Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/GC03.20250315
Published: February 2, 2026

Evaluating Azure Pronunciation Assessment for Mispronunciation Detection in Children Learning Arabic

Jalal Zainaddin
Computer Engineering Department
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia
s202154790@kfupm.edu.sa

Mohammad Amro
Interdisciplinary Research Center for Intelligent Secure Systems,
Information and Computer Science Department
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia
mamro@kfupm.edu.sa

Wasfi G. Al-Khatib
Information and Computer Science Department,
Interdisciplinary Research Center for Intelligent Secure Systems
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia
wasfi@kfupm.edu.sa

Abstract

We evaluate Microsoft Azure Pronunciation Assessment (PA) for Arabic-speaking children using three datasets: two word-level datasets (Arabic Speech Mispronunciation Detection Dataset (ASMDD) and Down syndrome) and one sentence-level dataset (Arabic Pronunciation Assessment for Saudi Arabian Students (APASAS)). Supplying reference text to PA stabilizes Accuracy, Fluency, Completeness, and Pronunciation scores, while no-reference mode inflates high scores since it grades recognition hypotheses rather than prompts. Errors cluster around difficult Arabic sounds (emphatics, pharyngeals, uvular /ق/, interdentals) for both readers and Azure PA. Standardized audio front-end processing (noise conditioning, voice-activity detection, pre-emphasis, loudness normalization) yields small but consistent improvements on word-level corpora. On APASAS, preprocessing reduces Azure's over-rating and better matches teacher labels. With intended text supplied, Azure PA is suitable for tracking intelligibility and formative feedback; no-reference is best for basic screening.

Keywords

Arabic children's speech, Mispronunciation detection and diagnosis (MDD), Azure Cognitive Services, Artificial intelligence (AI) in education, Speech technology for underrepresented languages.

1. Introduction

Reading fluency is crucial for children's academic and communication development. In the Middle East, Arabic reading practice is often limited to school and Qur'an recitation due to the prevalence of English media and content.

Traditional reading instruction no longer meets expectations for convenience and engagement. This research develops and evaluates an Arabic Mispronunciation Detection and Diagnosis system for young learners using AI-based pronunciation assessment. We test whether Microsoft Azure Pronunciation Assessment can reliably help primary school students in the Middle East improve Arabic reading skills through automatic speech assessment, alongside teacher guidance.

1.1 Objectives

An essential phase of this research involves benchmarking the Microsoft Azure Pronunciation Assessment (PA) tool for Arabic learning. The tool is available in multiple languages. However, little is known about its suitability for observing the phonetic nuances of Arabic, especially when dialect variation, child speakers, or environmental noise are present. This work evaluates PA using a carefully selected set of Arabic speech datasets from children. We used Python tools to extract PA scores and metadata across datasets recorded under varied acoustic conditions.

This work may serve as a technical benchmark for Arabic speech assessment using existing AI-based tools within an interactive, AI-supported educational platform. By incorporating teacher-based corrections alongside generative feedback, the system aims to encourage sustained reading practice outside the classroom, thereby advancing Arabic literacy among young learners.

2. Literature Review

Several Automatic Pronunciation Assessment (APA) systems have been developed for young learners, such as Krajka's *English for Kids* (Krajka, 2003) and other tools designed for early English language education. In contrast, research on computer-assisted language learning (CALL) technology-based tools to support language learning for Arabic remains in its formative stages (Ahmad Khan et al., 2013)**.** Azure PA is a cloud-based (runs over the internet and does not require local installation) Microsoft Azure service for evaluating spoken language. It generates objective scores for pronunciation Accuracy (how well the speech matches the expected pronunciation), Fluency (how naturally and smoothly the words are spoken), Completeness (whether all words and sounds are included), and Pronunciation (overall correctness of speech sounds) at the phoneme (the smallest sound unit in a language), word, and sentence levels. Azure PA is intended for CALL, teacher feedback, and automated oral-skill assessment. Its scalability (ability to handle large numbers of users) and API (application programming interface, which allows software programs to communicate) access enable integration into educational platforms focused on speech-driven learning analytics and pronunciation improvement. Reliable detection of Arabic learners' mispronunciation requires high-quality speech datasets (collections of recorded spoken language used for system evaluation). These must include specific speaker demographics (e.g., age and gender), varied recording conditions, and differences in pronunciation among young learners. This section reviews current Arabic speech datasets, summarizes their features, and outlines their limitations.

2.1 Arabic Speech Mispronunciation Detection Dataset (ASMDD)

ASMDD is a purpose-built dataset for mispronunciation benchmarking in Arabic for early learners. It comprises 100 Egyptian children aged 2-8 years who utter some of the 100 most commonly used Arabic words shown in Table 1. Some children recorded all 100 words; others recorded 50 words. Each audio file contains a single word, providing a clean word-level segmentation for evaluating isolated mispronunciation (Aly et al., 2021).

The uniqueness of this dataset lies in its emphasis on child speech, an underrepresented demographic in Arabic corpora. Its design closely aligns with the purpose of this research and thus constitutes a rich testbed for evaluating PA's functionality in detecting and scoring child mispronunciations in Modern Standard Arabic.

2.2 QASR

QASR is a multi-purpose, multi-dialect Arabic speech resource on an extremely large scale. With around 2,041 hours of speech, it was compiled from Al Jazeera television broadcasts across various genres, including news, interviews, and documentaries (Mubarak et al., 2021). These recordings are transcribed in detail, alongside speaker metadata and segmentation. This dataset is used in ASR, dialect detection, and punctuation restoration, among other applications. While QASR lacks child-language samples, it serves as a benchmark corpus for transfer learning, pronunciation modeling, and dialect adaptation in speaker-dependent MDD research.

2.3 KSU Arabic Speech Database

The King Saud University (KSU) Speech Database provides another resource for speech diversity, with over 300

speakers of Arab and non-Arab origins (Alsulaiman et al., 2013). Recordings are collected in different environments, such as quiet rooms, offices, and cafeterias, to simulate various noise types. While mainly geared toward speech recognition and speaker identification tasks, the KSU database shows pronunciation variability across nationalities and settings. The database includes a mix of sentence- and word-type utterances, focusing specifically on common Arabic words and numbers.

Table 1. 100 most frequently used Arabic words in Egyptian dialogue (ASMDD)

Index	Word	Index	Word	Index	Word	Index	Word	Index	Word
1	نعم	21	الطريق	41	للغاية	61	المدرسة	81	ولد
2	رجل	22	عمل	42	فتاة	62	الصباح	82	رسالة
3	بخير	23	الجميع	43	كبيرة	63	الماء	83	عائلة
4	شخص	24	جيدة	44	آسفة	64	التحدث	84	القائد
5	الوقت	25	المال	45	الأرض	65	الساعة	85	المرأة
6	اليوم	26	الذهاب	46	البيت	66	الليل	86	الطبيب
7	صحيح	27	أرجوك	47	صباح	67	نهاية	87	اسم
8	أستطيع	28	المنزل	48	ألم	68	حياة	88	النقود
9	شكرا	29	الحياة	49	لحظة	69	الواقع	89	الكلام
10	الناس	30	انتظر	50	بالضبط	70	الطفل	90	مدينة
11	أعلم	31	الرجال	51	رقم	71	دكتور	91	مساء
12	رائع	32	الله	52	طريق	72	الهاتف	92	الشمس
13	مرحبا	33	الباب	53	المدينة	73	الطعام	93	ارجوك
14	آسف	34	جميل	54	الرئيس	74	فريق	94	السماء
15	تعال	35	الشرطة	55	صديقي	75	القى	95	الزواج
16	بالطبع	36	السيارة	56	ساعة	76	اللقاء	96	أصدقاء
17	العالم	37	النار	57	غرفة	77	نظرة	97	مكتب
18	الحقيقة	38	عظيم	58	عام	78	النساء	98	البحر
19	الليلة	39	الخير	59	الأطفال	79	العشاء	99	الكتاب
20	أمي	40	حالك	60	سنة	80	الأسبوع	100	الشارع

Table 2. Phoneme-level segment type letters correspondence

Segment Type	Example Arabic Letters	Description
Emphatics	ص، ض، ط، ظ	"Heavy" consonants produced with pharyngealization
Pharyngeals	ع، ح	Produced deep in the throat
Uvulars	خ، غ	Produced near the uvula
Glottals	ء، ه	Produced in the glottis
Velars	ك، ق	Back of the tongue
Dentals/Alveolars	ت، د، س، ز، ن، ش،ل	Front of the tongue
Labials	ب، م، ف	Lips
Vowels	ا، و، ي	Basic vowels or glides

2.4 MGB-3

The MGB-3 Speech Dataset specifically covers Egyptian Arabic as spoken in online media contexts. It comprises approximately 16 hours of speech, manually transcribed from YouTube videos across many genres, including comedy, science, and talk shows (Ali et al., 2017). The dataset is split into adaptation, development, and evaluation sets for experimental purposes. Though primarily intended for ASR benchmarking, its spontaneous and in-the-wild speech enables testing pronunciation assessment systems in less controlled, informal settings, complementing child-oriented evaluations.

2.5 Children Arabic Utterances for Mispronunciation Detection

This corpus of Arabic children's utterances for detecting mispronunciation has been developed to study pronunciation errors in school-aged Egyptian children. It contains speech samples from 27 children (14 boys and 13 girls), aged 7 to 12 years, uttering 16 very carefully chosen Arabic words (Sadik, 2023). These words were selected for their phonetic variety and significance in the vocabulary of primary school children. The recordings were conducted in relatively quiet environments, such as classrooms designed to minimize background noise and maintain clear speech. Despite having a smaller vocabulary set than ASMDD, its narrower age range and controlled conditions make it useful for studying error patterns and trends in gender-based pronunciation performance among mid-childhood learners.

2.6 Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation

The Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation contains recordings of 89 elementary-age children. Each child pronounces each of the 28 Arabic phonemes (distinct sounds in a language) 10 times, resulting in 890 utterances per phoneme (Arafa et al., 2018). Acoustic features, such as formant frequencies (the resonant frequencies of the vocal tract), were extracted and used to train classifiers (algorithms that categorize data) to assess pronunciation correctness. While highly focused on phoneme-level modeling, word-level mispronunciation labeling is lacking and does not support natural word recognition tasks.

2.7 Arabic Voice Recognition (Down-syndrome Children)

The Arabic Voice Recognition (DownSyn) dataset collects isolated utterances of Arabic words, numbers, letters, and colors from children with Down syndrome from various Arab countries (Arabic Voice Recognition (Down-syndrome Children), n.d.). With 37 children participating, 27 with Down syndrome as a target group, and the other 10 with speech disabilities. It provides a valuable opportunity to test the robustness of speech systems across a range of articulation profiles.

2.8 Arabic Pronunciation Assessment for Saudi Arabian Students

A specifically designed corpus and platform for Saudi elementary school children (ages 6-15) were created to evaluate and enhance their ability to read Arabic aloud. The data were gathered using an online platform that facilitated listening to and recording each student as they read the provided curricular sentences and paragraphs, stored the audio along with additional information to aid future teacher evaluation, and provided a platform feedback mechanism. The data comprising 503 cleaned participants and with an approximate duration of 33 hours of validated audio, is the largest pronunciation assessment corpus of Saudi primary students that has been reported so far (Al-Khatib et al., 2024)(Fanoush et al., n.d.).

2.9 Common Voice Arabic Dataset

The Arabic part of this dataset is part of Mozilla's worldwide initiative, Common Voice, which gathers thousands of speech contributions from volunteers. But it does not annotate explicit mispronunciations, and it generally reflects adult speech. While an asset in terms of accents and contributors, its focus on more general needs makes it less appropriate for word-level mispronunciation benchmarking, except after significant preprocessing or annotation (Ardila et al., 2019).

2.10 Summary

The comparative study of Arabic speech resources reveals that most collections lack child speech, explicit error labels, word-level segmentation, and teacher-verified targets. The ASMDD dataset works well for pinpointing sound errors in learners aged two to eight, but it is limited to Egyptian Arabic and lacks sentences. QASR, KSU, MGB-3, Common Voice, and the phoneme-pronunciation set aid ASR and coverage, but mainly feature adults, omit child mispronunciations, or have too few words for detailed analysis. Fair testing must include speakers with limited abilities, such as those with Down syndrome. APASAS excels at assessing child speech and lets teachers verify automated scores. We focus on three datasets: ASMDD (children, words), Down syndrome (special population, words), and APASAS (children, sentences with teacher feedback).

3. Methods

This section selects the data sets to be used for Microsoft Azure PA's evaluation. ASMDD is chosen because it aligns with the target demographic and uses the 100 most commonly used Arabic words, making it ideal for testing word-level segment-specific errors in early learners. Arabic Voice Recognition includes children with Down syndrome, which is useful for testing fairness and determining whether it works only for typical children's speech, especially at the word-level. The third dataset is Arabic Pronunciation Assessment for Saudi Arabian Students (APASAS), unlike others, it provides sentence-level recordings and teachers' evaluations, allowing comparison between Azure PA and teacher judgments. Together, the three datasets offer complementary coverage word vs. sentence, typical vs. special populations, and both automatic and expert references, aligned with our MDD benchmarking goals. working on the selected three datasets: All experiments implemented the recognition settings using Arabic Egyptian tone ar-EG locale; for APASAS, we used the ar-SA locale; 100-point grading, phoneme granularity, and miscue detection were enabled.

In with-reference mode, Azure PA aligns the audio to the supplied prompt (forced alignment) and computes: Accuracy (match quality of recognized phones/words to the prompt), Fluency (timing/flow, pauses, rate, rhythm), Completeness (coverage of the prompted content), and Pronunciation (aggregate phone-quality score normalized across the utterance). In no-reference mode, PA first recognizes speech (ASR hypothesis) and then grades that hypothesis; substitutions or deletions may be masked if they are internal to the hypothesis rather than mismatches to an external prompt. Scores are returned per file and (when available) per word, enabling both utterance- and token-level analyses. Throughout this paper, we explicitly separate with-reference (prompt-anchored) from no-reference (hypothesis-anchored) results. Therefore, this reflects Azure's scoring rubric; future work will be replicated with at least one additional engine to assess generality.

4. Data Collection

4.1 ASMDD

From the corpus, 20 children's voice folders were selected, each one giving approximately 50 isolated words (more than 800 utterances). Because ASMDD ships only the WAV file, the test harness mapped each file index to a fixed list of high-frequency Arabic words (such as **نعم، شكرا، أستطيع، الحقيقة**). Each utterance was judged twice:

- With reference (target word supplied)
- Without reference (Azure recognizes and then self-assesses)

4.2 Down syndrome Dataset

A subset of tokens was prepared, in which each target word (e.g., **وردة، زيتون، زرافة، وحيد القرن**) appears with acoustic variants: slower, faster, louder, quieter, pitch_up, pitch_down, and noise (with plain recordings wherever available). A total of 91 words were evaluated; each was pronounced by two participants, yielding 7 audio variants. This totals 1,274 audio files tested with Azure PA. As with ASMDD, each file was evaluated with and without a supplied reference. The results were then summarized per token and per variant.

4.3 APASAS

The execution was run on eight audio files; two were of females, and the rest were of males. It used the Saudi tone (ar-SA) language code. Each track was processed with both the "with reference" and “without reference" methods, where the reference text was the spoken caption rather than the original story text, since many students were not reading the entire text. No trimming, noise reduction, or loudness normalization was applied to the audio. This is mainly focused on the Azure PA evaluation compared to the teachers.

APASAS teacher ratings (Beginner/Intermediate/Advanced) were treated as a single gold standard in this study. Inter-rater agreement was not measured; therefore, some mismatches between Azure PA and teacher labels may reflect natural assessor subjectivity. Future work will (i) collect dual-rater judgments per item, (ii) report agreement (e.g., Cohen’s κ) by level and item type, and (iii) analyze whether disagreement concentrates on specific phoneme classes or timing phenomena.

5. Results and Discussion

5.1 Numerical & Graphical Results

Figure 1 shows that the ASMDD dataset with no reference averaged 95.7% for Accuracy, 98.5% for Fluency, 100% for Completeness, and 96.7% for Pronunciation. While Down syndrome without references averaged 93.7% accuracy, 94.0% for Fluency, 100.0% for Completeness, and 93.4% for Pronunciation. For the APASAS, it had 97.38% Accuracy, 92.88% Fluency, 100.00% Completeness, and 94.67% Pronunciation.

The utterance in audio file 27.wav (intended **أرجوك**) was recognized as "**أرجو**" (where the last letter /k/ was dropped) yet received very high scores; the omission was not penalized because the hypothesis itself omitted /k/. The utterance in 20.wav (intended **عمي**) was recognized as "**أمي**," with confidence measures agreeing with the hypothesis, not the target. Looking at Figure 2: more than 70% of words generated from the Azure PA without reference to the original intended texts were not in the vocabulary.

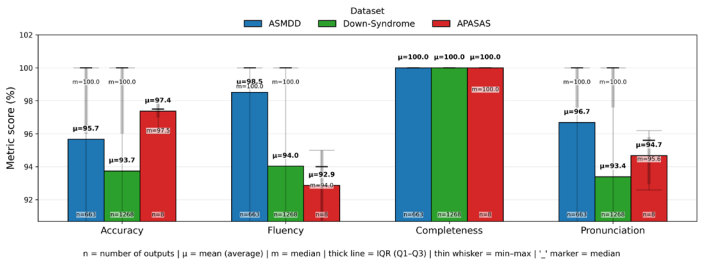

Figure 1. ASMDD & Children with Down-syndrome &

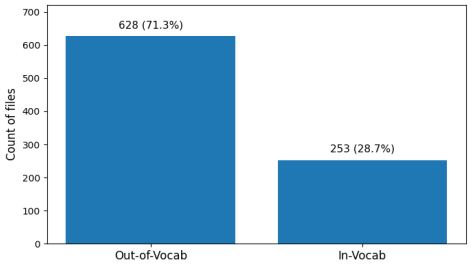

Figure 2. ASMDD dataset without reference scores in vs out of vocabulary recognition

APAPSAS datasets without reference combined metric scores

Figures 3 and 4 report results for Down-syndrome dataset without reference text. Figure 3 shows Accuracy, Fluency, Completeness, and Pronunciation scores (%), averaged per recognized Arabic word. To emphasize performance extremes, the worst 10 and best 10 words are shown. While Completeness remains consistently high across all words, substantial variability is observed in Fluency and Pronunciation, particularly among the lowest-performing words, indicating challenges in speech realization quality. Figure 4 further analyzes this behavior by comparing recognized words against the folder-level textual content. The PA identifies 55.8% of words as out-of-vocabulary like "**تحتاج**","**اتاحة**", "**وهو**","**يموت**", giving many false positives and also false negatives.

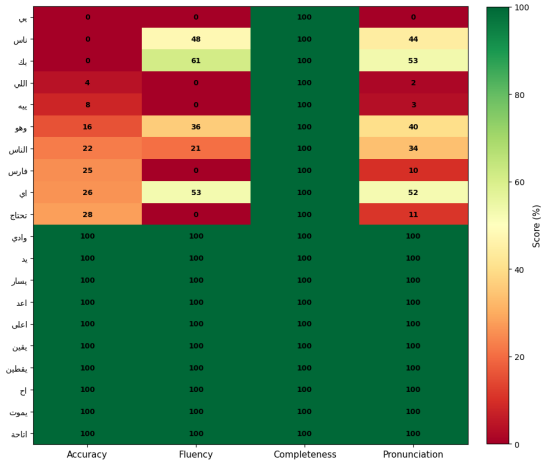

Figure 3. Heatmap of average recognition metrics per word for the Down-syndrome dataset (no reference)

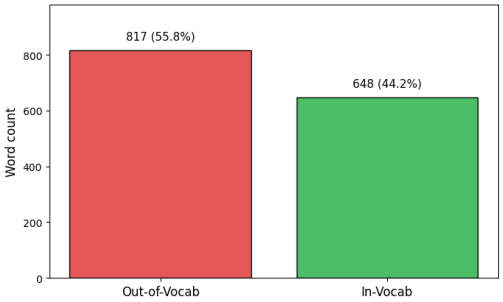

Figure 4. Down-syndrome dataset without reference vocabulary match

For the APASAS dataset, even with remarkably high scores, speech recognizers tended to wander off the mark semantically. One male Khobar file was meant to be "**الجَمَالُ وَحْدُهُ لا يَكْفِي**…," but the hypothesis contained nonsensical strings such as "**الوعي لو ... إنعكاس زورت ... ما الجدول**." Still, Azure PA metric returned 97/95/100/95.8, illustrating that PA highly grades the hypothesis when no target is supplied. Additional examples of hiding omissions, substitutions, and segmental distortions include the following:

- A Riyadh female file began with "**عمليات صح عن وبنصحهم**…," and scored a high 97/95/100/95.8 while giving unrelated words, again illustrating that, with No-Ref evaluation, high metrics are not incompatible with content errors.
- A male Dammam file showed the hypothesis combining parts of the prompt with "…**سيري سيرة ماء**…" for extraneous insertions and morphology errors such as "**الطحن**." The file nonetheless scored 98/94/100/95.6
- Several files devolved to ultra-short outputs. For example, an Ahsa male clip was reduced to " **صح عن وولده وحماره**." This both demonstrates the brittleness of the recognition system and the fact that PA maintains completeness at 100% for this output because it evaluates only the span of the returned hypothesis.
- Figure 5 shows that more than 37 percent of the recognized words generated by Azure PA ASR were not from the student's actual caption.

When the reference text is supplied to Azure PA, as shown in Figure 6, compared to results without a reference, both word-level datasets, ASMDD and Down syndrome, yielded lower results, with average differences of 2% to 11%. Azure PA generates many false positives, and an out-of-reference word may be inferred from the completeness scores. Without a reference, all datasets received full scores; whereas with a reference, they did not reach 90 percent.

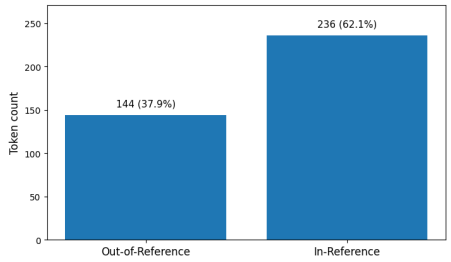

Figure 5. APASAS dataset without reference scores in vs out of reference (vocabulary) recognition

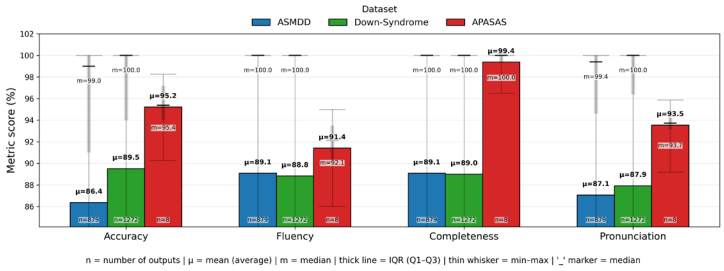

Figure 6. ASMDD & Children with Down-syndrome & APAPSAS datasets without reference combined metric scores

Here are some examples of the best-scoring words on ASMDD as shown in Figure 7:

- **صحيح**: Young speakers consistently articulated the vowels on both sides of /ħ/, so the system "hears" a steady vowel formant interval separated by sharp noise from /ħ/. The vowel-to-consonant transition is clear and provides robust phone boundaries, resulting in Accuracy/Pronunciation in the high nineties.
- **اللـه**: This word has a unique pattern of rhythms: an assimilated geminated /ll/ with tafkhīm (dark /l/) and a long open vowel followed by a soft final /h/. Children accurately imitate this pattern, thereby reducing variation and keeping all four metrics very high.
- **شخص**: Two fricatives (**ش**, then **خ**) mentioned in Table 2: surround a short vowel. Those high-frequency noise bursts provide clear markers for the recognizer, and most children hold them long enough to cleanly separate phones, which improves Accuracy and Pronunciation

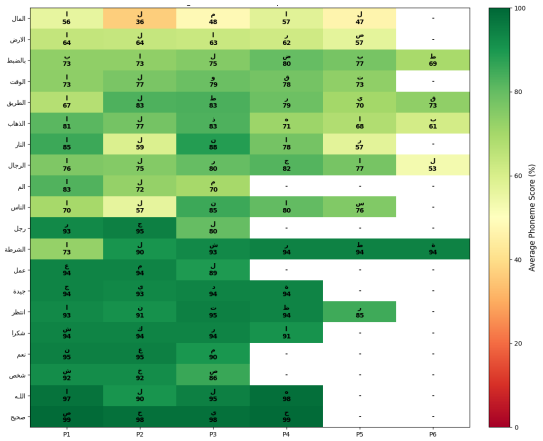

Figure 7. ASMDD dataset with reference best and worst 10 words average phoneme level scores heatmap

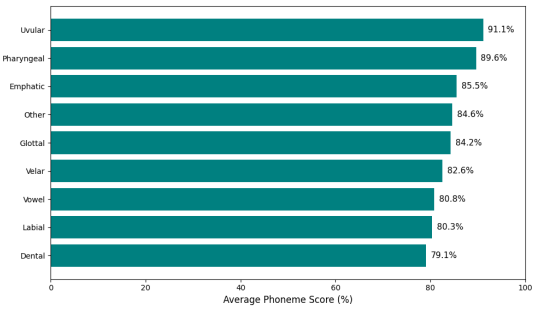

Figure 8. ASMDD dataset with reference average phoneme level segment type scores

Figure 7 shows the following as the worst-performing items in ASMDD with reference text:

- The word **المال** scored approximately 36-56 on word-level accuracy, with special phonemes such as /**ل**/ scoring below 40 in several files.
- **الوقت** contains complex codas of /qt/ and rising sonority of /wq/; mis-timed voice closures result in poor phoneme confidence and thus lower Accuracy.
- Low scores cluster for emphatics (/**ص، ط، ض**/), pharyngeals/glottals (/**ح، ع**/), uvular /**ق**/, final clusters, and interdentals (/**ذ**/), the segments and positions that are developmentally fragile in children.

Figure 8 illustrates the average phoneme-level pronunciation accuracy across different segment types in the ASMDD

dataset, using reference pronunciations. Uvular and pharyngeal segments achieve the highest accuracy, exceeding 89%, indicating relatively robust articulation for these categories. Emphatic, other, and glottal segments also show strong performance, with scores above 84%. In contrast, vowel, labial, and dental segments exhibit comparatively lower accuracy, with dental segments scoring the lowest at approximately 79%.

Results for the Down syndrome corpus items and variants are shown in Figures 9, 10, and 11, and can be summarized as follows:

- Back-of-the-mouth places earned good-to-best scores. Uvulars (**خ، غ**) and Velars (**ق، ك**) sit in the mid-90s. These segments come with strong friction/stop cues that ASR aligns reliably.
- Glottals (**ء، هـ**) and Dentals/Alveolars (**ت، د، س، ز، ن، ل**) are notably lower. Glottals suffer because /ʔ/ (**ء**) is often deleted or weakened in child speech and quietly realized in Arabic; /h/ (**هـ**) can be breathy and drift into noise.
- **أنانـاس**, which is the lowest in terms of average scores, especially on the letter "**س**" after the vowel "**ا**". This is probably not an issue with the children, but with Azure PA not processing Dentals or hisses like "**س**" accurately at the end of the audio file. Hence, it is considered an insertion or mispronunciation while it is not.
- **نهدي** got the second-lowest scores because of the last letters. The students pronounced the vowel "**ي**" well, but Azure PA considered all the readings to be mispronunciations (Figure 12).

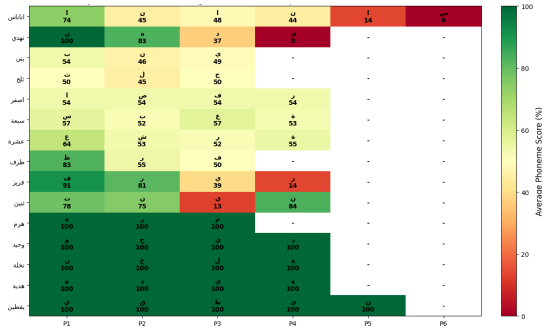

Figure 9. Down-syndrome dataset with reference per-phoneme average scores

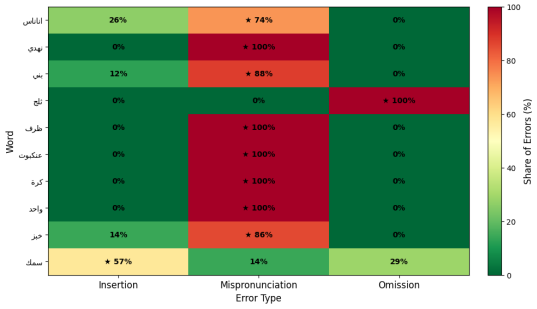

Figure 10. Down-syndrome dataset with reference lowest scored words error-type percentage

To estimate actual positioning on the APASAS corpus, we compare teacher evaluations to PA script thresholds: Beginner < 70%, Intermediate 70–95%, and Advanced ≥ 95% (applied to Accuracy). The Accuracy distribution in Figure 12 shows 0/4/4 (Beginner/Intermediate/Advanced) for Azure, compared to 2/3/3 for teachers, indicating that Azure tends to classify more speakers as "Advanced." The overall assessment was higher than human consensus for the file set.

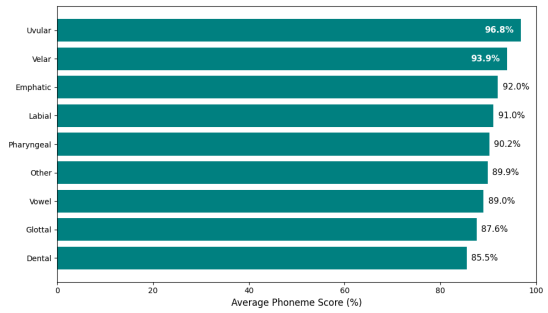

Figure 11. Down-syndrome dataset with reference per-phoneme average segments accuracy

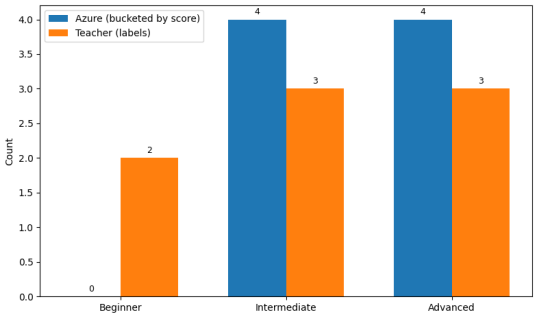

Figure 12. APASAS dataset category count scores with reference vs teacher evaluation

In summary, PA grades the ASR hypothesis rather than the intended prompt in no-reference mode. This can mask substitutions and deletions that would be penalized under forced alignment, inflating perceived performance when the hypothesis remains fluent but lexically incorrect. We, therefore, report paired examples and a simple sanity-check diagnostic to make this failure mode explicit. Hence, it produces high average scores across all datasets: ASMDD = 95.7 / 98.5 / 100.0 / 96.7, Down-syndrome = 93.7 / 94.0 / 100.0 / 93.4, and APASAS = 97.38 / 92.88 / 100.0 / 94.67 for (Accuracy / Fluency / Completeness / Pronunciation). However, the section highlights that these high metrics can be misleading, since >70% of the recognized words were out-of-vocabulary when no reference text is given.

5.2 Proposed Improvements

This section enumerates the five dominant acoustic/phonetic problems in children's speech recordings that degrade Azure's PA and introduces countermeasures for each issue. The first issue was broad, loudness-related variation, low SNR, and clipping, which frequently occurred with quiet speech, changes in microphone distance, and random overloads (Ardila et al., 2019). To mitigate this, we used EBU-R128 integrated loudness normalization to −23 LUFS with Pyloudnorm and a soft peak clamp (|sample| ≤ 0.999). When LUFS was unreliable (very short clips), RMS was used at −23 dBFS. Both methods provided consistent loudness suitable for endpointing and phonetic scoring, preserving roughly 14 dB of headroom.

Second, to address HVAC/handling rumble and 50/60 Hz hum, we used a fourth-order Butterworth band-pass (80–7000 Hz at 16 kHz) with gentle STFT spectral gating. All three quite effectively diminish the SNR of low-energy child consonants; hum splatter introduces spurious periodicity near the voicing detector passband (Kheir et al., 2022).

Third, we addressed the underestimation of high-frequency fricatives (/s, ʃ, θ, f/) in children's speech, which tend to display steeper spectral tilt and less turbulence (Hazan & Pettinato, 2014). The aperiodic energy was increased through pre-emphasis (α=0.97) in Equation (1). We implemented a fricative-safe VAD to retain the low-energy tails.

$$y[n] = x[n] - \alpha\, x[n-1] \qquad (1)$$

The system lifts 3–8 kHz aperiodic energies where Arabic fricatives dwell, thereby increasing the evidence at the phoneme level without precipitating significant changes to vowels.

Fourth, to prevent carving out micro-islands or brief pauses, which can harm the smooth joining of speech (gluing) and completeness (Hazan & Pettinato, 2014). In preprocessing script, WebRTC-VAD ran on 20 ms frames with aggressiveness=2. The settings also included +120 ms hangover, a minimum speech island length of 200 ms, and gap merging under 150 ms.

Fifth, to deal with temporal or pitch variations—such as higher F0, wider formants, or faster/slower tokens, all conditioning was fixed to 16 kHz. Also, the Azure PA timeouts increased (InitialSilenceTimeoutMs=4000) to prevent short child utterance onsets from being rejected too early.

5.3 Validation

On the ASMDD dataset with reference mode, all four metrics see their averages improve with the input of the target word. In Figure 13, with vs. without preprocessing, Fluency was 89.1% vs. 90.0%, Completeness was 89.1% vs. 90.0%, Accuracy was 86.4% vs. 87.3%, and Pronunciation was 87.1% vs. 88.1%, i.e., a ~1-point lift and fewer low outliers.

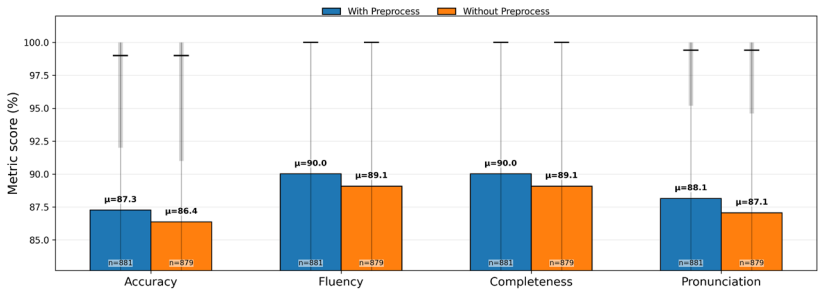

Figure 13. ASMDD with preprocess vs without preprocess with reference metric scores

In the no-reference mode, vocabulary-match auditing, as shown in Figure 14: illustrates a drop in the out-of-vocabulary (OOV) rate from 71.3% to 62.2%. This means the in-vocab hits increased from 28.7% to 37.8% post-processing. Therefore, hypothesized words of the recognizer are more likely to fall within the ASMDD vocabulary when the audio is clearer and end pointing is stable.

On the Down-syndrome dataset with reference mode in Figure 15, metric scores compared to before preprocessing were Accuracy ≈89.5% to 89.8%, Fluency ≈88.8% to 90.8%, Completeness from 89.0% to 91.0%, and Pronunciation ≈87.9% to 89.4%. An improvement of +0.5-2 points to alignment that is more dependable.

Unlike ASMDD, the Down syndrome no-reference audit on Figure 14, indicates an increase in Out of Vocabulary words from already high 86.9% to 89.6% after preprocessing. But looking at average metric scores in Figure 16: scores compared to before preprocessing were: Accuracy ≈93.7% to 95.1%, Fluency ≈94.0% to 96.6%, Completeness at 100.0%, and Pronunciation ≈93.4% to 95.4%. A similar improvement of +1-2 points despite guessing incorrect words without reference text.

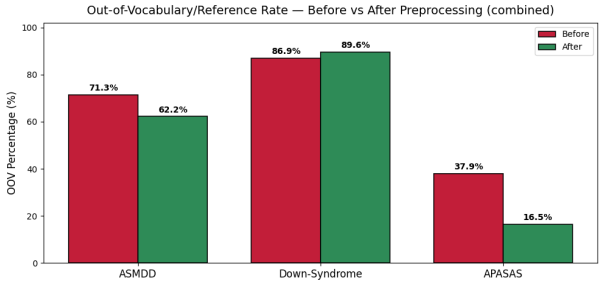

Figure 14. ASMDD, Down-syndrome and APASAS datasets before preprocess vs after preprocess without reference vocabulary match

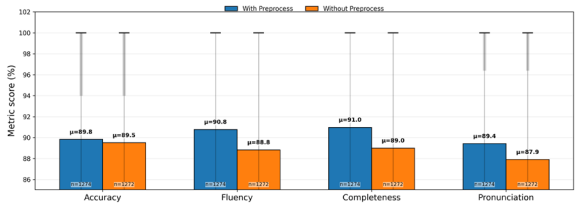

Figure 15. Down-syndrome dataset without preprocess vs with preprocess with reference metric scores

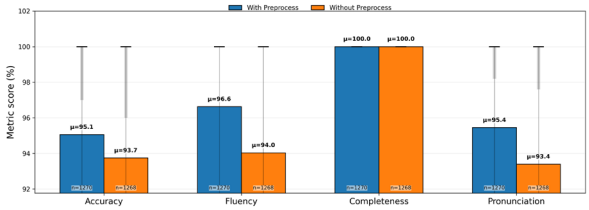

Figure 16. Down-syndrome dataset without preprocess vs with preprocess without reference metric scores

Reviewing the APASAS paragraph-length analysis, the with-reference indicators fall after preprocessing on Figure 17: Accuracy ≈ 85.2% compared to 95.2%, Fluency ≈ 82.0% compared to 91.4%, Completeness ≈ 89.2% compared to 99.4%, and Pronunciation ≈ 83.9% compared to 93.5%. The roughly consistent lower 9–10 percentage points changes indicate that in the lengthy sentences, enhancement alters timing and energy in a manner that negatively impacts forced alignment, even if segmental clarity might be improved.

Looking at Figure 18: without reference mode, while Accuracy scores went far below, 46.9% Accuracy compared to 97.4% previously, and other metrics fell between mid-40s to around mid-50s, compared to without preprocessing, which were around the full mark of all of the metrics. In Figure 14: the out-of-reference percentage decreases from 37.9% to 16.5% and the in-reference percentage increases from 62.1% to 83.5%. Therefore cleaned signal provides more words that are part of the intended passage when the recognizer is able to generate hypotheses without restrictions.

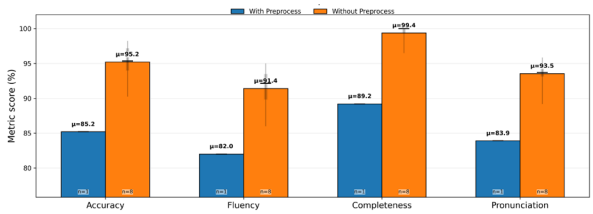

Figure 17. APASAS dataset without preprocess vs with preprocess with reference metric scores

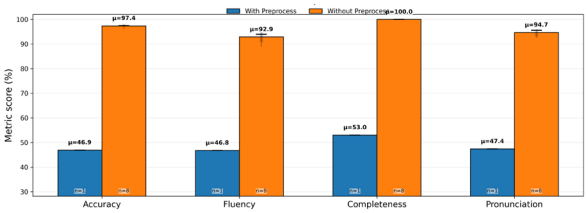

Figure 18. APASAS dataset without preprocess vs with preprocess without reference metric scores

In Figure 19, when no preprocessing is applied, Azure's scores reflected 0/4/4 (Beginner/Intermediate/Advanced), while the teacher's scores were 2/3/3. This suggests Azure tended to skew higher and was never assigned a Beginner level. With preprocessing, Azure adjusted to 1/4/3 to get closer to the teacher at 2/3/3.

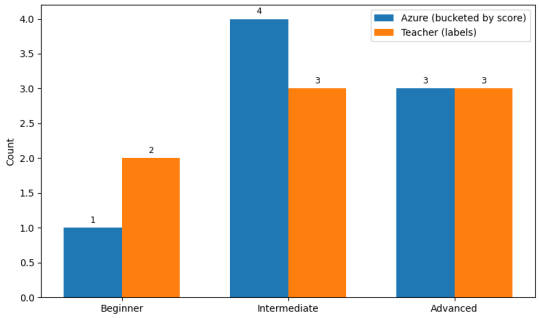

Figure 19. APASAS dataset without preprocess vs with preprocess with reference vs teacher evaluation

Preprocessing audio stopped the system's tendency to "guess through the noise." With higher SNR and sharper onsets/offsets, the alignment process became less forgiving: true deletions, truncated syllables, and over-long pauses were no longer masked and were penalized accordingly. In free recognition, the same clarity reduced spurious insertions and curtailed overconfident, self-reinforcing hypotheses, thereby removing an unintended source of score inflation. In short, preprocessing reduced artificial smoothing and increased diagnostic fidelity. The lower numbers, therefore, indicate stricter measurement of what speakers produced, not a loss of intelligibility.

The results confirm that preprocessing can reduce score inflation and improve recognition validity/alignment even though it may significantly lower some PA scores.

6. Conclusion

Azure PA was evaluated on the ASMDD, Down syndrome and the APASAS datasets, with and without a reference text. Preprocessing improved score reliability and reduced over-scoring on APASAS, aligning better with teacher ratings. No-reference scoring inflated false positives by grading the ASR hypothesis rather than the intended text. Residual errors concentrate in emphatics, pharyngeals, uvular /ق/, and interdentals. PA is suitable for reference-anchored formative feedback; no-reference use is best for coarse screening. Future work includes enhancing sentence-level alignment and integrating teacher priors in configuration.

Acknowledgements

We gratefully acknowledge King Fahd University of Petroleum & Minerals (KFUPM) for institutional support, access to facilities, and an enabling research environment throughout this work. We also thank the Undergraduate Research Office at KFUPM for their guidance, coordination, and continuous support, which materially contributed to the successful execution and dissemination of this study.

References

Ahmad Khan, A. F., Mourad, O., Amirul, M. K. B. M., Dahan, H. B. A. M., and Abushariah, M. A. M., "Automatic Arabic Pronunciation Scoring for Computer Aided Language Learning," *1st International Conference on Communications, Signal Processing and Their Applications (ICCSPA)*, pp. –, 2013. https://doi.org/10.1109/ICCSPA.2013.6487246

Ali, A., Vogel, S., and Renals, S., "Speech Recognition Challenge in the Wild: Arabic MGB-3," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 316–322, 2017. https://doi.org/10.1109/ASRU.2017.8268952

Al-Khatib, W. G., Amro, M. I., Alzahrani, A. B. S., Fanoush, T., and Elshafei, M., "Arabic Pronunciation Assessment for Saudi Arabian Students: Corpus Development and System Architecture," *Lecture Notes in Networks and Systems*, Vol. 862, pp. 28–40, 2024. https://doi.org/10.1007/978-3-031-71429-0_3

Alsulaiman, M., Ali, Z., Muhammed, G., Bencherif, M., and Mahmood, A., "KSU Speech Database: Text Selection, Recording and Verification," *UKSim-AMSS 7th European Modelling Symposium on Computer Modelling and Simulation (EMS)*, pp. 237–242, 2013. https://doi.org/10.1109/EMS.2013.41

Aly, S. A., Salah, A., and Eraqi, H. M., "ASMDD: Arabic Speech Mispronunciation Detection Dataset," Mendeley Data, 2021. https://doi.org/10.17632/x54dg53rmr.1

"Arabic Voice Recognition (Down Syndrome Children)," Kaggle Dataset, Retrieved August 31, 2025, from https://www.kaggle.com/datasets/alialhalabi3/arabic-voice-recognitiondown-syndrome-children

Arafa, M. N. M., Elbarougy, R., Ewees, A. A., and Behery, G. M., "A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation," *International Journal of Image, Graphics and Signal Processing*, Vol. 4, pp. 31–38, 2018. https://doi.org/10.5815/ijigsp.2018.04.04

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G., "Common Voice: A Massively-Multilingual Speech Corpus," *12th International Conference on Language Resources and Evaluation (LREC)*, pp. 4218–4222, 2020. https://arxiv.org/pdf/1912.06670

Fanoush, T., Al-Khatib, W. G., Amro, M. I., Alzahrani, A. B. S., and Elshafei, M., "Mispronunciation Detection and Diagnosis for Young Arabic Learners Using Transfer Learning," *IEEE Access*, 2025. https://doi.org/10.1109/ACCESS.2025.11186198

Hazan, V., and Pettinato, M., "The Emergence of Rhythmic Strategies for Clarifying Speech: Variation of Syllable Rate and Pausing in Adults, Children and Teenagers," 2014.

Kheir, Y. El, Chowdhury, S. A., Ali, A., Mubarak, H., and Afzal, S., "SpeechBlender: Speech Augmentation Framework for Mispronunciation Data Generation," *SLT/SLATE Workshop*, pp. 26–30, 2023. https://doi.org/10.21437/slate.2023-6

Krajka, J., "ENGLISH+KIDS," *Computer Assisted Language Learning Journal*, Vol. 20, No. 2, pp. 393–404, 2003. https://doi.org/10.1558/CJ.35173

Mubarak, H., Hussein, A., Chowdhury, S. A., and Ali, A., "QASR: QCRI Aljazeera Speech Resource – A Large Scale Annotated Arabic Speech Corpus," *59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Vol. 1, pp. 2274–2285, 2021. https://doi.org/10.18653/V1/2021.ACL-LONG.177

Sadik, M., and S., M., "Children Arabic Utterances for Mispronunciation Detection," *IEEE Dataport*, 2023

Biographies

Jalal Ali Zainaddin is a senior Computer Engineering student at King Fahd University of Petroleum & Minerals (KFUPM) and a former exchange student at Georgia Tech. His focus areas include AI-driven speech technologies and embedded systems. Jalal leads FloodSense Hybrid, a senior design project that integrates machine-learning flood-risk mapping with embedded sensing to deliver early, site-specific alerts through ESP32-based sensor nodes and API-driven dashboards. He co-develops the project's risk modeling pipeline and visualization tools for decision makers. Previously, he trained at Saudi Electricity Company (SEC), developed an Arabic-English Contracting Copilot Agent using Microsoft Copilot Studio and SharePoint, and contributed to WATHIQ (helmet/worker safety using LoRaWAN) and CateX (underground cable excavation detection). His interests include applied AI, edge computing, and socially impactful systems that bridge technology and real-world needs.

Mohammad Amro received the M.Sc. degree in Computer Science and the Ph.D. degree in Computer Science and Engineering from King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2009 and 2017, respectively. He has more than 18 years of academic and industrial experience. From 2006 to 2009, he served as a Lab Supervisor at Palestine Polytechnic University, Hebron, Palestine. He joined KFUPM in 2009, progressed

from a Research Assistant to Lecturer, and has served as a Technology Consultant and Advisor in the University's IT Department since 2017. He chairs the IT Strategy and Digital Transformation, is a member of the Smart Campus consultancy initiative, and leads KFUPM's collaboration with OpenAI to introduce ChatGPT Edu across the university and join the OpenAI Champion Network. He has taught courses in software testing, Python, C programming, and data structures. In 2023, he became a Guest Affiliate with KFUPM's Interdisciplinary Research Center for Intelligent Secure Systems, focusing on the assessment of Arabic mispronunciations and the identification of Arabic poetry meters. His research interests include speech recognition, machine translation, and software engineering.

Wasfi Al-Khatib received the B.S. degree in Computer Science from Kuwait University, Kuwait, in 1990, the M.S. degree in Computer Science from Purdue University, West Lafayette, IN, USA, in 1995, and the Ph.D. degree in Electrical and Computer Engineering from Purdue University in 2001. From 2001 to 2002, he was an Assistant Professor at Wright State University, Dayton, OH, USA. He is currently an Assistant Professor at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. His research interests include Arabic computing, multimedia computing, content-based retrieval, AI, and software engineering. He led funded research projects, supervised numerous graduate students, and contributed to curriculum development and ABET accreditation. He is a member of the ACM and IEEE Computer Societies.