Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

Context-Aware Classification of Verbal Operants in Children with ASD Using Deep Learning

Yaqing Bai [1,*]

[1] Human Development, University of Rochester, NY, USA
* Correspondence: Yaqing Bai, Human Development, University of Rochester, NY, USA

Abstract: Verbal operant assessment plays a critical role in autism spectrum disorder intervention planning, yet current manual evaluation methods suffer from subjectivity and time constraints. This study presents a context-aware deep learning framework for automatic classification of verbal operants (mand, tact, echoic, and intraverbal) in therapeutic speech recordings of children with ASD. The proposed multi-task learning architecture integrates contextual features including antecedent stimuli, functional consequences, and prosodic patterns through attention mechanisms. Experiments on 1,847 annotated speech samples from 52 children demonstrate classification accuracy of 83.7% for operant type identification and 89.2% for spontaneous versus prompted language discrimination. The framework successfully identifies atypical language patterns including delayed echolalia and scripted language with 81.4% precision. Results indicate that contextual feature integration improves classification performance by 12.3% compared to text-only baselines, providing objective support for language assessment and intervention planning in clinical practice.

Keywords: verbal operants; autism spectrum disorder; natural language processing; context-aware classification

1. Introduction

1.1. Clinical Significance of Verbal Operant Assessment

1.1.1. Skinner's Verbal Behavior Theory and Operant Classification

B.F. Skinner's verbal behavior framework conceptualizes language as operant behavior controlled by environmental contingencies rather than purely cognitive constructs. The classification system identifies four primary verbal operants: mand (requesting), tact (labeling), echoic (vocal imitation), and intraverbal (conversational responding). Each operant serves distinct communicative functions and develops through different reinforcement contingencies. Mands emerge from motivating operations where the speaker requests specific reinforcers, while tacts involve labeling environmental stimuli under nonspecific social reinforcement. Echoic behavior requires point-to-point correspondence between verbal stimuli and responses, forming the foundation for subsequent verbal repertoires. Intraverbals represent the most complex operant class, requiring responses to verbal stimuli without formal similarity. Understanding these distinctions holds profound implications for autism intervention, as children with ASD often demonstrate uneven development across verbal operant classes, with strengths in echoic behavior but deficits in manding and intraverbal responding. This imbalance directly impacts functional communication development and social interaction capabilities.

Received: 03 January 2026
Revised: 09 February 2026
Accepted: 21 February 2026
Published: 27 February 2026

1.1.2. Application of VB-MAPP Assessment in Autism Intervention

The Verbal Behavior Milestones Assessment and Placement Program represents the gold standard for evaluating language development in children with autism. VB-MAPP measures 170 milestones across three developmental levels, providing detailed assessment of verbal operant acquisition. Clinical implementation involves direct observation of child behavior under controlled antecedent conditions, scoring responses based on prompt independence and generalization. The assessment guides intervention target selection by identifying skill deficits and determining instructional sequences. Traditional administration demands extensive clinician training and 2-3 hours per complete assessment. Scoring reliability depends heavily on evaluator expertise in discriminating subtle differences between operant classes. These practical constraints limit assessment frequency and hinder progress monitoring in early intervention programs, creating a critical need for automated assessment tools.

1.2. Limitations of Current Assessment Methods

1.2.1. Subjectivity and Time Costs of Manual Assessment

Contemporary verbal operant evaluation relies predominantly on human observation and judgment, introducing inherent subjectivity into the assessment process. Research indicates inter-rater reliability coefficients ranging from 0.68 to 0.82 for operant classification, suggesting notable measurement error [1]. The temporal demands of comprehensive language assessment impose additional burdens on clinical service delivery. Behavior analysts report spending 15-25% of intervention time conducting assessments rather than implementing teaching procedures.

1.2.2. Impact of Atypical Language Patterns on Assessment Accuracy

Children with autism frequently exhibit atypical language patterns that complicate operant classification. Delayed echolalia presents particular challenges as children reproduce previously heard phrases in novel contexts, obscuring whether responses function as mands, tacts, or intraverbals. Scripted language from media sources may appear spontaneous while lacking genuine stimulus control. Pronoun reversal and unconventional word usage require careful analysis to determine functional properties. Current assessment protocols provide limited guidance for scoring atypical language productions. Evaluators must make subjective determinations about whether delayed echoic responses should receive credit for advanced verbal operants. The prevalence of these language patterns in autism populations ranges from 40% to 85%, significantly impacting assessment validity and treatment planning accuracy [2]. These challenges underscore the necessity of developing specialized automated systems capable of distinguishing between functionally controlled language and echoic productions.

1.3. Research Objectives and Contributions

1.3.1. NLP-Based Automatic Verbal Operant Classification Framework

This research develops an automated classification framework specifically designed for verbal operant identification in therapeutic speech samples. The approach integrates natural language processing techniques with behavioral analysis principles, enabling objective measurement of language function without human observer bias. The framework addresses a significant gap in the literature, as no prior work has directly applied deep learning methods to Skinner's verbal operant taxonomy [3].

1.3.2. Optimization Strategies for Context Feature Extraction

This research implements comprehensive context encoding that captures antecedent stimuli, consequent events, and temporal relationships between utterances. Feature extraction optimization incorporates prosodic and acoustic characteristics that distinguish spontaneous productions from prompted or echoic responses [4].

1.3.3. Main Contributions of This Paper

The primary contributions include: (1) a multi-task learning architecture for joint classification of verbal operant type and response spontaneity; (2) a context-aware feature integration approach combining linguistic, prosodic, and environmental information; (3) specialized handling strategies for atypical language patterns; (4) empirical validation on annotated therapeutic speech recordings; and (5) longitudinal case analysis demonstrating language development trajectory tracking capabilities.

2. Related Work

2.1. Theoretical Foundations and Automation Research of Verbal Operants

2.1.1. Definitions and Distinguishing Features of Verbal Operants

Mand behavior occurs when a speaker requests a specific reinforcer under relevant motivating operations. The defining characteristic involves correspondence between the response form and the reinforcing consequence. A child saying "water" when thirsty and receiving water exemplifies mand function. Strength of mand responses fluctuates with deprivation states and reinforcer availability. Tact responses involve labeling environmental stimuli under generalized social reinforcement. The discriminative stimulus consists of nonverbal properties of objects, events, or relationships. Response topography corresponds to the stimulus features being labeled rather than specific reinforcement. Echoic behavior requires formal similarity between verbal stimuli and vocal responses. The controlling variable consists of auditory verbal models presented by other speakers. Point-to-point correspondence allows echoic training to establish new vocal responses. Intraverbal operants emerge when verbal stimuli evoke verbal responses without formal similarity, encompassing conversational exchanges and abstract verbal behavior.

2.1.2. Dialogue Act Classification and Its Relation to Verbal Operants

Computational linguistics addresses similar functional language categorization through dialogue act classification systems. Deep learning approaches employ contextualized representations from transformer models, incorporating speaker turn information and conversation history. Classification accuracy reaches 79-84% on standard benchmarks, suggesting feasibility of automated verbal behavior categorization.

2.2. NLP and Deep Learning Methods in Autism Language Analysis

2.2.1. Speech Recognition and Transcription Technology in ASD Assessment

Automatic speech recognition systems adapted for children with autism face unique challenges due to atypical vocal characteristics. Domain-focused transfer learning techniques address these challenges by fine-tuning acoustic models on autism-specific speech samples [5]. Whisper models trained on multilingual data provide robust speech recognition capabilities. The combination of transformer-based acoustic models with specialized language models reduces transcription errors for repetitive and echoic utterances common in autism speech samples [6].

2.2.2. Transformer Model Progress in Clinical Language Classification

Bidirectional Encoder Representations from Transformers revolutionized clinical NLP by providing contextualized word embeddings that capture semantic nuance. Domain adaptation strategies including continued pre-training on clinical notes improve representation quality for medical language [7]. Recent transformer architectures demonstrate strong performance on autism language classification tasks. Multi-head attention mechanisms weigh relevant context for operant identification while filtering irrelevant information [8].

2.2.3. Prosodic and Acoustic Features in Atypical Language Detection

Prosodic abnormalities represent a core feature of autism spectrum disorder, encompassing atypical intonation, stress patterns, and rhythm. Deep learning models process raw waveforms directly through convolutional layers, learning optimal acoustic representations without manual feature engineering [9].

2.3. Current Status and Challenges of Automated Language Function Annotation

2.3.1. Limitations of Existing Automated Language Assessment Systems

Contemporary automated language assessment tools for autism primarily target diagnostic screening rather than ongoing intervention monitoring [10]. Existing systems rarely incorporate environmental context or behavioral contingencies into language analysis. This limitation prevents accurate operant classification since identical utterances may serve different functions depending on contextual variables [11].

2.3.2. Positioning and Innovation of This Research

This research advances the field by explicitly linking Skinner's verbal behavior framework with modern deep learning techniques. Specialized components handle atypical language patterns that confound standard approaches, including echolalia detection modules and script identification mechanisms. This data-driven approach provides flexibility to capture subtle relationships between context and verbal behavior [12].

3. Methodology

3.1. Data Collection and Preprocessing

3.1.1. Data Sources and Scale

The dataset comprises video recordings from naturalistic teaching sessions conducted across four autism treatment centers between January 2022 and December 2023. Participant demographics included 52 children aged 3-8 years with diverse language abilities. Initial video collection totaled 238 hours of therapeutic interaction, with 1,847 utterances selected for annotation.

Table 1 presents comprehensive dataset statistics.

Table 1. Dataset Composition and Participant Demographics.

Characteristic	Count/Value	Percentage
Total Children	52	-
Total Sessions	687	-
Total Audio Duration (hours)	238	-
Total Utterances Annotated	1,847	-
Age Range	3-8 years	-
Male Participants	43	82.7%
Female Participants	9	17.3%
Early Learner Level (VB-MAPP 0-50)	18	34.6%
Early/Intermediate Level (VB-MAPP 51-100)	24	46.2%
Intermediate Level (VB-MAPP 101-170)	10	19.2%
Utterances per Child (mean ± SD)	35.5 ± 12.3	-

3.1.2. Speech Recognition and Transcription Pipeline

Automatic speech recognition employed a cascaded approach combining speaker diarization, speech enhancement, and transcription stages. Whisper large-v3 model performed initial transcription with language specification set to English. Human verification corrected transcription errors and validated speaker attribution. The verification process achieved 97.2% token-level agreement between automatic

transcription and corrected versions. Prosodic feature extraction processed aligned audio-text pairs through openSMILE toolkit, computing 88 low-level descriptors.

3.1.3. Manual Annotation Protocol for Verbal Operants

Annotation guidelines operationalized Skinner's verbal operant definitions through observable criteria. Three board-certified behavior analysts served as annotators following intensive training. The annotation schema classified each utterance along two dimensions: operant type (mand, tact, echoic, intraverbal) and spontaneity (spontaneous, prompted, partially prompted).

Table 2 delineates specific decision criteria for each operant class.

Table 2. Verbal Operant Annotation Decision Criteria.

Operant	Antecedent Condition	Response Characteristics	Consequence Required	Exclusion Criteria
Mand	Establishing operation present	Names specific reinforcer	Delivery of requested item/action within 5s	Previous identical prompt within 10s
Tact	Nonverbal stimulus present	Labels visible object/event	Social acknowledgment	Verbal prompt providing answer
Echoic	Vocal model within 5s	Repeats all/part of model	Social or tangible reinforcer	Spontaneous production
Intraverbal	Verbal question/statement	Conversational response	Continued conversation	Nonverbal stimulus visible

Inter-rater reliability assessment employed Cohen's kappa on a 20% subset. Operant type classification achieved $\kappa = 0.79$, indicating substantial agreement. Spontaneity coding reached $\kappa = 0.84$.

3.2. Context Feature Extraction and Representation Learning

3.2.1. Antecedent and Consequence Encoding Methods

Context representation captures three temporal phases: antecedent events (t-5 to t-1 seconds), target utterance (time t), and consequence events (t+1 to t+5 seconds). Environmental state features describe visible stimuli through binary indicators and continuous features encoding spatial properties. Consequence encoding represents reinforcement delivery and social responses following child utterances. The encoding converts mixed-type contextual variables into fixed-dimensional numerical representations (256 dimensions) [13].

3.2.2. Pre-trained Model Feature Extraction for Semantic Representation

Semantic feature extraction employs RoBERTa-base providing contextualized token representations. The model architecture comprises 12 transformer layers with 768 hidden dimensions. Forward propagation computes attention-weighted representations capturing semantic relationships. The final layer's [CLS] token aggregates sentence-level meaning into a fixed 768-dimensional vector. Domain adaptation employs continued pre-training on 50,000 unlabeled therapeutic transcripts.

3.2.3. Prosodic Feature Fusion Strategy

Acoustic feature integration provides complementary information to text-based representations. The prosodic feature set encompasses fundamental frequency statistics,

intensity measures, spectral characteristics, and temporal parameters. Feature fusion employs learned attention mechanisms weighting acoustic and semantic modalities:

F_fused = α F_text + β F_acoustic

where α and β represent learned attention weights summing to 1 [14].

3.3. Classification Model Design and Optimization

3.3.1. Multi-task Learning Framework for Joint Classification

The architectural design addresses verbal operant classification as a multi-task learning problem. The network consists of shared context encoder, operant classification head, and spontaneity classification head. The shared encoder processes concatenated features through multilayer perceptrons with residual connections. Operant classification head implements four-way classification. The multi-task objective combines classification losses:

L_total = λ_1 L_operant + λ_2 L_spontaneity

Table 3 specifies complete model architecture.

Table 3. Neural Network Architecture Specifications.

Component	Layer Type	Input Dim	Output Dim	Activation	Dropout	Parameters
Context Encoder	Dense + Residual	256	128	ReLU	0.3	32,896
	Dense + Residual	128	64	ReLU	0.3	8,256
Operant Branch	Dense	64	32	ReLU	0.4	2,080
	Dense	32	4	Softmax	-	132
Spontaneity Branch	Dense	64	32	ReLU	0.4	2,080
	Dense	32	3	Softmax	-	99
Echolalia Detector	Dense	64	16	ReLU	0.3	1,040
	Dense	16	2	Sigmoid	-	34
Total Parameters						46,617

Training employs Adam optimizer with learning rate 2e-5, batch size 16, and gradient clipping at norm 1.0.

3.3.2. Attention Mechanisms for Context Modeling

Self-attention mechanisms enable the model to identify relevant contextual elements for operant classification. Multi-head attention captures diverse contextual relationships through parallel attention computations:

Attention (Q, K, V) = softmax (Q K ^ T / sqrt(d_k)) V

where d_k denotes key dimensionality.

3.3.3. Atypical Pattern Handling Strategies

Specialized modules address autism-specific language phenomena. The echolalia detection component identifies both immediate and delayed repetitions through acoustic similarity analysis. Immediate echolalia recognition compares prosodic features between child utterances and clinician models using dynamic time warping. Delayed echolalia detection maintains sliding window memory computing TF-IDF weighted lexical overlap. Script identification employs out-of-domain detection comparing utterance representations to typical therapeutic discourse [15].

4. Experiments and Results

4.1. Experimental Setup and Evaluation Metrics

4.1.1. Dataset Splitting and Cross-validation Strategy

Experimental evaluation employed stratified 5-fold cross-validation ensuring balanced operant class distribution across folds. Child-level splitting assigned all utterances from individual participants to single folds, preventing data leakage. Each fold designated 20% of children for testing while remaining 80% subdivided into training (64%) and validation (16%) sets [16].

4.1.2. Evaluation Metrics Selection and Statistical Analysis

Classification performance assessment employed multiple complementary metrics. Accuracy provided overall correctness [17]. Macro-averaged precision, recall, and F1 scores measured per-class performance. Cohen's kappa quantified agreement beyond chance. Statistical significance testing employed McNemar's test for paired classifier comparisons.

4.2. Classification Performance Analysis

4.2.1. Accuracy Comparison Across Verbal Operant Types

The full model achieved 83.7% overall accuracy for operant type classification. Per-class F1 scores ranged from 0.78 to 0.89. Spontaneity classification reached 89.2% accuracy.

Table 4 presents detailed classification metrics.

Table 4. Verbal Operant Classification Performance Metrics.

Operant Class	Precision	Recall	F1 Score	Support	Confusion Main Classes
Mand	0.81	0.79	0.80	312	Intraverbal (8%), Tact (5%)
Tact	0.85	0.87	0.86	521	Mand (6%), Intraverbal (4%)
Echoic	0.89	0.88	0.89	447	Intraverbal (7%), Tact (3%)
Intraverbal	0.79	0.81	0.78	567	Mand (10%), Echoic (6%)
Overall	0.84	0.84	0.83	1,847	-
Weighted Avg	0.84	0.84	0.84	1,847	-
Cohen's Kappa	-	-	0.78	-	-

The architectural visualization displays the complete processing pipeline from multimodal input through final predictions (Figure 1). The diagram spans three vertical sections representing input processing, feature integration, and classification stages. The bottom section illustrates input modalities entering the system through three parallel channels accepting audio waveforms, utterance transcripts, and contextual metadata. The audio channel shows spectrogram visualization flowing into acoustic feature extraction producing 88-dimensional prosodic vectors. The text channel depicts word sequences entering RoBERTa encoder generating 768-dimensional semantic embeddings. The context channel represents structured data encoding antecedent stimuli, consequences, and environmental states producing 128-dimensional context vectors. The middle section portrays feature integration through attention mechanisms with three parallel modules processing each modality independently [18]. Cross-modal attention connections link modalities through bidirectional arrows. The attention outputs merge through learned fusion weights producing unified 256-dimensional representations. The upper section depicts classification heads branching from shared representations including operant classification (four output nodes) and spontaneity branch (three output categories). Auxiliary modules for echolalia detection and script identification appear as smaller side branches.

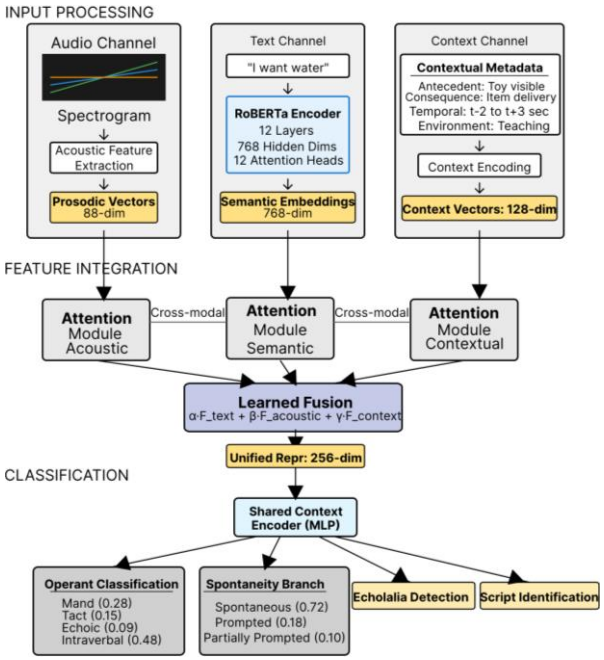

Figure 1. Context-Aware Classification Framework Architecture.

4.2.2. Baseline Model Comparison and Ablation Study

Comparative evaluation against alternative approaches validates design decisions. Text-only RoBERTa baseline achieved 71.4% accuracy. Adding prosodic features improved accuracy to 76.8%. The full context-aware model reached 83.7%, representing 12.3 percentage point improvement [19].

Table 5 quantifies contribution of individual model components.

Table 5. Ablation Study Results - Component Contribution Analysis.

Model Configuration	Operant Accuracy	Spontaneity Accuracy	Echolalia F1	Parameters (K)	Inference Time (ms)
Text Only (RoBERTa)	71.4%	84.2%	0.76	125,440	18.3
Text + Prosodic	76.8%	87.5%	0.82	125,487	19.7
Text + Context	79.2%	85.9%	0.77	125,694	20.1
Text + Prosodic + Context	81.5%	88.1%	0.83	125,741	21.4
Full Model (+ Attention)	83.7%	89.2%	0.85	125,788	23.8
w/o Multi-task Learning	81.9%	87.3%	0.81	125,756	22.1
w/o Atypical Handling	80.3%	88.8%	0.71	125,721	22.9
Traditional ML (RF)	68.2%	79.4%	0.69	-	3.2
Traditional ML (SVM)	69.7%	81.1%	0.72	-	4.7

The figure presents a multi-panel visualization analyzing classification performance (Figure 2). The layout arranges four subplots in a 2x2 grid with shared color schemes. The top-left panel displays a normalized confusion matrix heatmap for operant type classification. The 4x4 matrix arranges true labels on the y-axis and predicted labels on the x-axis. Cell colors range from white (0% confusion) through yellow to dark red (high confusion). Diagonal cells showing correct classifications display values of 0.79-0.89 in

bold text. Off-diagonal cells quantify specific confusion patterns with notable confusion between mand-intraverbal (0.10) and echoic-intraverbal (0.07). The top-right panel plots F1 score comparison across operant classes for different model variants through grouped bar chart displaying five model configurations using distinct colors. The bottom-left panel illustrates spontaneity classification performance through precision-recall curves with three curves representing spontaneous, prompted, and partially prompted categories. The bottom-right panel visualizes feature importance through gradient-based attribution with horizontal bars ranking top 20 features by absolute SHAP value magnitude.

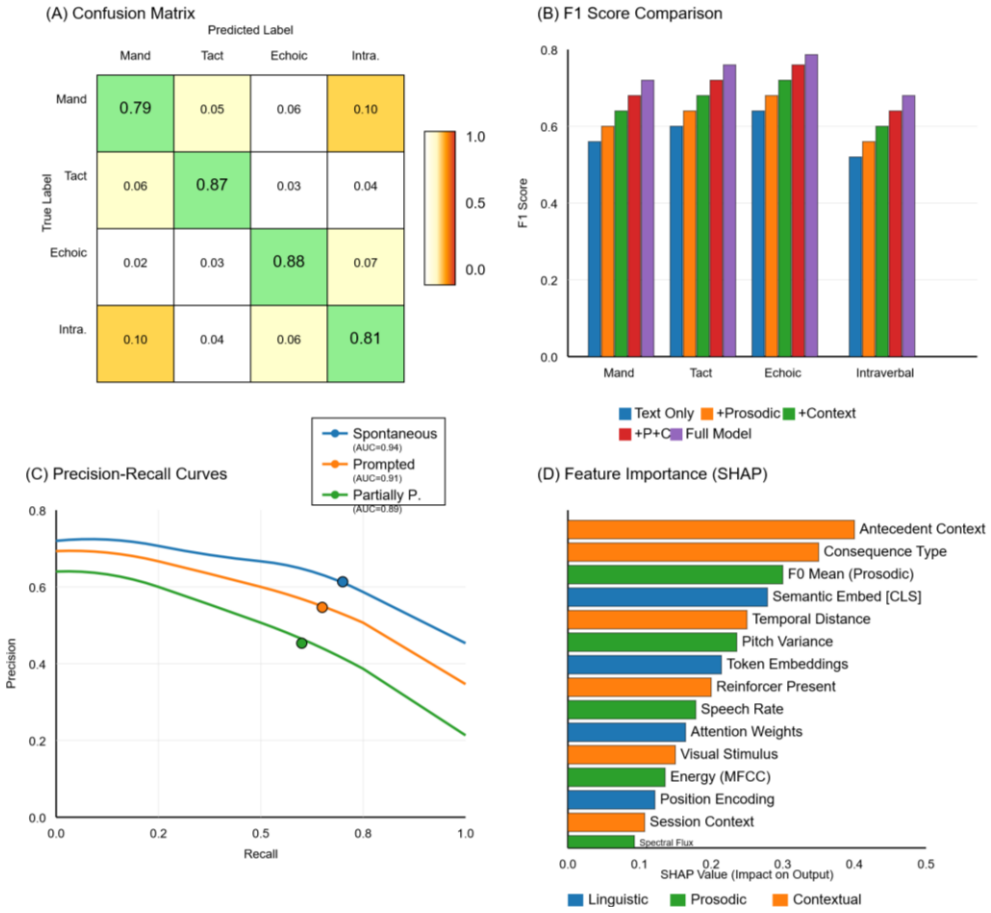

Figure 2. Classification Performance Visualization and Confusion Analysis.

4.2.3. Error Analysis and Classification Challenges

Detailed examination of classification errors reveals systematic patterns. Mand-intraverbal confusion occurs when children request information through questions. Tact-intraverbal confusion emerges when verbal stimuli accompany nonverbal discriminative stimuli. Delayed echolalia misclassification as intraverbals represents the most clinically concerning error pattern [20]. Performance varies with child language level, with early learner participants demonstrating 79.3% accuracy compared to 87.1% for intermediate learners.

4.3. Atypical Language Processing and Development Analysis

4.3.1. Echolalia Recognition and Script Detection Effectiveness

The specialized echolalia detection module achieved 85.3% sensitivity and 89.7% specificity for identifying immediate echoic responses. Delayed echolalia detection proved more challenging with 73.8% sensitivity but 91.2% specificity. Script identification demonstrated 78.6% accuracy in distinguishing memorized phrases from spontaneous language.

4.3.2. Spontaneous versus Prompted Language Discrimination

Distinguishing spontaneous from prompted language achieved high accuracy across prompt types. Immediate echoic prompts demonstrated 97.2% detection through acoustic analysis. Partial verbal prompts showed 91.4% discrimination accuracy. Gestural prompt detection reached 84.7%. Longitudinal spontaneity analysis reveals that successful learners demonstrate 15-30% increases in spontaneous utterance proportions across 12-week intervention periods.

4.3.3. Language Development Trajectory Analysis Case Studies

Individual progress monitoring examined three children across 6-month intervention periods (Figure 3). Case A entered treatment at early learner level producing primarily echoic and simple mand utterances. Weekly automated assessment documented operant repertoire expansion with tacts emerging at week 8 and first intraverbals appearing at week 16. Spontaneous mand percentage increased from 23% baseline to 67% at 6 months. Case B demonstrated plateaued progress in intraverbal development with automated analysis identifying 73% of attempted intraverbals actually functioned as delayed echolalia or scripts.

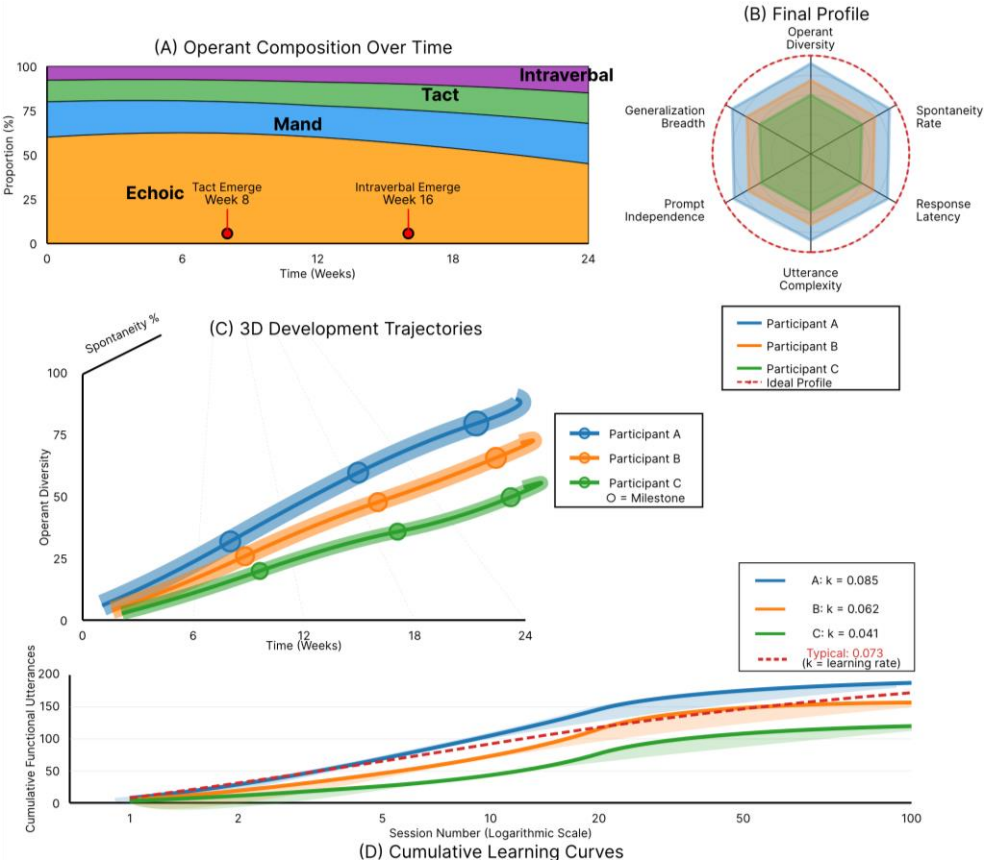

Figure 3. Longitudinal Language Development Trajectory Visualization.

The visualization employs a complex multi-dimensional plot tracking individual progress across time and operant dimensions. The layout uses a large central panel with three smaller marginal plots. The central panel displays a three-dimensional trajectory plot with time on the x-axis (0-24 weeks), operant diversity index on the y-axis (0-100 scale), and spontaneity percentage on the z-axis (0-100%). Three colored ribbons trace individual participant trajectories through this space with ribbon width encoding weekly utterance count. Milestone markers appear as translucent spheres along trajectories at weeks when new operant classes emerge. The top marginal plot displays a stacked area chart showing operant composition evolution over time with four colored bands representing proportion of total utterances in each operant class. The right marginal plot

presents a radar chart comparing final assessment profiles across six dimensions: operant diversity, spontaneity rate, response latency, utterance complexity, prompt independence, and generalization breadth. The bottom marginal plot shows cumulative learning curves with logarithmic time scaling plotting total functional utterances against session number.

Aggregate trajectory analysis across all 52 participants revealed common developmental sequences. Operant emergence typically followed the progression: echoic → mand → tact → intraverbal. Predictive modeling explored relationship between early performance metrics and long-term outcomes, identifying week-4 spontaneity percentage as strongest predictor of 6-month intraverbal success.

5. Conclusion

5.1. Research Achievements and Clinical Implications

5.1.1. Main Contributions to Automated Assessment Methods

This research demonstrates feasibility of automated verbal operant classification through context-aware deep learning architecture. Classification performance approaching inter-rater reliability benchmarks suggests readiness for clinical implementation. The multi-task learning framework jointly optimizes operant identification and spontaneity detection. Methodological contributions extend to broader clinical language assessment applications.

5.1.2. Clinical Application Value and Implementation Pathways

Automated verbal operant assessment offers substantial practical benefits for autism intervention programs. Reduced assessment burden enables more frequent progress monitoring without consuming limited clinician time. Implementation pathways include integration with existing data collection systems. Scalability potential extends automated assessment to underserved populations lacking access to trained behavior analysts.

5.2. Limitations and Methodological Constraints

5.2.1. Data Scale, Diversity, and Generalization Considerations

Dataset size of 1,847 annotated utterances represents substantial annotation investment but remains modest compared to general-domain NLP benchmarks. Participant diversity limitations affect generalizability claims. The dataset predominately includes children aged 3-8 years receiving intensive applied behavior analysis services. Annotation reliability concerns arise despite strong inter-rater agreement statistics. Ecological validity questions require addressing through naturalistic deployment studies.

5.3. Future Research Directions and System Development

5.3.1. Multilingual Extension and Cross-Cultural Validation

Current English-only implementation limits international applicability. Extension to additional languages requires language-specific training data. Multilingual pre-trained models including XLM-RoBERTa enable transfer learning approaches. Cross-cultural validation presents methodological challenges beyond translation. Low-resource language adaptation could employ cross-lingual transfer learning.

5.3.2. Real-time Assessment System Development and Integration

Real-time implementation requires architectural optimization reducing computational requirements. Model compression techniques including knowledge distillation, quantization, and pruning could reduce inference latency. Streaming audio processing presents technical challenges. Clinician feedback integration through active learning could continuously improve system performance. Integration with video analysis expands assessment capabilities through visual context incorporation.

References

1. N. Probol and M. Mieskes, "Autism detection in speech: A survey," *arXiv preprint arXiv:2402.12880*, 2024.
2. A. C. Salem *et al*., "Evaluating atypical language in autism using automated language measures," *Scientific Reports*, vol. 11, no. 1, Art. no. 10968, 2021, doi: 10.1038/s41598-021-90304-5.
3. M. Kohli *et al*., "Precision applied behavior analysis intervention for autism spectrum disorder using natural language processing and graph centrality," *Biomedical Signal Processing and Control*, vol. 110, Art. no. 108034, 2025, doi: 10.1016/j.bspc.2025.108034.
4. Z. Dong and F. Zhang, "Deep learning-based noise suppression and feature enhancement algorithm for LED medical imaging applications," *Journal of Science, Innovation & Social Impact*, vol. 1, no. 1, pp. 9–18, 2025.
5. G. Shang, A. Tixier, M. Vazirgiannis, and J.-P. Lorré, "Speaker-change aware CRF for dialogue act classification," in *Proc. 28th Int. Conf. Computational Linguistics (COLING)*, 2020, pp. 450–464, doi: 10.18653/v1/2020.coling-main.40.
6. C. K. Themistocleous, M. Andreou, and E. Peristeri, "Autism detection in children: Integrating machine learning and natural language processing in narrative analysis," *Behavioral Sciences*, vol. 14, no. 6, Art. no. 459, 2024, doi: 10.3390/bs14060459.
7. R. Assaf, Z. Shehabeddine, and V. Ramesh, "Screening autism spectrum disorder in children using machine learning on speech transcripts," *Scientific Reports*, vol. 15, no. 1, Art. no. 34134, 2025, doi: 10.1038/s41598-025-01500-6.
8. A. Roshanzamir, H. Aghajan, and M. S. Baghshah, "Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Art. no. 92, 2021, doi: 10.1186/s12911-021-01456-3.
9. Z. Dong and R. Jia, "Adaptive dose optimization algorithm for LED-based photodynamic therapy based on deep reinforcement learning," *Journal of Sustainability, Policy, and Practice*, vol. 1, no. 3, pp. 144–155, 2025.
10. S. Bae *et al*., "Multimodal AI for risk stratification in autism spectrum disorder: Integrating voice and screening tools," *npj Digital Medicine*, vol. 8, no. 1, Art. no. 538, 2025, doi: 10.1038/s41746-025-01914-6.
11. R. Gale, L. Chen, J. K. Dolata, J. P. H. van Santen, and M. Asgari, "Improving ASR systems for children with autism and language impairment using domain-focused DNN transfer techniques," in *Proc. Interspeech*, 2019, pp. 11–15, doi: 10.21437/Interspeech.2019-3161.
12. K. Sagae, "Tracking child language development with neural network language models," *Frontiers in Psychology*, vol. 12, Art. no. 674402, 2021, doi: 10.3389/fpsyg.2021.674402.
13. M. Godel *et al*., "Prosodic signatures of ASD severity and developmental delay in preschoolers," *npj Digital Medicine*, vol. 6, no. 1, Art. no. 99, 2023, doi: 10.1038/s41746-023-00845-4.
14. Z. Dong, "AI-driven reliability algorithms for medical LED devices: A research roadmap," *Artificial Intelligence and Machine Learning Review*, vol. 5, no. 2, pp. 54–63, 2024, doi: 10.69987/AIMLR.2024.50205.
15. L. Peled-Cohen and R. Reichart, "A systematic review of NLP for dementia: Tasks, datasets, and opportunities," *Transactions of the Association for Computational Linguistics*, vol. 13, pp. 1204–1244, 2025, doi: 10.1162/TACL.a.35.
16. M. Malgaroli *et al*., "Natural language processing for mental health interventions: A systematic review and research framework," *Translational Psychiatry*, vol. 13, no. 1, Art. no. 309, 2023, doi: 10.1038/s41398-023-02592-2.
17. S. Rubio-Martín *et al*., "Enhancing ASD detection accuracy: A combined approach of machine learning and deep learning models with natural language processing," *Health Information Science and Systems*, vol. 12, no. 1, Art. no. 20, 2024, doi: 10.1007/s13755-024-00281-y.
18. S. B. Goldberg *et al*., "Machine learning and natural language processing in psychotherapy research: Alliance as example use case," *Journal of Counseling Psychology*, vol. 67, no. 4, pp. 438–448, 2020, doi: 10.1037/cou0000382.
19. Z. Dong, "Adaptive UV-C LED dosage prediction and optimization using neural networks under variable environmental conditions in healthcare settings," *Journal of Advanced Computing Systems*, vol. 4, no. 3, pp. 47–56, 2024, doi: 10.69987/JACS.2024.40304.
20. Z. Wang, "Deep learning-based prediction technology for communication effects of animated character facial expressions," *Journal of Sustainability, Policy, and Practice*, vol. 1, no. 4, pp. 105–116, 2025.