Investigating Autism Patient Social Interaction Monitoring and Evaluation Through Audio Analysis Using Wearable Technology

***Abstract*— This study presents a novel, automated framework that leverages wearable technology and speaker diarization algorithm to objectively measure speaking time as a proxy for sociability, addressing the challenge of subjective data collection and analysis in assessing social interactions of individuals with autism spectrum disorder (ASD). A smartwatch app was developed to collect audio, heart rate in naturalistic settings to extract the patient's total speaking time and further context of their social patterns. Results from the algorithm were compared to manually labeled voice data as our ground truth to calculate accuracy. Clinical testing was done with 20 participants with ASD ranging from ages 7-24. Analysis of over 350 hours of voice data confirmed the algorithm's high accuracy in clear audio conditions, with minimal deviation from ground truth. Contextual data like heart rate provide additional insights into engagement patterns. Amazon Web Services (AWS) infrastructure was utilized for data processing and analysis to ensure reliable handling of the objective data. Patient data was visualized through a dashboard where physicians can potentially gain actionable insights into social development patterns. This multimodal, data-driven approach offers a scalable method to monitor social interaction in ASD, supporting personalized treatment planning and enhancing ASD behavioral research reliability.**

***Index Terms*— Autism Spectrum Disorder, Speaker Diarization, Social Communication, Wearable Technology, Machine Learning, Objective Measurement.**

I. Introduction

Autism Spectrum Disorder (ASD) is a developmental disability that presents unique challenges in how an individual assesses social interaction, communication, and behavioral patterns. Children with ASD often face significant learning challenges due to cognitive limitations. This includes sensory perception, information processing, attention span and memory retention, in particular attention playing a crucial role in determining learning success [1]. ASD is prevalent, with 1 in 36 children being diagnosed with ASD. This relatively high prevalence highlights the importance of ASD research, assessment tools and interventions [2]. Current assessments in ASD predominantly rely on subjective measures like questionnaires, interviews, or observational reports from the caregiver. These assessments introduce subjectiveness and bias that could complicate accurate treatment [3]. This issue is further exacerbated in the relationship between the caregiver, child and physician as the caregiver is not able to be present in the child's everyday social interaction, including educational settings, after-school activities, or other peer gatherings, leading to a gap in information to assess social development. This leads to difficulty for the physician to track meaningful social development over time and give accurate treatment recommendations, often leading to inconsistencies in evaluations, treatment plans, and outcome assessments [4]. Given the high prevalence of ASD, the development of objective measurement tools capable of collecting quantifiable data is urgently needed. Such data are crucial for enabling more accurate and objective analyses to better our understanding of ASD and inform treatment decisions.

To address the challenges of subjective self-reporting in social development among individuals with ASD, we developed an objective, technology-based strategy to capture and analyze real-time social interactions. This system integrates wearable technology with speaker diarization algorithms. Through a smartwatch, participants record their real-life conversations, which are analyzed to detect the user's speech segments and measure total speaking duration. This voice data, along with other health data such as heart rate readings, offers objective insights into patterns of social engagement. In our test involving participants with ASD, we assess algorithm accuracy by comparing the automatically estimated speaking time to manually labeled ground truth from recorded interactions, with overall speaking durations being reviewed manually and compared against the algorithm's outputs.

Our aim is to objectively track sociability in individuals with ASD by capturing everyday activity. By analyzing speaking patterns and physiological engagement, the system provides quantitative insights into social behavior. These insights are intended to assist clinicians in monitoring behavioral changes over time, including evaluating improvements before and after interventions

II. Related Work

Research on Autism spectrum disorder (ASD) now emphasizes digital health technologies which enhance traditional assessment techniques and treatment strategies. various technological approaches have emerged across the fields of behavioral monitoring, speech analysis, wearable tracking, and mobile-based observation. This section discusses existing research across five areas which relate to our research: contemporary ASD analytical technologies, speaker diarization in ASD, voice-based ASD behavioral patterns, wearable devices and mobile systems tracking social engagements and physiological signal analysis.

A. Current Digital Approaches in ASD Analysis

Earlier methods of evaluating social behavior within ASD individuals conducted their observations through professionals in controlled settings during face-to-face sessions [5]. While these approaches provided meaningful information, they were mainly dependent on subjective observations while lacking understanding of natural social behavior in typical everyday environment.

The search for better behavioral analysis led recent ASD research to focus on digital platforms which combine motion data with audio and physiological information. Machine learning plays a key role in ASD monitoring through wearable devices, acknowledging its ability to detect complex patterns across various behavioral domains [6]. Saghafi et al. [7] investigated the potential of humanoid robots to advance social connections between children with ASD during purposeful play activities. Their findings indicated robotically assisted therapy methods can enhance verbal interactions thus demonstrating that robot embodiments can work alongside passive sensing systems for ASD behavioral assessment. Smartwatches like WearSense can display excellent detection of real-time stereotypic motor behavior which showcases continuous behavior tracking from wearable devices [8]. These studies highlight the value of digital approaches like wearable and ML-based systems based on passive sensors and automated analysis for behavioral support.

Despite the progress made, many digital solutions for ASD monitoring encounter persistent challenges such as user compliance, data privacy concerns, and limited contextual awareness in communication tracking [6], [9]. Existing technologies typically track physiological indicators or movement behaviors but tend to overlook verbal exchanges and conversation patterns. Our study extends prior research by speaker diarization, heart rate data. Which enables a comprehensive perspective of social engagement of ASD individual.

B. Speaker Diarization and Voice-Based Behavioral Analysis in ASD Individuals

Speaker diarization has emerged as a valuable method which enables the analysis of verbal interacrtion patterns in autism spectrum disorder contexts. O'Sullivan et al [5]. The main purpose of this research focused on determining if automatic speaker diarization procedures could effectively break down natural home conversations occurring between ASD participants and their conversation partners. The research found effective results where they tried to record on home environment with minimal background noises, they pointed out few challenging conditions causes inconsistent diarization performance thus indicating a need for additional optimization. Recent advancements in speaker diarization, particularly through deep learning methods [10], have improved the accuracy and reliability of these systems.

Other studies have implemented diarization-based audio analysis methods on ASD participants to measure vocal engagement along with speaking times in diverse experimental conditions [11], [12]. The research showed how automated segmentation techniques could work yet their measurements primarily occurred in controlled environment settings or did not integrate multisensory data tracking. A notable development is the ASDSpeech algorithm, which analyzed 99,193 vocalizations from 197 ASD children using convolutional neural networks to quantify social communication difficulties during clinical assessments [13]. While ASDSpeech focused on clinical severity scoring in stationary control environments with microphones, our work emphasizes diarization accuracy in everyday environments, comparing algorithmic speaking time with manual labels to support objective behavioral tracking. Our work adopts Pyannote.audio [14]. along with reference participant samples to enhance diarization performance during real-world use. The proposed method uses speech diarization results to link heart rate data which creates an expandable system for ASD continuum social behavior observation.

III. Methodology

Including overall system workflow is shown in Fig. 1, the end-to-end workflow from data collection to analysis. Wear OS–based audio recorder application that prompts participants at scheduled intervals to record audio during natural conversations. Upon activation, the app captures audio (.m4a) and heart rate (.json) data, which are securely uploaded to an AWS S3 bucket. An AWS Lambda function monitors the bucket and triggers preprocessing, including conversion of audio files from .m4a to .wav format. Metadata from the uploaded files is sent as messages to an AWS SQS queue, The speaker diarization algorithm in AWS SageMaker obtains the audio file from the SQS queue of messages and processes it and stores it in the AWS RDS database.. A separate Lambda function, managed via CloudWatch events, handles the lifecycle of the SageMaker instance. Finally, the processed data is visualized through a web dashboard, providing clinicians and researchers with real-time insights into social engagement patterns in individuals with ASD.

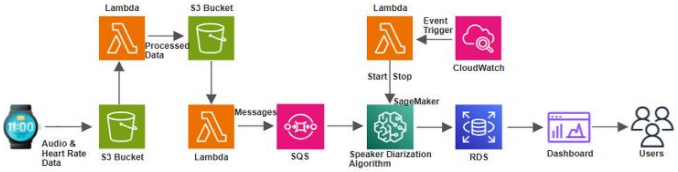

Fig. 1. System Architecture.

A. Wearable Audio Recorder Application

The Audio Recorder application named “Core Autism” was developed in Java for Wear OS devices. The app acts as a recording mechanism that collects audio data along with the user's health data like heart rate which transmits to an AWS S3 bucket for further processing.

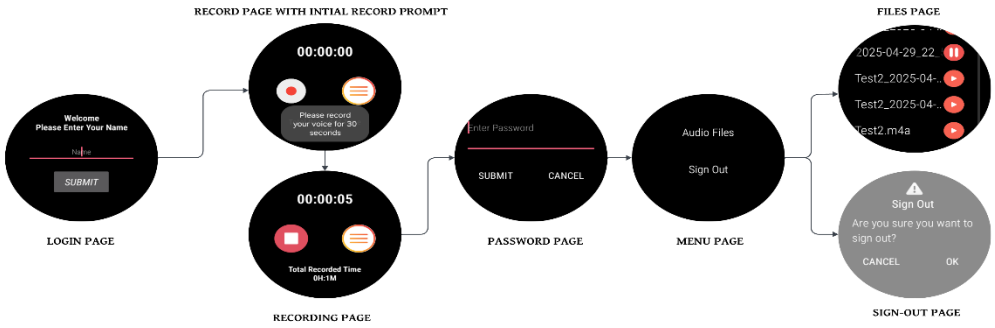

Fig. 2. Audio Recorder Application User Interface Flow

The application comprises six primary user interface screens: a Login Page, Recording Interface, Password Page, Menu Page, Audio Files Page, and Sign-Out Page. Upon launching the app, users are greeted with the Login Page, where they are prompted to enter their name. We didn't use a password for this study because each participant was manually onboarded by the research team and assigned a pseudonym for privacy reasons, eliminating the need for individual login credentials. The app was intentionally kept simple, allowing participants to focus solely on starting and stopping recordings. Administrative actions, such as signing out, were protected by an internal password not accessible to participants. This approach ensured ease of use while maintaining data security. In future public deployments, a secure login system may be introduced to support wider usage and enhance privacy safeguards. Following login, users are directed to the Recording Interface, which initially displays a pop-up prompting the user to record a 30-second audio clip used for speaker identification. The screen also displays the total recorded time for ongoing sessions. Access to the Menu Page requires authentication via the Password Page, providing a basic layer of security. The Menu Page allows users to navigate either the Audio Files Page or the Sign-Out Page. The Audio Files Page presents a list of previously recorded audio files, each with options to play or delete. Finally, the Sign-Out Page confirms the user's intention to exit the app. The interface is designed using Android's XML layout system, promoting separation between visual design and functional logic. As the entire flow shown in Fig. 2, the app ensures a user-friendly experience with clear navigation and guided prompts throughout the interaction flow.

The app primarily captures the audio data using the MediaRecorder APIs at a sampling rate of 44.1 kHz and 128 kbps, saved in .m4a format, Health data such as heart rate collects while recording and saved in Json format. All recorded data contains timestamps for precise alignment in downstream analysis that makes it possible to find correlations between physiological responses and social engagement. The application requests the user for essential permissions as shown in Fig. 3. to access microphones and sensors before allowing protected data transmission and storage through AWS Amplify using proper authentication methods, these permission requests are triggered only during the first login of a new user and remain in effect until the user logs out.

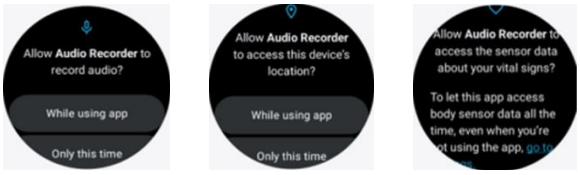

Fig. 3. Audio Recorder Application Notifications Layouts

B. Speaker Diarization for Voice Analysis.

In our study, we used a deep learning–based speaker diarization system using the pyannote-audio framework [14]. which is summarized in Fig. 4. This pipeline breaks down audio streams to identify "who spoke when" [15]. The diarization process starts with Speaker Activity Detection (SAD) which detects speech regions after which follows by Speaker Change Detection using Hidden Markov Models (HMMs) techniques [16]. Which analyze speaker transitions with probabilities derived from annotated datasets. The Viterbi algorithm is employed to compute the most likely sequence of speaker changes, annotating speech and silence regions with respect to speaker boundaries. This is followed by pulling out speaker embeddings using deep neural networks such as LSTMs and CNNs, which are trained to learn discriminative features from the acoustic signal. The embeddings are subsequently clustered using unsupervised models such as K-means and Gaussian Mixture Models (GMMs) to label segments with different speaker identities. The pipeline includes overlapped speech detection as a step to detect speaker overlap and performs re-segmentation for improving speaker identification which improves accuracy under poor audio conditions.

To ensure accurate identification of participant speech, our approach incorporated a short, pre-recorded voice sample collected during participant enrollment. This sample was prefixed to each subsequent conversation recording, allowing the diarization system to reliably recognize and isolate the participant's speech from other speakers and background noise. This setup facilitates a straightforward extraction of the target speaker's data by filtering the output to retain only the first speaker's segments. This strategy proved particularly effective in naturalistic environments characterized by overlapping conversations and ambient sounds. Diarized results were then time-aligned and compared against manually labeled results to compute the Diarization Error Rate (DER), accounting for false alarms, missed speech, and speaker confusion. This entire

process ensured robust, participant-specific speech extraction, supporting the downstream analysis of sociability metrics critical for ASD research.

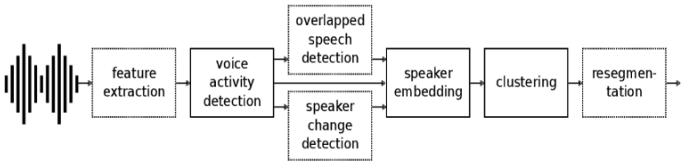

Fig. 4. Pyannote-Audio Speaker Diarization System [10].

C. Cloud Services Integration and Reporting Mechanism.

To enable secure, scalable, and efficient management of participant data, a fully cloud-based infrastructure was implemented. The system architecture integrates multiple AWS services to support seamless data transmission, real-time processing, reliable storage, and downstream analysis. Each service was strategically selected to optimize performance, ensure data integrity, and automate the entire workflow of data collection, storage, processing, and retrieval, The key components of the cloud infrastructure are described as follows.

1. AWS Amplify

AWS Amplify is a collection of development tools enabled for building safe and scalable applications. It allows a declarative interface for working with cloud services while providing better control to configure and manage cloud resources, through simplified resource management steps that significantly speeds up the development process. AWS amplify reduces much of the complex boilerplate code, it offers simple declarative APIs which makes it easily integrate with different AWS cloud services including authentication, analytics, and data storage functions. The Amplify CLI and Libraries along with Studio and UI Components and the Hosting segment allow developers to work with full-stack applications across iOS, Android, Flutter, Web and React Native platforms.

The Core Autism app uses AWS Amplify for secure data transfer from the app to AWS S3. The software efficiently handles uploading the collected audio and heart rate files to the designated S3 storage bucket, maintaining security and data integrity throughout the transfer process. The simple process of integration by AWS Amplify turns the data transfer workflow efficient while enabling timely and secure processing of all the user data. The Wear OS application depends on this feature for maintaining overall functionality and reliability along with securely managing the user data.

2. AWS S3 (Simple Storage Service)

Amazon S3 operates as an object storage service which provides high scalability to store and retrieve large amounts of data across the web from anywhere. This service provides a durability rate of 99.99% and 99.99% object availability over the course of a year. Amazon S3 provides multiple S3 storage classes which includes S3 Standard for active data retrieval and S3 Intelligent-Tiering for automated cost optimizing by shifting data between different tiers according to access patterns, and S3 Glacier and S3 Glacier Deep Archive for long-term data storage purposes.

In our project, S3 works as the main storage service to store both raw and processed audio files together heart rate data. The Core Autism app uploads audio files (.mp4) and heart rate files (.json) to an S3 bucket through AWS Amplify securely. After file upload to S3 the Lambda function triggers to handle the data and transforms .mp4 files to .wav format to meet processing requirements and then it moves the files into a different S3 storage space for further analysis.

3. AWS Sagemaker

AWS SageMaker is a fully managed service that simplifies the process of building, training, and deploying machine learning models, allowing developers and data scientists to quickly build and develop high-quality models with less effort and at a lower cost. SageMaker simplifies the workflow by providing a fully integrated Jupyter notebook instance to access data sources directly and avoids server management for the user.

This project uses an algorithm that analyzes audio as well as heart rate data, hosted on SageMaker. The speaker diarization algorithm is integrated on SageMaker and interacts with other AWS services. It polls messages from an SQS queue, downloads corresponding audio (.wav) heart rate (.json) files from S3 into temporary storage, then performs speaker diarization using the 'pyannote-audio' library, speaking time for the particular target speaker, and finally stores the result to an RDS database. It clears out the temporary storage after every operation and polls for messages until either the queue is empty or the runtime limit (initially set to 1 hour in order to cut costs) is reached.

4. AWS Queue

Amazon SQS is a fully managed message queuing service that enables decoupling and scaling of microservices, distributed systems, and serverless applications. In this project, SQS manages message queues between different components of the application, including communication between the database and backend, and handling asynchronous task processing.
For each audio file, a new message is created and placed in the SQS queue. The message contains the following details: the .wav audio file's S3 bucket location and its corresponding heart rate .json file location. SQS ensures reliable delivery of these messages even in the event of component failures, allowing for a fault-tolerant and scalable system. The decoupling achieved by SQS improves system resilience, maintains data integrity, and enables seamless scaling as the application grows.

5. AWS RDS

Amazon RDS was employed to manage the storage of processed audio and heart rate data in a structured format. RDS automates database setup, operation, and scaling by handling tasks like provisioning, patching, and backups.

In this project, RDS stores the results of the speaker diarization algorithm. For each user, a new row is created in the database, and if the user already exists, the results (including speaking time, start and end time, and heart rate file) are appended to their existing record. The processed data is then efficiently retrieved by the Flask backend and communicated to the Angular frontend, ensuring that the results are displayed on the user dashboard in a seamless and scalable manner. RDS's

cost-efficiency and resizable capacity allowed us to focus on data analysis and visualization without worrying about database management complexities.

IV. Experiments

A. Study Design

A total of 20 participants were selected for the trial of the app, two of which decided to drop from the study voluntarily. The 18 participates who completed the study include 11 male (61.1%) and 7 female (38.8%). Participants were between the ages of 7 and 24, diagnosed with ASD verbal, and willing to wear a smartwatch. All participants had a primary diagnosis of ASD and were in the higher functioning range. ASD, AD, and AS individuals were taken into consideration as well. Participants who were nonverbal were excluded from the study.

FSIQ (Full Scale IQ), VIQ (Verbal IQ), NVIQ (Nonverbal IQ), and IQ test scores were recorded, but no specific limit was set on IQ scores, as the goal was to include a diverse range of individuals with ASD. It was crucial that participants could consistently follow detailed instructions related to app usage for the 2–3-week period to ensure reliable verification.

All participants were onboard and fully informed about the nature, purpose and procedures of the study. The participant doesn't get a direct benefit, but this study allows us to build further along the app and future revisions with the algorithm to further test and add on features. This research may benefit society by providing better measurements of stress and social outcomes. All minors were consented by both the parent and child.

The privacy of participants was of the utmost consideration. Participants were given pseudonyms so that their names could not be traced back to them. All participants were given a participant number ranging from 1–20. All recordings are stored on secure Microsoft and Amazon servers which are limited to only certain members of the research team. People that didn't directly work with participant data did not have access to them. All emails referring to participants were referred to by their pseudonyms, not their actual names, in case of email leaks.

All participants in the study were onboard at the Thompson Center for Autism & Neurodevelopment. When participants arrived at the Autism Center, they were greeted and introduced to the entire study and our goal of testing the watch's algorithm for voice recognition. After, they were presented with an informed consent form. After completing the consent process, they were informed about the application's features, how to record their voice, and do initial setup of their voice. The setup of their voice was done by recording their voice while reading a short ~5 paragraph to ensure the algorithm gets a voice for it to look for.

Users would get a notification from 9am to 5pm every two hours to remind them to record. Any recording made will get a notification after one hour to stop the recording if they want to.

Participants were encouraged to record their social interactions with other people and were recommended not to record during situations that were not socially interactive. Examples include one-on-one conversation engaging in dialogue with teachers, friends, family, or any other person; group discussion engaging in dialogue in a group setting like a family dinner, team meeting, group of friends, or a classroom discussion. This also includes online conversation with voice chat using Discord, FaceTime, Zoom, and Microsoft Teams. Generally, we recommend not recording during moments that are not socially interactive, such as reading, watching TV alone, and non-social settings. Passive participation while participants might be in a social situation, where they are with other people, they might not be actively participating. We recommend not to record in these situations; an example of this is listening to a lecture in class.

We asked the participants for a goal of 14 hours of total recording time for the initial 2-week period. An additional week could be assigned if the goal of 14 hours was taking longer than expected and/or troubleshooting was needed. After completion of the study, participants were interviewed to provide feedback about the watch using a participant feedback form.

To assess the algorithm's accuracy against manually labeled data, we implemented a methodical comparison process. We aimed to review a total of 14 hours of recordings—the initial target set for participant involvement in the study. To achieve this, each file was assigned a sequential number from the total pool of recordings. We then used a random number generator to select specific files for manual labeling. This selection was repeated until we accumulated 14 hours of recordings. The alternative approach involved selecting audio files that featured clearer conversations with minimal overlapping voices and background noise.

V. Results

A. Individual Participant Analysis

Participant PA2 was enrolled in the study for a two-week period, during which she used the app consistently. Although the minimum required recording duration was 14 hours, PA2 contributed a total of 30.13 hours across 50 recordings, averaging over 2 hours per day. The total speaking time, as extracted by the diarization algorithm, was 0.96 hours, with noticeable day-to-day variability.

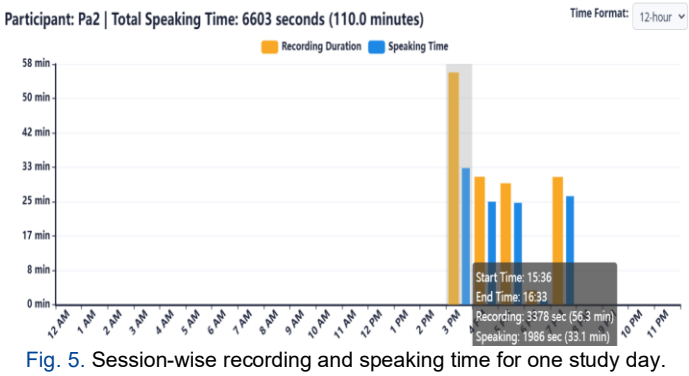

Fig. 5. Session-wise recording and speaking time for one study day.

To explore this highly active day in greater detail, Fig. 5. presents the session-level breakdown of recording and speaking durations on one day from their entire study. Each session is represented with yellow bars for total recording duration and blue bars for actual speaking time. For instance, one session from 15:36 to 16:33 lasted 56.3 minutes, of which 33.1 minutes

were identified as active speaking—a high engagement ratio. This session structure helped reveal not just total usage, but also how much of each session included verbal activity.

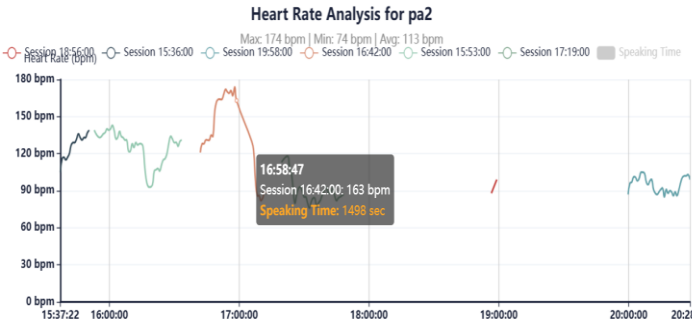

Fig. 6. Session-wise Heart rate trends with speaking time for one study day.

Fig. 6. shows the heart rate trends during the same day's sessions, with speaking intervals marked. The participant's heart rate showed notable fluctuations during active sessions, reaching a peak of 163 bpm during the 16:42 session, which included 1,498 seconds (~25 minutes) of speech. This alignment between heart rate activity and speaking time suggests a possible link between verbal engagement and physiological response, warranting further investigation.

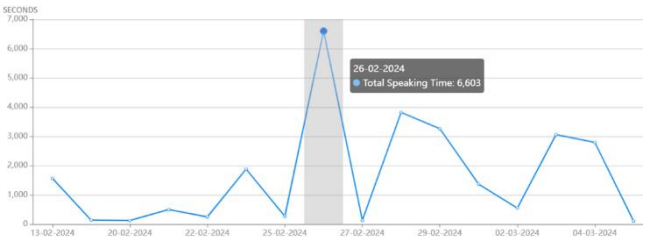

Fig. 7. Participant's Total Speaking Time Throughout the Study

Fig. 7. shows the participant's total speaking time per day throughout the study. A significant spike occurred on 26-02-2024, with 6,603 seconds (110 minutes) of speaking time, accounting for a major portion of her overall speech data.

Manual label analysis of 34% of PA2's files confirmed high diarization accuracy, ranging from 85.5% to 98.6%, with a mean of 92.52%. PA2 demonstrated active participation, including speaking during social contexts and singing when alone. In post-study feedback, she described the app as *"easy to use"* and noted that *"the reminders helped [her] record consistently."* Her consistent involvement and detailed physiological and behavioral data provided valuable contributions to the study's objectives.

TABLE 1
PARTICIPANT'S TOTAL AUDIO RECORDING TIME WITH ALGORITHM AND MANUAL SPEAKING TIME RESULTS FOR SELECTED FILES.

Participants	Total Recording Time (hrs)	Algorithm Results (hrs)	Manual Labels (hrs)
PA0	6.20	0.38	0.39
PA1	12.5	3.11	3.15
PA2	10.5	2.78	2.80
PA3	20.2	2.28	1.58
PA4	10.1	1.73	1.80
PA5	5.66	1.06	1.02
PA6	1.19	0.28	0.26
PA7	8.40	1.80	1.92
PA8	14.73	2.22	2.40
PA9	8.96	0.27	0.24
PA10	9.78	3.11	1.28
PA11	11.8	1.91	1.57
PA12	9.83	2.50	2.47
PA13	7.05	1.40	1.53
PA14	7.00	2.37	1.95
PA15	10.6	1.88	2.60
PA16	13.21	0.84	2.44
PA18	16.71	0.68	2.16

B. Overall Data Analysis

Table 1 summarizes the total audio recording time, algorithm-derived speaking time, and manually labeled speaking time for a selected subset of recordings from each participant. These selected files were used to evaluate the diarization algorithm's performance under varying conditions. Across all participants, there was considerable variability in speaking duration relative to the total recording time. This analysis was conducted under two specific evaluation conditions: a best-case scenario using carefully selected files with high audio quality, and a random selection scenario representing more naturalistic and uncontrolled audio environments.

B.1 Best-Case Scenario

In the best-case approach to selecting audio files for labeling and accuracy verification with the algorithm is methodical. From each participant's daily recordings, carefully choose one or two files that best represent the conditions. Specifically, prioritize recordings that are relatively clear, with minimal background noise, and feature more conversational content.

These files allowed the diarization algorithm to be tested under optimal conditions. As shown in Fig. 8 (left), the algorithm closely matched the manually labeled speaking time for most participants. Participants PA1, PA2, and PA12 achieved over 98% accuracy, with minimal difference between algorithm and manual outputs. The average accuracy across all cherry-picked files was 95.25%, reflecting strong algorithmic performance when audio quality is high. A notable exception was PA3, whose recordings included loud background conversations and overlapping speakers, resulting in lower accuracy.

B.2 Random Selection Scenario

In the randomly selected audio files approach, Files were chosen from each participant's recordings based on a chronological download order ranging from the earliest to the most recent. Then, each file numbered according to its position in the sequence. A random selection from these files was made until the total length of the selected recordings exceeded around

14 hours. This method introduced a mix of environments, including background noise, overlapping speech, and other real-world audio challenges. As shown in Fig. 8 (right), these factors led to a noticeable drop in algorithm performance. Participants such as PA10, PA16, and PA18 showed the largest discrepancies between algorithm and manual labels. The average accuracy across randomly selected files was 49.91%, highlighting the algorithm's sensitivity to complex and noisy conditions.

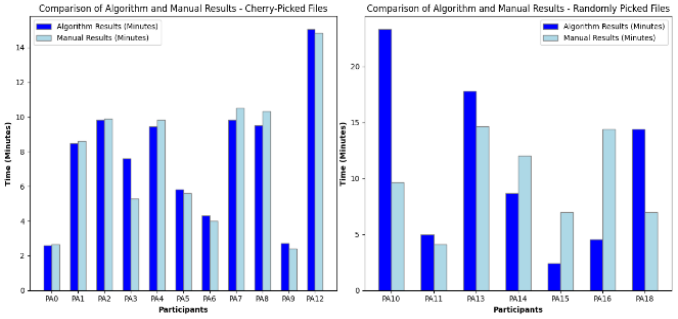

Fig. 8. Algorithm vs. manual speaking time for cherry-picked (left) and randomly picked (right) audio files.

E. Analysis of Total Recording and Speaking Time

Participants demonstrated substantial variability in both total recording duration and actual speaking time. The mean recording time was 20.92 hours, with a median of 16.45 hours and a standard deviation of 15.46 hours, reflecting diverse engagement levels. Some participants, such as PA3 (63.94 hrs) and PA8 (53.71 hrs), contributed extensive recordings but with proportionally lower speaking time—4.41 hrs and 12.90 hrs, respectively. In contrast, PA5, with a modest 16.98 hours of recording, yielded 3.49 hours of speaking time, indicating more active verbal participation during recordings.

These observations underscore the importance of evaluating not just the volume of recorded data, but the amount of meaningful speech content it contains. In some cases, lower speaking time may be attributed to background noise or environmental interference within the recordings, which can hinder the algorithm's ability to detect and isolate the participant's voice. The overall distribution is summarized in Table 2, and a visual comparison is presented in Figure 9.

TABLE 2
PARTICIPANTS TOTAL AUDIO RECORDING TIME AND ALGORITHM-DERIVED SPEAKING TIME ACROSS THE ENTIRE STUDY.

Participants	Total Recording Time (hrs)	Total Speaking Time (hrs)
PA0	12.30	0.96
PA1	24.43	5.90
PA2	30.13	0.96
PA3	63.94	4.41
PA4	15.72	3.86
PA5	16.98	3.49
PA6	1.56	0.39
PA7	26.47	4.13
PA8	53.71	12.90
PA9	15.72	0.92
PA10	10.74	3.80
PA11	15.91	2.05
PA12	26.56	6.57
PA13	8.86	1.96
PA14	8.03	2.57
PA15	11.58	1.44
PA16	17.30	1.69
PA18	22.05	1.32

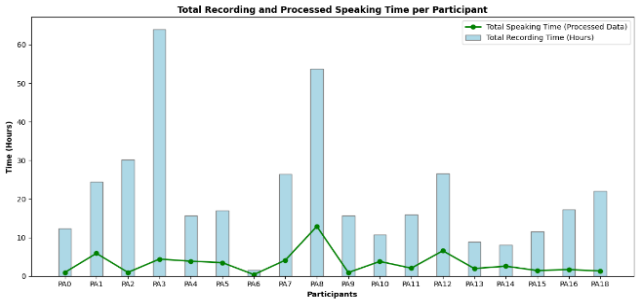

Fig. 9. Total recording vs. algorithm extracted speaking time per participant.

F. Participant Feedback

Post-study interviews were conducted to gather participants' feedback regarding the app and the smartwatch's usability. participants revealed both positive experiences and areas for improvement. Several users found the reminders helpful "Reminders were good they helped me remember to record" and described the app as "easy to use and clear." Some reported that "reminders came at the wrong time a few times" specifically during classroom lectures when speaking was not feasible. Suggested improvements like "I didn't know when it started recording — maybe a vibration would help." Hardware discomfort was also noted, with participants stating, "the watch felt heavy and made my wrist sweat" and "the strap was itchy I had to take it off during the day." Others experienced technical issues, saying "one time it said recording, but the file was not saved," and "battery ran out during recording, and it didn't save." A participant's parent suggested adding a total recording time display to "track how many hours they have recorded so far," which led to the implementation of a new progress-tracking feature. In contrast, one participant withdrew from the study, expressing that they were not really confident to talk while recording than usual. These insights were really helpful and directly influenced changes to improve usability, comfort, and recording reliability in future iterations.

G. Analysis of the Algorithm

PyannoteAudio diarization algorithm was evaluated under two distinct conditions: cherry-picked audio files with minimal background noise and randomly selected audio files with varying levels of noise and overlap. The algorithm demonstrated high performance on clear audio files, achieving a low Diarization Error Rate (DER). However, performance decreased significantly when background noise or overlapping speakers were present, highlighting a need for refinement in handling complex audio environments.

Participants demonstrated diverse levels of engagement during the study, with speaking times varying significantly across individuals. The following insights were observed based on the data:

1) Impact of Recording Prompts: Participants who followed the app's scheduled prompts and recorded in

environments with clear speech and minimal background noise showed significantly higher speaking times and improved data quality. This highlights the effectiveness of well-timed prompts and the importance of noise-free conditions for accurate data collection and better algorithm performance.

2) Environmental Factors: Participants recorded in quieter, controlled environments with minimal background noise, also recording made during group conversations where the participant's voice was clear without loud overlapping voices, generally provided high-quality data. Participants recording in noisier or more dynamic environments, such as public spaces with interruptions, heavy crowds, overlapping voices, or loud music, experienced difficulties capturing clear participant voices. These conditions posed significant challenges for the diarization algorithm.

VI. Discussion

A. Effect of Recording Quality on Diarization Accuracy

Comparison between randomly labeled files and selectively chosen recordings revealed several key insights. Files with minimal background noise and clear participant speech consistently produced higher-quality data and more accurate diarization outcomes. These recordings demonstrated lower Diarization Error Rates (DER), indicating optimal algorithm performance under favorable conditions.

In contrast, randomly selected files often included overlapping speech, background disturbances, and environmental noise, resulting in greater variability in accuracy and higher DER values. These observations highlight current limitations in diarization performance under real-world audio conditions.

Controlled recording conditions significantly enhance the quality of speaker recognition outcomes. Real-world challenges such as ambient noise and overlapping voices expose the need for improved noise-handling and voice separation techniques.

B. Suitable Audio Recordings

In our research, we looked for more conversational audio recordings captured from the participants' daily surroundings which are useful for testing the algorithm's effectiveness for users voice recognition, rather than just files filled with excessive background noise and loud music, In our scenario, a suitable audio file is simply a conversational recording without loud background noise or music that excessively interferes with the participant's voice. These findings highlight the importance of recording conditions and suggest strategies to improve both user engagement and algorithm performance.

C. Study Limitations and Participant Bias

A limited sample size of 18 may not be sufficient to generalize findings to include a broader population of individuals of ASD. We are looking at individuals with ASD that are more verbal and higher functioning. Demographic constraints from the age range of 7-24 years old with a requirement for participants to be higher function limits the generalizability of all individuals with ASD, particularly those who are younger, older, and lower functioning.

Reactivity Bias (Hawthorne Effect) must be taken into consideration as well. Participants were asked to wear a watch to self-record their social interactions. The bias occurs when the participant alters their behavior due to their awareness that they are going to be recording, being monitored and observed later. Wearing the smartwatch can influence the participant to speak more or less than they typically would.

VII. Conclusion

This study demonstrates the successful integration of wearable technology, cloud computing, and user-centered design to enhance communication monitoring for individuals with autism. By developing an Android-based wearable audio recorder, a scalable AWS backend, and an intuitive dashboard, the system offers reliable data capture, processing, and visualization. A novel speaker identification approach, using a sample of the target speaker's voice, improved diarization performance across diverse conditions. Despite technical challenges—such as achieving precise notification timing, automating cloud environment setup, and visualizing high-volume data, the system was refined through iterative development and validated through structured testing. The project not only meets technical goals but also emphasizes usability and accessibility, with future plans focused on scaling the platform, integrating real-time analytics, and expanding deployment through app store release and broader user testing. Future work will also explore the integration of sensor and patient activity data, enabling a more comprehensive understanding of behavioral context and communication patterns in naturalistic environments.

References

[1] M. Hajri *et al.*, "Cognitive deficits in children with autism spectrum disorders: Toward an integrative approach combining social and non-social cognition," *Frontiers in Psychiatry*, vol. 13, no. 1, Aug. 2022, doi: https://doi.org/10.3389/fpsyt.2022.917121.

[2] M. J. Maenner, "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020," *MMWR. Surveillance Summaries*, vol. 72, no. 2, pp. 1–14, Mar. 2023, doi: https://doi.org/10.15585/mmwr.ss7202a1.

[3] C. Lord *et al.*, "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of Autism and Developmental Disorders*, vol. 30, no. 3, pp. 205–223, 2000, doi: https://doi.org/10.1023/a:1005592401947

[4] S. L. Hyman, S. E. Levy, and S. M. Myers, "Identification, evaluation, and management of children with autism spectrum disorder," *Pediatrics*, vol. 145, no. 1, 2020, doi: https://doi.org/10.1542/peds.2019-3447

[5] O'Sullivan, J., Bogaarts, G., Schoenenberger, P. et al. Automatic speaker diarization for natural conversation analysis in autism clinical trials. Sci Rep 13, 10270 (2023). https://doi.org/10.1038/s41598-023-36701-4

[6] Francese, R., and Yang, X., "Supporting autism spectrum disorder screening and intervention with machine learning and wearables: a systematic literature review," Complex Intell. Syst., vol. 8, pp. 3659–3674, 2022. https://doi.org/10.1007/s40747-021-00447-1

[7] M. Del Coco et al., "Study of Mechanisms of Social Interaction Stimulation in Autism Spectrum Disorder by Assisted Humanoid Robot," in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 4, pp. 993-1004, Dec. 2018, doi: 10.1109/TCDS.2017.2783684.

[8] Amiri, A.M., et al., "WearSense: Detecting Autism Stereotypic Behaviors through Smartwatches," Healthcare, vol. 5, no. 1, p. 11, 2017. https://doi.org/10.3390/healthcare5010011

[9] Y. Koumpouros and T. Kafazis, "Wearables and mobile technologies in Autism Spectrum Disorder interventions: A systematic literature review," Res. Autism Spectr. Disord., vol. 66, p. 101405, 2019, doi: 10.1016/j.rasd.2019.05.005.

[10] D. S. Park, Y. Zhang, F. K. Soong, and J. Gao, "A review of speaker diarization: Recent advances with deep learning," arXiv preprint arXiv:2101.09624, 2021.

[11] A. Gorodetski, I. Dinstein and Y. Zigel, "Speaker diarization during noisy clinical diagnoses of autism," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 2593-2596, doi: 10.1109/EMBC.2019.8857247.

[12] T. Zhou, W. Cai, X. Chen, X. Zou, S. Zhang and M. Li, "Speaker diarization system for autism children's real-life audio data," 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 2016, pp. 1-5, doi: 10.1109/ISCSLP.2016.7918477.

[13] M. Eni et al., "Reliably quantifying the severity of social symptoms in children with autism using ASDSpeech," Translational Psychiatry, vol. 15, no. 1, Jan. 2025, doi: https://doi.org/10.1038/s41398-025-03233-6.

[14] H. Bredin et al., "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7124-7128, doi: 10.1109/ICASSP40776.2020.9052974.

[15] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 356-370, Feb. 2012, doi: 10.1109/TASL.2011.2125954.

[16] J. H. M. Wong, X. Xiao and Y. Gong, "Hidden Markov Model Diarisation with Speaker Location Information," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7158-7162, doi: 10.1109/ICASSP39728.2021.9413761.