SIRAJ: Benchmarking TTS Systems on Dialect-Diverse Arabic Text-to-Speech

Anonymous ACL submission

Abstract

Despite significant advances in speech processing technologies, a substantial number of the world's languages and their dialects remain under-resourced. This disparity contributes to a widening technological gap, thereby impeding broader technological inclusion. One obstacle to inclusion is the scarcity of comprehensive datasets for developing robust multi-dialectal Text-to-Speech (TTS) systems. In this work, we present SIRAJ, an extensive dataset of Arabic dialects for dialectal Text-to-Speech systems. SIRAJ contains audio, transcription, speaker gender, and dialect information for six Arabic dialects: Egyptian, Emirati, Jordanian, Saudi, Moroccan, and Syrian. We find that fine-tuning on SIRAJ affords human-like audio quality in generated samples for some dialects. By establishing strong baselines, we hope that SIRAJ can serve as a basis for future work in TTS for Arabic dialects.

1 Introduction

The Arabic languageis spoken by more than 450 million people across a rich array of dialects, presents a unique landscape for the development of speech technology. Although Modern Standard Arabic (MSA) is prevalent in formal written and broadcast media, everyday communication occurs predominantly in regional variations. In spite of this diversity of dialects, speech technologies for Arabic often collapse dialects or focus on MSA (Bouamor et al., 2018), or narrowly focus on one dialect, primarily Egyptian (e.g., Laouirine et al., 2024; Halabi, 2016; Azab et al., 2023), thereby serving only a small portion of Arabic speakers. This underscores a critical gap in speech research on Arabic: There is a scarcity of comprehensive resources across the breadth of Arabic dialects, which, in turn, hinders the creation of truly inclusive speech recognition and synthesis tools.

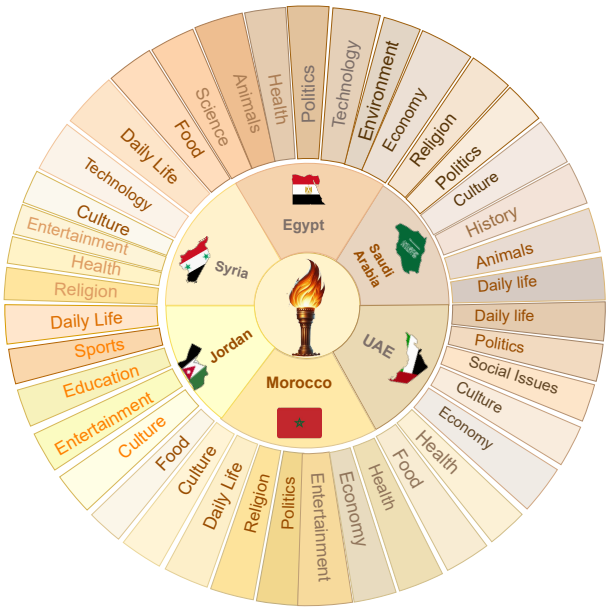

Figure 1: Overview of the core domains and sub-domains in SIRAJ. Our benchmark spans six major dialects and 16 sub-domains, providing a comprehensive evaluation framework for multi-Arabic dialects.

In this work, we address this gap in research by constructing SIRAJ (سراج or: torch), a dataset of speech recordings for the development of Text-to-Speech (TTS) systems for Arabic across six distinct dialects: Egyptian, Emirati, Jordanian, Saudi, Moroccan, and Syrian. For each dialect, we present the speaker audio, speech transcriptions, speaker gender identity, and dialect information. We then use SIRAJ to establish robust baseline models for Arabic dialectal TTS systems; for comparison for future work. Finally, we conduct extensive experiments to evaluate our models under different text-inputs, such as writing with and without diacritization, and examine the cross-dialect generalization performance of our models. We show that models are able to capture the phonetic features of the different dialects using around 3 hours of data for

finetuning, scoring as high as 4.38 MOS points in the Moroccan dialect. Moreover, the results show promising insights for cross-dialect transferability.

The contributions of this paper are (i) presenting SIRAJ: a carefully constructed, multi-dialectal dataset for Arabic dialects for Text-to-Speech research; (ii) developing state-of-the-art TTS models based on XTTSv2, FishSpeech, MMS-TTS for 6 dialects of Arabic which can serve as a benchmark for future work towards high-fidelity dialectally-aware Arabic speech synthesis; and (iii) extensive experiments to assess the synthesis quality for plain and diacritized input and evaluation of cross-dialect generalization of our models. Thus we provide insights into dialect transferability and the effects of phonetic disambiguation, and hope to enable future work on dialectal Text-to-Speech models for dialects of Arabic.

2 Related Work

2.1 Dialectal Arabic and Linguistic Variation

Spoken Arabic forms a rich dialect continuum of regional varieties that differ in phonology, lexicon, and prosody (Watson, 2002). This can produce an inability for models to identify the correct dialect for some words, such as كيف حالك؟ (How are you?) due to a lack of clear dialectal markers, while models are more easily able to identify the correct dialect for other words, e.g., the temporal adverb الآن (Now) appears as دلوقتي (Egyptian), هلق (Levantine), and دابا (Moroccan). As such, TTS systems may produce inconsistencies when generating speech from regional dialects (Algihab et al., 2019). Mutual intelligibility between distant dialects can drop below 70% (Čéplö et al., 2016), implying there is linguistic divergence which would complicate pronunciation modelling and prosody control in multi-dialect TTS. Yet most speech corpora collapse dialects or focus solely on MSA (Bouamor et al., 2018), limiting the community's ability to build and benchmark dialect-aware systems.

2.2 Arabic TTS Resources

Early research on Arabic TTS systems focused on small corpora for standardized Arabic. For example, the *Arabic Speech Corpus* (ASC) consists of 3.7 hours of diacratized MSA speech from a single Syrian speaker and has been used to develop end-to-end models (e.g., Halabi, 2016). Similarly, although ClarTTS (Kulkarni et al., 2023) provides more data–12 hours of Classical Arabic recordings from audio books–it also relies on a single speaker and dialect. The *QCRI Aljazeera Speech Resource* (QASR) presented 2000 hours by over 28,000 speakers but lacks human-verified dialect labels and has an imbalanced gender distribution of speakers (Mubarak et al., 2021). Attempts to adapt QASR for TTS have involved automatic dialect labeling, resulting in noisy annotations (Doan et al., 2024). To address the lack of dialect labels, more dialect-specific datasets have emerged, e.g., *TunArTTS*, a 3-hour manually diacritized corpus for Tunisian Arabic (Laouirine et al., 2024).

Recently, there has been an increase in the development of multi-dialect datasets for various NLP and Speech tasks (e.g., Talafha et al., 2024; Nafea et al., 2024; Alwajih et al., 2025), yet to the best of our knowledge, there exist no manually annotated datasets for multi-dialect Arabic TTS. SIRAJ aims to address this gap by collecting speech from six regional dialects, each represented by a carefully selected native speaker. All utterances are manually transcribed, annotated for the gender of the speaker, and labeled by dialect, thereby providing a robust foundation for dialect-aware Arabic TTS systems.

2.3 Multilingual and Zero-Shot TTS Models

Recent advances in TTS have emphasized multilingual and zero-shot architectures leveraging large-scale cross-lingual training data and base models. For example, Liao et al. (2024) present **FishSpeech**, a dual auto-regressive model with a Firefly-GAN vocoder trained on more than 720,000 hours of audio (including 20,000 hours of Arabic) to capture diverse phonetic and prosodic features; Casanova et al. (2024) present the **XTTSv2** model which is based on the Tortoise framework and offers zero-shot voice and dialect cloning from minimal audio inputs across 17 languages; and Pratap et al. (2023) release **MMS-TTS** which uses the VITS architecture (Kim et al., 2021) with language specific adapters for over 1,100 languages (including Arabic) using multilingual CTC objectives. Other efforts, such as **ArTST** (Toyin et al., 2023) further extend these methods by enabling few-shot dialect adaption using speech encoders and specialized vocoders, but their model is developed for MSA and is not expected to perform well in a multi-dialect setting. Additionally, commercial closed-source services such as **ElevenLabs** and **Speaktor** provide multilingual

Dataset	Hours	#Spkrs	Dialects	Gender	Domains	Annotation
ASC	3:42	1	MSA	Male	Audiobook	Fully manual
ClArTTS	12:00	1	Classical Arabic	Male	Audiobook	Fully manual
QasrTTS	1:00	1	MSA + Broadcast	Male	Broadcast	Manual review
TunArTTS	3:00	1	Tunisian Arabic	Male	Dictionary	Fully manual
SIRAJ (Ours)	**17:24**	**6**	**6 Dialects**	**Mixed**	**Multi-domain**	**Semi manual**

Table 1: Comparison of Arabic TTS datasets. M: Male, F: Female, U: Unknown.

TTS APIs for on-demand usage.

However, despite their multilingual capabilities or capabilities to Arabic, these models struggle with nuanced dialectal variation. To address this gap, our work introduces a multi-dialectal corpus specifically designed to benchmark and fine-tune state-of-the-art models with an aim to enhance their ability to generate authentic, intelligible speech across Arabic dialects.

3 Dataset

3.1 Data Collection

We construct SIRAJ by identifying YouTube channels by native speakers of six dialects (Egyptian, Saudi, Emirati, Jordanian, Syrian, Moroccan), where recordings are conducted in studio settings and feature monologue-style content to ensure high audio quality ($\geq$128 kbps, 44.1 kHz) and appropriate data for TTS.[1] We collect data from one speaker for each dialect, as prior work has shown that a small number of speakers is helpful to achieve more natural prosody and good pronunciation in data scarce settings (Neekhara et al., 2021).[2] We then apply a SpeechBrain CRDNN-based voice activity detection model (Bredin, 2021) to extract speech segments between 10 and 20 seconds long, filtering based on silence thresholds and energy-based heuristics. We align the speech segments with YouTube's automatic captions using timestamp overlaps ($\geq$30%).

3.2 Annotation Process

Our dataset is curated by a team of six trained annotators who are native speakers or fluent in the dialects they annotated. The annotators were selected for their strong linguistic familiarity with their own and other Arabic dialects.

We use the automatically generated captions from YouTube as the starting point for our annotation process. These transcriptions contains different types of errors, thus, we developed an annotation protocol to fix the transcriptions. A guiding principle in our annotation process was to emphasis phonetic accuracy over grammatical correctness or contextual coherence, i.e., annotators were instructed to transcribe speech as pronounced. The annotation workflow was carried out in a sequence of structured steps to ensure consistency and quality across the dataset[3]:

- **Transcript-Speech Alignment.** Annotators aligned transcripts with audio by deleting unspoken words, inserting clearly pronounced missing ones, and verifying the beginning and end of segments to prevent missing or extra words.
- **Flagging of Problematic Segments.** Non-speech segments (e.g., music, background noise) were flagged with the color red (unusable), while minor cutoffs were flagged with the color orange (reviewable). The flagged segments were not transcribed.
- **Use of Dialectal Orthography.** Words were transcribed exactly as spoken in the speaker's dialect. MSA forms were avoided unless explicitly used in the speech.
- **Correction of Pronunciation-Affecting Errors.** Spelling and morphological errors were corrected, and homographs were disambiguated to preserve accurate pronunciation.
- **Selective Diacritization.** Tashkīl (diacritics) were only applied to resolve ambiguity in pro-

[1] The videos were retrieved using the YouTube Data APIv3 (Google, 2016), downloaded with `yt_dlp` (YTDL Contributors, 2024), and converted to 16-bit PCM WAV at 16 kHz.

[2] We obtained consent from the owners of each channel to their data being collected and used for training.

[3] we will be releasing the guidelines with the dataset

Dialect	Total Dur	Avg Dur	AVT	U-Wds	Avg WPS / CPS	# Segments
Egypt	02:47:31	12.99	34.41	8,967	2.65/14.56	774
Syria	02:39:59	13.66	31.48	4,967	2.31/12.02	703
Saudi	02:55:57	14.15	27.96	8,246	1.98/10.86	746
UAE	02:53:49	13.44	23.19	7,029	1.73/09.28	776
Morocco	02:49:28	13.14	31.50	6,360	2.40/12.76	774
Jordan	03:17:04	14.42	42.57	9,638	2.95/15.18	820
Total	17:23:48	13.63	31.85	7,535	2.34/12.44	4,593

Table 2: Distribution of data in SIRAJ. **Total Dur:** total audio duration. **Avg Dur**: The average duration of each audio file (seconds). **AVT:** Average transcript length (words). **U-Wds:** # of unique words. **WPS:** words per second. **CPS:** characters per second. **# Segments:** total number of segments.

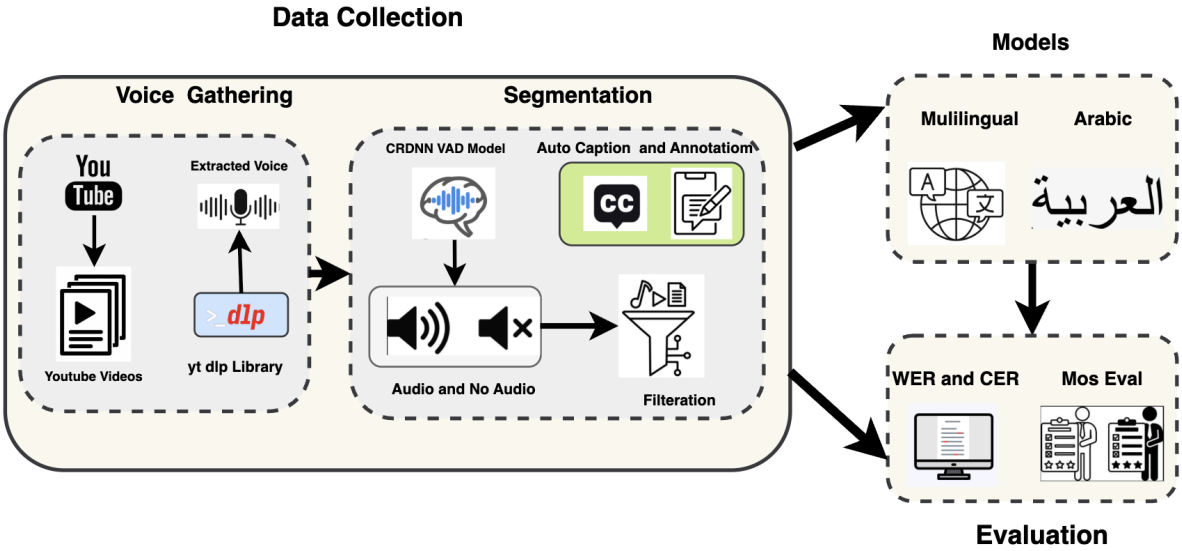

Figure 2: Overview of the pipeline from gathering the data to model training and evaluation.

Type of error	Before	After
Words written in MSA	التالته (MSA)	التالتة (Egyptian)
MSA vs. Dialect	شيء (MSA)	إشي (Jordanian)
Dialectal pronunciation	نشويهم	نشويهن (Syrian)
Taa Marbouta vs. Haa	وطنيه	وطنية
Missing word(s) added	تكون قادره على تاديه الواجبات	وده لأن الكلاب لازم تكون قادرة على تأدية الواجبات
Homograph disambiguation	زي	زَيّ / زِيّ
Number standardization	تسعة عشر	تسعتاشر

Table 3: Examples of types of errors that are corrected during the annotation process.

nunciation or meaning, while avoiding over-formalization of dialectal text.

- **Insertion of Prosodic Punctuation.** Commas, quotation marks, and colons were inserted to reflect natural pauses and phrasing, supporting accurate rhythm and intonation modeling in TTS systems.

Common Errors in YouTube Captioning

Throughout the annotation process, annotators reported several recurring issues in the raw YouTube captions that significantly affected pronunciation accuracy. The challenge most frequently reported challenge was the mismatch between some words written in MSA in the captions and the actual dialectal pronunciations in the audio. For example, annotators noted that standard terms like اتنين (two) in Egyptian Arabic were frequently written in the captions in the MSA form (اثنين), and شيء was commonly written in Jordanian Arabic (إشي). These mismatches, if not corrected, can lead to incorrect phoneme modeling and reduced intelligibility in TTS outputs. Another notable issue was the confusion between the Arabic letters Taa Marbouta (ـة) and Haa Marbouta (ـه), which are often indistinct in unvoweled orthography but produce different phonetic outputs. This issue may

lead to unnatural pronunciations during synthesis. Annotators emphasized the importance of orthographic precision in such cases to maintain accurate phoneme-to-grapheme alignment. These errors, along with other types of errors, are detailed in Table 3.

3.3 Dataset Analysis

Table 2 presents a detailed breakdown of the distribution of the dataset across the six dialect, including total audio duration, average segment length, and linguistic characteristics.

Across dialects, the total speech duration ranges from approximately 2 hours 40 minutes to over 3 hours, with the Jordanian dialect providing the longest total duration at 3 hours and 17 minutes. The average duration of audio segments is relatively consistent, typically around 13 to 14 seconds, while the number of unique words per dialect varies, reflecting dialectal lexical diversity, with the Syrian data the lowest number of unique words ≈5000, while the Jordanian data has the highest number of unique words with around 9600 word. The average words and characters per second metrics indicate the speech rates across dialects, which are consistent in the range of around 2-2.5 WPS, except for the UAE dialect, having a lower pace of speech with 1.73 WPS, and the Jordanian data having the highest WPS of 2.95.

Domain	Hours	Percentage (%)
Daily Life	3h 35m	20.61
Food	2h 36m	14.96
Culture	2h 25m	13.90
Politics	1h 58m	11.31
Religion	1h 34m	9.01
Education	1h 00m	5.75
Economy	0h 58m	5.56
Miscellaneous	3h 17m	18.89
Total	**17h 23m**	**100.0**

Table 4: Distribution of data in SIRAJ across different categories.

Figure 1 shows the domains of the dataset, with more details mentioned in Table 4, where it shows the distribution of the content across domains. The data spans 16 categories such as daily life, food, culture, politics, religion, and education, among others. It shows that narrative and conversational domains like daily life and food are more prominent, accounting for over 30% of the total content. This wide topical variety ensures the models are exposed to diverse linguistic registers, contributing to robustness in real-world synthesis tasks.

4 Experiments

To evaluate modern multilingual TTS systems on multi-dialectal Arabic, we tested three multilingual models FishSpeech, XTTSv2, and MMS alongside ARTST, an Arabic-specific model, on Siraj. We assessed each model under zero-shot inference, limited-data fine-tuning, and cross-dialect generalization scenarios. All audio was standardized to 16 kHz WAV format, with evaluation using both automatic metrics (WER/CER) and native speaker Mean Opinion Score (MOS) ratings.

4.1 Evaluation of Models Generation on Plain Text

This experiment aims to evaluate the quality of the output of Arabic TTS systems when presented with plain Arabic text. As such, we compare the three finetuned models against some other models from the literature- ArTST and some famous closed-source TTS systems- ElevenLabs and Speaktor. For this experiment, both the automatic and manual evaluation methods are used, having 10 samples from each model evaluated per dialect, which allows us to fairly compare the models, given that the models were not fine-tuned in the study and have different speaker characteristics.

4.2 Diacritization effect in finetuning

This experiment investigates the impact of the presence of diacritics in input text on the quality of synthesized speech. Most Arabic text available is written without diacritics, which introduces pronunciation ambiguity in some cases that can affect text-to-speech systems. To evaluate this, we prepared two parallel test and training sets: one with partially diacriticized transcripts to remove disambiguation and one with the same sentences stripped of diacritics. These inputs were used to synthesize speech using the fine-tuned versions of FishSpeech and XTTSv2, as these models exhibited the most consistent performance in previous experiments.

The primary goal of this experiment is to assess whether supplying phonetic disambiguation via diacritics leads to improved pronunciation naturalness, and overall intelligibility. We used the ASR models for WER/CER evaluation against the ground-truth transcript for this experiment.

	Gold	XTTS	Fishspeech	MMS	ElevenLabs	Speaktor	ArTST
Egypt	25/06	35/10	27/08	60/22	**21/05**	40/25	80/60
Syria	36/11	43/15	34/12	161/129	**20/05**	38/17	71/45
Saudi	28/09	37/11	33/11	65/25	**25/06**	32/08	68/47
UAE	32/09	41/15	28/10	93/54	**20/05**	31/13	55/34
Morocco	**36/10**	42/13	68/23	119/83	49/14	52/28	84/52
Jordan	55/21	57/30	53/26	101/69	**38/08**	42/15	84/67

Table 5: Results for all models on plain text generation using ASR models in WER/CER format. Gold is the original audio from the test set.

	XTTS	Fishspeech	MMS	ElevenLabs	Speaktor	ArTST
Egypt	4.18±0.34	3.88±0.22	1.94±0.43	4.19±0.27	2.09±0.67	1.70±0.31
Syria	2.57±0.86	3.52±0.72	1.86±0.52	3.95±0.46	2.19±0.33	1.48±0.47
Saudi	3.66±0.45	3.70±0.47	2.35±0.44	3.57±0.30	2.39±0.34	2.81±0.62
UAE	4.11±0.51	3.45±0.50	2.62±0.97	3.55±0.26	2.48±0.59	1.59±0.62
Morocco	4.38±0.57	2.85±0.18	2.71±0.22	3.04±0.88	2.06±0.27	1.56±0.17
Jordan	3.70±0.57	3.37±0.43	1.52±0.37	3.30±0.33	2.23±0.45	1.58±0.48
Average	3.76±0.55	3.46±0.42	2.17±0.49	3.60±0.42	2.24±0.44	1.79±0.45

Table 6: Results of manual evaluation (MOS) on plain text generation using all models recorded on a scale from one to five.

4.3 Cross-Dialect Generalization

We assess the ability of fine-tuned TTS representations to transfer across seven Arabic dialects by training on a single source dialect (e.g., Egyptian) and evaluating on the six unseen targets. Using our top two models (XTTSv2 and FishSpeech), we perform dialect-specific fine-tuning in isolation, then synthesize speech for all other dialectal texts without further adaptation. Repeating this for each dialect yields a cross-dialect performance matrix, enabling pairwise comparison of transfer effectiveness and identification of dialects that furnish the most universally useful features.

This multi-dialectal analysis addresses substantial phonological, lexical, and prosodic variation: dialects such as Egyptian and Levantine often share features broadly across regions, whereas Maghrebi dialects (e.g., Moroccan) diverge phonetically, challenging generalization. We quantify synthesis fidelity using automatic metrics, word error rate (WER) and character error rate (CER) and conduct manual evaluation of the single best-generalizing model to verify naturalness, dialectal authenticity, and intelligibility, given the prohibitive effort of exhaustive human assessment.

4.4 Metrics

In our study, we mix between manual and automatic evaluation methods used for TTS systems, for automatic evaluation, we use WER and CER, which are metrics commonly used to measure the quality of ASR systems, but they have been recently used in TTS evaluation and showed correlation with manual evaluation metrics (Laouirine et al., 2024; Jain et al., 2022; Alharthi et al., 2023). We use the latest whisper model [4] to compare the WER and CER of the original and generated audio against the manually annotated transcriptions for all dialects, but for the Moroccan dialect we use this finetuned wav2vec model instead[5], as whisper performed poorly on Moroccan.

For manual evaluation, we use MOS, which is the research standard for TTS systems, which is based on the concept of averaging the judgment of N judges. The judges score the quality of the audios in terms of naturalness and intelligibility using a scale ranging from 1 to 5, where 1 means "Bad" and 5 means "Excellent" (Minchan Kim and Myeonghun Jeong and Byoung Jin Choi and Sunghwan Ahn and Joun Yeop Lee and Nam Soo Kim, 2022). In our work, we assign 10 native Arabic speakers experienced across multiple dialects. (4 males and 6 females) to assess 10 audio samples generated from each model to assess the quality of audio generation. As MOS is an expensive process, it's only used to initially compare all the models across the different dialects in the plain

[4] https://huggingface.co/openai/whisper-large-v3-turbo
[5] https://huggingface.co/speechbrain/asr-wav2vec2-dvoice-darija

test experiment, and to verify the quality of the best cross-dialect model; the rest of the experiments are evaluated only using the automatic metrics.

4.5 Experiment setup

We split each dataset into training, development, and test sets using an 80/10/10 ratio, and all experiments were carried out on a machine equipped with three NVIDIA A100 40GB GPU. For each model, we applied grid search over the recommended hyperparameter ranges as suggested in their original implementations or associated literature. The best configuration was selected based on performance on the development set.

5 Results & Discussion

5.1 Finetuning on Plain text

Table 5 presents the automatic evaluation of the models presented in the WER and CER for plain text generation across the six Arabic dialects in using various TTS models. XTTS, ElevenLabs, and Fishspeech consistently outperform others in terms of accuracy, particularly on Egyptian and Saudi dialects, where ElevenLabs achieves WER/CER of 21/5 and 25/6, respectively. Fishspeech follows closely, with a notably low error rate. XTTS comes third with objectively low WER/CER scores, but outperform all other models on the Moroccan dialect with WER/CER of 42/13, which is a remarkable score considering the Gold samples scored 36/10 WER/CER. MMS, Speaktor, and ArTST on the other hand suffer from high error rates, especially for Morocco and Jordan, suggesting limited robustness for dialectal nuances despite MMS's broad applicability.

Manual evaluation through MOS, presented in Table 6, provides a more human-centric assessment of audio quality, focusing on naturalness, intelligibility.

Overall, XTTS achieved the highest scores across the majority of dialects, peaking at 4.18 ± 0.34 for Egyptian and 4.11 ± 0.51 for UAE. These results closely mirror the performance of ElevenLabs, which scored 4.19 ± 0.27 for Egyptian, underscoring the commercial system's strength in synthesis quality. Fishspeech maintained consistently good results as well, with scores in the range of 3.5 MOS points for all dialects, peaking at Saudi dialect with MOS score of Saudi 3.70. In contrast, MMS, Speaktor, and ArTST scored below 2.6 in most dialects, highlighting their inability to produce human-like, natural-sounding speech in Arabic dialects. The high standard deviation in some cases (e.g., 0.86 for XTTS on Syria) suggests variability in judge perceptions, possibly due to inconsistency in pronunciation or unnatural prosody.

Comparing the results from Tables 5, 6 we see that the ElevenLabs model scored by far the best results in terms of automatic evaluation, but XTTS showed to be similar, if not better in performance when it comes to human evaluation. Annotators reported that some systems, including ElevenLabs, generate some speech in MSA, not dialectic Arabic, which produces wrong pronunciation for some words as it stresses every letter, which can improve ASR transcription accuracy, even though the resulting pronunciation may be incorrect or unnatural.

These findings reinforce that quantitative metrics alone do not capture the full picture of TTS performance, especially in linguistically rich and prosodically variable languages like Arabic. Models may achieve low WER/CER but still sound robotic or mispronounce some words, while others may produce more natural prosody despite moderate transcription errors. Importantly, the MOS results confirm that dialect-aware synthesis is not just about intelligibility, it's also about expressiveness and regional appropriateness. For practical deployment, particularly in applications such as conversational agents, educational tools, or audiobooks, models must meet user expectations not just in accuracy, but in perceived fluency and cultural resonance. Manual evaluation remains an essential step in understanding these subjective dimensions of quality.

5.2 Cross-Dialect Generalization

We conduct this analysis to identify which Arabic dialects serve as the most effective source domains for cross-dialect TTS transfer, enabling efficient adaptation in low-resource settings. Table 7 shows cross-dialect generalization where columns denote the source dialect (fine-tuned) and rows the target evaluation dialect. Both XTTS and Fishspeech generalize well when trained on lexically and prosodically rich dialects like Egyptian and Saudi, achieving the lowest WER/CER in in-dialect evaluation, like Egyptian-to-Egyptian yields 35/10 (XTTS) and 27/8 (Fishspeech). For Syrian evaluation, XTTS performs best when trained on Saudi (34/11), while Fishspeech favors Egyptian (33/11), suggesting phonetic proximity. Saudi evaluation shows unexpected trends

	XTTS						Fishspeech					
	Egypt	**Syria**	**Saudi**	**UAE**	**Morocco**	**Jordan**	**Egypt**	**Syria**	**Saudi**	**UAE**	**Morocco**	**Jordan**
Egypt	35/10	45/15	38/14	45/17	53/20	50/21	27/08	32/11	32/10	28/07	29/09	36/16
Syria	39/13	43/15	34/11	42/16	42/24	57/30	33/11	34/12	33/12	47/24	36/15	43/21
Saudi	48/20	74/36	37/11	42/21	42/16	47/22	68/21	37/13	33/11	29/09	34/12	43/21
UAE	34/11	45/16	35/12	41/15	39/17	41/13	30/11	34/12	29/9	28/10	31/11	30/14
Morocco	68/23	63/21	69/24	71/27	42/13	74/30	71/25	69/24	70/24	77/27	68/23	74/25
Jordan	51/17	100/60	57/24	67/32	70/33	57/30	53/19	49/17	48/17	55/20	52/21	53/26
Average	46/16	62/27	45/16	51/21	48/21	54/24	47/17	43/15	41/14	44/16	41/18	68/23

Table 7: Cross-test results of models on different Arabic dialects reported in WER/CER format. Columns present the dialect the model was trained on, while rows present the dialect used for evaluation.

Source	**Zero-Shot**		**Plain FT**		**Diacritized FT**	
	Fishspeech	**XTTS**	**Fishspeech**	**XTTS**	**Fishspeech**	**XTTS**
Egypt	32/11	37/13	27/08	35/10	26/05	33/08
Syria	36/12	33/10	34/12	43/15	33/10	52/23
Saudi	35/12	39/23	33/11	37/11	34/15	39/14
UAE	29/10	29/10	28/10	41/15	26/08	43/16
Morocco	51/19	53/26	68/23	42/13	64/21	44/14
Jordan	59/27	48/24	53/26	57/30	50/23	52/28

Table 8: Comparison between zero-shot models, and finetuning the models with and without diacritization. Results are displayed in WER/CER format

with Fishspeech, as the lowest WER comes from the UAE model (29/9), while UAE is best modeled by XTTS trained on Egyptian (34/11) and by Fishspeech trained on UAE itself (28/10), indicating Fishspeech's stronger reliance on direct exposure. Moroccan is consistently difficult to generalize to, with acceptable results only when trained on Moroccan (42/13 XTTS, 68/23 Fishspeech), due to its phonological divergence; yet, Moroccan-trained models still transfer moderately to others. Overall, Saudi emerges as the most transferable source, with the best average WER in XTTS (45/16) and Fishspeech (41/14), highlighting the value of prioritizing annotation efforts on structurally balanced dialects like Saudi or Egyptian to optimize cross-dialect performance in low-resource settings. We've applied further visualization on the cross-dialect evaluation, and added analysis on using MOS that can be found in Appendix A (Figures 3 and 4, and Table 9)

5.3 Diacritization

Table 8 provides insight into how diacritization influences model performance and compares the zero-shot model to the fine-tuned models. Supplying partially diacritized inputs showed minor fluctuations in WER/CER than when feeding the model plain text only. It is also shown that Zero-shot models come close to fine-tuned models when comparing WER/CER, with the exception of Moroccan dialect, which shows significant improvement. For example, zero-shot XTTS scores 53% WER on the Moroccan test set, while the finetuned version scores 42% WER. This is possibly because we use a dedicated model for Moroccan Arabic ASR which might only perform well if the audio follows the Moroccan dialect; on the other hand, we are using whisper for all other dialects, which might be able to transcribe the audio correctly even if the audio does not follow the correct dialect.

6 Conclusion

In this paper, we have introduced SIRAJ, a new Arabic multi-dialectic TTS dataset spanning 6 different Arabic dialects and containing more than 17 hours of manually annotated data. We explore multiple state-of-the-art TTS systems, including open-source models such as Fishspeech, XTTSv2, MMS-TTS, ArTST, and commercial systems like ElevenLabs across various conditions such as zero-shot inference, fine-tuning, diacritization, and cross-dialect transfer. Our results from the automatic evaluation (WER and CER) and human evaluations (MOS) demonstrated that fine-tuned XTTS and Fishspeech models consistently outperformed other open-source systems, offering speech that comes close to human level, which is remarkable given that the size of the data is relatively small for TTS systems. Our analysis also revealed how some dialects present transferable features for other dialects, where models trained on Saudi and Egyptian data showed good generation results on other dialects.

Limitations

While the dataset and experimental benchmarks advance the state of Arabic TTS, several limitations warrant discussion. First, the dataset, though diverse in dialects and domains, includes only one speaker per dialect, which limits the generalizability of synthesized voices. This single-speaker setup may constrain the prosodic and expressive range of the generated speech, and a speaker who uses a lot of filler words may skew the results of automatic evaluation. Additionally, the annotation process, while manually verified, was limited to around 3 hours per dialect, which, although sufficient for basic fine-tuning, may not capture the full lexical and syntactic richness of each variety. Second, while MOS evaluations provide valuable subjective insights, they were conducted on a limited number of samples (10 per model per dialect) and raters (10 total). This small evaluation size, coupled with the inherent subjectivity of listener preferences and dialect familiarity, may lead to variability in the results. Furthermore, it would have been better for the study to perform manual evaluation on all experiments but this was not possible given how time intensive the process is. Moreover, despite our best efforts to fine-tune models uniformly, not all TTS systems allow the same level of control or customization (e.g., closed-source models like ElevenLabs), which introduces challenges in establishing truly fair comparisons.

References

Wajdan Algihab, Noura Alawwad, Anfal Aldawish, and Sarah AlHumoud. 2019. Arabic speech recognition with deep learning: A review. In *Social Computing and Social Media. Design, Human Behavior and Analytics*, pages 15–31, Cham. Springer International Publishing.

Dareen Alharthi, Roshan Sharma, Hira Dhamyal, Soumi Maiti, Bhiksha Raj, and Rita Singh. 2023. Evaluating speech synthesis by training recognizers on synthetic speech. *Preprint*, arXiv:2310.00706.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Ahmed Azab, Ahmed Zaky, Tetsuji Ogawa, and Walid Gomaa. 2023. Masry: A text-to-speech system for the egyptian arabic. In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO*, pages 219–226. INSTICC, SciTePress.

Houda Bouamor, Nizar Habash, and Mohammad Salameh. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Hervé Bredin. 2021. Pyannote.audio: Neural building blocks for speaker diarization. *Journal of Open Source Software*, 6(59):3819.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *Preprint*, arXiv:2406.04904.

Khai Doan, Abdul Waheed, and Muhammad Abdul-Mageed. 2024. Towards zero-shot text-to-speech for Arabic dialects. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 123–129, Bangkok, Thailand. Association for Computational Linguistics.

Google. 2016. Youtube data api v3. `https://developers.google.com/youtube/v3`. Accessed: 2024-05-14.

Nawar Halabi. 2016. *Modern Standard Arabic Phonetics for Speech Synthesis*. Ph.D. thesis, University of Southampton.

Rishabh Jain, Mariam Yiwere, Dan Bigioi, Peter Corcoran, and Horia Cucu. 2022. A text-to-speech pipeline, evaluation methodology, and initial fine-tuning results for child speech synthesis. *Preprint*, arXiv:2203.11562.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmonem Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. *Preprint*, arXiv:2303.00069.

Imen Laouirine, Rami Kammoun, and Fethi Bougares. 2024. TunArTTS: Tunisian Arabic text-to-speech corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16879–16889, Torino, Italia. ELRA and ICCL.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *Preprint*, arXiv:2411.01156.

Minchan Kim and Myeonghun Jeong and Byoung Jin Choi and Sunghwan Ahn and Joun Yeop Lee and Nam Soo Kim. 2022. Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus. In *Interspeech 2022*, pages 788–792.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI aljazeera speech resource a large scale annotated Arabic speech corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2274–2285, Online. Association for Computational Linguistics.

Youssef Nafea, Shady Shehata, Zeerak Talat, Ahmed Aboeitta, Ahmed Sharshar, and Preslav Nakov. 2024. Araoffence: Detecting offensive speech across dialects in arabic media. In *Interspeech 2024*, pages 4303–4307.

Paarth Neekhara, Jason Li, and Boris Ginsburg. 2021. Adapting TTS models for new speakers using transfer learning. *CoRR*, abs/2110.05798.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, and 8 others. 2024. Casablanca: Data and models for multidialectal Arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758, Miami, Florida, USA. Association for Computational Linguistics.

Hawau Olamide Toyin, Amirbek Djanibekov, Ajinkya Kulkarni, and Hanan Aldarmaki. 2023. ArTST: Arabic text and speech transformer. In *Proceedings of ArabicNLP 2023*, pages 41–51, Singapore (Hybrid). Association for Computational Linguistics.

J.C.E. Watson. 2002. *The Phonology and Morphology of Arabic*. Oxford linguistics. Oxford University Press.

YTDL Contributors. 2024. youtube-dl: A command-line program to download videos from youtube.com and other video sites. `https://github.com/ytdl-org/youtube-dl`. Version 2024.1.1.

Slavomír Čéplö, Ján Bátora, Adam Benkato, Jiří Milička, Christophe Pereira, and Zemánek Petr. 2016. Mutual intelligibility of spoken maltese, libyan arabic, and tunisian arabic functionally tested: A pilot study. *Folia Linguistica*, 50.

A Appendix

A.1 Cross-dialect analysis

Further analysis has been done on the Cross-dialect experiment to enhance the visualization and ensure the quality of the data. We've applied MOS evaluation using the country with the best scores from Table 7 (Saudi) using XTTS as it got better scores in the initial MOS experiment.

Country	XTTS_{Saudi}	XTTS_{FT}
Egypt	3.67±0.27	4.18±0.34
Syria	2.43±0.23	2.57±0.86
UAE	3.88±0.71	4.11±0.51
Morocco	3.74±0.37	4.38±0.57
Jordan	3.25±0.37	3.70±0.57

Table 9: Manual evaluation on cross-dialect generation. XTTS_{FT} means XTTS finetunned on the corresponding dialect

The results from Table 9 shows the results of the manual evaluation of the cross-dialect generated samples by the Saudi verson of XTTS in comparison to XTTS finetuned on the corresponding dialect. It's expected to see some decline in the MOS score as the model is trained on another dialect, thus it might lose the ability to correctly pronounce other dialect-exclusive words, and follow the speaking rythm of that specific dialect. The Saudi model however, showed very low decay of performance in specific countries such as the syrian dialect where the MOS only decreased by 0.14 points, and the UAE dialect where it decreased by 0.23 points.

We also include Figures 3, 4 to facilitate visualization of the results presented in Table 7. In the figures, we present each model with a line, where the model's training data is color-coded in the legend on the figure, and the x-axis includes dialects used for evaluation. From the visualization, we observe that some lines are following the same pattern (e.g., Morocco, Saudi, and Syria lines in Figure 4, which would indicate that these models are close in performance and follow the same trend in the different dialects. On the other hand, some other models show other inconsistencies that are easier to detect from the model; for example, the Syrian model in Figure 3 shows an increase in WER in the Saudi dialect, compared to the other models.

Variant	Sentence (Arabic, RTL)
Ground truth (spoken)	خلونا الآن ندفن الصندوق الذهبي فجام هنا كافور وجال لبخيت يا بخيت هن من الصباح جاعدين نخطط ونتكتك ونظبط واشتغلنا شغل كثير
Reference transcription	خلونا الآن ندفن الصندوق الذهبي فقام هنا كافور وقال لي بخيت يا بخيت هن من الصباح قاعدين نخطط ونتكتك ونضبط واشتغل نشغل كثير
ElevenLabs (generated)	خلونا الآن ندفن الصندوق الذهبي فجام هنا كافور وجال لبخيت يابخيت حنا من الصباح جاعدين نخطط ونتكتك ونزبط واشتغلنا شغل كتير

Table 10: Full-sentence example illustrating how ElevenLabs' synthetic speech normalises or modifies several dialectal tokens (هن → حنا, نضبط → نزبط, كثير → كتير), thereby lowering WER even though the ground-truth utterance is the authoritative source.

A.2 Error Analysis

A.2.1 Perceptual Quality vs. Automatic Scores

Although ElevenLabs attains the lowest WER/CER in Tables 5–6, human raters consistently prefer XTTS for overall naturalness. Careful listening reveals two main causes. **First**, ElevenLabs often defaults to Modern Standard Arabic, pronouncing every consonant and short vowel with equal stress; this produces exceptionally "clean" acoustics that maximise ASR-friendliness but can sound robotic or region-agnostic to listeners. **Second**, dialectal words are sometimes *over-normalised* (e.g., هنّ → حَنا, نِضبِط → نُزْبِط), removing substitutions that would otherwise inflate error counts. These

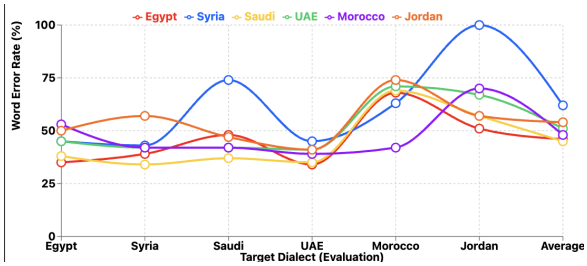

Figure 3: Cross-dialect results for XTTS model.

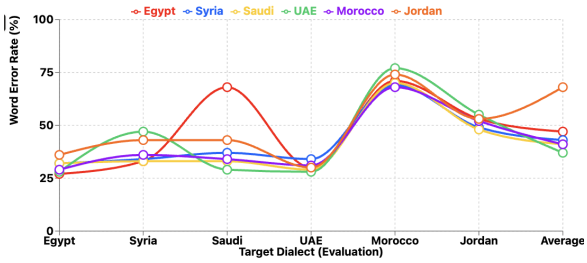

Figure 4: Cross-dialect results for FishSpeech model.

cases show that low WER/CER alone does not guarantee perceptual quality,especially in a prosodically rich language like Arabic, where intelligibility, expressiveness, and regional authenticity all matter. Manual MOS therefore remains indispensable: it captures fluency and cultural resonance that purely automatic scores miss, and it confirms that dialect-aware synthesis is more than *being understood*; it is about sounding *right* to native ears.

A.2.2 WER's orthographic fragility

While WER remains the de-facto metric for transcription quality, it is highly sensitive to surface variation, especially in Arabic, where multiple spellings of the same word are common. For instance, آه → اه, حضارة → حضاره, and بدأوا → بدوا are semantically identical yet counted as full substitutions by WER. Consequently, systems that merely regularise spelling can score better than more natural-sounding outputs. WER thus offers a useful quantitative *signal* for system comparison, but it cannot capture prosody, regional authenticity, or listener comfort. We therefore accompany WER/CER with manual MOS ratings: the MOS confirms that dialect-aware synthesis is not only about intelligibility but also about sounding *right* and culturally resonant to native ears.

A.3 Manual Evaluation (MOS)

Table 9 gives mean-opinion scores (MOS) for five systems across five dialects. Below, we summarise the qualitative reasons listeners gave for each score.

A.4 Systematic Failures in Models

A.4.1 FishSpeech

Parasitic Glide Insertion in Final Consonants

A recurrent issue is the mispronunciation of final ه as و, arising from confusion between the open ه (common in dialects) and the closed ة (تاء مربوطة). When the model encounters a word-final ه, it frequently appends an unintended vowel like sound a *parasitic glide* (e.g., /u/) resulting in a misrendering where ه is realized as و. For example, كثيفة ("very dense," dialectal) is erroneously pronounced as كثيفو, distorting both lexical meaning and dialectal identity. Such errors degrade *intelligibility* and *authenticity*, particularly in dialects where final vowels encode morphological or sociolinguistic cues (e.g., Gulf vs. Levantine Arabic).

Dialectal Identity Confusion and Prosodic Inconsistency

FishSpeech exhibits *dialectal identity confusion*, wherein an ambiguous or cross-dialectal token (e.g., لكن or إذا) triggers a mismatched prosodic pattern for the entire utterance. The model often defaults to MSA or a dominant dialect (e.g., Egyptian), even when subsequent words are unequivocally dialectal. For instance, a Levantine sentence beginning with بس ("but") may abruptly shift to MSA phonology, stripping the output of *regional authenticity* and undermining conversational naturalness. This reflects a systemic over-reliance on early-sentence cues for dialect classification, rather than holistic context.

Numeric Rendering Failures: Digits vs. Orthography

The model struggles to pronounce numerals in digit form (e.g., 1847 or ٣٠), resulting in hesitations, omissions, or garbled output. In contrast, fully spelled-out numerals (e.g., التسعتاشر, "the nineteenth") are rendered flawlessly. This disparity reveals a critical gap in grapheme-to-phoneme conversion for non-alphabetic characters, with severe implications for domains requiring precise numeric articulation (e.g., dates, prices, or measurements).

4. Voice-Style Incongruity

FishSpeech exhibits *vocal timbre mismatch*, where dialectally accurate pronunciations are delivered

through regionally inconsistent voice qualities. For instance, an Egyptian phrase like إزيك إنت؟ may retain correct Cairene phonology but be rendered with Gulf style vocal characteristics including mismatched pitch contours, glottal tension patterns, and formant structures creating perceptual dissonance despite technically accurate pronunciation. This suggests inadequate conditioning of speaker embeddings on dialect metadata, causing dissociation between linguistic content and vocal delivery style.

A.5 XTTS

Despite training all dialects on equal durations (around 3 hours each), XTTS tended to produce more fluent and natural-sounding speech in Egyptian and Saudi dialects. This suggests a potential adaptation bias toward certain phonetic or prosodic patterns that align better with the model's internal representations. While all dialects were learnable to some extent, the fluency gap in Moroccan and Jordanian outputs was more noticeable.

Also, XTTS occasionally failed to handle complex phrasing, especially when the same word appears in different forms within the same sentence. For example, in the phrase: (طحنته طحن بمطحنة التوابل) the model sometimes pronounces it incorrectly as: (طحنته طحنته طحنته بمطحنة التوابل)

This issue was not fully resolved by checkpoint tuning alone. More consistent behavior was only achieved with carefully selected checkpoints and a reasonable increase in training data for that dialect. These kinds of phrase-level pronunciation errors highlight the challenges of modeling morphologically rich and prosodically varied languages like Arabic in a single-speaker, low-resource setting.

A.5.1 ElevenLabs

While ElevenLabs achieved low WER/CER in our tests, it consistently defaulted to Modern Standard Arabic pronouncing some vowels and consonants equally and "normalized" dialectal words artificially lowering error rates but reducing dialect authenticity, which resulted into a lower MOS score than other models. Its overly "clean" prosody often sounds robotic in conversational contexts, and, as a closed-source system, it offers no fine-tuning or adaptations for regional variation.

A.5.2 MMS

The synthesized MMS output exhibits generally appropriate pacing, with natural pause placement and a speaking rate consistent with conversational Egyptian Arabic. Temporal alignment across utterances is stable, contributing to a rhythmically coherent prosodic structure. However, the system demonstrates substantial pronunciation inconsistencies. In many instances, the produced phoneme sequences deviate not only from Egyptian Arabic but also from any valid Modern Standard Arabic realization, resulting in phonetically implausible or unintelligible segments. While prosodic characteristics remain adequate, these pronunciation errors considerably degrade the intelligibility, naturalness, and dialectal fidelity of the synthesized speech, which is why the model under-performs in both automatic and manual evaluation metrics.

A.5.3 Speaktor

Speaktor struggles with dialectal vocabulary, consistently failing to pronounce regional words often substituting them with Modern Standard Arabic counterparts, or omitting them entirely. In many cases, even MSA words are mispronounced or left untranslated. As a closed-source system, it lacks customization options, adaptation capabilities, or dialectal tuning, resulting in poor handling of both colloquial and formal Arabic inputs.

Dialect	Before Annotation	After Annotation
Egyptian	باول مع تزايد حجم مشكلة المخدرات والمواد المتفجرة واحتدام الصراع بين أجهزة الأمن وعصابات التهريب، بدأ كل طرف منهم يلجأ لحيل ووسائل تمكنه من تحقيق أهدافه	مع تزايد حجم مشكلة المُخَدِّرات والمواد المُتفجِّرة واحتدام الصراع بين أجهزة الأمن وعصابات التهريب، بدأ كل طرف منهم يلجأ لحيل ووسائل تمكنه من تحقيق أهدافه
UAE	لين تهدأ نفسك تستوعب انت ايش قاعد تسوي شلون تتعامل معاه بطريقه ذكيه بحيث انك ما تخسر نفسك ولا تخسر عملك ولا تخسر وقتك ولا سمعتك ولا انه يصدر منك لفظ او فعل تندم عليه	لين تهدأ نفسك، تستوعب انت ايش جاعد تسوي، شلون تتعامل معاه بطريقة ذكية، بحيث إنك ما تخسر نفسك ولا تخسر عملك ولا تخسر وقتك ولا سُمعتك، ولا إنه يصدر منك لفظ أو فِعِل تندم عليه بعد
Moroccan	وقال لها فين ساكنين المراه ها السميه ديال الام ديالها والاب ديالها الراجل حتى هو السميه ديال الام دياله وديال الاب دياله وهاش كيدير وهاش كيدير ويفعلوا	وهو يبدا يعاود ليها وال ليها فين ساكنين المرا ها السمية ديال الأم ديالها والأب ديالها الراجل حتى هو السمية ديال الأم ديالو وديال الأب ديالو وهاش كيديرو وهاش كيديرو وها كيفعلو وها كيتركو
Syrian	انا بدي استخدمه شايفين كيف في منه كذا شكل والشكل اللي بتفضلوه انتم فيكم تستخدموه هلا بدي اخذه هذا احطه بالزيت سائل بالنسبه للقالب هذا هو القالب اللي	بالنسبة للقالب، هذا هو القالب اللي أنا بدي استخدمه، شايفين كيف في منه كذا شكل، والشكل اللي بتفضّلوه أنتوا فيكم تستخدموه، هلأ بدي آخده هذا أحطه بالزت شوي
Saudi	المنطلق ومن باب المسؤوليه اليوم حل قنا بتكون معلومات حول شخصيه الصحابي الجليل والكثير من المعلومات الخاطئه فمن هذا	من المعلومات الخاطئة، فمن هذا المنطل ومن باب المسؤولية، اليوم حلقتنا بتكون معلومات حول شخصية الصحابي الجليل

Table 11: Examples of dialectal Arabic text before and after annotation.