Using Speech Characteristics from Children's Story Narratives to Detect Autistic Tendencies through Deep Learning Methods

Yen Yu Lai
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
tonylai1204@gmail.com

Arbee L.P. Chen
Department of Computer Science and Information Engineering
Asia University
Taichung, Taiwan
arbee@asia.edu.tw

***Abstract*—The number of children diagnosed with Autism Spectrum Disorder (ASD) is continually increasing. However, diagnosing autism is not straightforward; the process is lengthy and complex. Previous research has indicated that children with autism exhibit deficits in using language within social contexts, such as storytelling skills. Additionally, children with autism may display distinct patterns in certain acoustic features compared to typically developing (TD) children. With the advancement of computational models, we aim to employ deep neural network models to rapidly analyze the acoustic features of children's narratives for detecting autism. In this study, we collected narrative data from 12 children with autism and 19 TD children using Module 3 of the standardized tool ADOS-2 (Autism Diagnostic Observation Schedule, Second Edition). We then represented the acoustic features using Mel-Frequency Cepstral Coefficients (MFCCs) and employed computational models for training and classification. Moreover, we identified 10 low-level descriptors (LLDs) in our dataset through t-tests that showed significant differences between ASD and TD children. We combined these 10 LLDs with MFCCs as inputs to the model and achieved an F1 score of 89.4%. Upon further analysis of these 10 LLDs, we found that they indeed represent the speech characteristic differences between ASD and TD children. This further provides interpretability of our model in identifying autistic tendencies based on the speech characteristic differences.**

***Index Terms*—Autism, Storytelling, Speech signal analysis, Deep learning, Model interpretability**

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder characterized by persistent deficits in social interaction and communication, along with restricted and repetitive patterns of behavior, interests, and activities [1]–[3]. ASD presents a wide range of symptoms and severities, therefore described as a spectrum. Some individuals with ASD may have profound intellectual disabilities, while others may have exceptional abilities in specific areas. These symptoms typically manifest in early childhood and can significantly impact an individual's social, educational, and occupational functioning.

According to recent estimates by the World Health Organization, approximately 1 in 100 children worldwide is diagnosed with ASD [4], indicating a substantial public health concern. The prevalence of ASD has been rising over the past decades, which may be attributed to increased awareness, improved diagnostic methods, and broader diagnostic criteria. This growing prevalence underscores the importance of early detection and intervention.

Diagnosing ASD is a complex and time-consuming process that requires a multidisciplinary approach. The diagnostic criteria, as outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [5], involve a thorough assessment of the individual's developmental history and behavior. This assessment typically includes structured interviews, behavioral observations, and standardized tests.

One of the primary challenges in diagnosing ASD is the heterogeneity of its symptoms. No two individuals with ASD are exactly alike, making it difficult to establish a one-size-fits-all diagnostic criterion. Additionally, the overlap of ASD symptoms with other developmental disorders complicates the diagnostic process.

The process is further complicated by the need for trained professionals, including psychologists, neurologists, and speech therapists, to conduct comprehensive evaluations. These evaluations are often resource-intensive, requiring significant time and expertise to accurately diagnose ASD. This complexity contributes to delays in diagnosis, which can postpone critical early interventions that are most effective in improving long-term outcomes for individuals with autism.

A critical area of difference between children with autism and typically developing (TD) children lies in their narrative skills. Narratives are fundamental to effective communication as they enable individuals to share experiences, convey information, and engage in social interactions. The ability to construct and comprehend narratives is crucial for social development and academic success.

Children with autism often exhibit notable deficits in narrative skills compared to their TD peers [6]–[10]. These deficits can manifest in several ways:

- Coherence and Structure: ASD children may struggle with organizing their narratives coherently. Their stories

may lack a clear beginning, middle, and end, and they may omit critical details that are necessary for the listener to understand the context.
- Use of Language: While TD children typically use varied and rich vocabulary to describe events and emotions, ASD children may use more repetitive and simplistic language. They may also have difficulty using pronouns correctly or understanding figurative language.
- Emotional Content: ASD children often find it challenging to incorporate emotional content into their narratives. They may not adequately describe characters' emotions or the emotional significance of events, making their stories less engaging and harder to follow.

In addition to narrative skills, there are significant differences in the speech characteristics of ASD children compared to their TD peers [11]–[14], particularly in areas such as prosody, speech rate, and other aspects of speech production:

- Prosody: Prosody refers to the rhythm, stress, and intonation of speech. ASD children often exhibit atypical prosody, including monotone speech, unusual pitch contours, and irregular stress patterns. These prosodic features make their speech sound robotic or sing-songy, and may impact the listener's ability to understand the emotional tone or intent behind the words.
- Speech Rate: The rate at which speech is produced can also differ between ASD and TD children. Some ASD children may speak more slowly, with frequent pauses and hesitations, while others may speak rapidly, making it difficult for listeners to follow. These variations in speech rate can affect the fluidity and clarity of communication.
- Articulation and Fluency: ASD children may have difficulties with articulation, leading to unclear or imprecise pronunciation of words. They may also exhibit disfluencies, such as stuttering or repetition of sounds, syllables, or words, which can disrupt the flow of speech.
- Language Complexity: While TD children generally progress to using more complex sentence structures as they develop, ASD children may remain at a more basic level of language use. They may rely heavily on simple sentences and exhibit limited use of grammatical constructs such as conjunctions and subordinate clauses.

In our study, we recorded narratives from ASD and TD children about the picture book "Tuesday" [15], a picture book for the storytelling activities in Module 3 of Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [16], and used Mel-Frequency Cepstral Coefficients (MFCCs) [17] as the primary acoustic features to train a classification models for ASD. Through this approach, we successfully developed a highly accurate models. Moreover, we identified 10 low-level descriptors (LLDs) [18] in our speech data that showed significant differences between ASD and TD children. By incorporating these LLDs, an even better-performing model was achieved.

Upon further analysis of these 10 LLDs, we found that they could represent differences in intonation, speech rate, articulation clarity, and language complexity between ASD and TD children. This indicates that our model is capable of detecting whether a child exhibits autism tendencies based on their speech characteristics. Additionally, we compared the impact of different data partitioning methods and data collection processes on the model performance. These performance results as well as our data collection and processing methods can serve as references in the diagnosis of ASD to expedite the overall diagnostic process. Our study's main contributions are as follows:

- We constructed a model capable of detecting autism tendencies from children's narrative speech data, achieving an F1 score of 89.4%.
- We used the storytelling activity from ADOS-2 to collect speech data. This collection process ensured the thematic consistency and quality of the data. The results demonstrated the effectiveness of this data collection method.
- We identified 10 low-level descriptors (LLDs) with significant differences between ASD and TD children. Incorporating these LLDs into the model improved its accuracy. By analyzing these LLDs, we were able to identify speech characteristic differences between ASD and TD children.

The remaining of the paper is organized as follows. In Section 2, we introduce the acoustic features used in our study, as well as related research work. In Section 3, we describe the dataset and the data collection process. Section 4 covers the model architecture, data processing, and feature extraction methods. Section 5 presents the experiments we conducted and their results. Finally, in Section 6, we conclude our work and suggest future work.

II. Related Work

In this section, we introduce the audio features used in our study and present related work that classified ASD using picture book narratives as well as speech recordings.

A. Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used features in the field of speech and audio processing. MFCCs capture the power spectrum of a sound signal in a way that approximates human auditory perception, making them particularly effective for various speech-related tasks.

The MFCC algorithm involves several steps to transform a raw audio signal into a set of coefficients that can be used for further processing [17]. These steps include:

1) Pre-emphasis: A high-pass filter is applied to the audio signal to amplify high-frequency components, which helps balance the spectrum of the signal and makes it more robust to noise.
2) Framing: The continuous audio signal is divided into overlapping frames, typically of 20-40 milliseconds in length, with a common overlap of 50%. This step allows for the analysis of short segments of the signal, which are assumed to be stationary.

3) Windowing: Each frame is multiplied element-wise with a window function to reduce *spectral leakage*, which occurs when signal discontinuities at the edges of a frame cause energy to spread across multiple frequencies, to smooth the edges of the frame.
4) Fast Fourier Transform (FFT): The windowed frames are transformed from the time domain to the frequency domain using FFT, producing a spectrum for each frame.
5) Mel Filter Bank: The frequency spectrum is passed through a series of filters spaced according to the *Mel scale*, which approximates the human ear's response to different frequencies. This step results in a set of Mel-scaled power spectrum coefficients.
6) Logarithm: The logarithm of the Mel-scaled power spectrum is taken to simulate the logarithmic perception of loudness by the human ear.
7) Discrete Cosine Transform (DCT): The log Mel spectrum is then transformed using DCT to produce the Mel-Frequency Cepstral Coefficients. The first few coefficients are retained as they represent the most significant features of the spectrum.

MFCCs remain a cornerstone in the field of speech and audio processing, providing a reliable and efficient way to extract meaningful features from audio signals. Their continuous use and development highlight their importance and effectiveness in a wide range of applications.

B. Low Level Descriptors

Low-Level Descriptors (LLDs) are fundamental features used in speech and audio processing to characterize various aspects of audio signals. These descriptors capture essential properties of the signal at a low level, providing a basis for more complex analysis. LLDs are crucial in various applications, including speech recognition, speaker identification, emotion recognition, and audio classification.

LLDs are extracted from short frames of the audio signal, typically ranging from 20 to 40 milliseconds in length. These frames are then analyzed to extract features that describe the basic properties of the signal within each frame. Commonly used LLDs include:

- Energy and Loudness: Measure the power or intensity of the audio signal. Energy is calculated as the sum of the squared amplitude values within a frame, while loudness is a perceptual measure that accounts for the human ear's sensitivity to different frequencies.
- Pitch (Fundamental Frequency): Represents the perceived frequency of the sound, corresponding to the rate of vocal fold vibrations in voiced speech. Pitch is crucial for identifying intonation patterns and speech melody.
- Formants: Represent the resonant frequencies of the vocal tract that shape the speech sound. The first three formants (F1, F2, F3) are particularly important for distinguishing vowel sounds.
- Harmonics-to-Noise Ratio (HNR): Measures the ratio of harmonic (periodic) components to noise (aperiodic) components in the signal, providing an indication of voice quality.
- Zero-Crossing Rate (ZCR): Represents the rate at which the signal changes sign (crosses the zero amplitude line) within a frame. ZCR is associated with the noisiness or spectral texture of the signal.
- Spectral Features: Include various measures of the signal's frequency content, such as spectral centroid (center of mass of the spectrum), spectral bandwidth (spread of the spectrum), spectral roll-off (frequency below which a certain percentage of the total spectral energy is contained), and spectral flatness (tonal vs. noise-like quality of the signal).

Cho et al. [19] collected data from 70 children, including 35 ASD children and 35 TD children, through a 5-minute "get-to-know-you" conversation with young adult confederate. The conversation data was divided into audio features and textual features, with the audio features being extracted using low-level descriptors (LLDs). Ultimately, using both audio and textual features, an accuracy of 0.76 was achieved with a Gradient Boosting Model [20].

C. Classifying Autism from Picture Book Narratives

Sun et al. [21] collected narratives from 7 ASD children and 16 TD children, using two picture books with different styles. These narratives were then transcribed into text and used to train three different computational models. The study incorporated language features such as verbal productivity and word usage. With these features, the TinyBERT [22] model achieved accuracy, sensitivity, and specificity rates all exceeding 90%. This result indicates significant differences in narrative abilities between ASD children and TD children.

Based on Sun et al.'s findings, we hypothesized that using the same data collection method but instead training the model with acoustic features from the narrative data might also yield excellent results. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) can capture subtle variations in speech signals, which may offer unique advantages in analyzing the language patterns of ASD children. Furthermore, analyzing acoustic features can provide perspectives and insights different from those based on text features, therefore enriching our research methodology and result interpretation.

D. Classifying Autism from Speech Recordings

Chi et al. [23] collected videos from 20 ASD children and 38 TD children using "Guess What?", a mobile game designed for prosocial play and interaction at home between 2-8 years old children and their parents. The children's voices were manually sampled from each video. Ultimately, 850 audio clips were obtained, including 425 clips from ASD children and 425 clips from TD children.

To prevent the models from learning from the individual speaking characteristics of the children, a speaker-independent data splitting method was implemented. The data was divided into 5 folds to ensure that clips from the same child appeared only in the same fold. This approach prevents the models

from encountering the same child's voice during both training and testing, therefore enhancing the models' generalization capability.

Additionally, spectrograms were used to provide a two-dimensional representation of the audio signal in terms of frequency and time, allowing the model to capture richer audio features. A Convolutional Neural Network (CNN) was then used to train a classification model. Through a series of experiments, the best accuracy of 0.793 was achieved.

III. Dataset

We utilized the narrative data collected in Sun et al.'s study. In the study, 7 ASD children and 16 TD children were recruited. Among these, one ASD child was moderately affected and one severely affected. These two children showed significant differences in speech compared to other children. Since our research focuses on identifying mildly autistic children who do not exhibit significant differences from TD children, we excluded the data from these two children
. Additionally, Sun et al.'s study faced issues of a small and imbalanced dataset between the number of ASD and TD children. To address this, we collected additional data from 7 ASD children and 3 TD children for our study to balance the dataset and enhance the reliability of our models.

Moreover, during this data collection, we made more efforts to guide the children by asking them questions related to the picture book content to encourage them to speak more. Specifically, in Sun et al.'s study, children only needed to roughly narrate the story, with guidance provided only if they spoke too little. In our data collection, we asked children about the emotions of characters and the possible reasons for events in the story to obtain more insights into their understanding. For example, if the story depicted a dog chasing a frog, we would ask the children why the dog was chasing the frog and what the emotions of the dog and frog might be at that moment.

Subsequently, we segmented the speech data of the total 12 ASD children and 19 TD children into sentences, effectively removing the examiner's voice and ensuring that only the children's voices remained. A small portion of the dataset with overlapping voices of the children and examiner was also removed.

In the end, we obtained a total of 672 clips for the ASD children and 618 clips for the TD children. Table 1 gives the mean and standard deviation number of the clips per child and the mean and standard deviation duration per clip for both ASD and TD children, while Table 2 details the data collected in the first (Sun et al. [21]) and second phases. As shown in Table 2, the number of clips obtained from each child in this collection is significantly higher than that in the first collection. This increase is due to the additional guidance provided during this data collection process, which effectively achieved our goal of increasing the data volume.

As shown in Table 2, the number of clips obtained from each child in this collection is significantly higher than that in the first collection. This increase is due to the additional guidance provided during this data collection process, which effectively achieved our goal of increasing the data volume.

TABLE I
THE MEAN AND STANDARD DEVIATION (SD) NUMBER OF CLIPS PER CHILD AND THE MEAN AND STANDARD DEVIATION (SD) DURATION PER CLIP FOR THE ASD AND TD CHILDREN (DURATION UNIT: SECONDS)

	TD	ASD
Mean number of clips (SD)	34.3 (20.2)	56 (39)
Mean duration per clip (SD)	2.81 (1.55)	2.23 (1.53)

TABLE II
THE MEAN AND STANDARD DEVIATION (SD) NUMBER OF CLIPS PER CHILD AND THE MEAN AND STANDARD DEVIATION (SD) DURATION PER CLIP FOR THE ASD AND TD CHILDREN (DURATION UNIT: SECONDS)

	TD in Sun et al.'s collection	ASD in Sun et al.'s collection	TD in this collection	ASD in this collection
Mean number of clips per child (SD)	27.7 (10.29)	20.8 (8.5)	67.3 (24.7)	81.1 (32.2)
Mean duration per clip (SD)	2.8 (1.41)	3.21 (1.65)	2.84 (1.8)	2.05 (1.4)

IV. Methods

In this section, we introduce the structure of our study, including data preprocessing, acoustic features, model selection, and other techniques used.

A. Data Preprocessing

Signal denoising and enhancement are crucial steps not only in speech tasks but also in image tasks. These processes eliminate noise that may affect the model performance and enhance the key features the model needs to focus on, thereby improving the overall performance and accuracy of the model. Signal denoising significantly reduces background noise and unnecessary interference, making the input signal clearer and purer. This is especially important for automatic speech recognition, speech synthesis, and other speech processing tasks. Signal enhancement focuses on improving the quality of the signal. For example, in speech processing, the speech enhancement techniques can improve the audio quality, clarity, and naturalness, making the generated speech easier to understand and communicate.

In our study, we utilized ESPnet [24] to perform signal denoising and speech enhancement on the raw data. ESPnet is an open-source, end-to-end speech processing tool based on deep learning technologies. In our task, signal denoising primarily involves removing non-speech background noise. Despite our efforts to minimize the background noise during data collection, it was challenging to avoid noises like page-turning sounds from children reading storybooks, which was

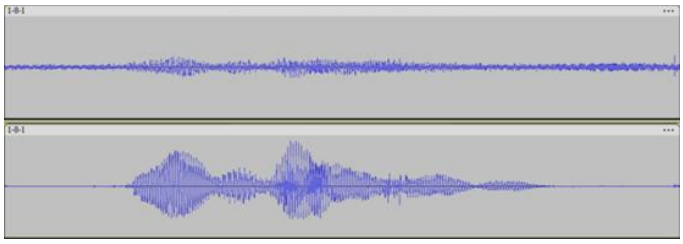

Fig. 1. Before preprocessing (Top) and after preprocessing (Bottom)

the most common noise in our data.

Speech enhancement aims to make the participants' spoken words more distinct and clearer. We leveraged ESPnet's speech enhancement technology to make the children's speech in the storytelling more prominent, thereby improving the accuracy of the subsequent model analysis.

As shown in Fig. 1, the original audio before denoising and enhancement contains page-turning sounds at the beginning and end, while the middle part features the child's speech. After processing, the audio shows almost no signal at the beginning and end, indicating the noise was successfully eliminated. Meanwhile, the middle part has become much clearer, demonstrating the speaker's voice was successfully enhanced.

Overall, the data preprocessing steps ensured high standards of quality and consistency in our data, providing a solid foundation for the subsequent training of the models.

B. Data Augmentation

Due to the limited amount of data available, data augmentation is essential to improve the robustness of the model and reduce the risk of overfitting. In our task, we applied additive Gaussian white noise [25] to our training dataset, a common data augmentation technique in speech tasks.

Specifically, we randomly selected a signal-to-noise ratio (SNR) between 15 and 30 for the noise addition. This approach aims to simulate varying levels of background noise found in real-world environments, ensuring that the model performs well under diverse noisy conditions. After this augmentation, the size of the training dataset was doubled, significantly enhancing the model's generalization ability.

C. Acoustic Features

In this subsection, we introduce the acoustic features we use, including Mel-Frequency Cepstral Coefficients (MFCCs) and Low-Level Descriptors (LLDs). Among these, MFCCs serve as the primary features, while LLDs the external knowledge to train our models. We experimentally verified whether adding LLDs improves the models' performance.

1) MFCCs: As previously mentioned, MFCCs are widely used features in speech-related tasks. In our study, we utilized 20-channel MFCCs to extract acoustic features, with a frame length of 25 milliseconds and an overlap of 10 milliseconds. We chose these parameter settings based on the experiment results, aiming to capture essential features of the speech signal while minimizing unnecessary noise and variability. The 20-channel MFCCs provide sufficient spectral resolution, and the selected frame length and overlap strike a balance between temporal resolution and computational efficiency.

These features not only enhance the classification performance of our models but also effectively reflect the speech characteristics and variations in children's voices, allowing us to more accurately analyze their speech patterns.

2) LLDs: We used the OpenSMILE ComParE16 configuration [26], setting the frame length to 25 milliseconds and the overlap length to 10 milliseconds, to extract a total of 65 LLDs. We performed independent sample t-tests on these LLDs for ASD and TD groups to identify which LLDs showed significant differences between ASD and TD children. The final results indicated that the following 10 LLDs showed significant differences:

1) *F0final_sma*
2) *jitterLocal_sma*
3) *jitterDDP_sma*
4) *shimmerLocal_sma*
5) *logHNR_sma*
6) *pcm_RMSenergy_sma*
7) *pcm_fftMag_spectralEntropy_sma*
8) *pcm_fftMag_spectralFlux_sma*
9) *pcm_fftMag_spectralRollOff25.0_sma*
10) *pcm_fftMag_spectralRollOff50.0_sma*

The specific meanings of these LLDs will be presented in the subsequent section of experiments.

D. Embedding Models

In this subsection, we introduce the embedding models used in our study, namely Time-Delay Neural Network (TDNN) [27] and Long Short-Term Memory (LSTM) networks [28]. We chose these models because they both perform exceptionally well in handling time-series data and are commonly used in speech-related tasks such as automatic speech recognition, speaker identification, and speaker verification. In the following, we provide an overview of each model used in our experiments.

TDNN is a feedforward neural network that processes sequential data through convolution operations applied at different time steps. Unlike traditional CNNs, which perform convolution along the spatial dimensions, TDNN performs convolution along the time dimension, enabling it to capture temporal features effectively.

The key feature of TDNN is its ability to process inputs from multiple time steps simultaneously, allowing the model to learn time-dependent patterns. This is achieved by applying a time delay to the inputs, which effectively spreads them over several time frames.

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) designed to address the limitations of traditional RNNs in modeling long-term dependencies. LSTMs have become a foundational component in various sequential data processing tasks, particularly in the fields of natural language processing (NLP) and speech

recognition. We employed a CNN-LSTM architecture to perform the experiments, i.e. before the LSTM, we used a CNN to extract local features from the speech signal. The LSTM then captures the temporal dependencies of these features. Additionally, our convolutional blocks utilize residual blocks with skip connections, which helps accelerate training and improves the model's accuracy.

V. Experiments

In this section, we present six experiments on 1) the performance of various classifiers, 2) the impact of different data partitioning strategies for the models, 3) the effect of different data augmentation techniques, 4) the model's performance with data collected through different processes, 5) the performance of the model after incorporating LLDs, and 6) a comparison with previous studies.

A. Performance of Different Methods

In the first experiment, we divided all the data into five folds and employed two widely used machine learning methods for binary classification tasks: Support Vector Machine (SVM) [29] and Random Forest [30]. These methods are commonly applied to speech-related tasks [31]. In addition to these traditional machine learning techniques, we compared the deep learning models, including Time-Delay Neural Network (TDNN) and Long Short-Term Memory (LSTM) networks. Beyond the standard TDNN, we also evaluated the ECAPA-TDNN [32], an extension of TDNN known for its excellent performance in speaker recognition. Similarly, besides the standard LSTM, we examined the LSTM enhanced with an attention mechanism. The attention mechanism helps the model focus on crucial parts of the input sequence, leading to a better performance in time-series tasks than standard LSTM [33]. We conducted 5-fold cross-validation on the methods, and the results are shown in Table 3.

The results indicate that traditional machine learning methods do not achieve satisfactory performance, while more complex deep learning models such as ECAPA-TDNN and LSTM with attention yield better results. Under basic architectures, TDNN outperforms LSTM. However, due to this data partitioning method, it is possible for speech segments from the same child to appear in both the training and testing sets. Therefore, we cannot guarantee the models' prediction results and further consider the data partitioning strategy as follows.

TABLE III
The results of the first experiment

	Sensitivity	Specificity	Accuracy	F1 Score
SVM	0.815	0.863	0.838	0.840
Random Forest	0.756	0.806	0.780	0.782
TDNN	0.917	0.914	0.901	0.902
ECAPA-TDNN	0.941	0.927	0.934	0.937
LSTM	0.879	0.876	0.863	0.865
LSTM with attention	0.919	0.911	0.915	0.919

B. Performance under Speaker-independent Data Partitioning

Previous research [34] has indicated that in speaker verification tasks, model tends to overfit to known speaker characteristics, leading to decreased performance when predicting unknown speakers. A similar issue arises in our task, which is undesirable for real-world applications. To address this problem, we implemented a speaker-independent data partitioning method in this experiment to ensure that all clips from the same child appear in the same fold. This approach guarantees that during cross-validation, clips from the same child do not appear in both the training and testing sets.
Since the ratio of the numbers of TD clips and ASD clips in our dataset is approximately 9:10, we maintained this proportion in each fold in the speaker-independent data partitioning. Table 4 shows the distribution of ASD and TD clips across the 5 folds.

Based on the results of the first experiment, traditional machine learning methods did not achieve satisfactory performance. Therefore, in this experiment, we focused solely on comparing deep learning models, namely TDNN, ECAPA-TDNN, LSTM, and LSTM with attention.
Table 5 shows the performance of the four models under speaker-independent data partitioning with 5-fold cross-validation. The results indicate that with this change, more complex models performed worse. This suggests that, given our task and the data size, overly complex models are not suitable. Combining this with the results from the first experiment, we can infer that the better performance of ECAPA-TDNN and LSTM with attention in the first experiment was due to their ability to capture subtle features, specifically those related to speakers. The data partitioning in the first experiment allowed segments from the same speaker to appear in both training and testing sets, enabling these models to leverage speaker differences for outstanding results, which indicates an overfitting to known speaker characteristics. In the comparison between TDNN and LSTM, we observed a significant performance drop for TDNN, indicating that it also relied on speaker differences in the first experiment. However,

TABLE IV
The distribution of ASD and TD clips across the 5 folds

	Fold1	Fold2	Fold3	Fold4	Fold5
TD	123	121	122	126	126
ASD	136	133	134	131	138

TABLE V
THE PERFORMANCE OF TDNN AND LSTM UNDER SPEAKER-INDEPENDENT DATA PARTITIONING, EVALUATED USING 5-FOLD CROSS-VALIDATION

	Sensitivity	Specificity	Accuracy	F1 Score
TDNN	0.798	0.809	0.804	0.809
ECAPA-TDNN	0.754	0.781	0.767	0.771
LSTM	0.822	0.826	0.824	0.828
LSTM with attention	0.791	0.787	0.789	0.796

LSTM's performance did not significantly decrease across both experiments, suggesting that LSTM made decisions based on speech characteristic differences between ASD and TD children. Therefore, LSTM is more suitable for practical applications and aligns better with the objective of our experiment. Compared to TDNN, LSTM is more adept at capturing longer temporal changes, implying that the speech characteristic differences between ASD and TD children require analysis over longer time.

This finding highlights the importance of selecting an appropriate model. In our tasks, it is crucial for the model to adapt to different speaker characteristics and capture long-term speech feature changes to achieve optimal classification performance. To ensure that the experiment results reflect real-world applications, we used the speaker-independent data partitioning and the LSTM model, which produced the best results in this experiment, in all subsequent experiments.

C. Impact of Different Data Augmentation Techniques

In this experiment, we compared the impact of different data augmentation techniques on model's performance. In addition to the additive Gaussian white noise, we employed two other commonly used data augmentation techniques for speech tasks: *speed perturbation* [35] and *frequency masking* [36]. The experiment results are shown in Table 6. From these results, we can see that only the addition of additive Gaussian white noise improved the model's performance.

Speed perturbation alters the speaking rate, and previous research indicates that ASD and TD children exhibit significant differences in speaking speed. Therefore, using speed perturbation as a data augmentation technique may disrupt this important feature, resulting in decreased model's performance. Frequency masking randomly removes or attenuates certain frequency components of the speech signal. Since ASD and TD children have significant differences in frequency fluctuations and range, frequency masking potentially disrupting the differences between ASD and TD children, thereby reducing model effectiveness.

The results of this experiment showed that using only additive Gaussian white noise as a data augmentation method produced the best results. Therefore, in all subsequent experiments, we applied data augmentation exclusively by adding additive Gaussian white noise.

D. Performance under Different Data Collection Phases

Our data collection was divided into two phases. In the first phase, we did not provide much guidance to the children. In the second phase, we attempted to guide the children more, asking them questions about the storybook to encourage them to speak more about its content.

In this experiment, we separated the data collected from these two phases and trained LSTM models on each set to compare whether the amount of guidance given to the children affects the model's performance. Specifically, we treated the data from the first phase as one group and the data from the second phase as another group, then trained and tested the models on these two groups separately. The data distribution is shown in Table 7. This design helps us understand whether providing more guidance during data collection can improve the model's accuracy and robustness. By comparing the results, we hope to draw valuable conclusions about whether more guidance should be employed in future data collection efforts.

The results of this experiment are shown in Table 8. From the results, it can be seen that under different evaluation metrics, these two phases have their own strengths and weaknesses, making it difficult to conclude which phase is more beneficial for the model. This outcome may be due to the imbalance in the number of ASD and TD children in both phases, with the first phase having more data from the TD children and the second collection having more data from ASD children.

However, training with data collected using a single method yields better results than mixing data from the two phases. This suggests that in future data collections, adopting a consistent guidance strategy may be more beneficial for the model's performance. Additionally, it highlights the importance of

TABLE VI
THE COMPARISON OF MODEL'S PERFORMANCE WITH DIFFERENT DATA AUGMENTATION TECHNIQUES

	Sensitivity	Specificity	Accuracy	F1 Score
Without data augmentatiom	0.799	0.801	0.800	0.806
Adding additive Gaussian white noise	0.822	0.826	0.824	0.828
Speed perturbation	0.787	0.788	0.788	0.794
Frequency masking	0.772	0.785	0.778	0.784

maintaining a balance between data from the ASD and TD children.

E. Performance Incorporating LLDs

During our research, we observed several differences in the raw speech data collected from TD and ASD children. These differences included variable speech rates, larger fluctuations in volume and pitch, less clear articulation, and more frequent pauses in the speech of ASD children. These variations may serve as important features for our model's decision-making process. In this experiment, we incorporated low-level descriptors (LLDs), which effectively represent the fundamental characteristics of speech. We trained the model by augmenting the original 20-channel MFCCs input with the 10 significant LLDs we identified through t-tests. The results are shown in Table 9.

The results indicated that incorporating LLDs with significant differences between ASD and TD considerably enhances the model's performance. Next, we gave the meaning of these LLDs to identify the speech differences between ASD and TD children, and then discussed how they practically affect speech characteristics.

1) *F0final_sma*
   - Fundamental frequency, representing the primary frequency component of speech.
   - The significant differences in the fundamental frequency between ASD and TD children indicate that there are many differences in their speech characteristics.
2) *jitterLocal_sma*
   - Local jitter, representing variations between consecutive periods of the fundamental frequency.
   - ASD children show higher jitterLocal, indicating that their fundamental frequency is more unstable, which also suggests that their pitch has greater fluctuations. This could be related to the speech control ability in ASD children. Additionally, difficulties in social communication may affect the stability and consistency of speech, which leads to the higher jitter observed in ASD children.
3) *jitterDDP_sma*
   - DDP jitter, representing relative variations between fundamental frequency periods. Compared to jitterLocal, jitterDDP captures subtler variations in the fundamental frequency.
   - ASD children show higher jitterDDP, which may reflect poorer fine control of voice production in ASD children.
4) *shimmerLocal_sma*
   - Local shimmer, representing variations in amplitude between consecutive sound periods.
   - ASD children show higher shimmer, indicating that the amplitude of their speech is more unstable. This also suggests that their volume has greater fluctuations and they have more pauses where the amplitude approaches zero. This could be related to the control of speech intensity and breath support in ASD children.
5) *logHNR_sma*
   - Log harmonics-to-noise ratio, representing the ratio of harmonic components to noise components and is commonly used to assess speech clarity and quality.
   - ASD children show lower HNR, indicating their speech contains more noise components and fewer harmonic components. This may reflect lower speech clarity and quality in ASD children.
6) *pcm_RMSenergy_sma*
   - Short-time energy, representing the energy of the speech signal within a short time window.
   - ASD children show lower RMSenergy, indicating their intensity of speech is generally lower. This may be due to the fact that ASD children have more pauses while speaking, which results in a lower overall energy level in the speech signal. It could be due to their lack of emotional variation or difficulties in social interaction.
7) *pcm_fftMag_spectralEntropy_sma*
   - Spectral entropy, representing the randomness or complexity of the spectral distribution.

TABLE VII
THE DATA DISTRIBUTION BETWEEN THE FIRST PHASE AND THE SECOND PHASE

	TD for training	ASD for training	TD for testing	ASD for testing
First phase	332 clips	82 clips	84 clips	22 clips
Second phase	156 clips	452 clips	46 clips	116 clips

TABLE VIII
THE COMPARISON OF THE MODEL'S PERFORMANCE UNDER DIFFERENT PHASES

	Precision	Sensitivity	Specificity	Accuracy	F1 Score
First phase	0.800	0.909	0.940	0.934	0.851
Second phase	0.932	0.828	0.848	0.833	0.877

TABLE IX
THE MODEL'S PERFORMANCE INCORPORATING LLDS

	Sensitivity	Specificity	Accuracy	F1 Score
w/o LLDs	0.822	0.826	0.824	0.828
w/ LLDs	0.886	0.896	0.891	0.894

- ASD children show higher SpectralEntropy, indicating their speech spectrum is more complex or random. This may be caused by the presence of more noise in the speech signals of ASD children. This may be related to the difficulties ASD children have in controlling their articulation and regulating their speech.

8) *pcm_fftMag_spectralFlux_sma*
   - Spectral flux, representing the degree of change in the spectrum over time.
   - ASD children show lower SpectralFlux, which may reflect less dynamic variation at the speech level, such as monotonous intonation and rhythm, lacking inflection.
9) *pcm_fftMag_spectralRollOff25.0_sma*
10) *pcm_fftMag_spectralRollOff50.0_sma*
   - 25% (50%) spectral roll-off point, representing the frequency position of the first 25% (50%) of the spectral energy.
   - ASD children show lower Mag_spectralRollOff25.0 (50.0), indicating fewer high-frequency components. When high-frequency components of speech are missing, it can significantly reduce the clarity and intelligibility of the speech, making it sound blurred and unclear [37].

From these LLDs, we can analyze the differences in speech between ASD and TD children. The results suggest that ASD children may have more monotonous speech, poorer fine control over their voice production, lower speech clarity and quality, greater fluctuations in pitch and volume, and more pauses in their speech. These differences align with our observations of the raw data and are consistent with the findings of previous studies that used traditional methods, such as statistical analysis, questionnaires [11]–[14]. The experiment results also demonstrate that these differences significantly aid the model's decision-making process.

F. Comparison with Previous Studies

In this experiment, we compare our results with previous studies. Table 10 presents a comparison between the results of our method and Sun et al. [21]. Although our results are slightly lower than Sun et al.'s, we both achieved commendable performance, demonstrating the success of our data collection method using the ADOS-2 storytelling task. This not only proves effective for text-based inputs but also confirms that using speech signals can also achieve good results. This flexibility allows us to choose the type of data based on the purpose or direction of the analysis in the future. Table 11 shows a comparison between our study and the research conducted by Cho et al. [19] and Chi et al. [23]. Our results outperform these studies in all aspects. We believe this is due to the thematic nature of our collected data, which enables the model to better predict based on the differences in speech between ASD and TD children. Additionally, we used the LSTM model, which is better at capturing temporal changes, and incorporated LLDs that show significant differences between ASD and TD children as external knowledge combined with MFCCs as input. These factors contributed to our model's superior performance.

TABLE X
OUR RESULT COMPARED WITH THE RESULT FROM SUN ET AL.

	Sensitivity	Specificity	Accuracy	F1 Score
Sun's	0.93	0.91	0.92	0.91
Ours	0.886	0.896	0.891	0.894

TABLE XI
OUR RESULT COMPARED WITH THE RESULTS FROM CHO ET AL. AND CHI ET AL.

	Sensitivity	Precision	Accuracy	F1 Score
Cho et al., 2019	0.76	0.76	0.76	0.76
Chi et al., 2022	0.793	0.804	0.793	0.790
Ours	0.886	0.896	0.891	0.894

VI. CONCLUSION AND FUTURE WORK

In this study, we used speech data from 12 ASD children and 19 TD children who narrated the story after reading the book "Tuesday." We first converted the speech data into MFCCs and then used a CNN-LSTM model to achieve an excellent performance in a speaker-independent data partition setting. Moreover, we identified 10 LLDs that showed significant differences between ASD and TD children. By using these 10 LLDs together with MFCCs as inputs to train the model, we achieved an F1 score of 0.894. These results indicate that our model successfully predicts based on the speech differences between ASD and TD children. Moreover, our results are comparable to those of a previous study that used the same data collection method but with text-based inputs, demonstrating the success of our data collection process. This suggests that in future practical applications, different data types can be chosen for training the model depending on the analysis purpose. We can also use both text and speech data as inputs to form a multimodal approach for the classification. This approach could potentially combine the features of ASD present in both speech and text, leading to better classification results.

Although our model showed satisfactory predictive results, our experiment to determine whether the amount of guidance given to the children during the data collection process affects the model's performance did not yield clear results. This is due to the data imbalance between the ASD and TD children in both data collection phases. If more data can be collected using this data collection method in the future to address the data imbalance issue, we believe a more definitive answer can be obtained. This would allow us to decide whether to provide children with appropriate guidance during the data collection

process in the future, making our data collection process more comprehensive and standardized.

REFERENCES

[1] D. American Psychiatric Association, D. American Psychiatric Association *et al.*, *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association Washington, DC, 2013, vol. 5, no. 5.

[2] C. Lord, S. Risi, P. S. DiLavore, C. Shulman, A. Thurm, and A. Pickles, "Autism from 2 to 9 years of age," *Archives of general psychiatry*, vol. 63, no. 6, pp. 694–701, 2006.

[3] S. L. Hyman, S. E. Levy, S. M. Myers, D. Z. Kuo, S. Apkon, L. F. Davidson, K. A. Ellerbeck, J. E. Foster, G. H. Noritz, M. O. Leppert *et al.*, "Identification, evaluation, and management of children with autism spectrum disorder," *Pediatrics*, vol. 145, no. 1, 2020.

[4] J. Zeidan, E. Fombonne, J. Scorah, A. Ibrahim, M. S. Durkin, S. Saxena, A. Yusuf, A. Shih, and M. Elsabbagh, "Global prevalence of autism: A systematic review update," *Autism research*, vol. 15, no. 5, pp. 778–790, 2022.

[5] F. Edition *et al.*, "Diagnostic and statistical manual of mental disorders," *Am Psychiatric Assoc*, vol. 21, no. 21, pp. 591–643, 2013.

[6] M. Losh and L. Capps, "Narrative ability in high-functioning children with autism or asperger's syndrome," *Journal of autism and developmental disorders*, vol. 33, pp. 239–251, 2003.

[7] L. Capps, M. Losh, and C. Thurber, ""the frog ate the bug and made his mouth sad": Narrative competence in children with autism," *Journal of abnormal child psychology*, vol. 28, pp. 193–204, 2000.

[8] L. Colle, S. Baron-Cohen, S. Wheelwright, and H. K. Van Der Lely, "Narrative discourse in adults with high-functioning autism or asperger syndrome," *Journal of autism and developmental disorders*, vol. 38, pp. 28–40, 2008.

[9] R. Landa, "Social language use in asperger syndrome and high-functioning autism," *Asperger syndrome*, vol. 18, pp. 125–155, 2000.

[10] T. L. Hutchins, P. A. Prelock, and W. Chace, "Test-retest reliability of a theory of mind task battery for children with autism spectrum disorders," *Focus on autism and other developmental disabilities*, vol. 23, no. 4, pp. 195–206, 2008.

[11] R. Paul, A. Augustyn, A. Klin, and F. R. Volkmar, "Perception and production of prosody by speakers with autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 35, pp. 205–220, 2005.

[12] S. Peppe´, J. McCann, F. Gibbon, A. O'Hare, and M. Rutherford, "Receptive and expressive prosodic ability in children with high-functioning autism," 2007.

[13] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and asperger syndrome," 2001.

[14] I.-M. Eigsti, L. Bennetto, and M. B. Dadlani, "Beyond pragmatics: Morphosyntactic development in autism," *Journal of autism and developmental disorders*, vol. 37, pp. 1007–1023, 2007.

[15] D. Wiesner, *Tuesday*. Clarion Books, 1991.

[16] C. Lord, "Autism diagnostic observation schedule," *(No Title)*, 1999.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[18] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[19] S. Cho, M. Liberman, N. Ryant, M. Cola, R. T. Schultz, and J. Parish-Morris, "Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations." in *Interspeech*, 2019, pp. 2513–2517.

[20] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[21] R. Sun, J. Wong, E. Chen, and A. Chen, "Using computational models to detect autistic tendencies for children from their story book narratives," National Tsing Hua University, Technical Report, 2024.

[22] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[23] N. A. Chi, P. Washington, A. Kline, A. Husic, C. Hou, C. He, K. Dunlap, and D. P. Wall, "Classifying autism from crowdsourced semistructured speech recordings: machine learning model comparison study," *JMIR pediatrics and parenting*, vol. 5, no. 2, p. e35406, 2022.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[26] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[27] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural net- work architecture for efficient modeling of long temporal contexts." in *Interspeech*, 2015, pp. 3214–3218.

[28] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[29] C. Cortes, "Support-vector networks," *Machine Learning*, 1995.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[31] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms," in *2019 Southern African Universi- ties Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*. IEEE, 2019, pp. 141–146.

[32] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[33] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[34] J.-W. Jung, H.-S. Heo, I. Yang, H.-J. Shim, and H.-J. Yu, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," *extraction*, vol. 8, no. 12, pp. 23–24, 2018.

[35] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *Interspeech*, vol. 2015, 2015, p. 3586.

[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[37] E. Jacewicz, J. M. Alexander, and R. A. Fox, "Introduction to the special issue on perception and production of sounds in the high-frequency range of human speech," *The Journal of the Acoustical Society of America*, vol. 154, no. 5, pp. 3168–3172, 2023.