Dialectal Coverage And Generalization in Arabic Speech Recognition

Anonymous ACL submission

Abstract

Developing robust automatic speech recognition (ASR) systems for Arabic requires effective strategies to manage its diversity. Existing ASR systems mainly cover the modern standard Arabic (MSA) variety and few high-resource dialects, but fall short in coverage and generalization across the multitude of spoken variants. Code-switching with English and French is also common in different regions of the Arab world, which challenges the performance of monolingual Arabic models. In this work, we introduce a suite of ASR models optimized to effectively recognize multiple variants of spoken Arabic, including MSA, various dialects, and code-switching. We provide open-source pre-trained models that cover data from 17 Arabic-speaking countries, and fine-tuned MSA and dialectal ASR models that include at least 11 variants, as well as multi-lingual ASR models covering embedded languages in code-swtiched utterances. Our open-source/open-weights models achieve the highest coverage and generalization for spoken Arabic and SOTA performance in all Arabic ASR benchamrks.

1 Introduction

The advent of large self-supervised acoustic models has transformed speech technology, enabling transfer learning and improving performance for both high-resource and low-resource languages. Prominent examples of such models include various versions of wav2vec (Schneider et al., 2019; Baevski et al., 2020), HuBERT (Hsu et al., 2021), and SpeechT5 (Ao et al., 2021), which have predominantly been trained on English datasets. Their multi-lingual variants, e.g. XLS-R (Babu et al., 2021) with 53 and 128 languages, in addition to many models that include both self-supervised and supervised pre-training, such as Whisper (Radford et al., 2023) with approximately hundred supported languages, MMS (Pratap et al., 2024) with thousands of languages, and UniSpeech (Wang et al., 2021), underscore the potential for cross-lingual transfer learning for more inclusive ASR. Yet, while these models indeed show great potential for transfer learning to new languages, even those unseen in training (Huang et al., 2013), they remain sub-optimal for specific target languages. A case in point is the Arabic Text and Speech Transformer (ArTST), a model pre-trained exclusively on Arabic, which has demonstrated superior performance for Modern Standard Arabic (MSA), surpassing larger multilingual models like Whisper and MMS in benchmark tests, in addition to establishing a new state-of-the-art (SOTA) performance compared to previous efforts for Arabic ASR. This highlights the advantage of monolingual pre-training when large amounts of unlabeled data for the target language are available. While the model showed some potential for dialectal coverage, it was trained and validated mainly on MSA data, which questions its applicability for spoken dialectal variants of Arabic. Evaluations on code-switched data also showed poor performance of ArTST compared to multi-lingual models (Kadaoui et al., 2024), demonstrating the delicate trade-off between monolingual and multilingual optimization. Arabic is a pluricentric language (Schuppler et al., 2024), diverse in regional variations, and models trained on MSA frequently struggle to adapt to these variations. This limitation is particularly acute given that many Arabic dialects are underrepresented and considered low-resource in speech technology research. Consequently, there is a need for optimized ASR systems that embrace, rather than overlook, the linguistic diversity of the Arabic-speaking world.

In light of these challenges, we conduct various investigations aimed at understanding and enhancing the dialectal diversity and performance of Arabic ASR systems. We focus on four inquiries aimed at optimizing potential strategies for integrating dialectal variation into ASR systems. First, we mea-

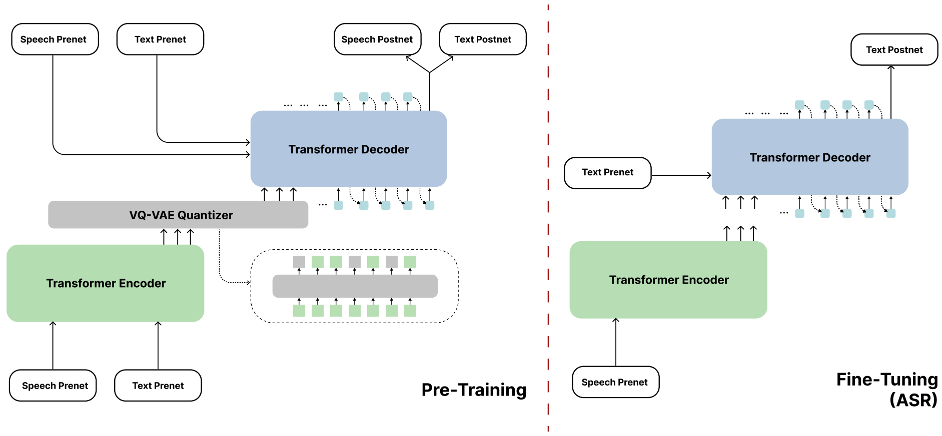

Figure 1: The architecture of SpeechT5/ArTST, which contains an encoder-decoder module and six modal specific pre/post-nets. During self-supervised pre-training (left), quantized tokens are shared across speech and text modalities. Hidden states and latent units are mixed up and used as the inputs of the cross-attention module in the decoder. The fine-tuning stage for ASR is shown on the right. Refer to Ao et al. (2021) for more details.

sure the impact of incorporating a broad collection of Arabic dialects during the model's pre-training phase. We hypothesize that a wider dialectal foundation could improve the model's performance across various dialects in the fine-tuning stage. Second, we quantify the comparative effectiveness of dialect-specific fine-tuning versus a more holistic, multi-dialectal fine-tuning strategy. The third question examines the model's capacity for zero-shot transfer to dialects not explicitly included in fine-tuning. Finally, we evaluate the model on code-switched utterances, and examine the effect of multilingual pre-training and fine-tuning on both monolingual and code-switched datasets. Our key findings from experiments spanning over 17 variants of spoken Arabic are: (1) pre-training with more data and wider dialectal coverage improves performance across most dialectal variants, including MSA, (2) multi-dialectal fine-tuning improves performance for low-resource dialects, but may not be optimal for high-resource dialects, (3) multi-dialectal pre-training and fine-tuning has higher potential for zero-shot transfer to unseen dialects, and (4) multilingual pre-training and fine-tuning greatly boosts performance on code-switching, while negatively impacting monolingual performance due to language interference. Our pretraining checkpoints and joint models were trained exclusively on open-source data and will be released as open-source, open-weights models.

2 Related Work

Recent research in Arabic speech recognition, as presented in Hussein et al. (2022), demonstrates the potential of contemporary deep learning techniques in decoding Arabic spoken language. However, this initial success was confined to Modern Standard Arabic (MSA), the formal variant predominant in news broadcasts and official communications. Large-scale multi-lingual ASR models, Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024), cover many languages within their scope, including Arabic. They utilize language embeddings or adapters to enhance language coverage and performance within the same model, but their performance across languages vary considerably. Toyin et al. (2023) demonstrated state-of-the-art performance in multi-task training for Arabic speech recognition and synthesis, improving over much larger multi-lingual models like Whisper and MMS, but their model was trained and evaluated predominantly on MSA. Kadaoui et al. (2024) evaluated ArTST, Whisper, and MMS, on the Mixat dataset (Al Ali and Aldarmaki, 2024), and showed that ArTST struggles with code-switching.

3 Methodology

Based on prior work, we start with the premise that monolingual training is more suitable for maximizing performance in Arabic ASR. However, the current Arabic SOTA models have limited coverage of spoken varieties and struggle with code-switching due to their monolingual training. Our objective is to maximize performance while also widening the coverage to include MSA, regional dialects, and instances of code-switching. To that end, we start with the current state-of-the-art model for Arabic, ArTST (Toyin et al., 2023), as the foundation model for our investigation. Figure 1 illustrates the high-level architecture of ArTST for self-supervised pre-training and fine-tuning. This model is based on the SpeechT5 approach (Ao et al., 2021), and supports multi-modal fine-tuning. The model was pre-trained on the MGB2 (Ali et al., 2016) dataset, which consists of newswire data, mainly in MSA, with a small subset of dialectal variants. In this work, we attempt to understand the factors that enable both high performance and wide coverage; we explore the following questions:

1. Is **pre-training** on dialectal data essential for improving down-stream dialectal performance, and would it negatively impact MSA performance?
2. Is it better to **fine-tune** ASR models jointly on multiple dialects or fine-tune on a specific target dialect?
3. Can we achieve reasonable **zero-shot** performance on unseen dialects?
4. Can we optimize performance in **code-switched** utterances using multilingual pre-training?
5. What is the effect of **multilingual** pre-training and fine-tuning on monolingual Arabic performance? (i.e. language interference).

The remaining sections detail our experimental settings and findings of these questions.

3.1 Terminology

For the rest of the paper, we will refer to Arabic variants using abbreviations. The categories below are based on regions and countries, and do not reflect any official classification of dialectal families:

MSA: Modern Standard Arabic. This is a common official variant of Arabic used in news, books, and education. **CA**: Classical Arabic. This is an old variant of Arabic found on religious texts and old books. It resembles MSA, but also contains outdated lexical items and structures.

GLF: A broad category of dialects spoken in the Arabian Peninsula, in particular the Gulf region, which, in our data sources, include **SAU**: Saudi, **KUW**: Kuwait, **UAE**, **OMA**: Oman, **QAT**: Qatar, **IRA**: Iraq, and **YEM**: Yemen.

LEV: Levantine dialects, which, in our data sources, include **SYR**: Syria, **JOR**: Jordan, **LEB**: Lebanon, and **PAL**: Palestine.

NOR: North African dialects, including **EGY**: Egypt, **TUN**: Tunisia, **MOR**: Morocco, **ALG**: Algeria, **MAU**: Mauritania, and **SUD**: Sudanese.

3.2 Pre-Training Data & Settings

To examine the effect of pre-training data coverage on downstream performance, we pre-trained ArTST from scratch[1] on both MSA and dialectal data. We sourced our data from various datasets: MGB2 (Ali et al., 2016), QASR (Mubarak et al., 2021) MGB3 (Ali et al., 2017), MGB5 (Ali et al., 2019), ClArTTS (Kulkarni et al., 2023), ASC (Halabi et al., 2016), and Common Voice (Ardila et al., 2019). We also used MADAR (Bouamor et al., 2018) and NADI (Abdul-Mageed et al., 2023) text datasets for pre-training. In our experiments, we compare the following:

- **ArTST**: This variant is as described in Toyin et al. (2023), pre-trained only on MSA.
- **Ours-D**: In this variant, we use a mixture of MSA and dialectal data in pre-training.
- **Ours-M**: In this variant, we use a mixture of MSA, dialectal, and multilingual data in pre-training.

See Table 11 in Appendix B for details of all the datasets used in pre-training.

3.3 Dialectal Fine-Tuning

The datasets we use for dialectal fine-tuning are shown in Table 1. We collected as many open-source data as needed to maximize coverage of dialects. Note that, for MGB5 and MGB3, as the data is based on YouTube videos, many of the originally referenced videos are no longer available, so at the time of our experiments, only 2.5 hours of training were available for MGB3 and 2 hours for MGB5. Furthermore, multi dialectal datasets, such as MASC (Al-Fetyani et al., 2021), have unbalanced representation of dialects. The high-resource dialects in our collection include SAU, SYR, EGY, and MSA; each has at least 200 hours of transcribed ASR data. UAE, MOR, JOR, LEB, IRA, and TUN have a medium amount of fine-tuning data between 10 and 50 hours. KUW and PAL are low-resource dialects with less than 10 hours of transcribed data in total. Finally, we left ALG, YEM, and SUD from the MASC dataset for zero-shot testing.

[1]We used the scripts and configurations provided in github.com/mbzuai-nlp/ArTST

Dataset	Dialect	Hours	Words
QASR	MSA	2000 hrs	13.33 M
MGB2	MSA	1000 hrs	7.31 M
MGB3[ASR]	EGY	2.83 hrs	18.93 K
MGB5[ASR]	MOR	6.74 hrs	56.97 K
SADA (Alharbi et al., 2024)	SAU	418 hrs	3.25 M
Mixat (Al Ali and Aldarmaki, 2024)	UAE	15 hrs	57.94 K
TARIC-SLU (Mdhaffar et al., 2024)	TUN	8 hrs	72.00 K
ParallelCorp (Almeman et al., 2013)	MSA	32 hrs	30.66K
	GLF	32 hrs	27.26K
	LEV	32 hrs	18.43K
	EGY	32 hrs	48.56K
MASC (Al-Fetyani et al., 2021)	MSA	612.28 hrs	3.80 M
	SAU	452.24 hrs	301.92 K
	SYR	211.33 hrs	1.06 M
	EGY	175.36 hrs	1.03 M
	JOR	42.21 hrs	330.83 K
	LEB	25.20 hrs	155.76 K
	IRA	17.37 hrs	121.12 K
	TUN	12.17 hrs	34.34 K
	Multiple	10.57 hrs	80.08 K
	UAE	9.87 hrs	6.42 K
	MOR	8.60 hrs	58.38 K
	PAL	6.17 hrs	45.35 K
	KUW	4.04 hrs	32.37 K

Table 1: Summary of Dataset Statistics for **Fine-Tuning**: Hours of Audio, Word Counts, and Associated Dialects. Multiple is mix of several dialects not neccessary from the listed dialects (no information from the source).

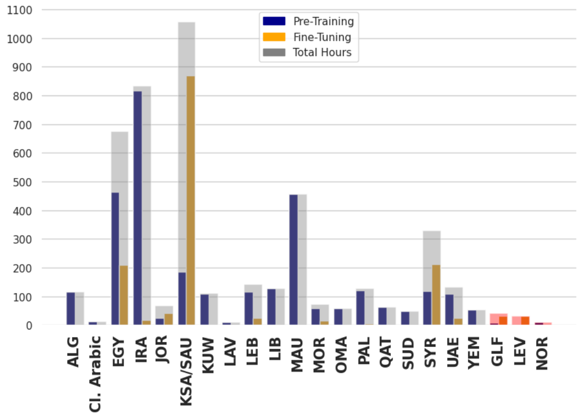

Figure 2: Distribution of dialectal speech data in pre-training and fine-tuning. MSA data are not shown.

Figure 2 illustrates the distribution of dialectal data we use for pre-training and fine-tuning our dialectal model. We exclude MSA from the figure as it has disproportionally more data than all dialects.

3.4 Multi-Lingual Fine-Tuning

In addition to the above dialectal data, we used the CommonVoice English, French, and Spanish sets for the multi-lingual fine-tuning and code-switching experiments described in section 7. English and French are commonly spoken in various Arabic-speaking countries, and to a lesser extent, Spanish is spoken in some parts of North Africa.

3.5 Experimental Settings

For partitioning the data into training, development, and test sets, we adhered to the official splits provided with each dataset. We followed the data preparation and training methodology established in the original ArTST implementation. For comprehensive details regarding the model architecture and data preprocessing, readers are directed to Toyin et al. (2023).

Computational Details The pre-training was executed on a cluster of 4 A100 GPUs over a duration of 14 to 21 days for each model. We used Adam optimizer with a learning rate of 2×10^{-4}, spanning 335K updates, and a warm-up phase of 64K updates. The maximum speech token length was set at 250K (equivalent to 15.625 seconds). Each fine-tuning experiment was run on one A100 GPU over a duration of 7 days (MGB2, QASR, MASC, SADA) or 2 days for smaller sets (MGB3, MGB5, etc.). We used Adam optimizer with a tri-stage scheduler with learning rate of 6×10^{-5}. The total computational budget for all experiments is estimated to be $\sim$6000 GPU-hours.

Normalization Prior to model training, we implemented the same data normalization steps outlined in Toyin et al. (2023). In addition, we applied post-prediction normalization steps before calculating Word Error Rates (WER), following standard practices in Arabic ASR. All reported results reflect post-normalization performance. The normalization script, sourced from a publicly available GitHub repository[2], performs orthographic standardization of *Alef*, *Yaa*, and *Taa* characters.

4 Effect of Pre-Training Data

We first examine the effect of pre-training on downstream ASR performance. As described in section 3.2, we compare a model pre-trained mainly on MSA (ArTST), and ours, trained on additional dialectal data (henceforth **Ours-D**). Note that pre-training does not utilize aligned speech and text; it incorporates un-aligned speech and text data for self-supervised learning. For these experiments, we use the same fine-tuning data, and only vary the pre-training sets.

[2] github.com/iamjazzar/arabic-nlp/blob/master/normalization/orthographic_normalization.py

System	WER(%)	CER(%)
From (Hussein et al., 2022):		
HMM-DNN	15.80	—
E2E, CTC + LM	16.90	—
E2E, Attention + LM	13.40	—
E2E, CTC , Attention + LM	12.50	—
ArTST + LM (Toyin et al., 2023)	12.78	6.33
Ours-D	12.49	6.44
Ours-D + LM	**12.39**	6.51

Table 2: Comparing our performance against models reported in Hussein et al. (2022) and Toyin et al. (2023), which include the best performing models previously reported on MGB2.

4.1 Benchmarking MSA Performance

We first report results on benchmark datasets to compare the performance of both models against the state of the art. MGB2 is the main benchmark for MSA speech recognition. We evaluated the performance of both ArTST and Our-D models fine-tuned in MGB2 in Table 2 compared to existing SOTA models. The results show that incorporating dialectal data in pre-training does not negatively affect the performance on MSA, as we achieve the best WER of 12.39%.

4.2 Benchmarking Dialectal Performance

Tables 3 and 4 show the performance of the models on the dialectal MGB3 (Egyptian) and MGB5 (Moroccan) benchmarks. Each of these benchmarks contain multiple references as dialectal speech has no standard spelling. We report the average and multi-refrence WER for our model variants, and compare against the best model in each challenge, as well as the SOTA model in each benchmark. Each model is first fine-tuned on MSA, then fine-tuned again on the target MGB train sets. We also report the results of the large multilingual models: Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024), fine-tuned on the same set. We refer to the Arabic data the models are previously fine-tuned on as *'Adaptation'* data. Starting with MSA data before fine-tuning on the target dialect has previously been established as an effective strategy for dialectal ASR (Ali et al., 2017).

In MGB3, dialectal pre-training (Ours-D) results in about 4% absolute reduction in WER, establishing a new SOTA result on this benchmark. Smaller improvement in terms of Average WER is observed for MGB5, where there is no clear advantage observed using our dialectal version. This difference

System	Adaptation	Fine-Tuning	AV-WER	MR-WER
Aalto	MGB2	MGB3	37.50	29.30
Whisper	ComVoice Fleurs Covost2	MGB3	39.04	24.92
MMS	BibleTrans NewTestamentRec	MGB3	100.04	99.92
ArTST	MGB2	MGB3	37.08	29.39
Ours-D	MGB2	MGB3	**33.20**	**25.28**

Table 3: WER(%) on MGB3 Egyptian ASR. Aalto is the best system in the MGB3 challenge (Ali et al., 2017)

System	Adaptation	Fine-Tuning	AV-WER	MR-WER
RDI-CU	MGB2	MGB5	59.40	37.60
Whisper	ComVoice Fleurs Covost2	MGB5	164.13	227.34
MMS	BibleTrans NewTestamentRec	MGB5	111.89	102.30
ArTST	MGB2	MGB5	49.39	**27.95**
Ours-D	MGB2	MGB5	**48.91**	28.02

Table 4: WER(%) on Moroccan ASR. RDI-CU is the best system in the MGB5 challenge (Ali et al., 2019)

is likely a result of our pre-training having a lot more Egyptian than Moroccan data (see Figure 2).

4.3 Zero-Shot & Fine-Tuning Results

To further quantify the effect of dialectal pre-training, we evaluate the performance of our model across different datasets. We first fine-tune models on MSA using MGB2 dataset. We test the model performance on dialects directly (zero-shot) and with dialectal fine-tuning. The results are shown in Table 5. On average, we see improvements in performance in both zero-shot and fine-tuning experiments using dialectal pre-training (Ours-D) compared to MSA-centric pre-training (ArTST). We also see that both models perform better than Whisper and MMS in zero-shot settings in most cases. There are some exceptions, such as in KUW, where Whisper performs better than all other models, including the fine-tuned models, but in most cases Ours-D performs best. This underscores the advantage of monolingual models compared to multilingual performance, as observed in Toyin et al. (2023) and Radford et al. (2023). In addition, the results underscore the importance of dialectal coverage in pre-training: the cases where Ours-D does not perform better than ArTST are all dialects for which pre-training data are limited, such as TUN (no pre-training data) and JOR (smallest dialect size in pre-training).

Dataset	Zero-Shot				Fine-Tuning	
	Whisper	MMS	ArTST	Ours-D	ArTST	Ours-D
TUN (TARIC-SLU)	138.14	93.54	107.56	106.46	**14.70**	14.80
MULT (ParallelCorp)	99.17	83.16	128.72	141.92	9.57	**9.31**
SAU (SADA)	82.16	78.28	39.41	**29.77**	39.24	29.91
MASC						
SAU	48.39	65.30	61.23	58.72	27.40	**27.33**
SYR	26.65	33.21	21.99	18.37	18.64	**17.42**
EGY	41.73	66.04	50.87	47.17	38.47	**36.43**
JOR	28.65	54.63	61.23	34.97	**19.72**	21.08
LEB	40.95	64.58	35.65	42.66	30.01	**28.05**
IRA	41.69	59.33	50.46	48.03	**31.10**	34.64
TUN	47.45	60.58	50.37	46.67	19.26	**18.52**
MOR	65.87	80.84	78.92	66.87	**47.59**	49.40
PAL	**53.20**	83.72	77.94	73.53	55.88	53.53
KUW	**36.00**	81.71	64.74	52.02	50.29	46.24

Table 5: WER (%) in zero-shot and fine-tuning settings. We compare zero-shot performance of Whisper, MMS, ArTST, and Ours dialectal pre-training (Ours-D). ArTST and Ours-D are fine-tuned on MGB2 (MSA), whereas Whisper and MMS are fine-tuned with multi-lingual data, including Arabic.

5 Joint Models & Dialect ID

So far, models were fine-tuned on MSA, followed by additional fine-tuning on the target dialect. This results in a separate model per dialect, which incurs memory costs and may have practical limitations as it requires deploying a specific model for each dialect.

In this section, we assess the relative effectiveness of individual dialectal fine-tuning compared with joint dialect fine-tuning, where we train a single model for all dialects. To that end, we joined multiple dialectal train sets as shown in Table 6. From MASC, we excluded ALG, YEM, SUD for zero-shot evaluation. The resulting joint corpus consists of 12 dialects including MSA, with approximately 1,501 hours in total. We fine-tune a single joint model using this data.

Dialect	Hours	Words	Source
MSA	612.28 hrs	3.80 M	MASC
SAU	452.24 hrs	301.92 K	SADA, MASC
SYR	211.33 hrs	1.06 M	MASC
EGY	175.36 hrs	1.03 M	MGB3, MASC
JOR	42.21 hrs	330.83 K	MASC
LEB	25.20 hrs	155.76 K	MASC
IRA	17.37 hrs	121.12 K	MASC
TUN	12.17 hrs	34.34 K	TARIC-SLU, MASC
UAE	9.87 hrs	6.42 K	Mixat, MASC
MOR	8.60 hrs	58.38 K	MASC
PAL	6.17 hrs	45.35 K	MASC
KUW	4.04 hrs	32.37 K	MASC

Table 6: Datasets used to train the joint model.

5.1 Dialect ID

We trained another model with the aforementioned joint dataset, but with the inclusion of explicit dialect identifiers. We augmented the dictionary with special tokens for dialect IDs, and used them to prepend the decoding string:

<S> DIALECT T_1 T_2 ... T_n </S>

For inference, we experimented with two strategies: (1) **Transcribing with dialect forcing**, where we manually add the dialect ID to condition the decoder output; the decoder is forced to start predictions with the tokens <S> DIALECT. (2) **Transcribing with dialect inference**, where we let the model predict the dialect token automatically. We use this approach for zero-shot ASR on unseen dialects.

The results of the models trained with joint dialects compared to models trained on MGB2 and QASR are shown in Table 7. Note that both MGB2 and QASR contain mostly MSA, but also a small amounts of various dialects, but their exact distribution is unknown. We also show the fine-tuning results from Table 5 for easy comparison. We see that joint modeling results in improvement for low-resource dialects, including: JOR, TUN, and KUW, but degrades performance of the high-resource SYR and EGY dialects. Interestingly, dialect forcing was worse on average than joint modeling with no dialect ID, while dialect inference resulted in the best performance overall. We surmise that the model learns dialectal patterns that do not perfectly align with the dialect ID as indicated in the training data. Since the dialect IDs are coarse country-level approximations, letting the model infer the dialect

Approach	Zero-Shot		Fine-Tuning	No Dialect ID	Dialect Forcing	Dialect Inference
Fine-Tuning Data	***MGB2***	***QASR***	***MGB2→ Target***	***Joint Mutli-Dialectal Set (Table 6)***		
SAU	58.72	43.41	**27.33**	29.41	30.56	29.94
SYR	18.37	**16.20**	17.42	19.20	22.41	20.30
EGY	47.17	**38.78**	36.43	45.17	61.06	46.79
JOR	34.97	25.42	21.08	**19.63**	21.49	20.11
LEB	42.66	40.51	28.05	28.22	29.43	**26.89**
IRA	48.03	40.27	36.10	**29.33**	31.75	30.83
TUN	46.67	45.93	**26.67**	37.23	28.47	27.74
MOR	66.87	55.42	56.63	57.49	53.89	**49.10**
PAL	73.53	45.59	53.53	46.22	**43.90**	44.48
KUW	52.02	45.09	46.24	**35.43**	39.43	37.71
MSA	21.09	16.78	15.34	**11.59**	12.66	12.09
Macro Average	46.37	37.58	33.17	32.63	34.09	**31.45**

Table 7: WER (%) of various models compared with joint dialectal fine-tuning with different dialect ID strategies

that best aligns with the speech is the best approach for most cases. Many dialectal sets, such as SYR and SAU, contain a lot of MSA utterances that are incorrectly identified as dialectal, and low-resource dialects, such as KUW, are predicted as their closest high-resource variant, such as SAU.

Figure 3 illustrates dialect inference errors. Note that the number of errors is proportional to the test data size. The overall dialect identification performance is around 90%.

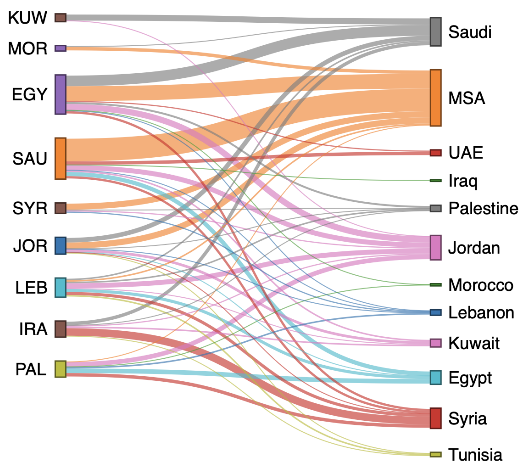

Figure 3: Dialect identification performance: true (left), predicted (right). All lines are proportional to their ratio over the total errors except for SAU→MSA, which is reduced 5 times for clarity.

6 Zero-Shot Performance

We show the zero-shot performance on the three held-out sets: ALG, SUD, and YEM. We compare the baseline, ArTST, with our multi-dialectal pre-training. We also compare models fine-tuned on MGB2, QASR, or our joint dialectal set. The results are in Table 8. Our model achives slighly lower error rates compared to ArTST, even when fine-tuned on the same MGB2 set. Better performance is achieved with QASR, which includes some dialectal data. The joint dialectal fine-tuning achieves the best performance on the held-out dialects. In general, performance in held-out sets is on a par with low-resource dialects, with WER above 30%. Table 13 in the Appendix shows the zero-shot performance after fine-tuning on a single target dialect.

System \ Dialect	ALG	SUD	YEM
ArTST→MGB2	73.18	69.20	41.64
Ours-D→MGB2	70.82	69.31	39.45
Ours-D→QASR	51.72	46.64	34.78
Ours-D →Joint	**45.20**	40.69	33.08
Ours-D →Joint (w. dialect inference)	47.12	**40.15**	**31.84**

Table 8: WER% of various models on held-out dialects.

7 Code-Switching Performance

The models analyzed so far were trained exclusively on Arabic data. While small amounts of code-switching (CS) exist in these sources, they are insufficient to learn the characteristics of the embedded languages. Large multi-lingual models are generally more effective on CS data (Kadaoui et al., 2024), even if they are less competent on monolingual Arabic. To make our models more inclusive, improving performance in the presence of code-switching is necessary. To that effect, we train a multilingual version of the model (we will refer to this as **Ours-M**). The pre-training data for this version are listed in Table 11 in the Appendix. We test **Ours-M** against **Ours-D** on available CS data for Arabic: ArZN and TunSwich. These sets cover Egyptian-English and Tunisian-French, respectively. In addition, we train a joint

Languages	Hours	Words	Source
EN	1601.92 hrs	10.35 M	CommonVoice
FR	732.02 hrs	5.03 M	CommonVoice
SP	408.34 hrs	2.79 M	CommonVoice
TUN-FR	10.89 hrs	70.86 K	TunSwitch
UAE-EN	8.97 hrs	57.82 K	Mixat
EGY-EN	5.61 hrs	52.00 K	ArzEn

Table 9: Additional datasets used to train the joint multilingual model.

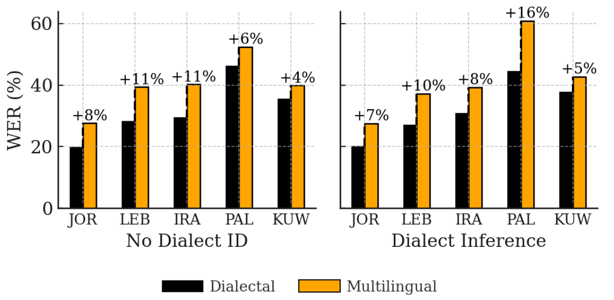

Figure 4: WER (%) and absolute difference on a subset of dialects, comparing our joint dialectal fine-tuning vs. joint multi-lingual fine-tuning on Arabic dialects.

multi-lingual model. In addition to the datasets described in Table 6, we add the multi-lingual and code-switching data shown in Table 9. The results are shown in Table 10. First, for models fine-tuned directly on the target set, we observe that multilingual pre-training significantly improves performance across all CS test sets, resulting in around 30% absolute reductions in WER for ArzEn and TunSwitch. This clearly illustrates the advantage of multi-lingual models in code-switching scenarios. We also evaluated models initialized from the joint models followed by target fine-tuning on the CS train sets, and this reduced error rates further. The effect is much larger on Ours-D with up to 40% WER reduction, which may be attributed to instances of code-switching in the joint set.

		Pre-training		Joint Fine-Tuning	
Dataset	**Language**	**Ours-D**	**Ours-M**	**Ours-D**	**Ours-M**
MGB2	MSA	**12.5**	13.0	-	-
ArzEn	EGY-EN	73.0	39.3	**32.2**	**32.2**
TunSwitch	TUN-FR	64.5	34.0	35.0	**29.2**

Table 10: ASR Results using our dialectal model (Ours-D) vs. multi-lingual model (Ours-M). We compare models trained directly from pre-trained checkpoint vs. starting with the joint model with no dialect/language ID.

Language Interference: We test the effect of multi-lingual pre-training on MSA performance. Language interference is known to negatively affect monolingual performance (Toyin et al., 2023), so we test our multi-lingual model on the MGB2 benchmark to quantify this effect (see Table 10). The model achieves 13.0% WER, which is indeed worse than the SOTA result we achieve with the Arabic-only model (see Table 2), but the difference at 0.5% absolute WER is rather small. When it comes to dialects, however, we find that language interference has a significant negative effect, resulting in 4% to 16% absolute increase in error rates, as shown in Figure 4.

8 Conclusions

We presented the largest study on dialectal Arabic ASR to empirically demonstrate the effect of various training paradigms on ASR performance. We compared models pre-trained with and without dialects, in high, low, and medium-resource settings, in addition to zero-shot. We find that overall, **dialectal pre-training improves performance in zero-shot and low-resource cases**, and mostly maintains performance on MSA and high-resource dialects. We also find that all dialects benefit from adaptation of models pre-fine-tuned on MSA, and this effect is most noticeable for low and medium-resource dialects. We experimented with multi-dialectal fine-tuning, where we joined the train sets of 12 dialects. We observe performance improvements on average, and at least the same performance as the target-dialect fine-tuning setting, and the best performance on held-out dialects. Interestingly, while using dialect ID in decoding is effective, **forcing the dialect ID results in worse performance compared to dialect inference**. While joint training results in improved performance for the medium and low-resource dialects, **target-dialect fine-tuning is more effective for high-resource dialects**. Finally, we experimented with multi-lingual pre-training and fine-tuning for improving performance on code-switched utterances, and achieved significant reductions in error rates on all available test sets. However, reductions in monolingual performance were also observed due to language interference. To enable easier adoption and further experiments, we will release the pre-trained dialectal and multilingual checkpoints, the fine-tuned MGB2 models, and the joint dialectal and multilingual models with dialect inference.

Limitations

One of the limitations in dialect-related work is the coarse classification of dialect IDs; dialects in our datasets are classified by regions or countries, whereas actual dialectal variations are far more fine-grained. For example, the Saudi dataset, SADA, covers a large geographical area and many dialects, but it is considered as one dialect based on our classification. Moreover, the way the datasets are collected do not guarantee that the data are indeed dialectal. For instance, with manual inspection of the Syrian test and dev sets from MASC, we observed that all instances are in MSA rather than Syrian dialects. In addition, Arabic dialects are spoken varieties that do not have a standard spelling system. This results in large variations in transcriptions, but standard WER does not account for these variations, resulting in more pessimistic results. With the exception of the MGB3 and MGB5 benchmarks where we report average and multi-reference WER across 4 references, all datasets have only a single reference.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, Singapore (Hybrid). Association for Computational Linguistics.

Maryam Khalifa Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-English speech. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. Masc: Massive arabic speech corpus.

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nawar Halabi et al. 2016. Arabic speech corpus. *Oxford Text Archive Core Collection*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7304–7308. IEEE.

Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2022. Arabic speech recognition by end-to-end, modular systems and human. *Computer Speech & Language*, 71:101272.

Karima Kadaoui, Maryam Al Ali, Hawau Olamide Toyin, Ibrahim Mohammed, and Hanan Aldarmaki. 2024. PolyWER: A holistic evaluation framework for code-switched speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6144–6153, Miami, Florida, USA. Association for Computational Linguistics.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. Clartts: An open-source classical arabic text-to-speech corpus. In *2023 INTERSPEECH*.

Salima Mdhaffar, Fethi Bougares, Renato de Mori, Salah Zaiem, Mirco Ravanelli, and Yannick Estève. 2024. TARIC-SLU: A Tunisian benchmark dataset for spoken language understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15606–15616, Torino, Italia. ELRA and ICCL.

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.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Barbara Schuppler, Martine Adda-Decker, Catia Cucchiarini, and Rudolf Muhr. 2024. An introduction to pluricentric languages in speech science and technology. *Speech Communication*, 156:103007.

Hawau 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.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR.

A Inference examples

فعشان كذا راح اخليكم تكملون ذاك الفلوق وبرجعلكم بعده فعشان كذا راح اخليكم تكملون ذاك الفلوق وبرجع لكم بعدها	(SAU)
وصف المجتمع الاسلامي من بعده في اخر الزمان بالمجتمع الكافر وصف المجتمع الاسلامي من بعده في اخر الزمان بالمجتمع الكافر	(SYR)
و ياريت الناس اللي بتكتب اسماء الشهور تكتب اسماء سهل ان هي تتحفظ وياليت الناس اللي تكتب اسماء الشهور تكتب اسماء سهل انها تتحفظ	(EGY)
ولادنا ذوي الاحتياجات الخاصه بدهم وقت اطول بالنسبه لهاي الاشياء ولادنا بالاحتياجات الخاصه بدهم وقت اطول بالنسبه لهالاشياء	(JOR)
و كل واحد بيامن فيه بيلاقي باب السما مفتوح على اخرو وكل واحد بيامن فيه بيلاقب بالسما مفتوح على اخرو	(LEB)
بعدين اجى حيوان قوي مغطي بفرو لونه برتقالي وخطوط سود عدين تجي حيوان قوي مغطى الفرو لونه برتقالي خطوط شود	(IRA)
و على خاطر انور و التوانسه الي كيفو يحبو يقدمو في خدمتهم ادارتي قربتلهم وعلى خاطر انور و توانس الي كيفو يحبو يقدمو في خدمتهم اداره قربتلهم	(TUN)
وبالضبط بعد بدايه الانتشار الاخير لايبولا وبالضبط بعده في تهيئه الانتشار الاخير لوباء ايبولا	(MOR)
يعقم نفسه انه كيف ما يتقدمش على الاطفال الثانيه او ممكن يقدم على الشخص الثاني نفسه انه كيف ما يتجزمش على الاطفال الثاني او ممكن يقدم على الشق الثاني	(PAL)
لكن الفيديو الاخير جدا قوي ورح يغير وجهه نظركم عن لكن الفيديو الاخير جدا قوي راح يغير وجهه النظر كم علي	(KUW)

Figure 5: Examples of dialectal recognition after targeted fine-tuning, following MGB2 adaptation.

B Pre-Training Dataset Statistics

Dataset	Dialect	Hours	Words	ArTST	Ours-D	Ours-M
ASC	MSA	3.7 hrs	20.58 K		✓	✓
ArzEn[cs]	EGY	5.61	52.00 K			✓
CommonVoice	Dialect Mix	133.24 hrs	494.83 K		✓	✓
CommonVoice	ENG	1601.92	10.35 M			✓
	FR	732.02	5.03 M			✓
	ES	408.34	2.79 M			✓
ClArTTS	CA	12 hrs	76.31 K		✓	✓
ESCWA[cs]	ALG					✓
	TUN					✓
	MOR					✓
MASC	EGY	175.36 hrs	1.03 M			✓
	IRA	17.37 hrs	121.12 K			✓
	JOR	42.21 hrs	330.83 K			✓
	KUW	4.04 hrs	32.37 K			✓
	LEB	25.20 hrs	155.76 K			✓
	MOR	8.60 hrs	58.38 K			✓
	MSA	612.28 hrs	3.80 M			✓
	PAL	6.17 hrs	45.35 K			✓
	SAU	452.24 hrs	301.92 K			✓
	SYR	211.33 hrs	1.06 M			✓
	TUN	12.17 hrs	34.34 K			✓
	UAE	9.87 hrs	6.42 K			✓
MGB2	Mostly MSA	1000 hrs	7.31 M	✓	✓	✓
QASR	Mostly MSA	2000 hrs	13.33 M		✓	✓
MGB3[ASR]	EGY	2.83 hrs	18.93 K		✓	✓
MGB3[ADI]	EGY	11.15 hrs	—		✓	✓
	GLF	8.92 hrs	—		✓	✓
	LAV	9.27 hrs	—		✓	✓
	MSA	9.39 hrs	—		✓	✓
	NOR	9.49 hrs	—		✓	✓
MGB5[ADI]	ALG	115.7hrs	—		✓	✓
	EGY	451.1 hrs	—		✓	✓
	IRA	815.8 hrs	—		✓	✓
	JOR	25.9 hrs	—		✓	✓
	KSA	186.1 hrs	—		✓	✓
	KUW	108.2 hrs	—		✓	✓
	LEB	116.8 hrs	—		✓	✓
	LIB	127.4 hrs	—		✓	✓
	MAU	456.4 hrs	—		✓	✓
	MOR	57.8 hrs	—		✓	✓
	OMA	58.5 hrs	—		✓	✓
	PAL	121.4 hrs	—		✓	✓
	QAT	62.3 hrs	—		✓	✓
	SUD	47.7 hrs	—		✓	✓
	SYR	119.5 hrs	—		✓	✓
	UAE	108.4 hrs	—		✓	✓
	YEM	53.4 hrs	—		✓	✓
Mixat	UAE	9.97 hrs	57.82 K			✓
ParallelCorp	EGY	32 hrs	48.56 K			✓
	GLF	32 hrs	27.26 K			✓
	LEV	32 hrs	18.43 K			✓
	MSA	32 hrs	30.66 K			✓
MADAR	MOR ALG TUN LIB EGY LEV IRA GLF YEM	—	532.37K		✓	✓
NADI	ALG BAH EGY IRA JOR KUW LEB LIB MOR OMN PAL QAT SAU SUD SYR TUN UAE YEM	—	702.67K		✓	✓
SADA	SAU					✓
TARIC-SLU	TUN					✓
TunSwitch[cs]	TUN	10.89 hrs	70.86K			✓

Table 11: Summary of Dataset Statistics for **Pre-Training**: Hours of Audio, Word Counts, and Associated Dialects. *[cs] refers to Code Switching datasets. *[txt] refers to textual datasets.

Fine-Tuning \ Adaptation	—	MGB2	QASR
MGB3	136.52	25.28	**19.53**
MGB5	94.84	28.02	**27.58**
TARIC-SLU	30.46	14.80	**14.48**
ParallelCorp	28.97	9.31	**9.08**
SADA	29.77	29.91	**29.55**
Mixat	100.0	**33.40**	35.21

Table 12: WER (%) of fine-tuned models on various datasets with different adaptation methods: —: no adaptation, MGB2, or QASR.

	True (Algeria):
الحمد لله انا لما اجيت يعني كنت خايف	Inference from **(SAU)** dialect
الحمدلله انا لما جيت يعني كنت خايف	Inference from **(SYR)** dialect
والحمد لله انا عندما اجيت عليكم بخير	Inference from **(EGY** dialect
لكن الحمد لله لما لا جيت يعني كنت خايف	Inference from **(JOR)** dialect
الحمد لله انا لما جيت يعني كنت خايف	Inference from **(LEB)** dialect
الحمد لله انا لما جيت ع لون الخيف	Inference from **(IRA)** dialect
الحمد لله انا لما جيت يعني كنت خيف	Inference from **(TUN)** dialect
الحمد لله انا لما جيت على كنت خايف	Inference from **(MOR)** dialect
الحمد لله انا لما جيت على كنت خايف	Inference from **(PAL)** dialect
الحمد لله انا لما جيت يعني كنت خايف	Inference from **(KUW)** dialect
الحمد لله انا لما جيت يعني كنت خايف	Inference from **(UAE)** dialect
الحمد لله انا لما جيت يعني كنت خيف	
	True (Sudan):
ما ادري ليش تجيني تعليقات زي كده	Inference from **(SAU)** dialect
ما ادري ليش جيت تعليقات زي كذا	Inference from **(SYR)** dialect
ما عادوا لي سجين تعليقات زي كدا	Inference from **(EGY** dialect
ما هذا ليس جيد تعليقات زي كده	Inference from **(JOR)** dialect
ما ادوا ليش جيت تعليقات زي كدا	Inference from **(LEB)** dialect
ما زالت جيده علاقات زايقتها	Inference from **(IRA)** dialect
ما ادوا لسجين تعليقات زي كذاا	Inference from **(TUN)** dialect
ماهو لاش جي تعليقات زي كداا	Inference from **(MOR)** dialect
ما هذا الشيء تعليقات زي كدا	Inference from **(PAL)** dialect
ما ادوا ليش جيت تعليقات زي كدا	Inference from **(KUW)** dialect
ما ذا ليش جيت تعليقات زي كذا	Inference from **(UAE)** dialect
ما دوا ليش جي تعليقات زي كذا	
	True (Yemen):
امال ايش الشيء الذي بيكون مهم طيب	Inference from **(SAU)** dialect
عمال ايش الشي الذي بيكون مهم طيب	Inference from **(SYR)** dialect
المهم طيب and this the the becom	Inference from **(EGY** dialect
العمال يشيش هي الذي بيكون مهم طيب	Inference from **(JOR)** dialect
عمال ايش الشيء الذي يكون مهم طيب	Inference from **(LEB)** dialect
عمالي شي الشي الذي يكون مهم طيب	Inference from **(IRA)** dialect
العمال يشيش هي الذي بيكون مهم طيب	Inference from **(TUN)** dialect
عمال ايش الشي آلذي يكون مهم طيب	Inference from **(MOR)** dialect
عمال ليس الشيء الذي يكون مهم طيب	Inference from **(PAL)** dialect
عمال ايش الشيء الذي بيكون مهم طيب	Inference from **(KUW)** dialect
عمال ليش الشي الذي يكون مهم طيب	Inference from **(UAE)** dialect
عمال ليش الشي الذي بيكون مهم طيب	

Figure 6: Example of ASR outputs on held-out dialects from dialect-specific fine-tuning.

C Effect of Data Adaptation

In the above experiments, we followed the strategy of initializing the models by first fine-tuning on MSA data. In most cases, we used MGB2 as the base model, following previously established results on Egyptian ASR (Ali et al., 2017). This adaptation approach is meant to enhance the performance on low-resource dialects, facilitating faster convergence with limited training samples. However, as pre-training on more diverse sets proved to be effective, adaptation on more diverse data is also likely to be fruitful. As observed in Table 7, models trained on QASR resulted in far better zero-shot performance, approaching the performance of joint-dialects models. This is attributed to the fact that QASR is both larger in size and known to have more dialectal data compared to MGB2 (but both have no documented statistics of dialectal coverage). To validate this observation, we experimented with dialectal fine-tuning adapted from two variants: one based on MGB2 and one based on QASR (Mubarak et al., 2021), followed by dialect-specific fine-tuning. Table 4 shows fine-tuning results with no adaptation (directly fine-tuning on the target dialect), compared with starting from MGB2 or QASR. First, our results corroborate the previous findings that adapting models from MSA results in large reduction in error rates. In all except the Mixat dataset, starting from QASR results in better performance compared to MGB2. However, the difference is negligible except on MGB3 Egyptian set (around 10% absolute WER reduction). There are two factors that we speculate underline this result: The small size of the MGB3 set, and the existence of Egyptian dialect in the QASR corpus more substantially than the other dialects. Overall, using the QASR dataset as a basis for adapting dialectal models is recommended as it improves or maintains performance.

Figure 5 lists examples of ASR outputs using the dialect-specific fine-tuned models. Note that the 'errors' in SAU, EGY, and JOR examples are in fact alternative spellings.

D Cross-Dialectal Performance

Table 13 shows the cross-dialectal performance, where models trained on a single target dialect are tested on other dialects, including the three held-out sets: ALG, SUD, and YEM. In most cases, the best performance is achieved by the model trained on the same target dialect (the diagonal in Table 13). However, for low-resource dialects, like KUW and PAL, the model trained on SAU achieved the lowest WER. This is likely a result of the large size of the SAU train set and the wide geographical area and dialectal variants it covers. Curiously, all models perform well on the SYR test set; upon close inspection, we found that the set consists mostly of MSA utterances, which explains the result since all models are adapted from MSA.

	Dialect	Fine-Tuning Set										
		SAU(%)	SYR(%)	EGY(%)	JOR(%)	LEB(%)	IRA(%)	TUN(%)	MOR(%)	PAL(%)	KUW(%)	UAE(%)
test set	**SAU**	**27.33**	53.88	63.24	43.95	56.64	46.10	49.36	50.02	45.48	46.85	46.14
	SYR	19.54	17.42	23.91	16.88	25.71	**16.85**	24.36	17.58	16.86	17.38	17.53
	EGY	38.07	53.43	**36.43**	40.08	51.61	41.37	45.58	43.88	40.41	41.88	40.38
	JOR	22.88	30.37	34.14	**21.08**	28.11	28.68	32.96	30.50	25.25	27.88	28.05
	LEB	41.03	42.53	53.23	39.07	**28.05**	41.23	48.53	42.34	41.55	42.14	40.97
	IRA	**35.18**	49.10	56.91	40.47	46.72	36.10	47.84	42.55	40.76	41.00	42.16
	TUN	44.44	57.78	51.85	46.67	45.19	45.19	**26.67**	44.44	47.41	46.67	47.41
	MOR	59.04	71.69	74.70	54.82	69.88	57.23	74.70	56.63	**56.02**	59.64	57.83
	PAL	**48.53**	66.18	62.65	52.35	60.59	58.53	60.00	63.24	53.53	58.53	60.59
	KUW	**26.59**	59.54	78.03	43.93	67.63	46.82	53.18	51.45	44.51	46.24	50.87
held-out	**ALG**	**50.21**	60.09	73.61	52.58	62.02	57.51	57.51	59.01	52.79	58.37	57.30
	SUD	**40.89**	64.97	64.32	53.15	64.86	54.12	56.51	58.79	52.39	56.18	55.53
	YEM	38.40	42.16	38.49	34.68	39.69	34.92	30.68	35.73	**33.06**	34.68	34.11

Table 13: (WER(%) for various models on unseen dialects. All models are adapted from v2→MGB2.