Whisper to Words: AI-Powered Speech Therapy Application

Tilila El Baissi, Andrea Lugo, Maiko Lum, Joscelynn Palen

Group 13

Advisors: Dr. Vyshedskiy (Biomedical Engineering, ImagiRation)

Department of Biomedical Engineering
Boston University

April 28, 2025

Abstract

Children with autism face barriers to effective speech therapy due to high costs, limited availability, and lack of accessible transportation. This project introduces an innovative, AI-driven solution that offers an accessible, real-time speech assessment application built in Unity and powered by OpenAI's Whisper model via Undertone. Users are prompted with a set of 370 target words. Their spoken responses are transcribed and scored using a custom-built 370 × 13,000 similarity matrix. Unlike existing tools that rely on amplitude of speech, this approach evaluates pronunciation accuracy based on phonetic similarity, providing immediate numerical feedback on spoken words, and reinforcing correct speech patterns. Early testing with native and non-native speakers demonstrated significant improvement after iterative matrix refinements to account for issues including homophones, pluralization, and dialectal differences. Refinement improved the app's scoring accuracy from 65% to 95% for perfect pronunciations, 13% to 80% for pronunciations characteristic of early speech development, and 58% to 73% for babbling articulation, exhibiting a consistent upward trend across tests. This project represents a significant step toward scalable, cost-effective therapy solutions that empower families and educators, while advancing inclusive technologies for neurodiverse populations. It lays the groundwork for broader adoption of AI in therapeutic settings, bridging the gap between clinical care and everyday learning environments.

1.0 Introduction & Background

According to the latest data from the Centers for Disease Control and Prevention (CDC), the prevalence of Autism Spectrum Disorder (ASD) has rapidly increased to 1 in 31 children in the United States.[1] This represents a notable rise from previous estimates of 1 in 150, in the year 2000[1], highlighting the growing importance of understanding and addressing this condition. As a complex neurodevelopmental disorder, ASD affects individuals differently and to varying degrees, but its impact is most commonly observed in the way individuals interact, learn, exhibit behaviors, and communicate with others. One of the most persistent and impairing features is difficulty in speech and language development. Many children on the spectrum experience delayed speech milestones, limited expressive and receptive vocabulary, and challenges in both verbal and nonverbal communication.[2] As the number of children diagnosed with ASD continues to rise, so does the demand for innovative, inclusive therapies.

Helping children with ASD to improve their communication skills is essential for them to reach their full potential. Previous research has shown that a child's functional speech language ability is closely associated with fewer maladaptive behaviors and better social outcomes.[3] This critical insight emphasizes the critical role of language development in the overall well-being and integration of children with ASD. Additionally, research has shown that early implementation of speech therapy for children with ASD has demonstrated favorable outcomes including boost of expressive language and enhanced communication skills.[4] Despite the recognized benefits of early intervention, many children with ASD do not receive the recommended level of treatment. Approximately two-thirds of US children on the autism spectrum under the age of 8 fail to meet the minimum recommended treatment guidelines,[5] a statistic that is cause for concern. Some common barriers to healthcare access include shortage of specialists, and cost of services.[6]

Furthermore, there are certain communities who are disproportionately affected. According to the National Rural Health care Association (NRHA), rural Americans "*face a unique combination of factors that create disparities in health care. Economic factors, cultural and social differences, [and] educational shortcomings ... all conspire to impede rural Americans in their struggle to lead a normal, healthy life."* [7] These challenges are particularly pronounced for families with children on the autism spectrum, where access to specialized care can be even more limited. Geographic isolation presents a significant barrier to healthcare for rural residents with autism. On average, individuals from these areas travel over twice the distance to access healthcare facilities compared to those in urban locales.[8] Additionally, an overwhelming 83.86% of U.S. counties lack sufficient autism diagnostic services, disproportionately affecting rural communities.[8]

The importance of our AI speech therapy application project, *Whisper to Words*, lies in its aim to bridge a well-documented gap between recommended therapeutic interventions and actual care received– especially among underserved communities.

Leveraging technology in therapy delivery shows promise in bridging the gap between the recommended therapy for children with ASD and the actual therapy they receive.[9] This project sits at the intersection of biomedical engineering, speech-language pathology, and artificial intelligence, aligning with current scientific efforts to leverage digital health technologies for accessible care for families. While prior research has explored the benefits of early intervention and remote therapy platforms, few solutions have implemented real-time, AI-driven feedback mechanisms that evaluate pronunciation accuracy rather than just vocal presence or volume.

Unlike other applications on the market, *Whisper to Words* implements a similarity score calculation that produces a score on the user's pronunciation, responding adaptively to a user's vocalizations, and reinforcing vocalizations that are more similar to the model word. Pronunciation accuracy is assessed in real-time using the application's similarity scoring algorithm, which allows for more insightful feedback compared to other applications that frequently only use amplitude– which means that those applications solely consider the vocalization's volume and not its accuracy. The real-time feedback component helps children to get personalized feedback that targets their specific speech difficulties. Furthermore, studies show that children with autism benefit greatly from early access to therapeutic programs, making this project a timely and valuable approach.[10]

Our app will allow children with ASD to obtain an early therapy intervention and make significant speech progress by tackling the most common barriers to healthcare access, which will directly increase accessibility to treatment. By providing a convenient and accessible platform for speech therapy, this app aims to ensure that all children, regardless of their geographical location or economic situation, have the opportunity to develop essential communication skills. Our app introduces a unique approach to speech therapy for children with ASD. It distinguishes itself by using voice recognition driven by AI to provide remote, real-time therapeutic interventions. The platform integrates a unique similarity scoring method to assess and enhance children's pronunciation.

As we continue to understand and address the complexities of ASD, innovative solutions like *Whisper to Words* will play a crucial role in improving the lives of children and their families. Our project introduces AI-powered solutions that have shown great potential in providing tailored, successful therapies for children, increasing their engagement and results.[6] Our app builds on this potential by incorporating simple, kid-friendly components to boost learning and motivation, which really helps children's learning, especially in autism treatment.[12]

Our hypothesis is that an AI-powered, home-based speech therapy tool with real-time pronunciation feedback can significantly enhance therapy outcomes for children with ASD, particularly when used to supplement traditional care. That means that our application is not designed to directly replace traditional therapy sessions but to enhance them by allowing children to practice independently at home. This type of therapy has a positive effect not only on the development of speech but also on the development of other aspects of behavior including self-esteem and socialization skills.[13] By empowering families with a user-friendly tool that can be used independently at home, this project promotes early, consistent, and cost-effective therapy, and challenges the barriers that have long limited access to care, while also offering a method that therapists find supportive in complementing their work.[14]

In conclusion, *Whisper to Words* represents a significant advancement in speech therapy for children with autism. Its innovative use of AI and voice recognition, coupled with a commitment to accessibility and therapeutic precision, positions it as a transformative tool in ASD treatment that holds promise for dismantling long-standing hurdles in autism care. As the field increasingly embraces digital solutions, this project exemplifies how technology can drive equitable healthcare and improved outcomes for neurodiverse children across all communities. Its unique similarity scoring algorithm goes far beyond what existing solutions offer, moving on from current generic volume-based models. What sets *Whisper to Words* apart is its commitment to both clinical efficacy and accessibility. In targeting the structural barriers that prevent many children from receiving early and consistent therapy, the application emerges as an equity-centered innovation. Our project provides families with the ability to access high-quality therapy from the comfort of their own homes, supplements and strengthens therapist-led interventions, and most importantly allows children to find their voice, fostering independence, and boosting confidence.

2.0 Methods

2.1 Participants

Four participants from the Senior Design team conducted preliminary testing: three non-native English speakers and one native English speaker. Participants were selected with the purpose of identifying correct transcriptions, early detection of scoring issues, and adaptability to different pronunciation styles.

2.2 Instruments

The apparatus used for testing our application was the Undertone transcription model and the algorithm itself. To create the algorithm, we initially designed it to loop through a 370 × 13,000 word matrix in order to match the prompted word (what the user *should* say) with the application's transcribed word (what the user *actually* said), thereby producing the score on the intersection. For example, Table 1 shows that if the user said "Ball" when they should have said "Table", then the resulting score would be 20, because of the low phonetic similarities between those words. The scoring matrix size was 370 columns by 13,000 rows. The number of columns was determined by the vocabulary list provided by our project advisor who carefully selected words that would accurately assess speech skills of children with ASD. The number of rows came from Undertone's built-in transcription library.

Table 1. Excerpt of similarity scoring matrix.

		Prompted to user		
		Table	**Chair**	**Ball**
Transcribed by app	**Table**	100	20	20
	Chair	20	100	0
	Ball	20	0	100

After designing our approach, we implemented the algorithm by writing and editing three scripts in C# with Visual Studio Code. These scripts were executed in Unity 2022.3.45f1 while our application ran. Because we were creating this app in Unity alongside a transcription model called Undertone, it was imperative to learn how to use Unity, as well as search through Undertone's files to find the script where spoken words were transcribed.

The first script we made, PromptController.cs, enabled the user to click through the list of 370 prompted words. We implemented this by reading and displaying words from an imported file based on when the user would click on the screen in a specific area.

The second script we made, DisplayTranscription.cs, was our main engine to utilize the matrix and calculate the score. In this script, we needed to access Undertone's transcription variable, which changed every time the user would press the "RECORD" button. To accomplish this, we made a reference to Undertone's RecordButtonUndertone.cs script inside of DisplayTranscription.cs to continuously extract the transcription variable and use it for score calculations.

We made our score matrix by inputting our list of 370 words and the transcription list of 13,000 words into generative AI[3] and prompting it to produce similarity scores for all combinations (370 × 13,000 = 4,810,000 scores). We imported this as a CSV file into DisplayTranscription.cs, and turned it into a list of string arrays. With our matrix implemented and access to the prompted and transcribed variables, we finally had a fully fledged program that prompted the user what word to say, transcribed their recording, and looped through a built-in matrix to output their score. This method of calculating a user's speech accuracy score is useful and efficient because it allows us to easily revise the similarity matrix when improvements are needed. Because our program is so simple, there are little to no bugs, and there is no need to be constantly editing the code.

After three iterations of testing, we began to alter our scoring matrix to include more accurate scores and new possible transcriptions. We did this by adding new rows onto a google spreadsheet, updating the content of the row with accurate scores, exporting the sheet as a csv file, and importing the new file into Unity.

Our testing results showed inconsistencies with monosyllable pronunciations. Since we were not able to find a pattern based on monosyllabic pronunciation or duration after 3 additional focused ‘monosyllable’ tests, another update to the program was made. Specifically, we needed to ensure that transcribed words did not contain repeating letters when a user prolonged the pronunciation of a monosyllable, so that their transcription would be found in the matrix. To address this, we modified the algorithm to compare Undertone’s transcribed output and check for instances where the same letter appeared more than twice consecutively. This adjustment was necessary because no word in Undertone’s library naturally contains more than two identical consecutive letters. For example, words like "moon" with two repeating letters needed to be preserved. The updated algorithm processes each transcribed word by counting consecutive occurrences of each letter and storing the counts in a dictionary. If a letter appears more than twice in a row, the algorithm removes the excess letters then concatenates the remaining characters to create a single word.

2.3 Procedures

Model Testing

To determine which transcription engine would serve as the foundation for the application, we conducted a controlled, identical, and comparative test between OpenAI’s Whisper model and the Undertone model. Both transcription engines were exposed to the same set of 370 prompted words, spoken in a quiet setting with standardized microphone placement. Each model’s output for each prompted word was manually recorded by a third team member in a standardized google spreadsheet Each word was recorded once per model, and no retries or revisions were permitted. By comparing the returned word to the requested word, we were able to determine the transcription accuracy and compare baseline performance for each engine under the same settings.

Initial Internal Testing

Two non-native English-speaking team members tested the application initially. Each participant completed a standardized testing protocol consisting of pronouncing the complete set of 370 prompted words, each in three specific pronunciation conditions:

1. Perfect pronunciation (Control Condition): Participants clearly and precisely articulated the word.
2. Child-like pronunciation (Experimental Variable): Participants intentionally mispronounced or simplified words to mimic a child’s speech patterns.
3. Monosyllable pronunciation (Experimental Variable): Participants spoke only a single syllable of the word (e.g., pronouncing “Apple” as “App”).

Participants used headphones and completed recordings in a quiet setting accompanied with another team member. They clicked the “RECORD” button on the application interface, pronounced the prompted word, while the other team member documented the app’s transcription and output pronunciation score on a spreadsheet.

A total of 1200 pronunciations were examined in tests 1-3.

Secondary Internal Testing

A native English-speaking team member conducted the next phase of testing rounds using a subset of selected words carefully chosen from the initial internal testing. The participant repeated each word in the same three categories as above: Perfect pronunciation (~100% accuracy), Moderate pronunciation (~50% accuracy), and Monosyllable or partial pronunciation. The participant was accompanied by another team member who was in charge of documenting the pronunciation, the transcription, and the score.

To demonstrate our methodology of choosing three or more pronunciations to test for each word, Table 2 shows examples of real words (‘grandma’ and ‘mirror’) and many possible pronunciations a child could have when trying to say those words. During testing, if ‘grandma’ was the prompted word, a native English-speaking team member would record all of the pronunciations listed in the ‘grandma’ row of Table 2. The team member would say ‘grandma’ with close-to-perfect pronunciation, and any returned score with a value less than 90% would be marked inaccurate and changed later. After saying ‘grand’, any returned score with a value outside of the range 40%-60% would also be marked inaccurate and changed later. Likewise, the same would happen to any returned score outside of the range 20%-40% after saying ‘gra’, or ‘ma’.

Table 2. Examples of categorized pronunciations.

Prompted word	90-100% Accuracy	50% Accuracy	Monosyllables
grandma	‘grandma’	‘grand’	‘gra’; ‘ma’
mirror	‘mirror’	‘mere’	‘mi’

A total of 524 pronunciations were examined in tests 4-6. The matrix was updated with new transcriptions and more accurate scores after each secondary testing round, for a total of three times.

Monosyllable Testing

To further understand the inconsistencies in the score accuracies of monosyllables, we conducted a focused investigation involving ten common syllables in the 370-word list (e.g., “ca,” “er,” “na”).

All four subjects participated in pronouncing selected monosyllabic words at four different durations. The first was a control condition, where the syllables were pronounced clearly and naturally. The second condition involved evenly prolonging the monosyllable for 1 second. This was repeated for two additional durations—2 seconds and 3 seconds—ensuring that the entire syllable was stretched evenly across each respective time period. Three trials were conducted per duration, totaling 120 cases.

This specific round of testing was not done to make alterations to the scoring matrix, rather, it was done to determine if there was a correlation between the length of one’s pronunciation of a monosyllable and the length of the resultant transcription. Therefore, the team was testing Undertone’s transcription capabilities instead of the algorithm.

2.4 Data Collection and Analysis

We documented preliminary test sessions using a structured google spreadsheet. We have completed six iterations of testing. Each test recorded the prompted word, the application’s transcribed word, the application’s pronunciation score, and the manual evaluation. For the last three iterations, we compared the manual score and the application’s score to refine our transcription dictionary and scoring matrix. If a spoken word/pronunciation (e.g., ‘place’) had an inaccurate score (e.g., 0% instead of 50%-60% when the prompted word was ‘fireplace’), that case was marked as incorrect. A score was deemed inaccurate if all four members of the team independently agreed that the returned score fell outside of a set acceptable range of scores.

We repeated this for a set of 125 pronunciations in our last few iterations of testing, and were able to analyze the application’s percent correctness for that set. The application’s percent correctness was calculated by dividing the number of pronunciations tested with correct scores by the number of total pronunciations tested. We were then able to feed that data back into our matrix to improve our algorithm. For example, if a pronunciation was not found in the matrix or a score was inaccurate, we added the pronunciation as a new transcribed word and updated the score.

Statistical analysis was performed using Fisher’s Exact Test to compare the accurate score rates of Test 4 and Test 6, with $p < 0.05$ being considered statistically significant. Using the Wilson score approach, which is well suited for binomial data and small to moderate sample sizes, 95% confidence intervals were computed for observed proportions in order to quantify uncertainty.

3.0 Results

3.1 Model Comparison and Selection

During Model Testing, Undertone successfully transcribed 316 of the 370 prompted words that were tested, while Whisper only returned 198 correct transcriptions. These correspond to accuracies of 85% for Undertone and 53% for Whisper.

Although latency was not quantitatively measured with timestamps, clear differences in performance were observed. Undertone consistently returned results within 1-2 seconds, while Whisper regularly experienced delays of several seconds or failed to generate a transcription altogether. Given the application's requirement for real-time responsiveness, latency differences were deemed highly consequential.

Undertone's significantly higher transcription accuracy and superior responsiveness led us to select it as the core transcription engine for our system. These results are summarized in Figure 1, which shows the raw count of correct transcriptions for each engine under identical testing conditions.

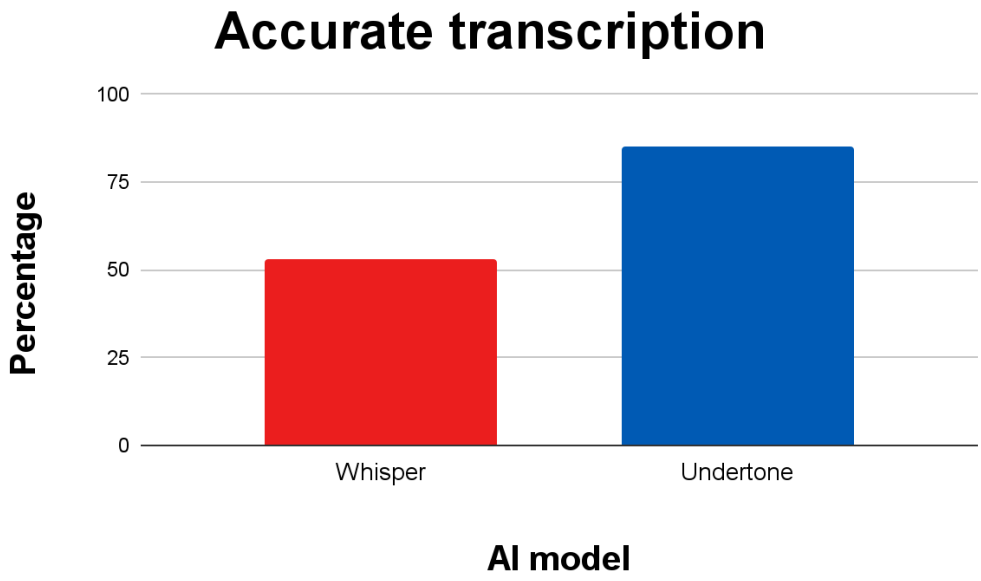

Figure 1. Transcription accuracy comparison between Whisper and Undertone models. Undertone shows superior performance.

3.2 Accuracy Across Testing Rounds

We performed six structured testing rounds to assess the impact of the scoring matrix modifications on speech accuracy scores. Each round included three different test types: Monosyllabic utterances, partial pronunciation and complete word pronunciation. The scoring algorithm compared user transcriptions to the words in the rows of the matrix. Each of the first three tests contained 370 pronunciations; the fourth test contained 125 pronunciations; the fifth test included 181 pronunciations; the sixth test contained 218 pronunciations. Over time, we updated the matrix by adding plurals and homophones. Non-dictionary phonetic strings returned by Undertone (e.g., "tuh," "gra," "nah") were also added to increase score coverage. These modifications significantly increased scoring performance, especially for partially pronounced words. Additionally, the refinements led to the scoring matrix nearly tripling in size after Test 6.

The complete word pronunciation group maintained strong performance, achieving scoring accuracy between 88% and 91% across all rounds. The impact of matrix updates was demonstrated by the partial pronunciation group improvements from 13% in Test 2 to 79% in Test 6, though it exhibited higher variability, reflected in larger error bars. The monosyllable group score accuracy increased from 58% to 72%, though results remained inconsistent. Figure 2's standard deviation error bars show how each group varies during test rounds, with the 50% pronunciation group showing the greatest range.

These findings show the matrix additions had the biggest impact in the 50% group, where the scoring algorithm was most responsive to change. On the other hand, the monosyllable group remained unreliable even after corrections, while the 100% pronunciation group was consistently scored right from the start.

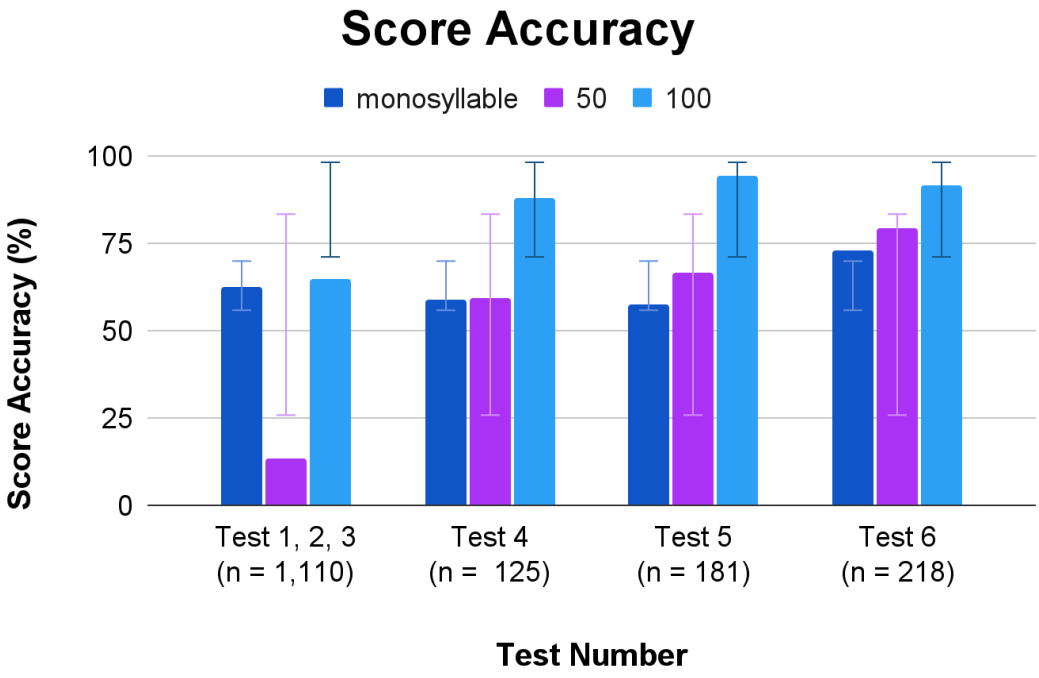

Figure 2. Pronunciation scoring accuracy across six testing rounds for 100%, 50%, and monosyllable conditions. Error bars represent +/-1 SD calculated from individual test values. N is the number of pronunciations.

3.3 Hypothesis Testing of Matrix Performance

To determine whether the improvements in scoring accuracy and the updates of the scoring matrix were statistically significant, we conducted hypothesis testing using Fisher's Exact Test for the three pronunciation categories. This approach was chosen because it works well for examining categorical results (right vs. wrong) in datasets with small or uneven sample numbers. Its use enabled us to assess the likelihood that the observed differences between Test 4 (before matrix updates) and Test 6 were likely to have occurred by chance.

We compared the null hypothesis (H_0) that the proportion of correct scores in Test 6 was equal to that in Test 4, against the alternative hypothesis (H_1) that the proportion in Test 6 was greater.

From 32 correct scores out of 54 tries (59%) in Test 4 to 68 out of 86 (79%) in Test 6, the 50% pronunciation group demonstrated a statistically significant improvement in accuracy ($p = 0.013$). The 100% pronunciation group, on the other hand, showed a slight increase in score accuracy from 88% to 91.5% (22/25 to 43/47), which was not statistically significant ($p = 0.720$). The monosyllable group improved from 58.7% (27/46) to 72.9% (62/85), but this difference also failed to reach statistical significance ($p = 0.118$), largely due to high response variability and inconsistency in matrix matches. These results show that the null hypothesis was rejected.

We also computed 95% confidence intervals (CIs) using the Wilson score interval to give an indication of the degree of uncertainty surrounding each observed accuracy rate. Based on the sample size and result, these intervals show the range that the true accuracy is most likely to lie within. The reported CIs allow for more robust interpretation of observed differences, particularly when comparing Test 4 and Test 6. Table 3 displays all of the confidence intervals, which were computed using a statistically appropriate method for proportions.

Table 3. Fisher's Exact Test results comparing Test 4 and Test 6 across pronunciation groups. Confidence intervals (95%) calculated using the Wilson Score method. Significance is defined by $p<0.05$.

Group	Test 4 (Correct/Total)	95% CI (Test 4)	Test 6 (Correct/Total)	95% CI (Test 6)	p-Value
100%	22/25	[70.0%, 95.8%]	43/47	[80.1%, 96.6%]	0.720
50%	32/54	[45.9%, 71.3%]	68/86	[69.3%, 86.3%]	0.013
Monosyllables	27/46	[44.3%, 71.7%]	62/85	[62.7%, 81.2%]	0.118

3.4 Monosyllabic Transcription Variability

The scoring matrix was iteratively improved, but the accuracy of the scores for monosyllabic utterances remained inconsistent. To further understand this limitation, we conducted a test solely to examine transcription of monosyllables, as described in section 2.3.

At constant duration, analysis showed that Undertone frequently generated inconsistent transcriptions for the same monosyllable. For instance, the monosyllable "na" produced transcriptions such as "na", "naa", "nah", "now", and "nuh", with no discernible pattern associated with timing. This inconsistency was observed across all syllables tested.

Similarly, "pa" resulted in transcriptions such as "pa", "paa", "paaah", or "paaaaa" often with excessive repeated letters at the end. These extended spellings were frequently unrecognized by the matrix, which contained only standard dictionary spellings or known alternate forms (e.g., american vs. british spelling). As a result, many of these cases failed to return a score. Additionally, in some trials Undertone interpreted the vocalization as a sound rather than a word, producing outputs like [groans], [sound], or [singing]. These were particularly common in cases where articulation was prolonged, so they happened primarily in the two and three seconds durations. Because these outputs do not correspond to any expected transcription, they contributed further to the monosyllables' scoring instability.

Across the full set of trials, 61.7% of transcribed outputs matched entries in the similarity matrix, rendering nealy 38% unscored. The considerable heterogeneity of the monosyllable group and the lack of statistical significance in performance improvement are likely explained by this mismatch rate. Although the matrix additions helped scores rise, a major barrier was Undertone's inconsistent transcription of short words. Table 4 shows representative transcription results for five of the monosyllables that were evaluated.

Table 4. Monosyllabic utterances and their inconsistent transcription outputs across different durations. Variants were often not recognized by the matrix.

Prompted Word	Monosyllable	Transcription Variants
Narrow	na	Naaaaa; nah; now; nuh; [sound]
Panda	pa	Pa; paaaaah; puh; paw; [singing]
Grandma	gra	Gra; grah; guh; graw
Book	boo	Boo; bo; [groans]; [sigh]
Carrot	ca	Ka; kaa; cawwww; [singing]

4.0 Discussion

The results of this project demonstrate strong alignment with our original hypothesis: an AI-powered, home-based speech therapy tool that is capable of delivering real-time pronunciation feedback and can meaningfully enhance therapy outcomes for children with ASD. Across six structured rounds of internal testing, we observed clear and consistent improvements in the application's performance. Scoring accuracy increased from 65% to 95% for perfect pronunciations, from 13% to 80% for early developmental pronunciations, and from 58% to 73% for monosyllabic utterances. These trends highlight the effectiveness of our iterative refinement process and validate the potential of our scoring matrix to evaluate pronunciation more accurately than conventional tools that rely solely on speech amplitude. The results suggest that real-time, phonetic-based feedback has the ability to reinforce correct speech patterns and support learning in a way that is both adaptive and accessible.

These findings are in line with existing literature emphasizing the benefits of early and consistent speech therapy for children with autism. Prior studies have underscored the importance of language development in reducing maladaptive behaviors and improving social outcomes[12]. Our results build on these insights by introducing a scalable tool that is designed to overcome barriers to care, particularly those affecting rural and underserved communities. Unlike many existing speech therapy platforms, our application goes beyond basic speech detection to assess the quality of pronunciation using a custom-built similarity matrix. This added layer of precision contributes to a more targeted therapeutic experience and aligns with recent efforts to improve digital health interventions using AI and natural language processing.

While our initial results are encouraging, there are limitations that should be addressed in future work. The participants in our study were internal team members, not children with ASD, the population for which the app is ultimately intended. Therefore, additional testing is needed to evaluate how well the tool performs with younger users who may present a wider range of speech characteristics. In addition, although the scoring matrix has been significantly expanded, it may not yet capture the full diversity of phonetic variation present across different dialects or speech disorders. Future development will focus on collecting broader speech data, incorporating additional phoneme patterns, and exploring adaptive scoring algorithms that personalize feedback based on individual progress.

Another limitation of our work is our algorithm's dependence on Undertone's transcription capabilities. Although our early tests proved the Undertone model to have the strongest ability of our options to transcribe audio accurately and efficiently, it is still not perfect. Undertone's large variance in transcribing monosyllables is what encouraged the team to dedicate a separate test for examining its transcription of monosyllables. Addressing the diverging monosyllabic transcriptions was the first step in updating our algorithm so that transcription errors could be fixed. Because any AI-driven speech therapy application will always rely on the capacity of its underlying model, it is imperative to acknowledge the model's limits. Future work will additionally involve adapting the algorithm to address transcription errors, all the while staying consistent and general.

Whisper to Words represents a promising step forward in the development of inclusive, technology-driven speech therapy tools. By combining real-time transcription with phonetic similarity scoring, the application delivers meaningful, immediate feedback that has the potential to accelerate language learning for children with ASD. Our work lays the foundation for broader adoption of AI-powered tools in therapeutic settings and reinforces the importance of designing with accessibility and equity in mind. With continued iteration and clinical testing, *Whisper to Words* can play a key role in transforming how speech therapy is delivered, bringing effective, personalized support into the homes of families who need it most.

5.0 References

[1] Shaw, Williams et al. “Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022”. *Surveillance Summaries*, vol. 74,2 (2025): 1–22. doi: http://dx.doi.org/10.15585/mmwr.ss7402a1.

[2] Mastermind Behavior Services. *The Connection Between Language Skills and Cognitive Development in ASD*.https://www.mastermindbehavior.com/post/the-connection-between-language-skills-and-cognitive-development-in-asd#:~:text=Autism%20Spectrum%20Disorder%20(ASD)%20significantly,comprehension%20and%20social%20conversation%20cues. (accessed April 23, 2025).

[3] Broome, Kate et al. “Speech Development Across Subgroups of Autistic Children: A Longitudinal Study.” *Journal of autism and developmental disorders,* vol. 53,7 (2023): 2570. doi: https://doi.org/10.1007%2Fs10803-022-05561-8

[4] Osman, Hafsa A et al. “A Systematic Review of the Efficacy of Early Initiation of Speech Therapy and Its Positive Impact on Autism Spectrum Disorder.” *Cureus*, vol. 15,3 (2023). doi: https://doi.org/10.7759%2Fcureus.35930

[5] Dunn, Rita & Vyshedskiy, Andrey. Mental Imagery Therapy for Autism (MITA) - An Early Intervention Computerized Brain Training Program for Children with ASD. *Autism Open Access,* vol. 5,3 (2015):153. doi: 10.4172/2165-7890.1000153

[6] Malik-Soni, Natasha et al. “Tackling healthcare access barriers for individuals with autism from diagnosis to adulthood.” *Pediatric research,* vol. 91,5 (2022):1030. doi:10.1038/s41390-021-01465-y

[7] N. Douthit, S. Kiv, et al. “Exposing some important barriers to health care access in the rural USA.” *Public Health* vol. 129, 6 (2015): 611-620. doi: https://doi.org/10.1016/j.puhe.2015.04.001

[8] Rising Above, ABA. *Autism in Rural Communities*. https://www.risingaboveaba.com/autism-blog/autism-in-rural-communities (accessed April 23, 2025).

[9] National Rural Health Association. *About Rural Health Care*. https://www.ruralhealth.us/about-us/about-rural-health-care (accessed November 23, 2024).

[10] Mohammadi Rouzbahani, Hossein, and Hadis Karimipour. "Application of Artificial Intelligence in Supporting Healthcare Professionals and Caregivers in Treatment of Autistic Children." *arXiv*, 2024, https://doi.org/10.48550/arXiv.2407.08902.

[11] Chiong, C., & Shuler, C. Learning: Is there an app for that? Investigations of young children’s usage and learning with mobile devices and apps. New York: The Joan Ganz Cooney Center at Sesame Workshop. (2010)

[12] Iannone, A., & Giansanti, D. Breaking barriers—the intersection of AI and assistive technology in autism care: A narrative review. J*ournal of Personalized Medicine*, vol. 14,1 (2024): 41. doi: https://doi.org/10.3390/jpm14010041

[13] Hryntsiv, Zamishchak et al. “Approaches to Speech Therapy for Children with Autism Spectrum Disorders (ASD)”. *International Journal of Child Health and Nutrition*, vol 14,1 (2025): 32-45. doi: http://dx.doi.org/10.6000/1929-4247.2025.14.01.05.

[14] Austin, Julianna, et al. "Perceptions of Artificial Intelligence and ChatGPT by Speech-Language Pathologists and Students." *American Journal of Speech-Language Pathology*, vol. 1, 1 (2024): 1. doi: https://doi.org/10.1044/2024_AJSLP-24-00218.