KFUPM
Journal of Undergraduate Research International

Diamond Open Access 65

Article Computer Engineering

Evaluation of Azure Pronunciation Assessment Tool for Arabic-Speaking Learners

Jalal Zainaddin and Mohammad Amro*

Cite https://doi.org/10.64589/juri/221322

Submitted: November 29, 2025 **Revised:** February 19, 2026 **Accepted:** March 11, 2026

ABSTRACT

Accurate reading aloud is crucial for Arabic-speaking learners because minor errors can alter meaning or hinder understanding. This study evaluated the Microsoft Azure Pronunciation Assessment (PA) tool for modern standard Arabic using the KSU Arabic Speech Database. We focused on male speakers recorded in silent rooms. We used the KSU database's varied prompts words, phrases, and sentences to test Azure PA at the word, phoneme, and sentence levels. Because Azure PA is a closed source, we approximated its scoring by deriving equations for the Accuracy, Completeness, and Fluency metrics. We showed that a simple weighted combination closely matched the Pronunciation scores. We also applied a standard audio preprocessing pipeline (channel selection, denoising, pre-emphasis, silence trimming, and loudness normalization) to measure its impact. With preprocessing, for male speakers in the KSU database under the no-reference mode, Accuracy increased from 91.74% to 95.33%, Fluency from 60.84% to 97.84%, Completeness from 97.16% to 100.0%, and Pronunciation from 73.64% to 96.26%. The number of decoded word tokens increased fivefold. After preprocessing, all the consonant classes reached 90% accuracy, although the glottal segments remained weak. Sentence-level tests showed that Azure PA originally captured only brief fragments of long recordings. After preprocessing, it covered a greater portion of each sentence and recovered more of the target text. These results demonstrate both the strengths and limitations of Azure PA as a diagnostic tool for Arabic pronunciation based on KSU data.

Keywords: Arabic, audio preprocessing, automatic speech recognition (ASR), computer-assisted language learning (CALL), pronunciation assessment, speech scoring

1. INTRODUCTION

Advances in AI and technology have rendered traditional language schools less relevant. The focus has shifted to bridging language gaps, especially in speech, by using speech-to-text technology, an approach now adopted by companies such as Apple in their earphones. Automated speech and language technologies are being increasingly used to support language learning by providing fast and scalable feedback on spoken performance. A particularly important application is pronunciation assessment, in which learners need feedback not only on what they say (the transcript) but also on how accurately individual speech sounds are produced. Existing studies on Arabic speech recognition often report aggregate error metrics, such as word error rate (WER) and character error rate (CER); however, these measures are limited for diagnostic analysis because they treat recognition errors as equally important and do not reveal segment-level pronunciation behavior, such as performance on dentals, emphatics, and pharyngeals (see Table 1)[1]. A previous study tested Microsoft Azure Pronunciation Assessment (PA) on three children's datasets in Arabic and found that when automatic speech recognition (ASR) was used for evaluation without supplying a reference word to the tool, most of the recognized words were incorrect in the majority of datasets, and using audio preprocessing to fix some gaps led to higher and more accurate scores. This study used the Azure PA tool on the King Saud University (KSU) dataset to test the strengths and weaknesses of its ASR technology. The dataset features many young adult native Arabic speakers reading texts, answering questions, and pronouncing words with distinct phonemes. Identifying gaps will help improve Arabic speech-to-text technology.

2. RELATED WORK

Automatic pronunciation assessment (APA) systems, such as Krajka's English for Kids[2], exist for young English learners. However, computer-assisted language learning (CALL) tools for Arabic remain underdeveloped[3]. This section reviews Arabic speech datasets and summarizes their features and limitations. Reliable mispronunciation detection for Arabic learners requires high-quality speech datasets that include detailed demographics, varied recording conditions, and diverse pronunciations. Our previous study tested Azure PA, a cloud-based Microsoft service designed for CALL, teacher feedback, and automated oral-skill assessment. We evaluated its ability to recognize beginner-level

Table 1. *Phoneme-level segment types and corresponding letters*

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، هـ	Produced in the glottis
Velars	ك، ق	Back of the tongue
Dentals/Alveolars	ش,ت، د، س، ز، ن، ل	Front of the tongue
Labials	ب، م، ف	Lips
Vowels	ا، و، ي	Basic vowels or glides

reading in children by comparing the ASR scores with and without a reference text. Azure PA scores the pronunciation accuracy (correctness of sounds), fluency (smoothness and rhythm), completeness (coverage of expected words), and overall pronunciation at the phoneme, word, and sentence levels. It can be integrated into educational platforms through its application programming interface (API).

2.1. Objectives. To characterize Azure PA for Arabic beyond aggregate WER/CER metrics and quantify how input conditioning affects reliability, this study aims to

- benchmark Azure PA on the KSU Arabic Speech Database using its built-in ASR-based scoring across multiple prompt types;
- quantify the effects of standardized audio preprocessing/conditioning on the PA metrics (Accuracy, Fluency, Completeness, Pronunciation) and score stability;
- measure the ASR coverage improvements after preprocessing (e.g., increased decoded tokens/words and more complete hypotheses for long prompts);
- analyze segment/phoneme-category performance (e.g., dentals, emphatics, pharyngeals, glottals) to identify systematic strengths and weaknesses; and
- identify consistent failure patterns and summarize practical guidance for reliably deploying Azure PA for Arabic assessment/CALL tools.

3. METHODOLOGY

3.1. Dataset Review. Potential datasets were reviewed to select the most suitable one for this study. Table 2 summarizes all the reviewed datasets.

(a) Arabic speech mispronunciation detection dataset (ASMDD) The ASMDD contains single-word recordings from 100 Egyptian children (aged 2–8 years), supporting word-level mispronunciation benchmarking[4].

(b) QCRI aljazeera speech resource (QASR) QASR provides 2,041 h of multi-dialect Arabic broadcast speech with transcripts, mainly for transfer learning and dialect modeling[5].

(c) KSU arabic speech database The KSU Arabic Speech Database includes sentences and words of more than 300 speakers and recorded in diverse environments, thereby capturing pronunciation variability[6].

(d) Multi-Genre broadcast challenge 3 (MGB-3) MGB-3 covers 16 h of Egyptian Arabic from YouTube, spanning many genres[7].

(e) Children arabic utterances for mispronunciation detection This dataset includes 16 Arabic words pronounced in quiet settings by 27 Egyptian children (aged 7–12 years). It supports error-pattern studies[8].

(f) Dataset for speech recognition to support arabic phoneme pronunciation With 28 Arabic phonemes pronounced 10 times by 89 children, this dataset supports phoneme-level modeling[9].

(g) Arabic voice recognition (down syndrome children) This dataset features speech from 37 children (27 with Down syndrome and 10 with speech disabilities) and is useful for testing articulation diversity 10.

(h) Arabic pronunciation assessment for saudi arabian students (APASAS) This corpus for Saudi elementary children (ages 6–15) included 503 participants and 33 h of validated audio. It is the largest pronunciation assessment dataset for Saudi students[11,12].

(i) Common voice arabic dataset Common Voice Arabic contains volunteer speech, mostly from adults, but lacks mispronunciations and is less suitable for word-level benchmarking[13].

3.2. KSU Arabic Speech Database. In this study, the KSU Rich Arabic Speech Database was used as the main evaluation corpus. The KSU database provides recordings from hundreds of modern standard Arabic speakers and diverse prompts, including digits, words, sentences, and Q&A. Its comprehensive phonetic coverage enabled us to analyze the performance of Azure PA across challenging segment classes. The database uses clean pronunciations of young adult speakers to minimize confounding factors. Other datasets focus on children's speech, small vocabularies, or uncontrolled recordings, making them less suitable for precise analysis. By contrast, KSU database's controlled Multi environment and multichannel recordings provide consistent, high-quality data, supporting a robust evaluation and establishing it as an ideal choice for our study. Therefore, the other datasets were used as supplementary references.

3.3. Mispronunciation Detection in Children with Azure Pronunciation Assessment. Our earlier study systematically evaluated Microsoft Azure's pronunciation assessment system for Arabic-speaking children using three datasets:

Table 2. *Comparison of databases*

Database	Speakers	Dialect	Prompts	Sampling rate	Environment
ASMDD	100 children (2–8)	Egyptian	100 words	44.1 kHz	Quiet
QASR	– (2000 h)	Multi-dialect	Broadcast news	16 kHz	Studio / field
KSU Arabic Speech DB	>300	MSA, mixed accents	Digits, words, sentences	48 kHz	Quiet, office, cafeteria
MGB-3	– (16 h)	Egyptian	YouTube shows	16 kHz	Noisy, real-world
Children Arabic Utterances	27 children	Egyptian	16 words	–	Quiet
Phoneme Pronunciation DB	89 students	Arabic learners	28 phonemes ×10	16 kHz	Classroom
(Down-syndrome)	37 children	Mixed Arabic	Letters, words, numbers	–	Controlled
APASAS	503 students	Saudi MSA	Sentences, paragraphs	–	School
Common Voice Arabic	>1000 adults	Mixed Arabic	Read sentences	16 kHz	Home / mobile

Egyptian children's word-level recordings, recordings from children with Down syndrome, and Arabic pronunciation assessments for Saudi students. We compared reference scoring (aligned to the prompt) with no-reference scoring (grading Azure's own ASR output), and found that the reference text and phoneme-level scoring produced clearer and more accurate results, especially for early readers. By contrast, the no-reference mode often assigned high scores to incorrect or out-of-vocabulary words.

We also introduced standardized audio preprocessing, which improved the word-level scores and brought Azure's proficiency ratings closer to teacher assessments, particularly for beginners. However, even after conditioning, Azure PA struggled with Arabic-specific contrasts, such as emphatics, pharyngeals, uvular /ق/, interdentals, and final clusters. Therefore, this study serves as a key reference for comparing Azure PA with other engines and highlights its capabilities and limitations in cloud-based Arabic reading support for pronunciation assessment. It can improve mispronunciation detection for diverse groups and guide interventions for difficult segment types.

3.4. Overview of the Evaluation Pipeline. This study evaluated Microsoft Azure PA, a tool that scores how spoken words match the expected pronunciation, for modern standard Arabic using recordings from the KSU Arabic Speech Database. The end-to-end workflow is shown in Fig. 1, where KSU database audio is selected and submitted to Azure PA, and the returned ASR (software converting speech to text) hypothesis and scoring outputs are aggregated for downstream analyses (examining how specific sound groups perform, how much of each prompt is recognized, and the types of errors made).

3.5. Dataset and Audio Selection (KSU Arabic Speech Database). We used the male subset of the KSU Arabic Speech Database (Sessions 2 and 3; LDC2014S02), which is organized as session → speaker → environment. Each speaker/environment includes multiple prompt categories (e.g., word lists, fixed sentences, paragraph reading, and responses), enabling evaluation across short and long utterances and diverse phonetic content.

To reduce confounding from channel quality and recording variability, we focused on high-quality Yamaha_Mixer recordings from controlled environments (e.g., Silent_Room). All selected .flac files were batch converted to mono WAV files while preserving the dataset folder structure for consistent automated processing.

3.6. Azure Pronunciation Assessment Scoring Model. Azure PA is a closed-source cloud service built on Azure Speech-to-Text. It runs the ASR and then computes the Accuracy, Fluency, Completeness, and Pronunciation scores at the sentence, word, and phoneme levels. Because its implementation is not public, we treated the PA as a black box and approximated its behavior using Microsoft's official metric descriptions.

(a) Accuracy: GOP-style scoring from phonemes upward Microsoft states that "Syllable, word, and full text accuracy scores are aggregated from phoneme-level accuracy scores." Therefore, we modeled Accuracy as a goodness-of-pronunciation (GOP)-type score in Equation 1:

$$\mathrm{GOP}(p) = \frac{1}{T_p} \sum_{t \in T_p} \log \left(\frac{P(q_t = p | O_t)}{\max_{p''}} \right). \quad (1)$$

The simplified equation is expressed in Equation 2:

$$\text{Accuracy} = \left(\frac{\text{Number of Correct Phonemes}}{\text{Total Expected Phonemes}} \right) \times 100. \quad (2)$$

(b) Completeness: coverage ratio over the reference text Completeness only measures how the reference script is pronounced, and not its clarity (Equation 3):

$$\text{Completeness} = \left(\frac{\text{Number of Matched Words}}{\text{Total Reference Words}} \right) \times 100. \quad (3)$$

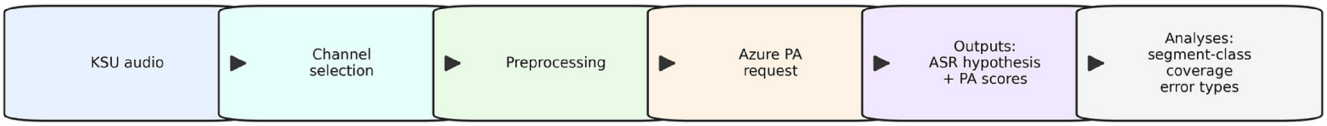

Figure 1. Complete workflow of data ingestion, Azure PA inference, and analysis

(c) Fluency: timing-based penalties Fluency penalizes unnatural rhythms, pauses, hesitation, and irregular rates. We approximate this because Azure's penalty system is not public (Equation 4).

$$\text{Fluency} = 100 - (W_1 \times \#Pauses + W_2 \times \#Hesitations + W_3 \times RateDeviation) \quad (4)$$

(d) Overall Pronunciation score: weighted combination of sub-scores Pronunciation is a combined score, heavily weighted toward the lowest sub-metric, based on a weighting of 0.6, 0.2/0.2 weighting (equation 5):

$$\text{PronScore} = 0.60s_0 + 0.20s_1 + 0.20s_2. \quad (5)$$

3.7. Audio Preprocessing (Overview). To improve robustness and reduce variability owing to recording conditions and silent artifacts, we applied a standardized audio conditioning pipeline before submitting the recordings to Azure PA (see Fig. 2). Each file was processed using bandpass filtering and noise gating, pre-emphasis, voice activity detection (VAD) based trimming, loudness normalization, and resampling. This preprocessing was applied uniformly across the evaluated subset to increase the decoding stability and improve the reliability of phoneme-level scoring. Full parameter settings and an ablation-style comparison (before and after pre-processing) are reported in the Experiments section.

4. EXPERIMENTS

4.1. Test Setup. We used the male subset of the KSU Rich Arabic Speech Database (Sessions 2 and 3, LDC2014S02), totaling 29.5 GB of audio. The data were organized by session, speaker, and environment (Office or Silent_Room), each containing text-category folders. Each folder contained recordings from Mic_CreativeSB.flac (medium quality) and Yamaha_Mixer.flac (high quality). We analyzed only the high-quality YamahaMixer recordings from controlled environments.

For this study, we focused on the high-quality mixer channel under controlled conditions, and therefore, only used the Yamaha_Mixer track from the Silent_Room (or office) environment.

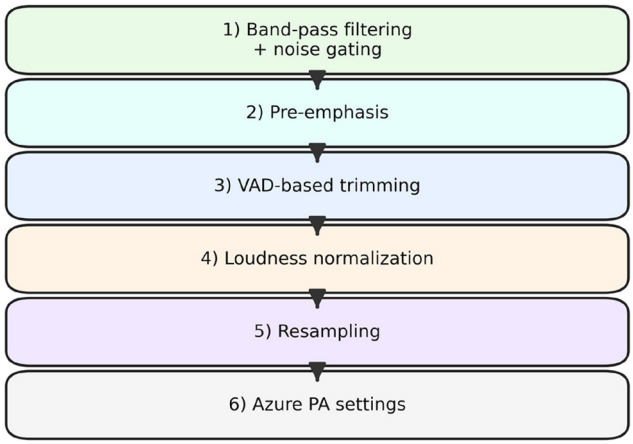

Figure 2. Preprocessing pipeline applied prior to Azure PA

Each environment included 10 text-category folders designed to ensure phonetic richness. Each speaker/session/environment had the following characteristics:

- SAAB_sentences_1 and 2: Continuous readings of sublists. SAAB_sentences_1 is common across speakers, and SAAB_sentences_2 is randomly assigned, enabling the assessment of long, phonetically dense material.
- Fixed_Sentences: Two short, accent-sensitive sentences for dialect discrimination.
- Numbers: Recordings of digits 0–9 were included because these numbers are common in real-life scenarios.
- Common_Words_1 & 2: Ten high-frequency familiar words per list covering most Arabic letters.
- Paragraphs_1 & 2: Two long paragraph readings—one prose and one Qur'anic—both letter-complete.
- Phonetic_Distinctive_Words_1 & 2: Lists of words emphasizing nasals, fricatives, and vowels to probe difficult contrasts.
- Common_List: Shared, phonetically balanced sentence prompts for most speakers.
- Random_List: Additional phonetically balanced sentences assigned to subsets of speakers.
- Questions_1 & 2: Recordings of volunteers answering simple spoken prompts and providing semi-spontaneous conversational speech.

For Azure integration, each Yamaha_Mixer.flac file was batch converted to a mono WAV format, preserving the folder structure. A Python script scanned these files, extracted metadata, and ran two Azure Speech passes: (1) ASR to obtain a hypothesis transcript, and (2) Pronunciation Assessment using the ASR result as a reference. All scores and breakdowns were aggregated for a systematic evaluation, thus eliminating the need for manual transcription. All experiments used Azure Speech locale Arabic-Saudi Arabia (ar-SA) with consistent PA settings:

- 100-point grading scale
- Phoneme-level granularity (to support segment-level diagnostics)
- Miscue detection enabled
- InitialSilenceTimeoutMs = 4000 ms to reduce early cutoff and improve decoding continuity on longer files

We evaluated Azure PA under two scoring modes:

- With-reference mode: Prompt/reference text was provided (for tasks with a target script), enabling alignment-based scoring and word/phoneme feedback.
- No-reference (ASR-driven) mode: Azure's ASR hypothesis was used as the reference, enabling transcript-free large-scale processing, but potentially inflating scores when ASR produced incorrect words.

4.2. Experimental Conditions. We ran two main conditions: (C1) Raw audio → Azure PA and (C2) Preprocessed audio → Azure PA, where preprocessing followed the steps in Fig. 2 (band-pass/noise gating, pre-emphasis, VAD trimming, loudness normalization, resampling). Azure was configured consistently across runs (local ar-SA, 100-point scale, phoneme granularity, miscue detection, and extended initial-silence timeout).

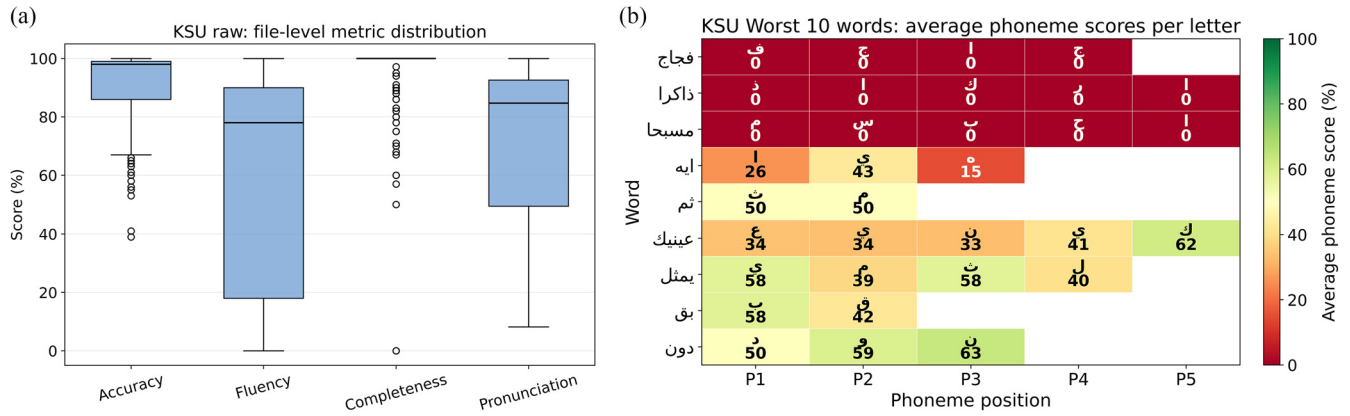

Figure 3. Azure Pronunciation Assessment results: (a) metric scores; (b) worst 10 words, average phoneme (per letter) scores

4.3. Procedure and Outputs. For each file, the pipeline (Fig. 1) submitted audio to Azure and logged (i) the ASR hypothesis and (ii) PA scores (Accuracy, Fluency, Completeness, Pronunciation) with word/phoneme-level details. We aggregated the results per condition and analyzed the performance according to prompt type and Arabic segment classes (Table 1).

5. RESULTS

5.1. Numerical & Graphical Results. As shown in Fig. 3 (a), without reference, the KSU database mean scores were 91.74% Accuracy, 60.84% Fluency, 97.16% Completeness, and 73.64% Pronunciation. Accuracy clustered above 90%, reflecting clean acoustics rather than perfect script adherence.

Fluency varied significantly (from 0 to 100), averaging 60.84%, likely because of Azure's sensitivity to pauses and hesitation, especially in long or conversational passages.

Completeness reached nearly 100% across most files, with the ASR hypothesis serving as the reference for evaluation.

Pronunciation averaged 73.64%, between the Accuracy/Completeness and Fluency scores, closely matching the predicted score distribution. This means that the distribution of these three variables is similar to what is suggested, but not at the decimal point level.

In Fig. 3(b), words like "فجاج" always received 0 Accuracy and an Omission error, meaning Azure could not align any token to the reference. The same occurred for "ذاكرا" and "مسبحا".

For "ثم", the phoneme scores were low (15%–40%), which is usually labeled as Mispronunciation, as shown in Fig. 4 (a). This highlights the Azure PA's weakness in glottal-heavy interjection.

"ثم" scored near 100%, reflecting Azure's strength on dental, labial, and vowel segments, although performance can decline in fast speech due to merging with adjacent words.

"عينيك" contains pharyngeal, vowel, nasal, and velar sounds, revealing gaps in Azure's handling of pharyngeal–vowel–nasal patterns.

The word "بق" mostly came from a wrong ASR assumption in "أصابكم بقذيفة"; an example of ASR output is "أصابكم بق؟". The full audio text was "رأيت قذيفا، على واستعظم الإنسان، لماذا أخطأ ، وآل يرث ملكنا، كن هنا فهل كانت حلب إمارة ، أدم، سيشكروننا في ندوة، من غضب لعلمهم، قابلهم وأصبر، لا لم يبقى زوجها", but the ASR output cut it by a huge margin, reducing it to only "أصابكم بق؟". This behavior was observed in many, if not most, cases, suggesting that ASR struggles with long audio texts.

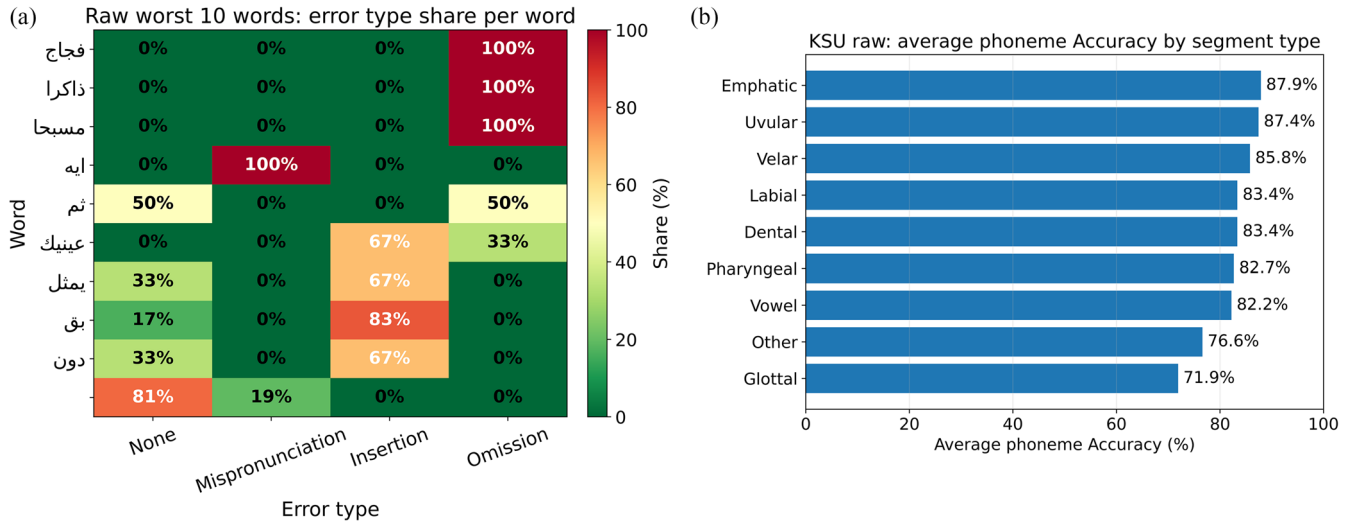

Figure 4. Azure Pronunciation Assessment results: (a) error type for the worst 10 words; (b) average pronunciation accuracy scores by phoneme segment type

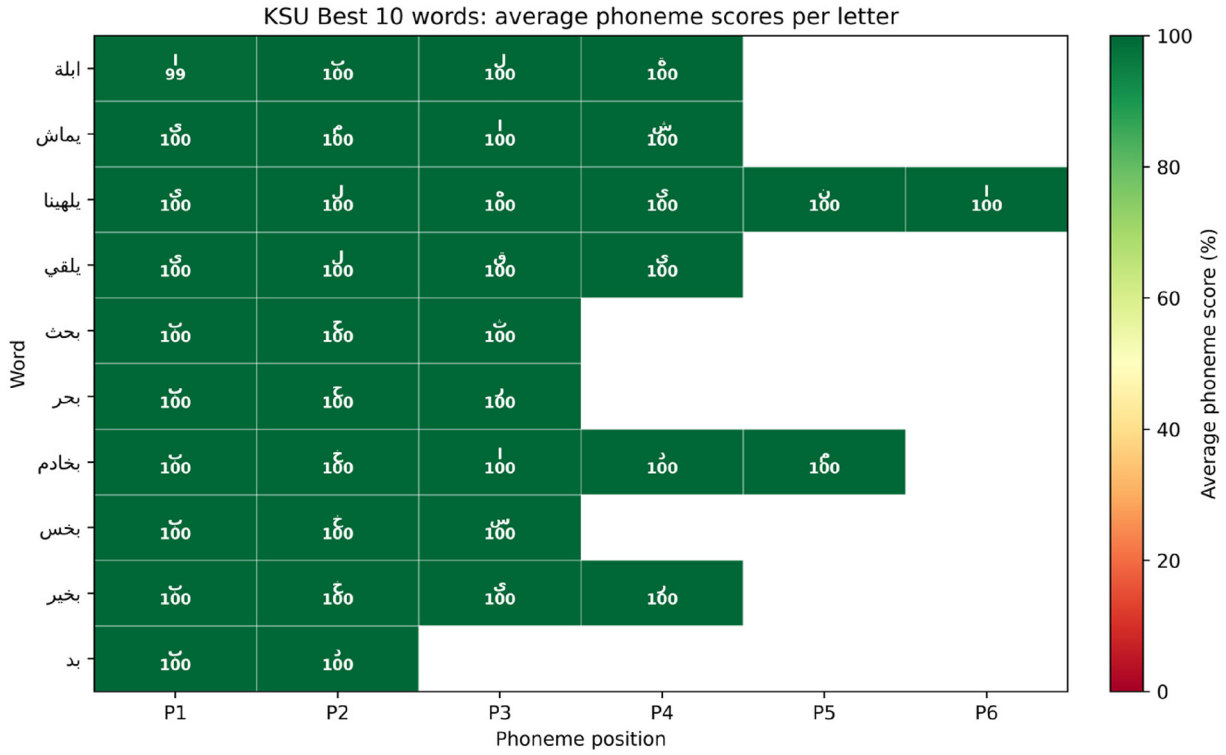

Figure 5. KSU database average phoneme scores (per letter) for the best 10 words

This example is unique because it consists of many unrelated sentences with short pauses. Consequently, ASR took the first sentence it recognized and stopped too early, after a small pause from the reader. Moreover, as it presented the output of the same text, it always left out the first sentence, "فهل كانت حلب إمارة," and took the second one, "رأيت قذيفا.".

Figure 5 shows the words with the highest phoneme scores, but most are false positives (e.g., "يماش ابلة") that do not exist in the original sentences. This indicates that Azure PA is unreliable for long sentences without a reference text. Some correct words, such as "بحر" (/b/, vowel, /h/, /r/), received 100% scores, showing that Azure PA can reliably recognize all phonemes in simple words, even for difficult consonants such as /h/.

In summary, Azure PA performs best on lexically salient and stressed words, avoiding glottally heavy or very short structures, and achieves high phoneme scores even for difficult consonants. Many errors in the "worst words" are due to alignment and prosodic reduction rather than the inability of the model to recognize consonants.

5.2. Proposed Improvements. Five main acoustic/phonetic issues degrade Azure PA for children's speech. We addressed loudness variation, low signal-to-noise ratio (SNR), and clipping by applying loudness normalization (−23 LUFS/dBFS RMS) and peak clamping. The HVAC/handling

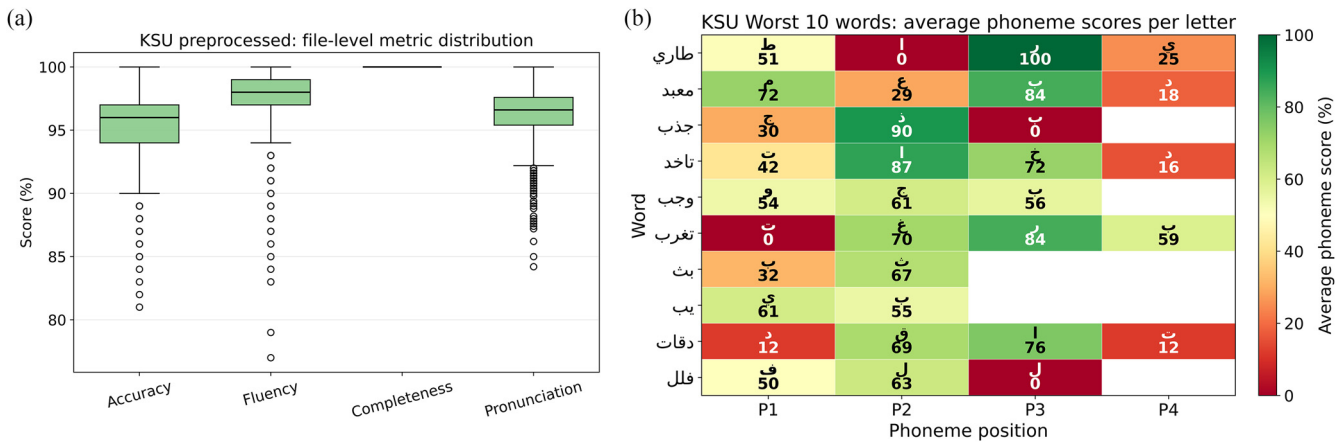

Figure 6. Azure Pronunciation Assessment results with preprocessing: (a) metric scores; (b) average phoneme (per letter) scores for the worst 10 words

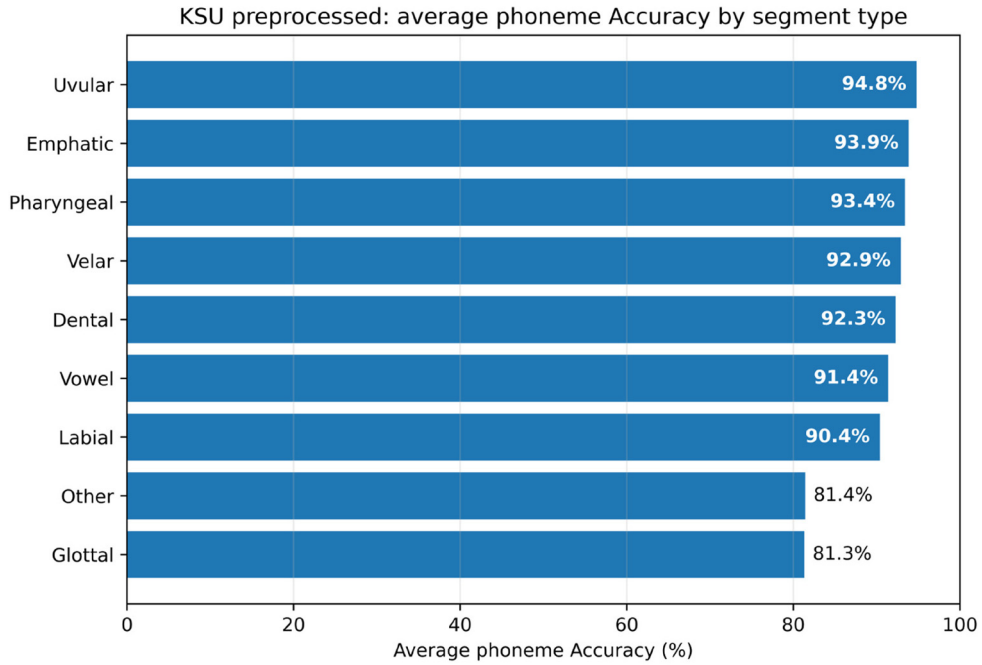

Figure 7. KSU database with preprocessed average pronunciation accuracy scores by phoneme segment type

noise was reduced using a Butterworth band-pass filter (80–7000 Hz) and spectral gating, thereby improving the SNR for low-energy consonants. To correct high-frequency fricative underestimation (/s, ʃ, θ, f/), we applied pre-emphasis ($\alpha = 0.97$) and used a fricative-safe VAD to retain low-energy sounds (Equation 6).

$$y[n] = x[n] - \alpha x[n-1] \tag{6}$$

Audio preprocessing steps included pre-emphasis ($\alpha = 0.97$, 3–8 kHz), a 4th-order Butterworth band-pass filter (80–7000 Hz), spectral gating, loudness normalization (−23 LUFS/dBFS RMS, peak clamp), WebRTC-VAD (20 ms, aggressiveness = 2, +120 ms hangover, 200 ms min, <150 ms gap), 16 kHz resampling, and setting Azure PA InitialSilenceTimeoutMs to 4000 ms.

5.3. Validation. After preprocessing, the scores improved to 95.33% for Accuracy, 97.84% for Fluency, 100.00% for Completeness, and 96.26% for Pronunciation, with Fluency showing the greatest gain. This confirms that the earlier disfluency was mostly due to recording artifacts and silences rather than speaker issues. Fig. 6 (a) shows that Accuracy and Fluency clustered near the top. Completeness was always nearly 100%, and Pronunciation closely tracked the other metrics. These results validated the KSU database corpus and our preprocessing. Azure PA consistently achieved high scores with conditioned audio, confirming that the subsequent analyses reflected Azure PA's true performance.

The lowest-scoring words in Fig. 6 (b) highlight specific Azure PA challenges. For "طاري", consonants were recognized reasonably ط (51%), ر (100%), but the long vowel ا (0%) and final ي (25%) were under-scored, likely due to vowel reduction in natural speech. In the word "معبد", ر (72%) and ب (84%), but the pharyngeal ع (29%) and final د (18%) were problematic, showing Azure's general difficulty with pharyngeals and word-final stops, which can be blurred or devoiced. For "تغرب", the initial ت was missed (0%), غ (70%), ر (84%) and ب (59%) were detected, pointing to challenges with word-initial stops, especially if spoken quickly or joined to a previous word.

Figure 7 shows that after preprocessing, the phoneme accuracy improved for all segment types. Most classes now exceeded 90% accuracy, with the uvular, emphatic, and pharyngeal segments showing substantial gains. Glottal remained the weakest but also improved. This suggests that the previous variability was mainly due to channel and silence artifacts, and not model limitations.

For the sentence-based KSU database material, Fig. 8 (a, b, c, d) demonstrates that after preprocessing, Azure PA processed many more audio files than with the raw recordings. The total decoded word count increased from approximately 14,000 to over 74,000 after preprocessing because Azure analyzed each file more deeply and produced more complete transcripts. The ASR coverage charts confirmed that more files yielded more usable hypotheses. For the "target sentence" figures, the improvement was clear in both key scripts. For Text 1 (the long chain "كن هنا فهل ... ، كانت حلب إمارة ، رأيت قذيفا، على واستعظم الإنسان، لماذا أخطأ أدم"), the raw ASR often collapsed the entire waveform into a brief fragment like "أصابكم بق؟". After preprocessing, a dramatic jump in detections occurred, and nearly every clause in Text 1 ("كن هنا", "فهل كانت حلب إمارة" "رأيت قذيفا" "قابلهم وأصبر", etc.) was now recovered in many more files. Azure followed the reader through pauses and extracted most of the scripted material. For Text 2 ("لحظة من فضلك السلام عليكم" "لا" "نعم" "لا أدري" "شكرا" "الله أعلم" "قريبا" "أخبار" "زيارة"), the improvement was even more pronounced. Detection counts increased from a few hundred to over 3,000, with big gains for nearly every prompt, especially "السلام عليكم", "نعم", "لا", and "لحظة من فضلك". The previously missed or partially decoded tokens now appeared reliably and repeatedly.

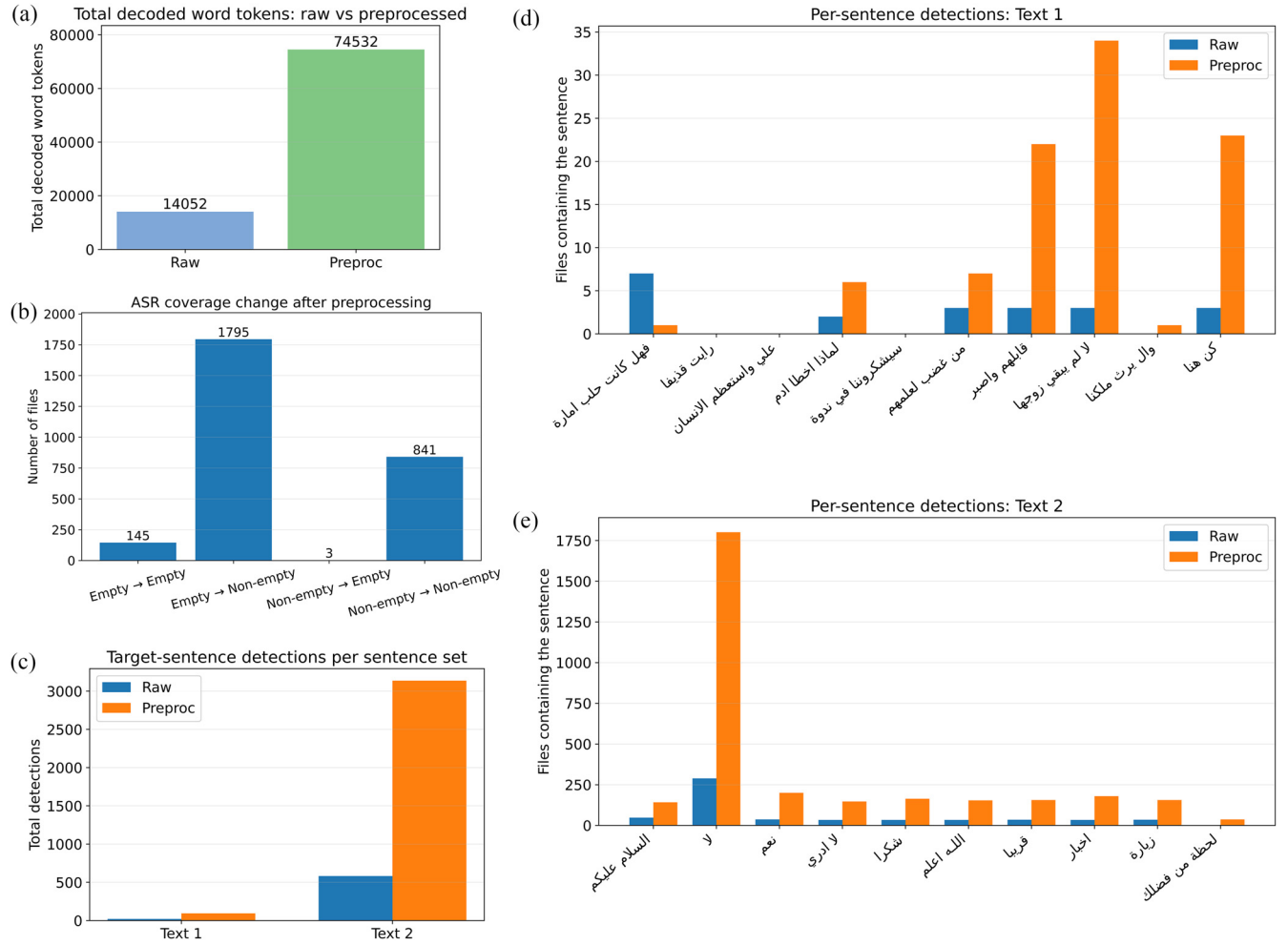

Figure 8. Azure Pronunciation Assessment results on APASAS dataset with and without preprocessing: (a) with reference; (b) without reference

Overall, these results show that preprocessing transforms Azure PA performance, from capturing only short fragments to analyzing and transcribing a much larger portion of each recording, resulting in outputs that more closely match the full KSU database sentence prompts.

6. DISCUSSION

Interpretation of preprocessing gains. The consistent improvement after preprocessing suggests that Azure PA reliability depends strongly on input conditioning. Because Azure is trained primarily on fluent adult speech, normalizing the bandwidth, silence, and loudness reduces mismatches and improves decoding continuity, particularly for longer utterances.

No-reference (ASR-driven) scoring limitation. In the no-reference mode, Azure PA scores can be inflated when the ASR hypothesis is incorrect because the assessment is partially aligned to the system's own transcript. Therefore, high Completeness/Pronunciation values should be interpreted alongside coverage and error-type indicators rather than as direct evidence of correct lexical recognition.

Segment-class implications. The segment-class analysis (Table 1) provides diagnostic insight beyond WER/CER-style reporting by highlighting which Arabic sound groups remain challenging (e.g., emphatics/pharyngeals/glottals), supporting targeted improvements in Arabic pronunciation feedback and CALL settings.

Generalizability. These findings are limited to a controlled subset (KSU database, adult male speakers, selected channels/environments, and ar-SA locale) and may not be generalizable to children, disordered speech, noisy mobile recordings, or other dialect locales/engines.

7. CONCLUSION

This study evaluated Microsoft Azure PA using the KSU Arabic Speech Database, focusing on high-quality male recordings to isolate the system behavior. We reverse-engineered Azure PA's scoring, modeling Accuracy, Completeness, and Fluency, and found a simple weighted combination that closely matched the reported Pronunciation scores. Audio preprocessing was essential; on raw audio, Azure PA often decoded only short fragments, resulting in low Fluency and Pronunciation scores. With preprocessing, the decoded words increased fivefold, scores stabilized in the high-90% range, and the system captured more target sentences. Segment-level analysis showed strong performance for uvular, emphatic, and pharyngeal consonants, but persistent

weaknesses for glottal consonants and certain word types. Overall, Azure PA with preprocessing is a valuable tool for segment-level diagnostics in Arabic pronunciation, although users must be aware of its limitations in certain phoneme segments, reduced words, and long passages. Future work should extend this analysis to other speaker groups and explore hybrid systems that combine Azure PA with custom Arabic phoneme models to obtain more robust feedback.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Jalal Zainaddin – *Department of Computer Engineering, King Fahd University of Petroleum & Minerals, Dhahran 31261, Saudi Arabia*; 0009-0006-3811-4232
Email: s202154790@kfupm.edu.sa

Corresponding Author

Mohammad Amro – *Research Mentor, Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), King Fahd University of Petroleum & Minerals (KFUPM), Dhahran 31261, Saudi Arabia*; 0009-0007-7624-1491
Email: mamro@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors acknowledge funding support from the Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS) at King Fahd University of Petroleum & Minerals under Project No. INSS2522. We also thank the Computer Engineering Department and the Undergraduate Research Office at KFUPM for their guidance, coordination, and support.

REFERENCES

(1) Miner, A. S. *et al.* Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digit. Med.* **3**, (2020).

(2) Krajka, J. ENGLISH+KIDS. https://doi.org/10.1558/cj.35173 **20**, 393–404 (2003).

(3) Ahmad Khan, A. F., Mourad, O., Amirul, M. K. B. M., Dahan, H. B. A. M. & Abushariah, M. A. M. Automatic Arabic pronunciation scoring for computer aided language learning. *2013 1st International Conference on Communications, Signal Processing and Their Applications, ICCSPA 2013* https://doi.org/10.1109/ICCSPA.2013.6487246 (2013) doi:10.1109/ICCSPA.2013.6487246.

(4) Aly, S. A., Salah, A. & Eraqi, H. M. ASMDD: Arabic Speech Mispronunciation Detection Dataset. https://doi.org/10.17632/x54dg53rmr.1 (2021) doi:10.17632/x54dg53rmr.1.

(5) Mubarak, H., Hussein, A., Chowdhury, S. A. & Ali, A. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* **1**, 2274–2285 (2021).

(6) Alsulaiman, M., Ali, Z., Muhammed, G., Bencherif, M. & Mahmood, A. KSU speech database: Text selection, recording and verification. *Proceedings - UKSim-AMSS 7th European Modelling Symposium on Computer Modelling and Simulation, EMS 2013* 237–242 (2013) doi:10.1109/EMS.2013.41.

(7) Ali, A., Vogel, S. & Renals, S. Speech recognition challenge in the wild: Arabic MGB-3. *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017 - Proceedings* **2018-January**, 316–322 (2017).

(8) Sadik, M. and M. S. Children Arabic utterances for mispronunciation detection. *IEEE Dataport* (2023).

(9) Arafa, M. N. M., Elbarougy, R., Ewees, A. A. & Behery, G. M. A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation. *Image, Graphics and Signal Processing* **4**, 31–38 (2018).

(10) Arabic voice recognition(Down syndrome children). https://www.kaggle.com/datasets/alialhalabi3/arabic-voice-recognitiondown-syndrome-children.

(11) Al-Khatib, W. G., Amro, M. I., Alzahrani, A. B. S., Fanoush, T. & Elshafei, M. Arabic Pronunciation Assessment for Saudi Arabian Students: Corpus Development and System Architecture. *Lecture Notes in Networks and Systems* **862 LNNS**, 28–40 (2024).

(12) T. Fanoush, W. G. Al-Khatib, M. Amro, A. Alzahrani and M. Elshafei, "Mispronunciation Detection and Diagnosis for Young Arabic Learners Using Transfer Learning," in IEEE Access, vol. 13, pp. 175047–175068, 2025, doi: 10.1109/ACCESS.2025.3616335 (2025).

(13) Ardila, R. *et al.* Common Voice: A Massively-Multilingual Speech Corpus. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings* 4218–4222 (2019).