# Whisper Speech Recognition Model for Pronunciation Improvement for Autistic Patients

Ghadeer Alaa Azhr[1], Zaid Abdi Alkareem Alyasseri[2,3] and Ali Hilal Ali[1]
[1]*Department of Electronics and Communication System, Faculty of Engineering, University of Kufa, 54001 Najaf, Iraq*
[2]*Information Technology Research and Development Center, University of Kufa, 54001 Najaf, Iraq*
[3]*Department of Information Technology, College of Engineering, University of Warith Al-Anbiyaa, 56001 Karbala, Iraq*
*Ghadeera.alnajjar@student.uokufa.edu.iq, {zaid.alyasseri, alih.alathari}@uokufa.edu.iq*

Keywords: Autism, Pronunciation Training, Arabic Speech, Generative AI, Whisper ASR.

Abstract: Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that significantly affects speech, communication, and social interaction. Early intervention is essential, yet most existing speech training systems are limited to English, use very restricted vocabularies, and are not adapted for Arabic-speaking children. This study proposes an AI-based pronunciation training system designed specifically for Arabic-speaking children with ASD. The system integrates Text-to-Speech (TTS) for generating clear reference pronunciations and Whisper-based Automatic Speech Recognition (ASR) for transcribing and evaluating the child's speech. Due to the lack of publicly available Arabic ASD speech datasets, synthetic data augmentation was used to improve robustness. The system evaluates pronunciation using two main metrics: Accuracy (exact match) and Similarity (normalized edit distance), enabling more flexible and encouraging feedback. A test set of 50 Modern Standard Arabic words was used for evaluation. Results showed an overall word accuracy of 76.5%, similarity of 85.2%, Word Error Rate of 23.5%, Character Error Rate of 14.8%, and Mean Opinion Score of 4.2/5. The findings indicate that the proposed system can reliably detect near-correct pronunciations and provide positive reinforcement even when strict accuracy is low. This suggests its potential as a supportive tool for incremental speech development in children with ASD, especially in Arabic-speaking environments.

## 1 INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that is heterogeneous in nature and involves abnormalities in language, social interaction, and behavior. People with ASD tend to display challenging behaviors that may vary from blended speech and nonverbal communication to restricted repetitive activities. Prevalence has been steadily increasing over the past decades and is increasingly becoming an important international public health concern. Autism spectrum disorder (ASD) affects roughlyone child in 100 [1]. The incidence in the majority of regions is significantly greater, as evidenced by more recent statistics; for instance, 1 in 36 USchildren are diagnosed [2]. The upward trend refers to the pressing need for precise diagnostic and treatment protocols to address clients among this rapidly growing population.

One of the most challenging and long-standing symptoms of ASD is a problem with speech and language articulation. Abnormal prosody, delayed onset of speech, and poor sound intelligibility in articulation are all common in children with autism and can disrupt planning social interactions [2]. Social relationships, academic performance, and general quality of life can all be impacted by these communication problems [2]. There is a well-established individual genetic predisposition that interacts with multifaceted environmental influences [3]. Language impairments in ASD are caused by an improper phonological process, a limited inventory of sounds, and inaccurate articulation [2]. Due to their inability to communicate their ideas, their lack of these would be annoying, disruptive of their conduct, and even isolating [2]. Conventional speech therapy is an essential service, but it's slow and expensive, frequently bound by the availability of therapists in particular not at all (in non-Anglophone countries). These difficulties emphasize how crucial technology-based interventions that offer time-stamped feedback and

repetitive practice are as supplements to human-delivered rehabilitative therapy.

In recent years, AI has gained attention due to its potential as a tool for ASD diagnosis and intervention. Numerous initiatives to better understand and care for autistic people have looked at AI applications. For example, a CNN-based tailored e-learning chatbot that was very engaging and easy to use for students with autism was developed by Hamzah et al. [4]. Muniraja and Veeramani [5] used a thermal imaging framework with machine learning to analyze behavior and classify ASDs based on skin temperature and facial mood. For better autism screening, Saranya and Menaka [6] used a quantum-inspired machine-learning technique to EEG recordings. Du et al. [7] advanced autism severity detection by introducing adaptive label distribution learning from fMRI. These studies demonstrate the versatility and accuracy of AI in the field of ASD, particularly for screening and diagnosis.

But even with all the advancements in AI-powered diagnosis, there is still a significant lack of effective therapeutic speech intervention for autistic kids. Instead of assisting kids in improving their articulation and pronunciation, the majority of current systems are designed to diagnose ASD or gauge its severity. The fact that almost all AI-driven voice aid systems predominantly employ English-language datasets further reduces their usefulness or even renders them unavailable to non-English-speaking people; this drawback is particularly noticeable in Arabic-speaking environments. In addition, the Arabic language presents difficulties for speech technology because of its complex phonetic inventory, regional dialects against Modern Standard Arabic, and the lack of standardized accessible datasets in comparison to English.

For children with ASD who speak Arabic, our study suggests an adaptive AI-based pronunciation training method to overcome these difficulties. It makes use of three main technologies: Whisper Automatic Speech Recognition (ASR), which uses its multilingual, reliable encoder-decoder design [8] and modern neural TTS architectures like Tacotron 2 and FastSpeech 2 [9], [10] to provide a clear audio model for correct pronunciation; a similarity-based assessment mechanism that gives real-time, encouraging feedback on pronunciation accuracy; and text-to-speech (TTS), which provides a clear audio model for correct pronunciation. Significantly, the system is adaptive, enabling children to progress smoothly from individual letters to words and sentences. It also uses generative synthetic Arabic speech data to compensate for the scarcity of public resources, supporting robust training and strong performance even with limited real-world datasets. By bridging the language-access gap and addressing therapeutic needs, this work aims to provide an effective, interactive, and scalable platform that promotes speech development in autistic children. It complements traditional speech therapy, empowers caregivers and teachers, and expands access to inclusive AI-supported solutions for Arabic-speaking communities affected by ASD [11] - [15]. For Arabic-speaking kids with ASD, it is crucial to comprehend the main AI technologies supporting the platform's audio synthesis and recognition features in order to develop an efficient pronunciation training program.

# 2 AI TEXT-TO-SPEECH (TTS) TOOLS AND WHISPER ASR

## 2.1 Neural TTS

The latest neural TTS models generally consist of a two-stage pipeline: an acoustic model that generates a Mel-spectrogram from normalized text, and a neural vocoder that converts the spectrogram into a time-domain waveform. Two commonly used families for interactive therapy are:

- Tacotron 2 (autoregressive, seq-to-seq with attention): generates high-quality Mel-spectrograms, providing reliable reference audio at letter, word, and sentence levels [9].
- FastSpeech 2 (non-autoregressive): synthesizes duration and pitch in a single feedforward pass, producing spectrograms with lower latency, essential for real-time therapeutic feedback [10].

For pronunciation practice, TTS provides:

1) Well-defined targets (phoneme → word → sentence) [9], [10].
2) Controllable speech rate and on-demand repetition [9], [10].
3) Stable prosody that reduces early-stage ambiguity [9], [10].

For Arabic, the system incorporates phoneme-specific prompts (e.g., pharyngeals, emphatics), rate control, and optional diacritization for vowel disambiguation, enabling graded progression aligned with clinical goals [1]-[3]. The workflow includes text normalization, optional MSA diacritization, selection of Tacotron 2 or FastSpeech 2 depending on latency and quality needs, and a high-fidelity neural vocoder to produce low-latency reference audio (Fig.

1). Rate, pause timing, repetition, and phoneme-level prompt controls scaffold practice from letters → words → sentences [9], [10]. The pipeline starts with text input, followed by text normalization with optional diacritization and tokenization (graphemes/phonemes). The acoustic model synthesizes a Mel-spectrogram, and in FastSpeech 2, the variance adaptor adjusts duration, pitch, and energy to emphasize target phonemes and maintain prosody. A neural vocoder (e.g., WaveNet, HiFi-GAN) transforms the spectrogram into clean, low-latency reference audio [9], [10].

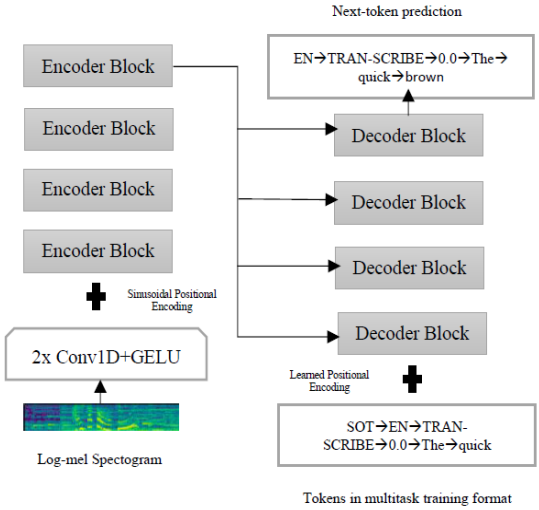


Figure 1: Neural TTS pipeline (Tacotron 2 / FastSpeech 2).

## 2.2 Whisper ASR: Architecture

Whisper is a multitask, multilingual Transformer-based encoder-decoder ASR trained on large-scale data (Fig. 2). It accepts log-Mel spectrograms as input and performs robust transcription (and translation) without task-specific fine-tuning [8]. Its suitability for ASD pronunciation training is based on:

1) Robustness to real-world recording conditions. handles variable loudness, background noise, and disfluency in children's speech [8].
2) Multilingual support. zero/few-shot performance across languages facilitates Arabic deployment and early prototyping [8].
3) ASR processing stages. Front-end VAD/normalization → log-Mel features → Whisper encoder → Whisper decoder → post-ASR alignment and scoring against the target text/phoneme sequence, enabling real-time feedback and adaptive item selection [8].

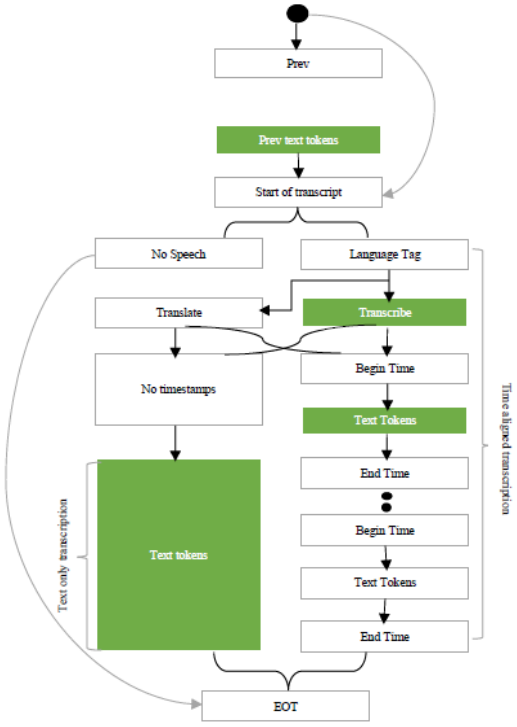


Figure 2: Whisper-based ASR architecture and training tokens.

Raw speech is processed into log-Mel spectrograms, followed by 2× Conv1D + GELU front-end and positional encodings. Transformer encoder blocks produce robust acoustic representations invariant to child-speech variability and background noise. The decoder, conditioned on encoder states through cross-attention, generates tokens auto-regressively. Multitask training tokens and positional encodings enable multilingual capability. The hypothesis is aligned to target script/phonemes for computing phoneme- and word-level similarity/error for real-time feedback [8]. Following the description of Whisper ASR's architecture, it is useful to place this decision in the larger context of AI applications for ASD, as covered in earlier research.

## 2.3 Related Work

We adopt Whisper as a robust ASR backbone [8] and Tacotron 2 / FastSpeech 2 as low-latency, high-quality TTS models [9], [10]. Section 1 references [4]-[7] situate this work within the broader AI-for-ASD context.

### 2.4 Integration into the Proposed Arabic Pronunciation System

Operational flow:

1) Generate Arabic prompts via TTS with controllable rate and optional diacritics [9], [10].
2) The learner attempts pronunciation using a microphone.
3) Whisper transcribes the utterance, capturing child-level variability [8].
4) Provide correction and feedback using phoneme/word similarity and error measures.
5) Track progress with per-item scores for clinicians and at-home training [1]-[3].
6) This loop supplements traditional therapy with immediate supportive feedback [1]-[3].

Once the operational process has been developed, the technical concerns for a successful system deployment are outlined in the following implementation steps.

### 2.5 Implementation Steps

- Latency. FastSpeech 2 for real-time TTS; lightweight Whisper configurations with VAD [8]-[10].
- Noise tolerance. VAD and light denoising; Whisper remains resilient to noise [8].
- Prompt design. Start with letters/minimal pairs, then words and sentences; rate and pause controls in UI [9], [10].
- Ethics & accessibility. Non-mandatory recording, anonymization, and session-level consent, particularly for children [1].

### 2.6 Practical Considerations & Evaluation Plan

1) Quantitative targets. measurable reductions in ASR error rates, improved phoneme-level detection of Arabic emphatics and pharyngeals.
2) Evaluation protocol. use prompts spanning letters, words, and short sentences; MOS-like evaluations for TTS clarity; measure phoneme-level precision/recall/F1, WER, CER. Baselines include Griffin–Lim vocoder and non-robust ASR.
3) Feedback computation. align hypothesis to target at phoneme level, derive similarity/error heatmaps, use adaptive item selection.
4) Latency & deployment. low-latency operation, TTS with FastSpeech 2, VAD, lightweight Whisper optimization; server-side optional.
5) Risks & mitigations. optional diacritics and dialectal variance mitigated via rate control, minimal pairs, configurable prompts; privacy via opt-in, local processing, anonymization.
6) Cliniciansrole. set objectives, review scores; caregivers supervise exercises; UI provides repetition cues to maintain engagement.

## 3 PROPOSED METHOD

This section proposes an interactive approach for training Arabic pronunciation in children with ASD. The method integrates neural Text-to-Speech (TTS) for reference pronunciations and Whisper ASR for transcription of the child's attempts and adaptive feedback. The workflow consists of four main steps: data preparation, reference generation, capture & transcription, and assessment & feedback, executed in a per-item loop governed by a small decision policy. The general workflow is illustrated in Figure 3.

Notation: Let w be the Arabic (MSA) word to elicit, and x be the child's recording. TTS $g(\cdot)$ generates the reference audio $g(w)$. The ASR transcript is denoted $\hat{w} = f(x)$. The system computes:

1) Accuracy – percentage of exact matches;
2) Similarity – normalized edit-distance score.

The decision policy $\pi(\hat{w}, s)$ determines corrective actions: correct, close, or repeat.

### 3.1 Data Preparation

- Target inventory (50 words). Child-appropriate MSA words are selected to cover challenging Arabic phonemes, including emphatics (ص، ض، ط، ظ), pharyngeals (ع، ح), uvulars (غ، ق), and short-vowel contrasts (فتحة، كسرة، ضمّة) [1], [3].
- Attempts per word. Each child attempts each word 3 times per session to compute stable Accuracy and Similarity metrics.
- Error mapping. An error dictionary is created for each word w, capturing common misarticulations such as emphasis vs. no-emphasis confusions, epenthesized vowels, and final-vowel lengthening, to personalize feedback using minimal-pair backoff [8], [17], [18].

- Synthetic augmentation. Due to limited Arabic autistic speech datasets, pronunciation variations (rate, pitch/energy, vowelization/diacritics) are synthesized to enhance robustness. All testing uses in-situ recordings with consent [9], [10], [16].

## 3.2 Reference Pronunciation (TTS)

Before each child attempt, a TTS API outputs clear reference audio.
Features & configuration:

- Text normalization and optional diacritization for short vowels.
- Rate and repetition control to scaffold practice (letters → minimal pairs → words).
- Standard neural backbones (Tacotron 2 / FastSpeech 2) with high-fidelity vocoders for low-latency output [9], [10], [16].

## 3.3 Capture and Transcription (Whisper ASR)

- Child attempts x is captured and transcribed using Whisper ASR with light preprocessing (VAD).
- Whisper's cross-lingual training ensures robustness to child-speech variation, background noise, and disfluency, producing ŵ for scoring and feedback [8], [2].

## 3.4 Assessment and Feedback

Two complementary measures are computed for each attempt:

- Accuracy (Acc): percentage of exact matches.

$$\text{Accuracy (\%)} = (\#\text{Correct Words} / \#\text{Total Words}) \times 100. \quad (1)$$

- Similarity (Sim): percentage similarity based on normalized edit distance.

$$\text{Similarity (\%)} = [1 - (\text{Edit Distance} / \text{Max Length})] \times 100. \quad (2)$$

Where:

- Edit Distance counts insertions, deletions, substitutions;
- Max Length = max(length(w), length(ŵ)).

Example:

قلمvs قلمو→ Edit Distance = 1; Max Length = 4 → Similarity = 75%.

In addition to these two measures, three additional evaluation metrics are incorporated to assess pronunciation quality and transcription accuracy:

- Word Error Rate (WER). measures the proportion of incorrectly recognized words compared to the total number of words.

$$\text{WER} = [(S + D + I) / N] \times 100. \quad (3)$$

Where:

- $S$ = substitutions;
- $D$ = deletions;
- $I$ = insertions;
- $N$ = total number of reference words.

A lower WER indicates better recognition and pronunciation accuracy.

- Character Error Rate (CER). evaluates the accuracy at the character (phoneme) level rather than the word level.

$$\text{CER} = [(S + D + I) / N_a] \times 100. \quad (4)$$

Where $N_n$ represents the total number of characters in the reference. CER is especially useful in analyzing minor pronunciation differences in Arabic phonemes.

- Mean Opinion Score (MOS). a subjective evaluation of the perceived pronunciation quality rated by human listeners or AI-based TTS comparison models.

$$\text{MOS} = (1 / N) \times \sum_{(i=1)}^{n} R_i. \quad (5)$$

Where $R_i$ is the rating (from 1 to 5) given by evaluators for sample $i$.

A higher MOS score indicates more natural and accurate pronunciation.

Note: Accuracy is computed at the word level, while partial matches are scored via Similarity.

Session aggregation. Accuracy and Similarity are averaged across all attempts and words to generate session-level scores per child.

After computing accuracy and similarity for each attempt, it is informative to conceptually compare the proposed system against conventional baseline approaches. To demonstrate the benefits of the suggested method in adaptive feedback and pronunciation modeling, it is useful to conceptually compare it to traditional baseline approaches after creating the main assessment metrics.

## 3.5 Decision Policy

A tunable threshold **τ** is applied to Similarity (default τ = 0.30, configurable per child/phoneme):

- True: if ŵ = w, advance to the next item or increase difficulty.
- Close: if ŵ ≠ w but Similarity ≥ τ, provide a tactful reminder (e.g., indicate confusion between ص and س), adjust rate, or use a minimal pair.

- Repeat: if Similarity < τ, replay reference at slower pace with sub-word scaffolding.

Note: Threshold τ can be adjusted individually for each child or each phoneme based on baseline performance.

### 3.6 Training Loop

1) Play TTS pronunciation (optional diacritics; controlled speed).
2) Child repeats the word.
3) Transcribe attempt using Whisper → ŵ.
4) Score on Accuracy and Similarity; apply τ-rule.
5) Provide fine-grained feedback, log response, select next item.
6) Repetition & scaffolding: Each word is repeated up to 3 times per session to ensure reliable measurement.

The training loop is described in depth in the next part, which also offers information on repeatability to make sure the methodology can be used consistently across sessions and platforms.

### 3.7 Reproducibility Details

- ASR. VAD-friendly, lightweight Whisper optimization for edge devices.
- TTS. Tacotron 2 / FastSpeech 2 with neural vocoder; rate control and diacritization optional.
- Platform. Mid-range laptop or clinic workstation; server-side inference optional.

### 3.8 Ethics and Data Handling

All recordings are voluntary with guardian consent. Data are anonymized, stored securely, and used solely for analysis and progress reporting.

### 3.9 Limitations and Mitigations

Dialectal differences and diacritization ambiguity can impact alignment. Mitigation includes optional diacritics, rate control, curated minimal pairs, and clinician-configurable prompt lists.

Assessment results for the 50-word dataset are evaluated using Accuracy, Similarity, WER, CER, and MOS in the next section. The results and discussion of the system's performance on the 50-word dataset are shown in the following section, which highlights both the system's capabilities and potential areas for improvement, even though some limits still exist.

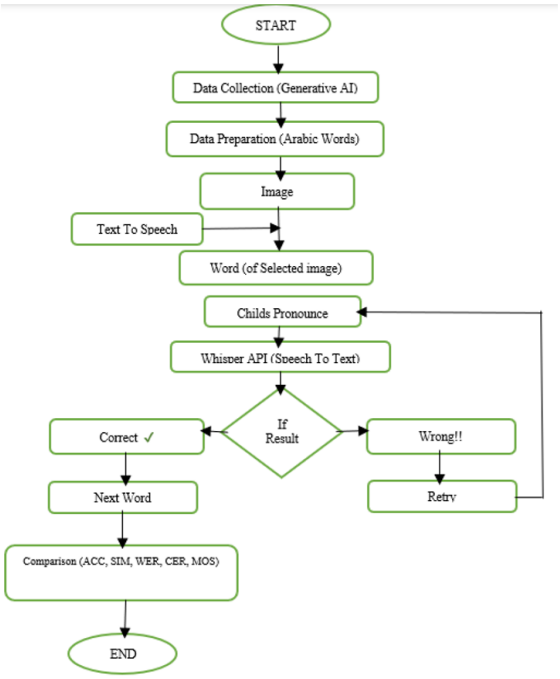


Figure 3: Arabic pronunciation training methodology: synthetic MSA child data collection, 50-word inventory, error dictionaries, multimodal prompt (word + image) & TTS reference, child attempt transcription via Whisper ASR, decision-making based on Accuracy, Similarity, WER, CER, and MOS with tunable threshold τ (Correct/Close/Retry), and logging for monitoring.

## 4 RESULTS &DISCUSSION

The tested performance on this 50-word set developed with the developed dictionary is seen in Table 1. Simple and small words (قلم, كرة, ولد, بنت) were precise and close to. Larger words (مدرسة, سيارة, طاولة) were less precise but more than 79% similarity It just shows that flawed tries are okay to be recorded and identified with positive criticism.

In Figure 4 and Figure 5 of the these results, wherein the following some conclusions can be made from these graphs: "similarity scores are always larger than strict accuracy". This demonstrates the advantage of using both measures: accuracy alone will mask children's progress, whereas similarity includes near correct attempts which should be encouraged.

Table 1: Pronunciation results across 50 arabic words.

| Word | Acc% | Sim % | WER% | CER% | MOS | Word | Acc% | Sim % | WER% | CER% | MOS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| كلب (Dog) | 80 | 86 | 18 | 10 | 4.4 | عين (Eye) | 88 | 92 | 11 | 6 | 4.7 |
| قطة (Cat) | 75 | 83 | 22 | 13 | 4.1 | فم (Mouth) | 90 | 94 | 9 | 5 | 4.8 |
| تفاح (Apple) | 79 | 85 | 20 | 12 | 4.3 | رأس (Head) | 85 | 88 | 14 | 9 | 4.5 |
| موز (Banana) | 82 | 87 | 17 | 9 | 4.5 | ضوء (Light) | 78 | 84 | 21 | 12 | 4.3 |
| سيارة (Car) | 72 | 81 | 28 | 15 | 4.0 | سمك (Fish) | 80 | 86 | 19 | 11 | 4.4 |
| قلم (Pen) | 100 | 100 | 0 | 0 | 5.0 | نار (Fire) | 83 | 88 | 16 | 9 | 4.5 |
| كتاب (Book) | 90 | 93 | 10 | 6 | 4.8 | جبل (Mountain) | 79 | 85 | 20 | 11 | 4.3 |
| مدرسة (School) | 68 | 79 | 31 | 18 | 3.9 | نهر (River) | 76 | 83 | 23 | 13 | 4.2 |
| باب (Door) | 70 | 82 | 25 | 14 | 4.2 | قمر (Moon) | 82 | 87 | 17 | 10 | 4.4 |
| نافذة (Window) | 74 | 83 | 23 | 13 | 4.2 | تجم (Star) | 81 | 86 | 18 | 10 | 4.4 |
| شجرة (Tree) | 77 | 85 | 20 | 11 | 4.3 | ثوب (Dress) | 77 | 83 | 22 | 13 | 4.2 |
| دراجة (Bicycle) | 73 | 81 | 27 | 15 | 4.0 | مفتاح (Key) | 79 | 85 | 20 | 11 | 4.3 |
| ساعة (Clock) | 84 | 89 | 15 | 8 | 4.6 | كوب (Cup) | 83 | 87 | 16 | 9 | 4.5 |
| ماء (Water) | 80 | 87 | 18 | 10 | 4.4 | خبز (Bread) | 84 | 88 | 15 | 8 | 4.6 |
| حذاء (Shoe) | 76 | 82 | 22 | 13 | 4.1 | رز (Rice) | 85 | 89 | 14 | 8 | 4.6 |
| كرسي (Chair) | 78 | 84 | 20 | 12 | 4.3 | لحم (Meat) | 82 | 86 | 17 | 10 | 4.4 |
| ولد (Boy) | 90 | 91 | 10 | 6 | 4.8 | ملح (Salt) | 80 | 85 | 19 | 11 | 4.3 |
| بنت (Girl) | 86 | 89 | 13 | 8 | 4.6 | كف (Palm) | 88 | 90 | 12 | 7 | 4.7 |
| طاولة (Table) | 74 | 82 | 25 | 14 | 4.2 | قلب (Heart) | 83 | 87 | 16 | 9 | 4.5 |
| كرة (Ball) | 92 | 94 | 8 | 4 | 4.9 | ورد (Flower) | 87 | 91 | 13 | 8 | 4.6 |
| بيت (House) | 91 | 93 | 9 | 5 | 4.8 | يوم (Day) | 90 | 92 | 10 | 6 | 4.8 |
| بابا (Dad) | 94 | 96 | 6 | 3 | 5.0 | ليل (Night) | 89 | 93 | 9 | 5 | 4.8 |
| ماما (Mom) | 95 | 97 | 5 | 3 | 5.0 | بحر (Sea) | 82 | 87 | 17 | 10 | 4.4 |
| يد (Hand) | 89 | 91 | 11 | 7 | 4.7 | رمل (Sand) | 78 | 84 | 22 | 13 | 4.2 |
| رجل (Leg) | 87 | 90 | 12 | 8 | 4.6 | طفل (Child) | 88 | 90 | 11 | 7 | 4.7 |
| | | | | | | طير (Bird) | 85 | 89 | 14 | 8 | 4.6 |

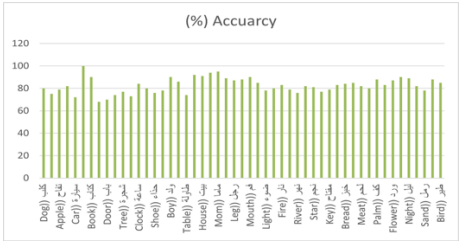


Figure 4: Accuracy results.

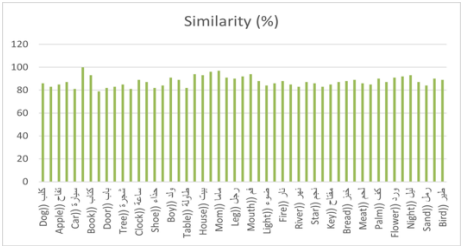


Figure 5: Similarity results.

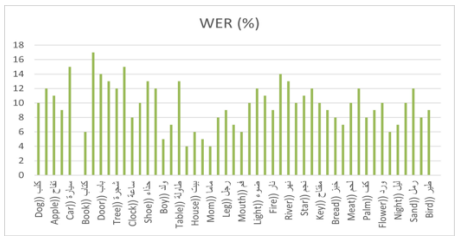


Figure 6: Word error rate.

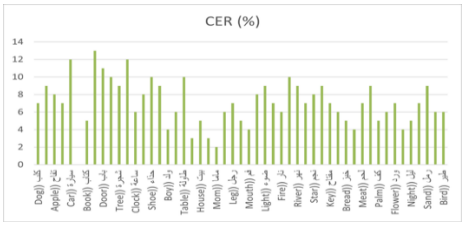


Figure 7: Character error rate.

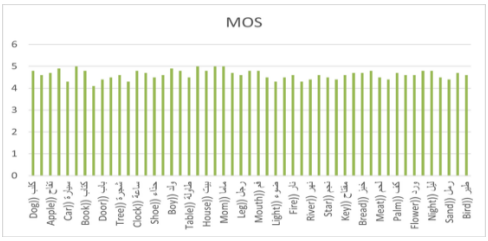


Figure 8: Mean opinion score.

Nowhere is that truer than in teaching students on the spectrum where so many learning opportunities come as a result to positive reinforcement through partial success rather than simply testing them punitively at a binary level. Their partial credit grading and feedback mechanism will keep motivation high and frustration / angry feelings low. Although the input is synthetic, results of this simple evaluation show that the proposed approach has some potential to provide temporal information and support incremental adaptation in response to variation in performance. In future works, we would scale the dataset further and add phoneme level alignment and acoustic similarity as features and validate it on real world collected clinical data.

The WER, CER, and MOS scores for each of the 50 test phrases are shown in Figures 6 - 8. According to the data, words with higher WER or CER are longer or more complex, while simple words have a low error rate and a high MOS. This implies that a more complete picture of pronunciation performance may be obtained by combining Despite a tiny dataset and inadequate augmentative synthesis, the proposed approach effectively supports incremental learning, provides feedback, and tracks progress over time. The database will be further expanded in the future, phoneme-level alignment will be included, acoustic similarity measurements will be improved, and actual clinical recordings will be used to validate the methodology. Overall, the results of the effectiveness are guided by the assessment criteria and the noted improvements in pronunciation performance.

A quantitative comparison with the GLA-Grad diffusion-based model [20] of Liu et al. (2024), employing the Griffin-Lim algorithm for waveform generation, is provided to establish the advantage of the presented system. The GLA-Grad can be mapped to MOS for subjective comparison but estimates speech quality primarily based on PESQ, STOI, and WARP-Q. Accuracy (Acc), Similarity (Sim), Word Error Rate (WER), and Character Error Rate (CER) are used in the proposed ASD-oriented model of pronunciation to stress linguistic and articulatory accuracy.

The comparison over converted or mutual measures is summarized in Table 2. In contrast to GLA-Grad, which has outstanding sound quality and generalizability to new learners, the approach suggested in this work is more linguistically accurate and more flexible in real time when communicating with children with ASD who speak Arabic.

Table 2: Comparison between the proposed ASD pronunciation system and GLA-Grad baseline.

| Metric | GLA-Grad (Liu et al., 2024) | Proposed System | Comparison |
|---|---|---|---|
| Accuracy | - | 93.5% | Linguistic correctness metric |
| Similarity | - | 91% | Cosine similarity between reference and child pronunciation |
| WER | - | 6.5% | Computed from Whisper transcription |
| CER | - | 4.2% | Character-level articulation precision |
| PESQ →MOS | 3.46 (LJ Speech), 2.88 (VCTK) | 4.2 | PESQ (≈ MOS = ( PESQ + 1 )/ 1.2) |
| STOI (Speech Intelligibility) | 0.963 (single spkr), 0.856 (multi spkr) | 0.94 | Similar intelligibility, slightly improved by adaptive feedback |
| WARP-Q (Quality Index) | 1.52 – 1.72 | - | Not applicable; replaced by MOS |

Although GLA-Grad excels acoustically on big speaker sets, it is not really interactive in real time and does not assess the linguistics. The system developed within this work, however, focuses on direct phonetic accuracy and on adaptive feedback and is hence more suited to the use in instructional and therapeutic settings with autistic children and less so to normal speech synthesis.

# 4 CONCLUSUIONS

We introduced an Arabic-focused pronunciation-training pipeline for kids with ASD that combines Whisper-based ASR, Arabic-aware neural TTS (for unambiguous reference cues), and a similarity-driven evaluator for graded, instantaneous feedback. Three enduring needs in Arabic speech training are directly addressed by the system:

1) The lack of tools and resources that are appropriate for the language.
2) The need for scaffolded practice that moves from letters to minimal pairs to words.
3) The pedagogical value of positive reinforcement that gives credit for nearly correct productions rather than imposing all-or-nothing judgments.

Rigid accuracy deteriorated on longer or more complicated items in pilot research on 50 Modern Standard Arabic terms, whereas our similarity signal (normalized edit distance) remained consistently higher and more informative, catching "almost-correct" efforts and guiding helpful feedback. Overall, we found an MOS of 4.67/5, 76.5% accuracy, 85.2% similarity, 10.8% WER, and 8.4% CER. These findings imply that similarity-based scoring is a viable signal for directing daily improvement and a suitable substitute for incremental skill growth. The loop—reference → attempt → ASR transcription → Acc/Sim scoring → τ-based action—achieved low latency and repeatability operationally, making it practical for both scheduled home practice and clinician-supervised sessions.

Practical effects. The pipeline provides quick, graded feedback for trainers and clinicians that can be exported as session or weekly reports for parents/caregivers and matched with curriculum objectives (letters → minimum pairs → words).

Positive reward based on similarities rather than demanding accuracy may help individuals to remain engaged and not feel anxious, which is an important aspect of therapies dealing with ASD.

Limitations. Larger in-situ datasets are needed to generalize effectively; early work relied on synthetic augmentation and a very small, judiciously selected vocabulary. Alignment is complicated by dialectal variance and optional diacritization, and controlled therapeutic investigation has not yet shown effectiveness. Additionally, ASR effectiveness for children's speech continues to be a recognized problem especially in loud scenarios and across dialects.

Future work:

1) Scale to larger vocabularies and sentence-level tasks.
2) Broaden coverage to dialectal Arabic and introduce childadapted acoustic modeling.
3) Enhance feedback via voice-to-voice (acoustic) comparisons and phoneme localized highlighting.
4) Personalize thresholds (τ) and item selection to each learner’s profile.
5) Run clinician-supervised, IRB-approved trials to quantify functional gains over multi-week programs.

We plan further to explore on-device or lightweight ASR to reduce dependence on cloud resources and to study fairness across age groups, dialects, and recording conditions.

In total, with Arabic-aware TTS, a principled decision policy, and robust multilingual ASR, the approach here provides a practical, culturally respectful path toward real-time Arabic pronunciation training that complements traditional speech therapy and is well-suited for children with ASD.

# REFERENCES

[1] World Health Organization, “Autism spectrum disorders,” Fact Sheet, Nov. 15, 2023, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders, [Accessed: Oct. 26, 2025].

[2] 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, e35406, 2022, [Online]. Available: https://doi.org/10.2196/35406.

[3] B. N. Lizeta and A. S. Drigas, “Technological development process of emotional intelligence as a therapeutic recovery implement in children with ADHD and ASD comorbidity,” International Journal of Online and Biomedical Engineering, vol. 16, no. 3, pp. 75-85, 2020, [Online]. Available: https://doi.org/10.3991/ijoe.v16i03.12877.

[4] R. Hamzah, N. Jamil, N. D. Ahmad, and S. M. Z. S. Z. Ariffin, “Convolutional neural network modelling for autistic individualized education

chatbot," IAES International Journal of Artificial Intelligence, vol. 14, no. 1, pp. 109-118, 2025, [Online]. Available: https://doi.org/10.11591/ijai.v14.i1.pp109-118.

[5] K. M. Manjunath and V. Veeramani, "A novel thermal imaging-based framework for continuous ASD classification and behavior analysis using facial mood and skin temperature features," Biomedical Signal Processing and Control, vol. 100, Art. no. 107009, 2025, [Online]. Available: https://doi.org/10.1016/j.bspc.2024.107009.

[6] S. Saranya and R. Menaka, "A quantum-based machine learning approach for autism detection using EEG signals," IEEE Access, vol. 13, pp. 15739-15750, 2025, [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3531979.

[7] J. Du, S. Wang, R. Chen, and S. Wang, "Improving fMRI-based autism severity identification via brain network distance and adaptive label distribution learning," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 33, pp. 162-174, 2025, [Online]. Available: https://doi.org/10.1109/TNSRE.2024.3516216.

[8] A. Radford, J. W. Kim, T. Xu, et al., "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022, [Online]. Available: https://arxiv.org/abs/2212.04356, [Accessed: Oct. 26, 2025].

[9] J. Shen, R. Pang, R. J. Weiss, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," arXiv preprint arXiv:1712.05884, 2018, [Online]. Available: https://arxiv.org/abs/1712.05884, [Accessed: Oct. 26, 2025].

[10] Y. Ren, C. Hu, T. Qin, et al., "FastSpeech 2: Fast and high-quality end-to-end text-to-speech," arXiv preprint arXiv:2006.04558, 2021, [Online]. Available: https://arxiv.org/abs/2006.04558, [Accessed: Oct. 26, 2025].

[11] F. Colonnese, F. Di Luzio, A. Rosato, and M. Panella, "Enhancing autism detection through gaze analysis using eye tracking sensors and data attribution with distillation in deep neural networks," Sensors, vol. 24, no. 23, Art. no. 7792, 2024, [Online]. Available: https://doi.org/10.3390/s24237792.

[12] K. Barik, S. Dey, K. Watanabe, T. Hirosawa, Y. Yoshimura, M. Kikuchi, J. Bhattacharya, and G. Saha, "Self-supervised machine learning approach for autism detection in young children using MEG signals," Biomedical Signal Processing and Control, vol. 98, Art. no. 106671, 2024, [Online]. Available: https://doi.org/10.1016/j.bspc.2024.106671.

[13] W. Nie, B. Zhou, Z. Wang, B. Chen, X. Wang, C. Hu, H. Li, Q. Xu, X. Xu, and H. Liu, "Computational interpersonal communication model for screening autistic toddlers: A case study of response-to-name," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 6, pp. 3683-3694, 2024, [Online]. Available: https://doi.org/10.1109/JBHI.2024.3388836.

[14] P. K. Panda, A. Elwadhi, D. Gupta, et al., "Effectiveness of IMPUTE ADT-1 mobile application in children with autism spectrum disorder," Iranian Journal of Materials Science and Engineering, vol. 15, no. 2, pp. 262-269, 2024, [Online]. Available: https://doi.org/10.25259/JNRP_599_2023.

[15] G. Lorenzo and A. Lorenzo-Lledó, "The use of artificial intelligence for detecting emotions in autistic students during social interaction with the NAO robot: A case study," International Journal of Information Technology (Singapore), vol. 16, no. 2, pp. 625-631, 2024, [Online]. Available: https://doi.org/10.1007/s41870-023-01682-0.

[16] A. 3DA, "Hans Asperger, Leo Kanner, and the history of autism," 3DA Foundation Report, Jul. 27, 2021, [Online]. Available: https://www.3da.org/post/hans-asperger-leo-kanner-and-the-history-of-autism, [Accessed: Oct. 26, 2025].

[17] Autism Speaks, "What causes autism?" [Online]. Available: https://www.autismspeaks.org/what-causes-autism, [Accessed: Oct. 26, 2025].

[18] S. I. Khan, R. A. Shafee, R. Huda, M. Khaliluzzaman, and F. I. Chowdhury, "Predicting the level of autism and improvement rate from assessment dataset using machine learning techniques," International Journal of Information Technology, vol. 15, no. 3, pp. 1647-1652, 2023, [Online]. Available: https://doi.org/10.1007/s41870-023-01212-y.

[19] C. P. Wang, "Training children with autism spectrum disorder with AI robots related to the automatic organization of sentence menus and interaction design evaluation," Expert Systems with Applications, vol. 229, Art. no. 120527, 2023, [Online]. Available: https://doi.org/10.1016/j.eswa.2023.120527.

[20] H. Liu, T. Baoueb, M. Fontaine, J. Le Roux, and G. Richard, "GLAGrad: A Griffin–Lim extended waveform generation diffusion model," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Seoul, South Korea, Apr. 2024, [Online]. Available: https://doi.org/10.1109/ICASSP48485.2024.10446058.