Original Research Article

Limits of Linguistic and Zipfian Features for ASD Detection: A Neural Network Replication on ASDBank

Krishna Pabbu

William Fremd High School, 1000 S. Quentin Road, Palatine, IL 60067-7018, United States

ABSTRACT

Early screening for autism spectrum disorder (ASD) is still difficult. Current assessments take time, depend heavily on clinicians, and are hard to scale. Speech transcripts are easier to collect and avoid many privacy concerns. This study aims to evaluate whether simple linguistic features and Zipfian word-frequency metrics can distinguish children with ASD from typically developing children using short speech transcripts. Prior work reports classification accuracies of 75–80% using linguistic features, and some studies suggest that ASD speech may deviate from Zipfian word-frequency patterns. This study tests those claims using the Eigsti corpus, a balanced subset from ASDBank, which includes 16 autistic participants and 16 typically developing participants. Neural networks were trained on linguistic features, Zipf-based metrics, and their combination across 100 randomized train-test splits with data augmentation. Performance remained near chance level, approximately 50% accuracy, across all conditions. Feature analysis showed that ASD participants used fewer conjunctions, while other linguistic and Zipfian features did not show consistent separation or predictive value. Overall, these results do not support earlier reports. Linguistic and Zipfian features alone are not enough for reliable ASD classification from short transcripts. The results also highlight reproducibility challenges in small datasets and suggest that larger datasets and multimodal approaches will be needed.

Keywords: Autism spectrum disorder; ASDBank; speech transcripts; machine learning; Zipf's law; linguistic features; neural network; reproducibility

INTRODUCTION

Autism Spectrum Disorder (ASD) is a developmental condition affecting about 1% of the global population (1), including approximately 1 in 54 children in the United States (2). Early diagnosis and intervention can significantly improve long-term outcomes for children with ASD (3). However, current diagnostic practices are time consuming, labor intensive, and subjective, often requiring extensive evaluations by clinicians and repeated observations of the child (3). As a result, an estimated one-fourth of children with ASD go undiagnosed in early childhood (4). There is a clear need for more efficient and objective screening methods to identify ASD as early as possible.

In recent years, researchers have turned to machine learning (ML) and artificial intelligence to support ASD detection (3). Various modalities have been explored, including computer vision analysis of facial images, eye-gaze tracking, and neuroimaging data. These approaches have revealed patterns associated with ASD. Studies show that children with ASD produce facial expressions that are more ambiguous and less socially meaningful than those of typically developing peers (5). Eye-tracking

Corresponding author: Krishna Pabbu, E-mail: krishnapabbu7@gmail.com.

Accepted July 14, 2026
https://doi.org/10.70251/HYJR2348.44283292

research has also found that children with ASD tend to focus less on human faces and more on background objects during social interactions (6). In addition, ML models can differentiate individuals with ASD using brain MRI or EEG features (6).

While these approaches show promise, they raise practical and ethical concerns. Collecting images of a child's face, retina, or brain over time can be invasive and raises privacy issues, especially for minors (3). In contrast, analyzing a child's natural speech provides a rich and less intrusive data source. Unlike images, speech can be collected more easily and anonymized while preserving content (3). Every child's speech patterns are unique, and language differences are a well known characteristic of ASD (3). Children with ASD often show unusual speech patterns. For example, they may repeat words and phrases. This is a common symptom of ASD called echolalia. Additionally, they tend to use a more limited or simplified vocabulary and syntax compared to neurotypically developing peers (3). These traits make speech transcripts a strong candidate for automated ASD screening while preserving privacy.

Prior studies in computational linguistics show that analyzing child speech can help distinguish ASD from typical development. For example, Ramesh and Assaf (3) used transcribed child speech from TalkBank's ASDBank corpus (7) and extracted over 50 linguistic features, such as mean utterance length, vocabulary richness, and part-of-speech frequencies, to train machine learning classifiers (3). Their models achieved up to 75–80% accuracy in predicting ASD versus typical development, with logistic regression and random forests performing best (3). Their feature analysis also identified several linguistic indicators of ASD. One important metric was utterance length. An utterance is a single spoken unit in a transcript, such as one response, phrase, or sentence produced by the child denoted by a stop on speaking. Children with ASD tended to produce shorter utterances, differed in conversational turn length, and showed different usage of certain parts of speech compared to typically developing children (3). Metrics such as Mean Length of Utterance (MLU), mean turn length ratio, and the frequency of pronouns, conjunctions, and negations were among the strongest predictors (3). Together, these results suggest that measurable transcript-level features, such as utterance length, turn length, and part-of-speech usage, can capture language patterns that differ between ASD and typical development. Because these features can be extracted computationally, they provide a useful starting point for automated ASD detection from speech transcripts.

In addition to traditional linguistic features, recent work highlights the role of Zipf's law in word usage. Zipf's law describes a common pattern in natural language where word frequencies follow a power-law distribution. In other words, a small set of words appears very often, while most words appear only occasionally. When word frequencies are ranked from most common to least common, typical language tends to show a predictable decline in frequency. This makes Zipf's law useful for studying the overall structure of word use in speech. Zipfian features were included in this study because they capture a different aspect of language than standard linguistic metrics. While features such as utterance length and part-of-speech usage measure sentence structure and grammar, Zipf-based metrics measure the distribution of words across an entire transcript. This is relevant because speech from children with ASD may be more repetitive or less lexically diverse, which could cause it to deviate from the expected power-law pattern (9). In one study, transcripts from neurotypical children closely followed a power-law distribution, while transcripts from children with ASD did not (9). Statistical tests such as Kolmogorov–Smirnov goodness-of-fit and likelihood ratio tests supported this difference (9).

Word-frequency differences may reflect less diverse or more repetitive language in ASD. For example, a child with ASD may use a small set of words very frequently, which changes the overall frequency pattern. These differences in word distribution could serve as measurable indicators of ASD-related language patterns (9). De Palma *et al*. suggest that Zipfian metrics may complement traditional behavioral indicators in ASD diagnosis (8). In other words, how closely speech follows or deviates from Zipf's law may be a useful signal for automated detection.

Motivated by these findings, this work proposes a neural network approach that combines linguistic features with Zipfian statistical features for ASD detection from speech transcripts. It tests whether combining these feature sets improves performance compared to using linguistic features alone. The model is evaluated on child speech data from the ASDBank corpus (7). The inclusion of Zipf-based features aims to capture distributional differences in ASD speech that are not reflected in standard linguistic measures. The hypothesis was that combining linguistic complexity measures with lexical distribution metrics would improve the ability to distinguish children with ASD from typically developing

children. The next sections describe the methodology and present the results.

METHODS AND MATERIALS

Database Extraction

This study used the ASDBank English Eigsti Corpus dataset from TalkBank (7, 9). This dataset consists of transcribed spontaneous speech from children in a controlled play interaction setting (30-minute free play sessions). The corpus includes three groups of children ages 3–6: ASD (children diagnosed with autism spectrum disorder), DD (children with non-ASD developmental delays), and TD (typically developing children), with 16 children in each group. This study focused on distinguishing children with ASD from typically developing children, so only transcripts from the ASD group (n = 16) and TD group (n = 16) were used, for a total of 32 child transcripts. Each transcript contains the first 100 utterances from the child in the play session, interacting with an adult experimenter. The transcripts are formatted in CHAT notation and include both the child's and the adult's utterances. Prior to feature extraction, basic preprocessing was applied by removing annotations and non-speech elements and standardizing the text (for example, lowercasing). No additional normalization was required beyond what was already provided in the TalkBank transcripts.

Linguistic Feature Extraction

From each child's transcript, linguistic features identified in previous work (3) as useful for distinguishing ASD and TD speech were extracted. These features measured three main areas: utterance length, part-of-speech usage, and overall verbal output. An utterance refers to one spoken unit in the transcript, such as a phrase, response, or sentence produced by the child. The child's mean length of utterance (MLU), defined as the average number of words per utterance, was calculated as a measure of sentence length and linguistic complexity (3). The mean length of turn (MLT) ratio was also computed. This ratio was defined as the child's mean utterance length divided by the interviewer's mean utterance length in the conversation (3). This feature captures how the child's responses compare in length to the person guiding the interaction. A lower MLT ratio may indicate shorter responses or reduced conversational engagement.

In addition, the child's use of certain parts of speech was quantified by measuring the percentage of words in the transcript that were pronouns, conjunctions, or negations, such as "no" and "not." These categories were chosen based on prior findings of atypical language patterns in ASD, including pronoun differences and reduced use of connective words (3). These counts were normalized by the total number of words and reported as rates per 100 words to account for transcript length. Finally, as a simple measure of verbal output, the total number of words produced by the child in the 100-utterance sample was recorded (3).

All feature values were calculated using custom scripts that parsed the transcripts. The resulting linguistic feature vector for each child consisted of six features: MLU, MLT ratio, pronoun rate per 100 words, conjunction rate per 100 words, negation rate per 100 words, and total word count.

Zipfian Feature Extraction

Simultaneously, features describing the word-frequency distribution of each child's speech were extracted. For each child's transcript, the frequency count of every word produced by the child was obtained, and the distribution was analyzed to determine whether it followed a Zipfian power-law pattern. To quantify this, the statistical approach described by Clauset *et al*. (10), as used in De Palma *et al*. (8), was applied. Specifically, a maximum-likelihood power-law fit was performed for the upper tail of the word-frequency distribution using a power-law analysis software package (11). This fitting step estimates the power-law curve that best matches the observed word-frequency pattern and produces parameters that describe how quickly word frequencies decline. From this fit, the estimated power-law exponent (α) was recorded. This value reflects how quickly word frequency drops off for less common words. A higher α indicates a steeper drop and may suggest heavier repetition of a smaller set of words. The minimum frequency rank above which the power-law model was fit was also determined. This cutoff is chosen by the fitting procedure to optimize the fit (10).

Next, each child's word-frequency distribution was evaluated to determine how well it conformed to Zipf's law. A goodness-of-fit p-value was calculated using a Kolmogorov–Smirnov test with bootstrap simulation (10). A high p-value, typically greater than 0.1, suggests that the data are consistent with a Zipfian distribution, while a low p-value indicates a poor fit, meaning a significant deviation from Zipf's law. In addition, the power-law fit was compared to a log-normal alternative using a likelihood ratio (LR) test (10). This provided an

LR statistic and an associated p-value. If the log-normal model fits significantly better, such as in a significant LR test favoring the log-normal model, it suggests that the word frequencies are not well described by a power law. This is consistent with a non-Zipfian pattern in the child's speech (9). From these analyses, a set of Zipf-based features was derived for each transcript. In particular, the fitted power-law exponent α, the goodness-of-fit p-value, and the significance result of the power-law versus log-normal LR test were used as features. The minimum frequency rank and raw LR statistics were initially considered as well, but these were less informative. In summary, each child's Zipf feature vector captures the degree to which the lexical distribution follows or deviates from Zipf's law.

After extracting all features, data cleaning and normalization were performed. Any undefined or infinite feature values were removed or replaced with neutral values. For example, if a feature such as the MLT ratio was undefined due to zero adult utterances, which did not occur in the data, or if a transcript contained anomalies, that sample was excluded. An outlier filter was also applied, removing any transcript where a feature value was more than 3 standard deviations from the group mean. After outlier filtering, three transcripts were removed: two from the ASD group and one from the TD group. The final analytic sample therefore included 14 ASD transcripts and 15 TD transcripts, for a total of 29 transcripts. This clarification is included to show that outlier filtering only slightly affected class balance. Finally, all feature values were standardized using z-scores before modeling so that features on different scales could be combined.

Model Training

ASD detection was framed as a binary classification problem between ASD and TD transcripts. Given the relatively small size of the dataset, 32 transcripts reduced to 29 after data cleaning, data augmentation was used to expand the training set. A Bayesian generative approach was applied to generate additional training examples in feature space. For each training fold, the distribution of features in the training set was modeled using probabilistic inference. Each feature dimension, after z-score normalization, was assumed to follow an approximately normal distribution. Using a Bayesian sampling method through the PyMC library, the mean and standard deviation for each feature were estimated based on the training data. Markov Chain Monte Carlo (MCMC) simulations were then used to sample from the posterior distribution of these parameters, capturing uncertainty due to the limited dataset.

Bayesian data augmentation was used because the analytic dataset was very small, and a neural network trained only on the real training samples would be highly unstable across random splits. The goal of this augmentation was not to create new independent participants, but to regularize training by generating feature-space examples that reflected the estimated distribution of the training data. This approach allowed the model to be evaluated repeatedly across randomized train-test splits while preserving class balance. However, this choice has important limitations. Because only 29 real transcripts were available after cleaning, the estimated feature means and standard deviations were uncertain. As a result, synthetic samples generated from these posterior distributions may not fully represent the true ASD or TD population distributions and may amplify noise or artifacts in the original data. Simpler alternatives, such as leave-one-out cross-validation or models without augmentation, could also be considered for datasets of this size. Therefore, the Bayesian augmentation procedure should be interpreted as an exploratory modeling choice rather than as evidence that the synthetic samples accurately represented the broader ASD and TD populations. Using these learned feature distributions, synthetic data points were generated to augment the training set. In each training iteration, feature means and variances were sampled from the posterior and used to generate new feature vectors. In total, 1,000 synthetic feature vectors were created for each training split. Labels were assigned to maintain class balance, with approximately half labeled ASD and half labeled TD, matching the real data. Next, a feed-forward neural network classifier was trained on the combined dataset of real and synthetic examples. The network was a multilayer perceptron with an input layer equal to the number of features, two hidden layers of 64 and 32 neurons with ReLU activation, a dense layer of 16 neurons, and a final sigmoid output for binary classification. Dropout regularization, with rates of 0.3 and 0.2, was applied during training to reduce overfitting. The model was trained using the Adam optimizer with binary cross-entropy loss, and early stopping was used based on validation performance to prevent over-training.

To evaluate performance, repeated random hold-out validation was used. A total of 100 independent trials were run. In each trial, the data were randomly split into an 80% training set and a 20% test set. Feature modeling, synthetic data generation, and model training were

repeated for each trial, and performance was evaluated on the unseen test set. This produced 100 independent accuracy measurements. This approach was used instead of a fixed split or standard k-fold cross-validation to make better use of the limited data and to observe how results varied across different splits. The primary evaluation metric was classification accuracy. The next section presents the results and analyzes which features were most informative for distinguishing ASD and TD speech.

RESULTS

Across all three feature configurations, linguistic only, Zipf only, and their combination, the classifier's performance stayed around chance, with large variability across splits. This instability suggests weak generalization from short transcripts in a small-sample setting. Figure 1 shows the classification accuracy obtained using only the linguistic features.

The model trained using only the Zipfian features also showed unstable performance, as shown in Figure 2.

Combining the linguistic and Zipfian features did not improve classification performance, as shown in Figure 3.

In addition to classification accuracy, individual features and group differences were analyzed to better understand these results. Two-sample t-tests were conducted to compare ASD and TD groups on each feature. Out of all linguistic measures, only one showed a

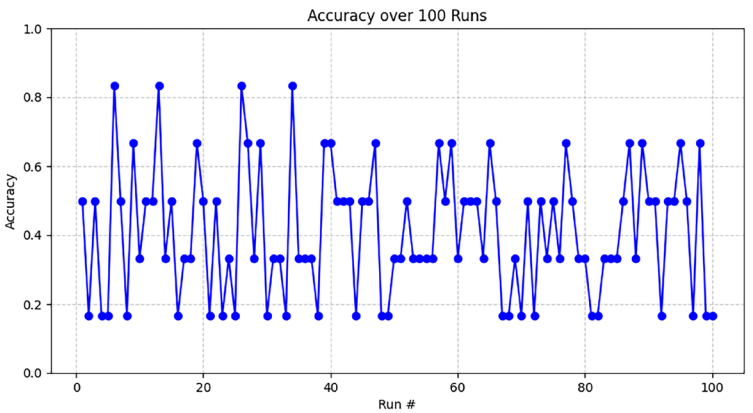

Figure 1. *Classification accuracy across 100 randomized 80/20 train-test splits using six linguistic features: mean length of utterance, mean length of turn ratio, pronoun rate, conjunction rate, negation rate, and total word count. The horizontal axis represents the run number, and the vertical axis represents test accuracy. Accuracy ranged from 0.167 to 0.833, with a mean of 0.412.*

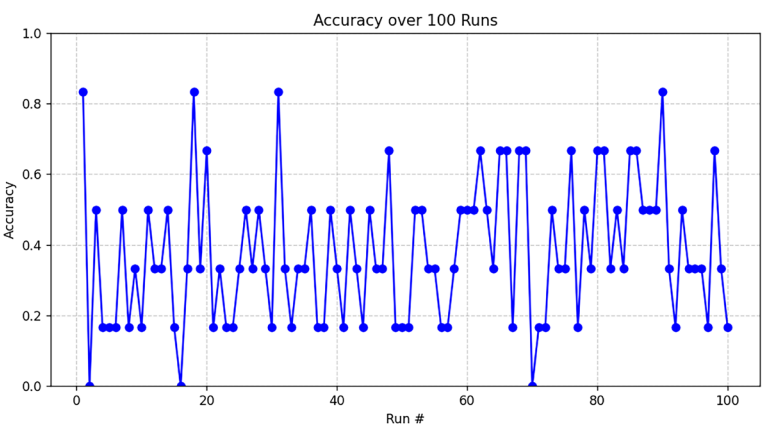

Figure 2. *Classification accuracy across 100 randomized 80/20 train-test splits using the Zipfian features described in the Materials and Methods section. The horizontal axis represents the run number, and the vertical axis represents test accuracy. Accuracy ranged from 0 to 0.833, with a mean of 0.380.*

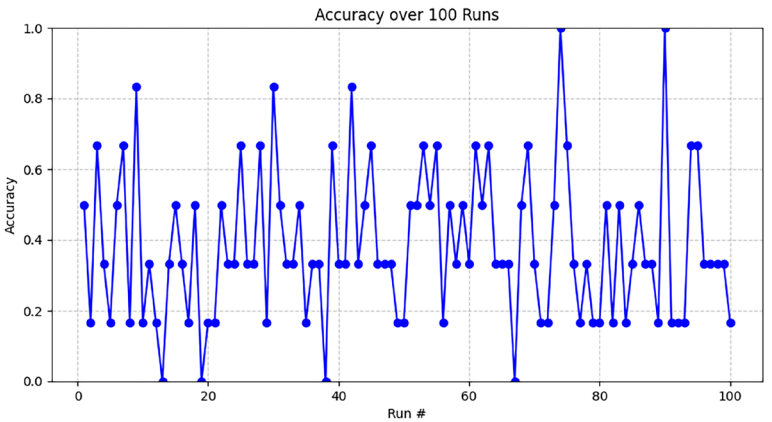

Figure 3. *Classification accuracy across 100 randomized 80/20 train-test splits using the combined linguistic and Zipfian feature set. The horizontal axis represents the run number, and the vertical axis represents test accuracy. Accuracy ranged from 0 to 1.000, with a mean of 0.385.*

significant difference: conjunction usage was lower in the ASD group. Children with ASD used fewer conjunctions per 100 words on average than typically developing children, ASD mean = 2.00, TD mean = 3.21, t = –2.293, p = 0.0313. This suggests that children with ASD in this sample tended to produce simpler, less connected sentences, using fewer words such as "and," "but," and "because." None of the other linguistic features showed statistically significant differences, with all p-values greater than 0.1. For example, the ASD group had slightly lower mean utterance length (MLU) and slightly lower pronoun usage than the TD group, but these differences were not reliable given the variability in the data. The mean length of turn ratio, negation rate, and total word count also did not differ significantly between groups.

Similarly, none of the Zipfian distribution metrics differed significantly between ASD and TD groups, with all p-values greater than 0.14. In this dataset, no consistent deviation from Zipf's law was observed in ASD speech. Some ASD transcripts showed low Zipf fit p-values, indicating non-Zipfian distributions, but this pattern was not consistent across the group. Some TD transcripts also showed poor Zipfian fits. The power-law exponent α did not differ significantly between groups, and the likelihood ratio test results also showed no consistent differences. These findings match the classification results, where Zipf features alone did not provide reliable separation between ASD and TD.

To better interpret the model, feature importance in the combined classifier was examined. Zipf-related features were generally not strong contributors in the neural network. When the trained models were inspected, for example by looking at weight magnitudes or feature importance measures, Zipf features tended to rank in the middle to lower range in terms of influence. In contrast, some linguistic features, especially conjunction usage rate, appeared among the more influential inputs. This aligns with the earlier analysis. Since conjunction usage was the only feature showing a significant group difference, it likely contributed some predictive signal, while Zipf metrics did not. Overall, the results show that neither feature set provided a strong or reliable signal for ASD, and combining them did not improve performance. This points to clear limitations in using these features alone, especially with a small dataset.

DISCUSSION

The findings of this study show that, contrary to the expectations and prior reports, simple linguistic features and Zipfian word-frequency features are not enough to reliably distinguish children with ASD from typically developing children using short speech transcripts. All models performed at or below chance level (around 50% accuracy), which contrasts with the 75–80% accuracy reported by Ramesh and Assaf (3) using similar data. This failure to replicate suggests there may be important differences in methodology or data that explain the gap. Table 1 compares the present study with Ramesh and Assaf (3) across dataset selection, sample size, feature set, model type, validation approach, and main findings.

Among these differences, feature set size and model complexity are likely especially important. Ramesh and Assaf used a much broader feature set, which may have

Table 1. *Methodological comparison between Ramesh and Assaf (3) and the present study. The table compares the dataset, sample size, feature set, model type, validation approach, and main findings of the two studies.*

Category	Ramesh and Assaf (3)	Present Study	Likely Effect on Results
Dataset	Used speech transcripts from ASDBank	Used the Eigsti subset of ASDBank, limited to ASD and TD groups	A smaller or different subset may reduce generalizability and make classification harder
Sample size	Reported results using a broader transcript-based dataset	Began with 32 transcripts and used 29 after outlier filtering	The smaller sample increases variance and makes model performance less stable
Feature set	Used more than 50 linguistic features	Used a smaller feature set focused on MLU, MLT ratio, pronouns, conjunctions, negations, total words, and Zipfian features	Fewer linguistic features may miss signals captured by the larger feature set
Models	Used simpler classifiers such as logistic regression and random forest	Used a neural network with Bayesian feature-space augmentation	Neural networks may overfit when the dataset is very small
Validation approach	Reported high accuracy across their modeling setup	Used 100 randomized 80/20 train-test splits	Repeated random splits showed high instability and weak generalization
Main finding	Reported approximately 75–80% classification accuracy	Found performance near chance across linguistic-only, Zipf-only, and combined models	The discrepancy may reflect methodological divergence rather than a direct contradiction

captured syntactic, lexical, or conversational patterns not included in the present study. In contrast, this study used a smaller set of features and added Zipfian metrics, which did not show consistent group differences. The use of a neural network may also have contributed to instability because the model had very few real samples to learn from. Therefore, the lower performance observed here should be interpreted as evidence that this reduced feature set and modeling pipeline did not reproduce the earlier level of accuracy, rather than as proof that transcript-based ASD classification is impossible.

The one significant result observed was lower conjunction usage in ASD speech. This suggests that children with ASD may produce simpler, less connected sentences. Conjunctions such as "and," "but," and "because" help link ideas, so lower usage points to reduced sentence complexity. This aligns with prior findings that ASD language can be less syntactically complex (3). However, other expected patterns did not appear clearly. For example, no significant difference was found in pronoun usage, even though some studies report pronoun-related differences in ASD. This may be due to the young age group, ages 3–6, or the short transcript length, where such patterns are harder to measure reliably. Similarly, mean length of utterance (MLU) was slightly lower in ASD on average but was not statistically significant. This may reflect high variability within the ASD group, where some children have near-typical language while others show delays. Overall, aside from conjunction usage, linguistic differences in this dataset were small and inconsistent, which helps explain the poor classification performance.

Limitations

Several factors likely contributed to the poor performance and lack of significant findings. One possibility is that the earlier results were overly optimistic or specific to their dataset and feature set. Ramesh and Assaf used over 50 features and may have relied on a larger or more diverse set of transcripts, including additional linguistic cues or multiple data sources (3). The present study focused on a smaller, controlled subset, 16 ASD and 16 TD children from the same corpus, and used a reduced feature set based on their strongest predictors. With this smaller sample and limited feature set, the classification task became extremely difficult. Even with a neural network and data augmentation, the model could not learn a stable, generalizable pattern. This highlights a key reproducibility issue. Results from small datasets may not hold when tested independently. These findings

suggest that high accuracy numbers reported on limited data should be interpreted carefully, since models can overfit or rely on patterns that do not generalize.

Sample size is a major limitation. With only 32 children (29 after filtering) and 100 utterances per transcript, the statistical power to detect real differences is low, and machine learning models have very little data to learn from. The linguistic-only model remained near chance and showed substantial variation across splits (Figure 1). The Zipf-only model showed even greater instability, with accuracy ranging from 0% to 83.3% across trials (Figure 2). These results depend heavily on which children appear in the training and test sets. In some splits, the model may perform well by chance, while in others it fails completely. This variability shows that the model is not learning a consistent pattern but is instead reacting to small differences in the data.

Another limitation is the short length and limited context of the transcripts. Only the first 100 utterances from each child's session were used. While this kept the data consistent, it may not have been enough to capture meaningful language patterns. Many linguistic metrics, including Zipf's law behavior, become more reliable with larger samples. Zipf's law is a large-scale property of language, and short transcripts may not clearly show this pattern. These results support this limitation. Many transcripts appeared to follow Zipf's law, even for ASD children. De Palma *et al.* reported deviations from Zipf's law in ASD speech, but that analysis may have used larger or aggregated samples (8). At the individual transcript level, any deviation may be too subtle to detect. This suggests that Zipf-based features may require more data per subject to be useful. Future work could test longer transcripts or combine multiple samples from the same child.

The modeling approach also has limitations. Bayesian data augmentation with MCMC sampling was used to generate synthetic training data, combined with a neural network classifier. This assumes that the feature distributions are well estimated, but with so few real samples, this assumption may not hold. The synthetic data may not reflect the true population, and the model may overfit to artifacts in the generated data. Interestingly, prior work using simpler models such as logistic regression reported better performance (3). This suggests that more complex models are not always better, especially with small datasets. Neural networks require large amounts of data, and in this case, the model likely learned noise rather than meaningful patterns. The combined model's accuracy around 38–40% (Figure 3) suggests that it may have defaulted toward weak or biased predictions. Without strong features, the model had no reliable signal to learn from.

Beyond sample size, transcript length, and modeling choices, several broader factors may also explain the difference between these results and earlier studies. First, many successful machine learning studies in autism research use larger datasets or multiple modalities, which may capture signals that a small transcript-only dataset cannot (12). Second, the feature set was limited to a small number of linguistic and Zipf-based features. Real differences in ASD language are likely more complex and may involve pragmatic features such as conversational appropriateness, topic maintenance, or turn-taking patterns beyond length. Prosodic features such as tone, rhythm, and intonation are also not captured in transcript data. These may be important but require different methods, such as audio analysis. Third, the dataset context is very uniform. All children were recorded in similar play sessions, which may limit variation in language use. External factors, such as the behavior of the adult or the structure of the interaction, may introduce noise into features like total word count.

Implications

These limitations point to a broader challenge for speech-based ASD screening because ASD is highly heterogeneous. Within the 16 ASD children, language ability likely varies widely. Some children may have near-typical speech, while others show significant delays or atypical patterns. With such a small sample, this variation makes it difficult for a classifier to learn a consistent signal. One child might show repetitive language, another might have strong vocabulary but pragmatic differences, and another might speak very little. Grouping them together assumes a shared pattern that may not exist at this scale. Larger datasets would allow for subgroup analysis or more personalized modeling.

These results have important implications for speech-based ASD screening. On one hand, they challenge the idea that simple transcript-based features can reliably detect ASD. With short transcripts and basic features, performance remains near chance, which would not be acceptable in a real screening setting. On the other hand, the idea of using speech as a non-intrusive and privacy-preserving signal is still promising. These results suggest that stronger approaches are needed. This may include combining multiple data sources or developing more advanced language features that better capture the

complexity of ASD communication.

One possible direction is the use of more advanced natural language processing methods. These features were relatively simple, focusing on counts, ratios, and distribution properties. More advanced approaches could analyze semantic content, coherence, or topic structure, or use language models to detect subtle differences in usage patterns. For example, embedding-based methods or large language models could capture differences that surface-level features miss. However, these approaches would still require larger datasets to avoid overfitting.

In summary, several key points emerge. First, the results do not support earlier claims of high classification accuracy using simple transcript-based features. Second, both linguistic and Zipfian features appear insufficient for reliable classification in small samples. Third, building a useful ASD screening tool from speech will likely require larger datasets and more comprehensive features, possibly including multiple modalities. Finally, any proposed method must be tested on independent data to ensure it captures a real, generalizable signal rather than dataset-specific patterns. Although the results of this study are negative, they provide a useful check on earlier findings and highlight the work still needed in this area.

CONCLUSION

Early identification of autism from natural speech remains a challenging goal. This study attempted to replicate and extend previous work on automated ASD detection from child speech transcripts by incorporating both traditional linguistic features and Zipf's law-based lexical distribution features. The results did not confirm the high accuracies reported by prior research. Instead, all models performed around chance level when distinguishing 3–6 year-old children with ASD from typically developing peers using transcripts of 100 utterances. The addition of Zipfian features, which were intended to capture differences in word usage distributions, did not improve performance. These features showed no significant differences between ASD and TD groups and were not heavily used by the classifier. The only statistically significant linguistic marker was lower conjunction usage in ASD speech, suggesting simpler sentence structure. However, this single feature was not enough to support reliable classification.

Overall, these findings suggest that the linguistic and Zipfian features examined, at least on short transcripts, are not sufficient for reliable ASD screening on their own. This study highlights the limits of reproducibility from earlier work and the need for larger datasets and more informative features. It may be necessary to combine multiple data sources, such as longer language samples, acoustic or prosodic features, or other behavioral signals, to build a more reliable screening system. While speech remains a promising and non-intrusive source for ASD detection, simple transcript-based analysis did not produce a dependable classifier in this setting. Future work should focus on expanding both the dataset and the feature space. This includes exploring more advanced language modeling techniques, multimodal approaches, and careful validation on independent samples. With more comprehensive efforts, more accurate and objective early screening methods for ASD may become possible. This study highlights both the challenges and the work still needed to reach that goal.

ACKNOWLEDGEMENTS

The author would like to thank Ellie Fassman from Cornell University for guidance during the development of this research paper.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

CODE AND DATA AVAILABILITY

All code for feature extraction, Zipf analysis, and neural-network training, including scripts to reproduce the 100-split experiments and Figures 1-3, is available at: https://github.com/Krishnapabbu17/ASD_Project

REFERENCES

1. World Health Organization. Autism. Geneva: World Health Organization; 2025. Available from: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (accessed 2026-07-12).
2. Interagency Autism Coordinating Committee. Summary of Advances in Autism Research 2020. Bethesda (MD): U.S. Department of Health and Human Services; 2020. Available from: https://iacc.hhs.gov/publications/summary-of-advances/2020/ (accessed 2026-07-12).
3. Ramesh V, Assaf R. Detecting autism spectrum disorders with machine learning models using speech transcripts [preprint]. arXiv. 2021. Available from:

https://arxiv.org/abs/2110.03281 (accessed 2026-07-12).
4. Verbanas P. One-fourth of children with autism are undiagnosed. Rutgers Today. 2020 Jan 9. Available from: https://www.rutgers.edu/news/one-fourth-children-autism-are-undiagnosed (accessed 2026-07-12).
5. Grossard C, Dapogny A, Cohen D, Bernheim S, Juillet E, Hamel F, *et al*. Children with autism spectrum disorder produce more ambiguous and less socially meaningful facial expressions: an experimental study using random forest classifiers. *Mol Autism*. 2020; 11: 5. https://doi.org/10.1186/s13229-020-0312-2.
6. Colonnese F, Di Luzio F, Rosato A, Panella M. Enhancing autism detection through gaze analysis using eye-tracking sensors and data attribution with distillation in deep neural networks. *Sensors (Basel)*. 2024; 24 (23): 7792. https://doi.org/10.3390/s24237792.
7. Eigsti IM. ASDBank English Eigsti Corpus [dataset]. TalkBank. https://doi.org/10.21415/Z0Z7-WW89. Available from: https://talkbank.org/asd/access/English/Eigsti.html (accessed 2026-07-12).
8. De Palma P, Garcia-Camargo LA, Kilfoyle J, VanDam M, Stover J. Speech tested for Zipfian fit using rigorous statistical techniques. *Proc Linguist Soc Am*. 2021; 6 (1): 394–402. https://doi.org/10.3765/plsa.v6i1.4975.
9. Eigsti IM, Bennetto L, Dadlani MB. Beyond pragmatics: morphosyntactic development in autism. *J Autism Dev Disord*. 2007; 37 (6): 1007–1023. https://doi.org/10.1007/s10803-006-0239-2.
10. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. *SIAM Rev*. 2009; 51 (4): 661–703. https://doi.org/10.1137/070710111.
11. Alstott J, Bullmore E, Plenz D. powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS One*. 2014; 9 (1): e85777. https://doi.org/10.1371/journal.pone.0085777.
12. Al Futaisi ND, Schuller BW, Ringeval F, Pantic M. The Noor Project: fair transformer transfer learning for autism spectrum disorder recognition from speech. *Front Digit Health*. 2025; 7: 1274675. https://doi.org/10.3389/fdgth.2025.1274675.