threshold, correlation removal, univariate, mutual info, RFE, RF importance, PCA-50

70/30 train-test split at participant level to prevent data leakage

Results: Human Perception

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

80 %
Accuracy

0.820
AUC-ROC

90 %
Specificity
HC correctly classified

73.3 %
Sensitivity
SSD correctly classified

Results: Human Perception

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

r = -0.17

Clinical expertise

r = 0.01

Research experience

p = 0.39

Inter-rater agreement

Results: Machine Learning Classification

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

80 %

Logistic Regression (all features)

Matches human perception level

Conclusions

01

Prosodic abnormalities are independently perceptible

80% accuracy via both human and ML approaches confirms prosodic cues alone carry substantial diagnostic information

Conclusions

01

Prosodic abnormalities are independently perceptible

80% accuracy via both human and ML approaches confirms prosodic cues alone carry substantial diagnostic information

02

Expertise doesn't necessarily help – general social perception does

No advantage for clinical or phonetics training

Conclusions

01

Prosodic abnormalities are independently perceptible

80% accuracy via both human and ML approaches confirms prosodic cues alone carry substantial diagnostic information

02

Expertise doesn't necessarily help – general social perception does

No advantage for clinical or phonetics training

03

Prosody may be trait-level, not severity-dependent

BPRS and SANS scores uncorrelated with prosody ratings

Conclusions

01

Prosodic abnormalities are independently perceptible

80% accuracy via both human and ML approaches confirms prosodic cues alone carry substantial diagnostic information

02

Expertise doesn't necessarily help – general social perception does

No advantage for clinical or phonetics training

03

Prosody may be trait-level, not severity-dependent

BPRS and SANS scores uncorrelated with prosody ratings

04

Distributed acoustic signal

Logistic Regression with all 108 features beat PCA and subset methods

Future Directions

- Explainable AI to identify driving prosodic features
- Cross-linguistic validity across language families
- Longitudinal tracking of symptom change via prosody
- Pragmatic language interventions – can training normalize prosody?

Thank You!