Casablanca: Data and Models for Multidialectal Arabic Speech Recognition

Bashar Talafha[1] * Karima Kadaoui[2] Samar M. Magdy[2] Mariem Habiboullah[9] Chafei Mohamed[11]
Ahmed O. El-Shangiti[2] Hiba Zayed[3] Mohamedou Cheikh Tourad[6] Rahaf Alhamouri[4] Rwaa Assi[3]
Aisha Alraeesi[2] Hoor Mohamed[2] Fakhraddin Alwajih[1] Abdelrahman Mohamed[2] Abdellah El Mekki[2]
El Moatez Billah Nagoudi[1] Saadia Benelhadj[7] Hamzah A. Alsayadi[8] Walid Al-Dhabyani[12] Sara Shatnawi[2]
Yasir Ech-chammakhy[5] Amal Makouar[10] Yousra Berrachedi[1] Mustafa Jarrar[3] Shady Shehata[2,13]
Ismail Berrada[10] Muhammad Abdul-Mageed[1,2,13] *

[1]University of British Columbia, [2]MBZUAI, [3]Birzeit University, [4]JUST, [5]INSEA, [6]Université de Nouakchott, [7]ESI,
[8]Ain Shams Univ., [9]Technische Hochschule Mittelhessen, [10]UM6P, [11]TEK-UP, [12]Cairo University, [13]Invertible AI

Abstract

In spite of the recent progress in speech processing, the majority of world languages and dialects remain uncovered. This situation only furthers an already wide technological divide, thereby hindering technological and socioeconomic inclusion. This challenge is largely due to the absence of datasets that can empower diverse speech systems. In this paper, we seek to mitigate this obstacle for a number of Arabic dialects by presenting *Casablanca*, a large-scale community-driven effort to collect and transcribe a multi-dialectal Arabic dataset. The dataset covers eight dialects: Algerian, Egyptian, Emirati, Jordanian, Mauritanian, Moroccan, Palestinian, and Yemeni, and includes annotations for transcription, gender, dialect, and code-switching. We also develop a number of strong baselines exploiting *Casablanca*. The project page for *Casablanca* is accessible at: https://www.dlnlp.ai/speech/casablanca.

1 Introduction

Self-supervised learning (SSL) has significantly advanced the field of speech processing, impacting everything from speech recognition to speech synthesis and speaker verification. However, the success of these methods heavily relies on the availability of large datasets, which are primarily available for a select few languages. This bias towards resource-rich languages leaves behind the majority of the world's languages (Bartelds et al., 2023; Talafha et al., 2023; Meelen et al., 2024; Tonja et al., 2024). In this work, we report our efforts to alleviate this challenge for Arabic—a collection of languages and dialects spoken by more than 450 million people. We detail a year-long community effort to collect and annotate a novel dataset for eight Arabic dialects spanning both Africa and Asia. This new dataset, dubbed *Casablanca*, is rich with various layers of annotation. In addition to speech transcriptions, we include speaker gender, dialect, and code-switching information. Notably, to the best of our knowledge, some of the dialects included in *Casablanca* have not been featured in any prior speech or broader NLP research. In addition to describing our dataset, we develop baseline systems for automatic speech recognition (ASR). To summarize, our contributions are as follows:

1. We introduce *Casablanca*, the largest fully supervised speech dataset for Arabic dialects, labeled with transcriptions, code-switching, dialect, and gender.
2. We evaluate SoTA multilingual ASR models and four Arabic-centered Whisper models across the eight dialects in *Casablanca* to assess their adaptability and performance, particularly in handling the linguistic nuances of Arabic dialectal variation.
3. We assess the performance of the best-performing model in code-switching scenarios, analyzing the segments using both the original Latin characters and their transliterated counterparts.

2 Related Work

Arabic. Arabic encompasses a diverse array of linguistic varieties, many of which are nearly mutually unintelligible (Watson, 2007; Abdul-Mageed et al., 2024). This diversity includes three primary categories: Classical Arabic, historically used in literature and still employed in religious contexts; Modern Standard Arabic (MSA), used in media, education, and governmental settings; and numerous colloquial dialects, which are the main forms of daily communication across the Arab world and often involve code-switching (Abdul-Mageed et al., 2020; Mubarak et al., 2021). The significant differences between these varieties pose challenges in adapting technologies from one variety to another

* Corresponding Authors: btalafha@mail.ubc.ca, muhammad.mageed@ubc.ca

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21745–21758
November 12-16, 2024 ©2024 Association for Computational Linguistics

(e.g. MSA to the Yemeni dialect) (Habash, 2022; Talafha et al., 2023).

Arabic ASR data. Early efforts to develop Egyptian Arabic speech datasets began in 1996 with the *CallHome* task (Pallett, 2003) under the National Institute of Standards and Technology's (NIST) evaluations, focusing on the Egyptian and Levantine dialects. In 2006, the DARPA-led Global Autonomous Language Exploitation (GALE) (Soltau et al., 2009) and the Spoken-Language Communication and Translation System for Tactical Use (TRANSTAC) programs (Weiss et al., 2008) aimed to develop Iraqi dialect dataset, driven by U.S. military needs (Olive et al., 2011). The Multi-Genre Broadcast (MGB) Challenge ~~has~~ later introduced several datasets aimed at advancing speech recognition, speaker diarization, alignment, and dialect identification using content from TV and YouTube. *MGB-2* (Ali et al., 2016) provides 1,200 hours of speech with lightly supervised transcriptions, derived from Aljazeera Arabic news broadcasts with MSA making up 78%[1] of the total content. *MGB-3* (Ali et al., 2017) compiles video clips from Egyptian YouTube channels while MGB-5 (Ali et al., 2019) focuses on Moroccan Arabic ASR. Additionally, the *QASR* project (Mubarak et al., 2021), sourced from Aljazeera's archives between 2004 and 2015, features over 4,000 episodes across various topics, including extensive code-switched transcriptions from multiple dialects. Further details of the MGB and QASR datasets are provided in Table 1.

Non-Arabic ASR data. Similar efforts exist for collecting diverse speech datasets across various language varieties and dialects. For instance, STT4SG-350 (Plüss et al., 2023) introduces a Swiss German corpus divided into seven dialect regions, annotated with Standard German transcriptions. AfriSpeech (Olatunji et al., 2023) also offers 200 hours of Pan-African English speech, featuring 67,577 audio clips from speakers across 13 countries, encompassing 120 indigenous accents for both clinical and general ASR applications. The ManDi Corpus (Zhao and Chodroff, 2022) provides a detailed spoken database of regional Mandarin dialects and Standard Mandarin, with 357 recordings totaling about 9.6 hours from 36 speakers across six major regions.

Additional information on Arabic ASR can be found in Appendix A.1.

***Casablanca* in comparison.** *Casablanca* is the largest fully supervised Arabic dialects dataset with 48 hours of human-transcribed data, surpassing MGB-3 and MGB-5. Although MGB-2 and QASR are larger in size, they utilize light supervision (using ASR systems for transcribing and aligning human transcripts) rather than manual transcriptions. This light supervision method accounts for potential inaccuracies in human transcripts, such as omissions, errors, and variations from factors like corrections, spelling errors, foreign language use, and overlapping speech, leading to possible mismatches between the transcriptions and actual spoken content (Mubarak et al., 2021). *Casablanca* is also the most fine-grained and diverse corpus available: while datasets such as MGB-2 and QASR focus on broad regional dialects like the Gulf, the Levant, and North Africa (including Egypt), *Casablanca* targets country-level variation focusing on eight countries belonging to different areas in the Arab world. To the best of our knowledge, our dataset is also the first to introduce zero-resourced dialects in addition to the low-resource ones (specifically the Emirati, Yemeni, and Mauritanian dialects), thus filling a significant need in the research landscape. Furthermore, *Casablanca* is rich with several layers of annotation: beyond *speech transcription*, each segment is also labeled with speaker *gender* and *country*, which provide valuable demographic information and can be exploited for downstream tasks involving gender and dialect identification. Table 1 provides a comparison between *Casablanca* and a number of notable Arabic datasets. Finally, with *Casablanca*, we are advancing the benchmarking efforts to encompass eight dialects and include evaluations on four multilingual models: Whisper (Radford et al., 2023) (both versions 2 and 3), SeamlessM4T (Barrault et al., 2023), and MMS (Pratap et al., 2023) under zero-shot and Arabic-enhanced[2] settings. This expansion strengthens our analysis by incorporating advanced models, offering a comprehensive evaluation of their capacity to handle diverse dialects.

3 Corpus Collection

3.1 Data Selection

We assembled a team of 15 native speakers (each with a research background) and assigned them the task of manually curating a list of YouTube

[1] The updated version of MGB-2 reported 78%, while the old one reported 70% (Mubarak et al., 2021).

[2] Further finetuned on Arabic data.

	MGB-2	MGB-3	MGB-5	QASR	*Casablanca*
Hours	1,200	16	14	2,000	48
Dialects	(MSA: 78%+) GLF, LEV, NOR, EGY	EGY	MOR	(MSA: majority) GLF, LEV, NOR, EGY	ALG, EGY, JOR, MOR, UAE, PAL, MAU, YEM
Dialect Label	✗	N/A	N/A	✗	8 labels
Segmentation	lightly	test: fully	test: fully	lightly	fully
Transcription	lightly	fully	fully	lightly	fully
Code-switching	✗	✗	✗	EN+FR	EN+FR (+transliteration)
Gender	✗	✗	✗	≈82% data	100% data

Table 1: *Casablanca* in comparison to notable Arabic speech datasets. **Lightly**: lightly supervised (labeling is performed using a pre-trained model). **Fully**: fully supervised (all annotations are carried out manually by humans). **Test: fully**: only the test set is labeled manually .**✗**: does not support. **N/A**: not applicable as those datasets have one dialect only. **EN**: English. **FR**: French. **+transliteration**: code-switching words are written in both Latin and Arabic scripts.

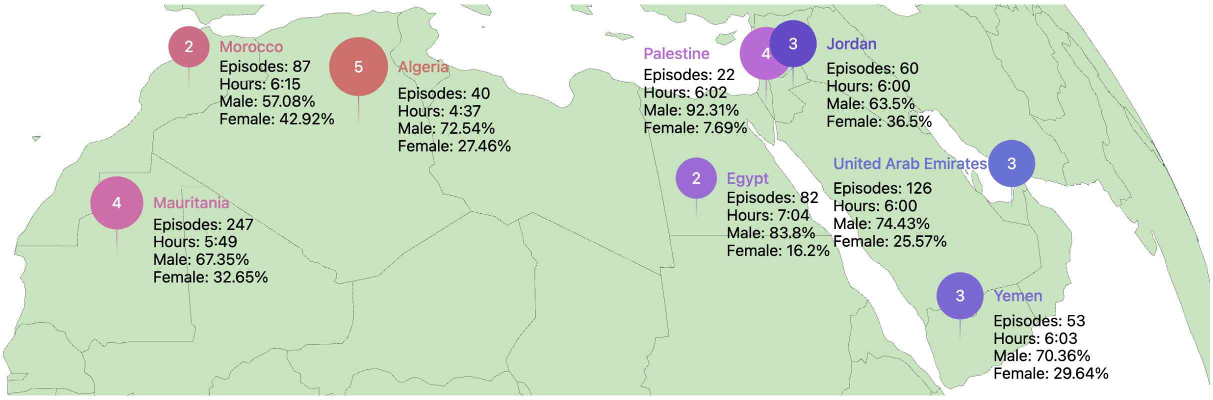


Figure 1: Geographic distribution of participants and data in *Casablanca*. **Pins** on each country represent the number of participants per dialect. **Episodes** denotes the number of selected episodes. **Hours** refer to the total hours of transcription per dialect. **Male** and **Female** are percentages of male and female speaker coverage over dialects.

episodes from TV series that represent the dialects of their countries. To ensure diversity, we instruct them to include a variety of actors and geographical settings[3]. We manually verified that each episode is over 15 minutes in length and removed introductory videos, such as trailers, to eliminate redundancy. Due to copyright restrictions on the original YouTube videos, we follow the approach by Uthus et al. (2024); Ali et al. (2019, 2017) and do not provide them directly. Instead, we make available the YouTube URLs, timestamps, and annotations. The copyright remains with the original video owners and data we release will be exclusively for research purposes.[4]

3.2 Data Segmentation

We segment the episodes into shorter utterances, thereby simplifying transcription and enabling task distribution among annotators for a more streamlined process. We use the voice activity detection model (VAD) of Bredin and Laurent (2021); Bredin et al. (2020), available through the *pyannotate*, to detect speech and remove non-speech segments such as music[5]. We then use *AudioSegment*[6] to extract the identified speech segments. We refer to these extracted audio segments as '**snippets**'. It is important to note that an output snippet may contain multiple utterances, often involving various speakers. We put the snippets on the LabelStudio platform (Tkachenko et al., 2020) for annotation. See more details about annotation in Appendix A.2.

[3] This involves diverse genders, ages, speaking styles, and locations reflecting various sub-dialects within the country.

[4] The project page for *Casablanca* is accessible at: https://www.dlnlp.ai/speech/casablanca.

[5] We utilize the model with its default hyperparameters (*onset*: 0.8104, *offset*: 0.4806, *min_duration_on*: 0.055, *min_duration_off*: 0.097).

[6] https://github.com/jiaaro/pydub

4 Data Annotation

4.1 Annotators

Our community-driven dataset, *Casablanca*, is created with the help of 27 annotators from the Arab world, each annotating their respective dialects. All annotators either have or are pursuing graduate degrees in natural language processing, making them well-positioned for the task. We involve at least two annotators per dialect, each coming from a different region within the respective country for an enhanced knowledge of sub-dialects[7], which adds a layer of linguistic richness and diversity to the orthographic representation of each dialect. Table 8 (Appendix A.4) illustrates lexical variation within the eight dialects in *Casablanca*, showcasing its linguistic diversity.

4.2 Tasks

We provided annotators with written guidelines explaining the annotation tasks. During weekly meetings with team members, we discussed, improved, and iteratively extended these guidelines. Annotators are also able to communicate with one another and ask questions through a Slack channel dedicated to the project. The main annotation tasks are.

Task 1: Segment Selection We introduced three annotation options as shown in Figure 3: *Dialect* for dialect-specific content, *MSA* for Modern Standard Arabic, and *Other* for segments containing non-verbal sounds. Selected segments, whether dialectal or MSA, are required to be "clear segments". They must feature only one speaker to avoid voice overlap, be audibly clear and transcribable despite potential background noise, and contain a minimum of three words without surpassing 30 seconds in length. Moreover, each segment must capture the complete utterance, from beginning to end, accurately representing every phoneme component of the first and last words to preserve speech boundaries.

Task 2: Transcription Given the absence of a standardized orthographic system for Arabic dialects, we asked annotators to transcribe in the manner they usually write in their daily lives. Furthermore, for a faithful representation of the speech signal, we encouraged the incorporation of Tanweens and Hamzat[8] in the transcriptions. We also asked annotators to render numbers in alphabetical format (e.g., انا عاوز عشرين بطاقة) instead of numerical symbols (e.g., انا عاوز ٢٠ بطاقة), since this allows for reflecting inflections these numbers can have (e.g., عشرين vs. عشرون). For code-switching (CS), we asked annotators to provide two versions of the transcript, one with the foreign words in Arabic script (e.g., بروفيشينال) and another in Latin script (e.g., "professional"); see Table 9 in Appendix A.5.

Task 3: Gender Annotators label speaker gender based on perceived biological sex[9] from the set {*male, female*}. This makes our dataset suited for studying gender-specific speech patterns across dialects.

Task 4: Validation In this task, each team engages in a peer validation process, with annotators reviewing and ensuring the accuracy of one another's transcriptions, focusing on correcting spelling errors while preserving dialectal orthographic variations.

Our annotation process utilized an agile methodology (Cohen et al., 2004) with work divided into weekly sprints, allowing for focused objectives and regular review sessions to refine strategies. We also gave annotators a guideline document[10] and a document on special cases to standardize dialect scenarios and document linguistic variations. See Appendix A.6 for examples. Overall, the annotation project ran for a total duration of six months.

5 Dialects Description

Casablanca is a detailed collection of around 48 hours of data covering eight Arabic dialects from regions like the Levant, Gulf, Yemen, and North Africa, including Algerian, Egyptian, Emirati, Jordanian, Mauritanian (Hassaniya), Moroccan, Palestinian, and Yemeni. *Casablanca* involves sub-dialects from these countries as well. In addition, to the best of our knowledge, we are among the first to offer annotated data for the less-represented Emirati, Mauritanian, and Yemeni dialects, addressing a gap in linguistic research.

[7] In the literature, these sub-dialects are sometimes referred to as “micro-dialects" (Abdul-Mageed et al., 2020).

[8] Tanween refers to the doubling of a vowel at the end of a word, indicated by diacritic marks, enhancing the noun's indefinite status in Arabic. Hamza represents a glottal stop, marked by its diacritic, crucial for words disambiguation (El-Imam, 2004).

[9] This acknowledges differences between biological sex and gender identity.

[10] Our annotation guidelines are available at the project page: https://www.dlnlp.ai/speech/casablanca.

6 Corpus statistics

Episode Coverage. As spelled out earlier, we annotate approximately 48 hours of content across eight dialects. The average annotation duration per episode is about four minutes, constituting roughly 14.71% of the average episode length. Dialects represented by a larger number of episodes typically exhibit lower per-episode annotation durations. This distribution allows annotators to engage with a more diverse range of content. For instance, Mauritanian episodes, totaling 247, feature an average of only one minute and 25 seconds (8.23%) of annotation per episode. Conversely, the Palestinian subset, with 22 episodes, averages 16 minutes and 30 seconds per episode, which is about 53.72% of the total episode length[11].

Average Duration. As detailed in Table 2, the average duration of segments across all dialects stands at 4.24 seconds, with the Moroccan having the shortest average duration and the Palestinian the longest. We define the speed rate as the average number of words per second (WPS) and the average number of characters per second (CPS). Interestingly, based on our analysis of the episodes, the Moroccan dialect stands out as the fastest spoken dialect in *Casablanca*, both in terms of WPS and CPS with 3.2 WPS and 15.7 CPS, respectively. Conversely, Jordanian dialect is the slowest in our dataset, yielding 1.2 WPS and 6.14 CPS[12].

The average transcript length across all dialects is 8.64 words, with Jordanian transcripts being the shortest and Palestinian the longest. These differences, even between closely related dialects, stem from episode script lengths and annotator preferences for word separation, including prefixes and suffixes. For instance, in the Jordanian dialect, the phrase ("*I sent it to her*") transcribed by some annotators as a single word: ("بعثتلهاياها"), while others split it into two: ("بعثتلها اياها") or even three words: ("بعثت الها اياها"). This highlights the subjectivity among annotators across the various dialects that influence word count and segment length differences. This subjectivity, in addition to the episodes' topic diversity, influence the unique word count per dialect as detailed in Table 2. For all dialects combined, the unique word count is 85,176 words. On a country level, the Morrocan dialect has the highest number of unique words per hour with 4,458 words, while the Algerian dialect has the smallest at 3,518 words. This indicates that, besides Moroccan being the fastest dialect, it also has the greatest word diversity compared to other dialects.

Code-Switching. Among all dialects in *Casablanca*, Algerian and Moroccan demonstrate a notably high usage of code-switching. Namely, as Table 2 shows, these dialects feature 500+ segments with code-switching. These North African dialects, in addition to Mauritanian, uniquely blend French into their code-switching. Other dialects in our dataset, such as Egyptian and Jordanian, involve switching into English. This linguistic diversity mirrors the historical colonial impact on languages in these regions. Overall, *Casablanca* includes 234 English code-switching segments (totaling $\approx$ 22 minutes) and 1,220 French code-switching segments (one hour and 44 minutes). Examples are shown in Table 10 in Appendix A.5. Conversely, we observe less code-switching in the other dialects. We suspected this may be due to episodes from other countries being relatively older as use of code-switching has become more prevalent among younger Arab generations (Brown, 2005). To test this hypothesis, we manually labeled the episodes for their time coverage. We found the following: Egypt (1997-2018), Jordan (1985-2000), and UAE (1995-2009) with 72, 52, and 59 code-switching instances, respectively. In contrast, newer episodes show higher instances: Algeria (2004-2017), and Morocco (2016-2018) with 586 and 598 cases, respectively. To summarize, our analysis shows that (i) French code-switching is more common than English and, even within the same dialect, (ii) newer episodes involve more code-switching than older ones.

Gender Bias. Despite our efforts to balance gender representation, a clear male dominance is observed across all dialects as demonstrated in Figure 1. The disparity is most notable in the Palestinian dialect, where male voices constitute 92.31%, leaving a mere 7.69% for female representation. In contrast, the Moroccan dialect exhibits a more gender balanced setup (with 57.08% male and 42.92% female). We now describe baseline models we developed exploiting our dataset.

[11] Despite our efforts, we could not acquire more episodes where the Palestinian dialect is not mixed with other dialects.

[12] Fastest to slowest: Morocco > Egypt > Algeria > UAE > Palestine > Mauritania > Yemen > Jordan. Although these observations are useful, we acknowledge they may be particular to our own dataset and hence should not be generalized.

Dialect	Total Dur	Avg Dur	AVT	U-Wds	Avg U-Wds/hr	Snippets	Segments	Skips	Avg WPS / CPS	CS
Algeria	4:37:35	4.15	8.41	11,085	3,518	2,537	4,013	769	2.662 / 10.723	586
Egypt	7:04:16	4.29	10.67	16,080	3,981	2,962	5,937	715	2.858 / 13.165	72
Jordan	6:00:16	4.23	5.71	13,145	3,653	4,255	5,105	5,257	1.286 / 6.142	52
Mauritania	5:49:40	3.67	5.83	12,835	3,605	3,099	5,325	5,556	1.631 / 7.170	36
Morocco	6:15:02	3.54	10.83	15,469	4,458	4,119	6,358	504	3.206 / 15.728	598
Palestine	6:02:59	5.30	11.30	13,405	3,628	2,543	4,107	720	2.264 / 10.612	50
UAE	6:00:06	4.25	9.57	13,067	3,565	2,780	5,087	853	2.362 / 10.954	59
Yemen	6:03:26	4.49	6.85	16,140	4,175	2,991	4,861	3,825	1.517 / 7.393	1
Total	**47:53:20**	4.24	8.64	85,176	3,822.9	25,286	40,793	18,199	2.223 / 10.235	1,454

Table 2: Distribution of data in *Casablanca*. **Total Dur**: total duration for each dialect. **Avg Dur**: total duration divided by number of segments. **AVT**: average transcript length. **U-Wds**: number of unique words. **Avg U-Wds/hr**: average number of unique words per hour. **Skips**: number of skipped snippets. **WPS**: words per second. **CPS**: characters per second. **CS**: Number of code-switching segments. For ***Total***, we take the average for average columns and sums for other columns.

7 Baseline models

We split *Casablanca* into Train, Dev, and Test, keeping the latter two splits each at one hour of the data per country. We perform a number of ASR experiments on the Dev and Test splits of *Casablanca*[13]. First, we evaluate general speech models under a zero-shot condition. Then, we evaluate models that were finetuned on MSA or other dialects. Finally, we report experiments on our code-switched data only. We report results in WER and CER, both with and without preprocessing of the data. Details of our preprocessing pipeline are in Appendix A.7.

7.1 Evaluation of General Models

We evaluated SoTA multilingual speech models on each dialect to understand their generic adaptability and performance across the eight dialects. Particularly, we evaluated two versions of Whisper (Radford et al., 2023) (*whisper-large-v2*[14] and *whisper-large-v3*[15], 1550M), SeamlessM4T (Barrault et al., 2023) (*seamless-m4t-v2-large*[16], 2.3B), and MMS (Pratap et al., 2023) (*mms-1b-all*[17], 1B)[18]. For this scenario, we report WER and CER of four different multilingual models on the eight novel dialects, which we hypothesize may not have been incorporated into the training data of these models. As shown in Table 3, all models exhibited high WER and CER across each dialect, indicating their inability to effectively generalize to entirely novel conditions. On average, *whisper-large-v3* recorded lower WER and CER compared to other models, both with preprocessing (63 WER and 28.17 CER) and without (69.49 WER and 31.16 CER). In terms of dialects, without any preprocessing, only on the Jordanian dialect we achieved a WER of less than 50, as recorded by both Whisper models and SeamlessM4T. After preprocessing, the Palestinian and Egyptian dialects approached a WER of around 50 with these models. On average, *mms-1b-all* yielded the lowest performance compared to others, which can be attributed to the significant difference in domains between MMS data, a closed domain focusing on religious texts in MSA, and the Youtube series, an open domain featuring dialectal content.

7.2 Evaluation of Dedicated Models

Here we evaluate models that were finetuned by Talafha et al. (2023) on MSA, Egyptian, and Moroccan. Since the models were not released, we follow the same approach in Talafha et al. (2023) and regenerate[19] four Arabic Whisper models based on whisper-large-v2: *whisper-msa* on Common Voice 11.0[20] (CV11) for MSA, *whisper-mixed* on MGB-2 targeting a blend of MSA and dialects, *whisper-egyptian* on MGB-3 focused on the Egyptian dialect, and *whisper-moroccan* on MGB-5 for the Moroccan dialect. Then, we evaluate these models on all dialects in *Casablanca*. As reported in Table 4, *whisper-egyptian* is notably superior for all dialects except Moroccan and Algerian. The superior performance of *whisper-egyptian* can be at-

[13] In this work, we do not use the Train splits in any experiments.

[14] https://huggingface.co/openai/whisper-large-v2

[15] https://huggingface.co/openai/whisper-large-v3

[16] https://huggingface.co/facebook/seamless-m4t-v2-large

[17] https://huggingface.co/facebook/mms-1b-all

[18] We could not evaluate Google USM model (Zhang et al., 2023) since it was not available as of the time of our writing this paper.

[19] Regenerate here means that we did the same finetunings in (Talafha et al., 2023)

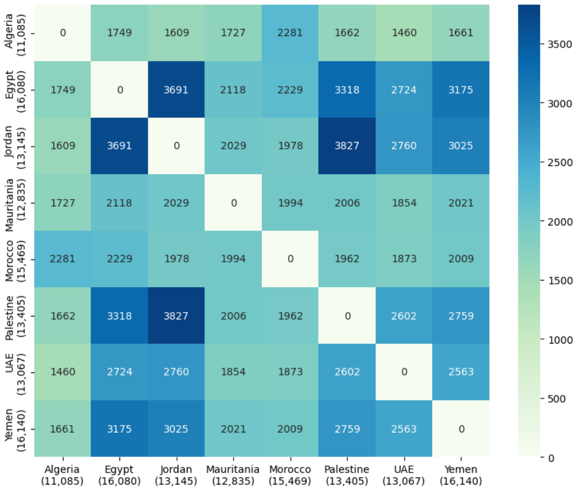
[20] https://huggingface.co/datasets/mozilla-foundation/common_voice_11_0

	whisper-lg-v2		whisper-lg-v3		seamless-m4t-v2-large		mms-1b-all	
	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc
Algeria	82.61 / 38.95	**80.47 / 36.82**	83.49 / 40.47	84.14 / 39.99	101.18 / 58.58	94.18 / 53.56	93.01 / 43.68	92.55 / 42.62
Egypt	61.99 / 26.38	52.38 / 21.71	59.11 / 24.77	**48.95 / 19.86**	61.82 / 29.83	49.75 / 24.47	88.54 / 43.59	85.84 / 40.58
Jordan	49.47 / 16.34	41.13 / 13.64	48.44 / 16.18	39.68 / 13.47	47.94 / 15.84	**39.24 / 13.12**	81.46 / 33.02	78.54 / 31.03
Mauritania	87.85 / 52.34	85.74 / 49.76	87.44 / 50.19	**85.68 / 48.08**	91.57 / 55.41	88.39 / 51.59	94.36 / 50.25	93.71 / 48.99
Morocco	88.55 / 46.57	84.52 / 44.02	87.2 / 44.41	**83.05 / 42.09**	95.18 / 58.29	91.01 / 54.97	96.91 / 49.01	95.45 / 47.34
Palestine	57.06 / 20.02	**48.64 / 17.24**	58.02 / 21.05	50.2 / 18.38	56.78 / 20.74	48.92 / 18.13	83.14 / 33.07	80.18 / 30.82
UAE	61.82 / 22.93	**52.03 / 19.15**	62.31 / 24.04	52.88 / 20.37	63.94 / 26.22	54.76 / 22.71	85.4 / 36.81	82.11 / 34.18
Yemen	71.31 / 29.8	60.65 / 24.49	69.94 / 28.17	**59.45 / 23.19**	73.65 / 32.55	62.72 / 27.43	86.73 / 38.55	81.64 / 34.36
AVG	70.08 / 31.66	63.195 / 28.35	69.49 / 31.16	**63.00 / 28.17**	74.00 / 37.18	66.12 / 33.24	88.69 / 40.99	86.25 / 38.74

Table 3: Results for dialect evaluation, scenario-1 on the Test set. Results are reported in WER and CER (/ separated). **pre-proc:** preprocessing (+ with, - without).

	whisper-msa		whisper-mixed		whisper-egyptian		whisper-moroccan	
	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc	- pre-proc	+ pre-proc
Algeria	87.86 / 48.31	87.82 / 48.20	129.63 / 79.63	129.77 / 79.68	86.68 / 35.80	86.75 / 35.70	**74.39** / 29.50	74.40 / **29.42**
Egypt	67.68 / 35.22	67.56 / 35.22	97.31 / 63.87	97.24 / 63.79	49.58 / 19.33	**49.49 / 19.24**	74.82 / 34.83	74.78 / 34.80
Jordan	61.18 / 23.43	51.93 / 20.43	78.15 / 40.34	68.89 / 37.84	56.11 / 18.15	**46.45 / 15.02**	72.79 / 27.12	64.87 / 24.32
Mauritania	88.02 / 47.5	88.02 / 47.44	114.39 / 78.02	114.43 / 78.09	**87.08 / 43.32**	87.11 / 43.35	89.93 / 45.16	89.93 / 45.17
Morocco	88.06 / 46.37	88.03 / 46.37	120.59 / 77.44	120.61 / 77.45	84.85 / 37.22	84.85 / 37.20	61.58 / 21.25	**61.57 / 21.24**
Palestine	68.06 / 28.90	59.78 / 26.00	76.92 / 36.81	67.90 / 34.25	63.70 / 22.31	**54.13 / 19.13**	76.83 / 30.15	69.42 / 27.36
UAE	74.24 / 35.37	64.54 / 31.79	104.60 / 60.20	96.95 / 57.99	67.45 / 24.48	**56.58 / 20.27**	78.37 / 31.51	70.41 / 27.95
Yemen	74.71 / 36.08	69.55 / 33.15	96.01 / 54.81	91.58 / 53.19	70.49 / 28.07	**64.96 / 24.83**	79.13 / 33.89	75.09 / 31.00
AVG	76.225 / 37.6475	72.15 / 36.08	102.20 / 61.39	98.42 / 60.29	70.74 / 28.58	**66.29 / 26.84**	75.98 / 31.68	72.56 / 30.16

Table 4: Results for dialect evaluation, scenario-2 on the Test set. Results are reported in WER and CER (/ separated). **pre-proc:** preprocessing (+ with, - without).



	Algeria (11,085)	Egypt (16,080)	Jordan (13,145)	Mauritania (12,835)	Morocco (15,469)	Palestine (13,405)	UAE (13,067)	Yemen (16,140)
Algeria (11,085)	0	1749	1609	1727	2281	1662	1460	1661
Egypt (16,080)	1749	0	3691	2118	2229	3318	2724	3175
Jordan (13,145)	1609	3691	0	2029	1978	3827	2760	3025
Mauritania (12,835)	1727	2118	2029	0	1994	2006	1854	2021
Morocco (15,469)	2281	2229	1978	1994	0	1962	1873	2009
Palestine (13,405)	1662	3318	3827	2006	1962	0	2602	2759
UAE (13,067)	1460	2724	2760	1854	1873	2602	0	2563
Yemen (16,140)	1661	3175	3025	2021	2009	2759	2563	0

Figure 2: Vocabulary intersection in *Casablanca*. "0" denotes no intersection with the dialect itself. Numbers under the country name denote the vocab size.

tributed to its enhanced likelihood of predicting dialectal words, a result of its fine-tuning, compared to *whisper-msa*. Additionally, *whisper-egyptian* is closely aligned with conversational domains that focus on everyday topics, a characteristic shared across all dialectal datasets. In comparison with *whisper-moroccan*, from a vocabulary perspective, as shown in Figure 2, the Egyptian dialect shares more vocabulary with Yemen, Jordan, UAE, Egypt, Palestine, and Mauritania than with the Moroccan dialect. Conversely, the Moroccan and Algerian dialects demonstrate a closer vocabulary alignment since these two North African dialects share more linguistic similarities than with other dialects. This correlation is consistent with the patterns observed in our experimental results. Therefore, *whisper-moroccan* performed better for Moroccan and Algerian compared to other models. Despite having the most extensive Arabic content (MGB-2 1200hrs), *whisper-mix* model showed the weakest performance overall. This is attributed to two main reasons: firstly, the data was recorded in studio settings (*Aljazeera.net*); and secondly, the content domain of the MGB-2 dataset (which includes politics, economy, society, culture, media, law, and science) differs significantly from daily conversation topics. This suggests that even though over 70% of the MGB-2 data is MSA, the remainder in dialects also does not accurately represent everyday speech, leaning more towards these specific close-domains. The evidence from the dialectal models supports the argument, showing that the MGB-3 and MGB-5 datasets, which were collected from YouTube (not including TV series), represent a wider range of real-life domains. Although these datasets are smaller in size compared to MGB-2, the relevance of the domain directly influenced their performance.

This effect is also noticeable in the comparison of the whisper-msa and whisper-mixed models. Both performed well with MSA, as reported in Talafha et al. (2023), yet *whisper-msa* yielded better outcomes on dialects than *whisper-mixed*, even though MGB-2 (1200hrs) has a much larger volume of data than CV11 (89hrs). This is also related to the domains covered by CV11 being more open than MGB-2. To further investigate the domain's effect, we juxtaposed the outcomes of *whisper-lg-v2* from scenario-1 with those of *whisper-msa* and *whisper-mix* from scenario-2. It was observed that *whisper-lg-v2* outperformed both models across all dialects, despite being the foundational model for the latter two. However, in the case of *whisper-egyptian* and *whisper-morrocan*, each surpassed *whisper-lg-v2* within their respective dialects as well as in Algerian with the Morrocan model. These findings highlight the significance of incorporating models that are both open-domain and dialect-specific. Moreover, they highlight a clear gap between the current multilingual and SOTA Arabic models on one hand, and actual world dialects on the other. We hope that *Casablanca* contributes to bridging this gap.

To further explore the effectiveness of *Casablanca*, we fine-tune Whisper-v3 using combined training splits from each dialect (*Whisper-Casablanca*) and conducted an evaluation on the Algerian dialect as a case study. We compare this model to *Whisper-lg-v3* as the baseline, *Whisper-mixed*, which was pre-trained on the largest dataset, and *Whisper-Moroccan*, the top-performing model for the Algerian dialect. The results displayed in Table 5 demonstrate a notable performance improvement over previous models. In comparison with *Whisper-Moroccan*, *Whisper-Casablanca* shows a 14.06 point reduction in WER before preprocessing and a 16.55 point reduction after preprocessing.

Model	**- Pre-proc**	**+ Pre-proc**
Whisper-lg-v3	83.49 / 40.47	84.14 / 39.99
Whisper-mixed	129.63 / 79.63	129.77 / 79.68
Whisper-Morrocan	74.39 / 29.50	74.40 / 29.42
Whisper-Casablanca	**60.33 / 26.92**	**57.85 / 25.38**

Table 5: Results for evaluating different Whisper models on the Algerian Test set. Results are reported in WER and CER (/ separated). **pre-proc:** preprocessing (+ with, - without).

7.3 Evaluation on Code-Switched Data Only

For ***code-switching evaluation***, we specifically focused on *whisper-large-v3*, selected for its overall superior performance compared to other models, as aforementioned (See Table 3). We conducted evaluations first on the original segments containing code-switching with Latin characters, and subsequently on their transliterated counterparts. Due to the relatively small number of code-switching segments, we consolidated all instances into one collective set for this focused evaluation. In the experiments, we evaluated Whisper's performance with inputs featuring either code-switching *(CS-)* or transliteration *(Transliterated-)*, under three distinct decoding scenarios: (1) decoding without specifying the language *(-Auto)*, (2) decoding with English identified as the language *(-EN)*, and (3) decoding with Arabic recognized as the language *(-AR)*. As reported in Table 6, the WER/CER scores are high in all settings, however identifying the target language makes the prediction worse. For a deeper comprehension of these findings, Table 12 and Table 13 detail the outputs for each condition, specifically for inputs involving code-switching and transliteration, respectively. With code-switched inputs, Table 12, Whisper failed to produce any code-switched words in all scenarios. Notably, even when the decoding language was set to English, Whisper performed a translation task even when specifying the task as *"transcription"*. For the Auto and Arabic settings, Whisper outputted only transliterations. This issue is also observable with the transliterated inputs, see Table 13. This highlights a limitation in Whisper's capacity to transcribe data containing code-switching.

Condition-predefined	**WER / CER**
CS-Auto	90.89 / 56.72
Transliterated-Auto	90.39 / 52.79
CS-EN	131.54 / 108.07
Transliterated-EN	133.48 / 115.56
CS-AR	103.57 / 67.58
Transliterated-AR	100.47 / 58.35

Table 6: Evaluation results for *whisper-lg-v3* on the segments with code-switching (Latin characters [CS]), and on the transliterated versions (Transliterated). Prefix **CS**: reference written with code-switching. Prefix **Transliterated**: reference written with Arabic letters. Postfix **Auto**: results without identifying the decoding language. Postfix **EN**: results with identifying the decoding language as English. Postfix **AR**: results with identifying the decoding language as Arabic.

7.4 Evaluation on Other Tasks

In addition to the main ASR evaluations, we also performed a zero-shot benchmark on two additional tasks: Arabic dialect identification (ADI) and gender recognition. For ADI, we use the best-performing HuBERT-based model from (Sullivan et al., 2023) and perform a zero-shot evaluation on Casablanca's eight dialects. The results in Table 7 reflect similar challenges observed in their study, where the model underperformed on the "YouTube Dramas" domain. In addition to providing dialect labels, Casablanca also includes gender information, as mentioned in Section 4.2. This allows for an evaluation of the gender recognition task. Therefore, we fine-tuned XLS-R (Babu et al., 2021) on Librispeech-clean-100 (Panayotov et al., 2015), as an out-of-domain dataset[21], and subsequently evaluated its performance on our dataset.

Task	Accuracy	Precision	Recall	F1 Score
ADI	36.44	54.68	36.44	39.24
Gender Rec.	83.56	89.23	83.56	84.32

Table 7: Zero-shot results of ADI and gender recognition tasks on *Casablanca*.

8 Conclusion

In this paper, we introduced *Casablanca*, the largest supervised dataset for Arabic dialects, featuring a diverse representation across eight dialects. *Casablanca* includes underrepresented dialects such as Emirati, Yemeni, and Mauritanian. Encompassing 48 hours of data, the dataset also involves detailed annotations on transcriptions, speaker gender, and code-switching. Initial experiments with SoTA models demonstrate the *Casablanca*'s utility for enhancing Arabic speech processing, especially in ASR, gender identification, and dialect identification. A subset of *Casablanca* is publicly available, aiming to support further research and innovation in both speech processing as well as linguistic research targeting dialects.

9 Limitations

While we believe *Casablanca* will have a significant impact on a wide range of tasks in Arabic speech, it is important to acknowledge some limitations. Although *Casablanca* includes eight dialects, substantially more than previous datasets, the Arabic language comprises several other dialects that we do not cover. In addition to dialects, there is also diversity within each dialect.[22] Therefore, we hope to expand the dataset to encompass a broader range of dialects in the future. Furthermore, as Figure 1 illustrates, for all dialects, the majority of speakers in *Casablanca* are male (over 60%, except for Morocco), potentially introducing gender biases. We recommend caution when working with gender-sensitive tasks. Finally, we provide only a YouTube URL for the source videos instead of the videos themselves due to copyright considerations. This could lead to availability issues if the videos are removed by their authors.

10 Ethical Considerations

In developing *Casablanca*, we adhere to ethical principles to ensure responsible and respectful use of data. Our dataset, sourced from publicly available TV series episodes on YouTube, is curated with careful consideration for privacy, omitting any personal identifiable information beyond what is publicly accessible. We try our best to ensure diverse representation in terms of gender and dialects to mitigate biases and promote inclusivity in ASR systems. All annotations and evaluations were conducted with linguistic and cultural sensitivity. While aiming to share the dataset to advance research, we implement access policies that require responsible use and proper citation. Our commitment to ethical standards is ongoing, and we welcome community feedback to continuously improve our practices.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada[23], and UBC Advanced Research Computing-Sockeye[24].

[21] Read-out books also trained on different language (i.e., English).

[22] If we go by country level, we can talk about 22 dialects. However, Abdul-Mageed et al. (2020) also introduce the concept of micro-dialects to describe sub-country variation.

[23] https://alliancecan.ca

[24] https://arc.ubc.ca/ubc-arc-sockeye

References

Muhammad Abdul-Mageed, Amr Keleg, Abdelrahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. Nadi 2024: The fifth nuanced arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728.

Muhammad Abdul Abdul-Mageed, Chiyu Zhang, Abdelrahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876.

Abdelrahman Ahmed, Yasser Hifny, Khaled Shaalan, and Sergio Toral. 2019. End-to-end lexicon free arabic speech recognition using recurrent neural networks. In *Computational Linguistics, Speech And Image Processing For Arabic Language*, pages 231–248. World Scientific.

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)*, pages 279–284. 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)*, pages 1026–1033. 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)*, pages 316–322. IEEE.

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*, 33:12449–12460.

Mohammed Bakheet. 2021. Improving speech recognition for arabic language using low amounts of labeled data.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, Brno, Czech Republic.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Keith Brown. 2005. *Encyclopedia of language and linguistics*, volume 1. Elsevier.

David Cohen, Mikael Lindvall, and Patricia Costa. 2004. An introduction to agile methods. *Adv. Comput.*, 62(03):1–66.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Yousif A El-Imam. 2004. Phonetization of arabic: rules and algorithms. *Computer Speech & Language*, 18(4):339–373.

Nizar Habash. 2022. Arabic natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 9–10.

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

Sameer Khurana and Ahmed Ali. 2016. Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 292–298. IEEE.

Marieke Meelen, Alexander O'neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of nepal. In *Proceedings of the*

Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 83–93.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource–a large scale annotated arabic speech corpus. *arXiv preprint arXiv:2106.13000*.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, et al. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

David S Pallett. 2003. A look at nist's benchmark asr tests: past, present, and future. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 483–488. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, et al. 2023. Stt4sg-350: A speech corpus for all swiss german dialect regions. *arXiv preprint arXiv:2305.18855*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

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.

Hagen Soltau, George Saon, Brian Kingsbury, Hong-Kwang Jeff Kuo, Lidia Mangu, Daniel Povey, and Ahmad Emami. 2009. Advances in arabic speech transcription at ibm under the darpa gale program. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):884–894.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the robustness of arabic speech dialect identification. *arXiv preprint arXiv:2306.03789*.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition. *arXiv preprint arXiv:2306.02902*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label studio: Data labeling software. *Open source software available from https://github. com/heartexlabs/label-studio*, 2022.

Atnafu Lambebo Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. Nlp progress in indigenous latin american languages. *arXiv preprint arXiv:2404.05365*.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Janet CE Watson. 2007. *The phonology and morphology of Arabic*. OUP Oxford.

Brian A Weiss, Craig Schlenoff, Gregory A Sanders, Michelle Potts Steves, Sherri L Condon, Jon Phillips, and Dan Parvaz. 2008. Performance evaluation of speech translation systems. In *LREC*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Liang Zhao and Eleanor Chodroff. 2022. The mandi corpus: A spoken corpus of mandarin regional dialects. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1985–1990.

Taha Zouhair. 2021. Automatic speech recognition for low-resource languages using wav2vec2: Modern standard arabic (msa) as an example of a low-resource language.

A Appendix

A.1 Arabic ASR

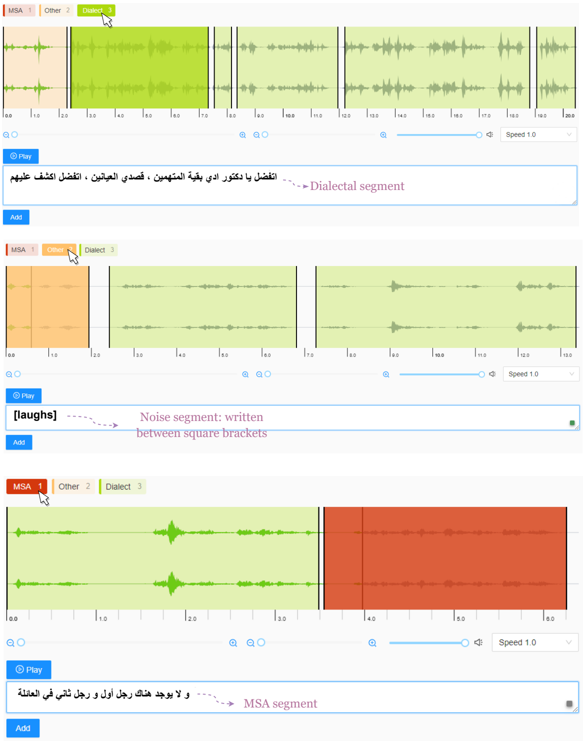
Historically, the Hidden Markov Model (HMM) combined with Gaussian Mixture Models (GMM) has been the dominant approach for achieving top results in large vocabulary continuous speech recognition (LVCSR). The first HMM-DNN hybrid for LVCSR was introduced by Dahl et al. (2011), outperforming traditional HMM-GMM systems. In the MGB2 challenge, Khurana and Ali (2016) utilized a combination of TDNN, LSTM, and BLSTM models, achieving a notable word error rate (WER) of 14.2%. End-to-end (E2E) models, mapping speech directly to text, gained popularity, simplifying ASR pipelines. Ahmed et al. (2019) introduced an E2E ASR model for Arabic, leveraging BRNNs with CTC for alignment. The introduction of an E2E transformer model addresses the morphological complexity and dialectal variations inherent in Arabic using self-attention mechanism and sub-word tokenization. Hussein et al. (2022) advanced Arabic ASR by employing a transformer-based encoder-decoder with a TDNN-LSTM language model, using Mel filter banks for acoustic features and training on MGB3 and MGB5 corpora, achieving leading performance with WERs of 27.5% for MGB3 and 33.8% for MGB5. In the era of large speech models, Arabic speech is still in its early stages. The XLS-R model (Babu et al., 2021), a large-scale model designed for cross-lingual speech representation learning, utilizing the wav2vec 2.0 framework (Baevski et al., 2020), was utilized on the Mozilla Common Voice dataset for MSA (Zouhair, 2021; Bakheet, 2021). The study of Ardila et al. (2019) benchmarks foundational models on Arabic ASR tasks, focusing on the performance of OpenAI's Whisper (Radford et al., 2023), Google's USM (Zhang et al., 2023), and the KANARI ASR model. These models were evaluated against a variety of datasets, emphasizing their efficacy across different Arabic dialects and speaking styles. Notably, USM typically surpassed Whisper, while KANARI demonstrated exceptional capability, especially in code-switching contexts between MSA and Egyptian dialect. The performance of Whisper across various Arabic dialects for ASR tasks was explored by Talafha et al. (2023). This evaluation spanned most publicly available datasets, utilizing n-shot (zero-, few-, full) fine-tuning approaches. The study also assessed Whisper's adaptability to novel scenarios, including dialect-accented MSA and previously unseen dialects. While Whisper demonstrated competitive results with MSA in zero-shot settings, its ability to adjust to different dialects was limited, showing inadequate performance and random output generation when encountering unfamiliar dialects.

A.2 Annotation Tool

We employed *Label-Studio*[25], a widely supported open-source labeling platform, as our choice for an annotation tool. We centrally hosted it on our servers and provided online access, allowing for remote and adaptable involvement from annotators across various locations. Within the tool we used the '*Automatic Speech Recognition using Segments*' template, enabling annotators to select multiple spans from each snippet and write their transcriptions accompanied by additional metadata. We also customized the tool to allow annotators to specify the gender of the speaker for each segment. We randomly shuffled the data to guarantee each snippet's independence, effectively reducing potential bias and sequencing effects that could impact annotators' perceptions during the annotation process.

A.3 Transcribing a segment

Figure 3 shows the process of transcribing a speech segment from a snippet based on its category (Dialect, MSA, and Other).



اتفضل يا دكتور ادي بقية المتهمين ، قصدي العيانين ، اتفضل اكشف عليهم

Dialectal segment

[laughs]

Noise segment: written between square brackets

و لا يوجد هناك رجل أول و رجل ثاني في العائلة

MSA segment

Figure 3: Example of transcribing a segment.

A.4 Inter-dialect diversity

Table 8 demonstrates how the same words can be written differently within the same dialect, showcasing the inter-dialect diversity and the rich nuances that this brings to dialectical expression.

A.5 Code-switching transcription

Table 9 shows the code-switching transcription process.

[25] https://labelstud.io/

Dialect	Var-1	Var-1	Var-3	MSA	English
Algeria	شوالا	واش	واشنو	ماذا	What
Egypt	بردو	برضه	برضو	أيضا	Also
Jordan	حكيتله	حكيتلو	حكيت له	قلت له	I told him
Morocco	عا قوتلو	هي گتليه	غير قلت ليه	قلت له فقط	I just told him
Mauritania	أمبجو	أمبيو	أمبدو	لحاف	Quilt
Palestine	هاض	هاد	هاظ	هذا	This
UAE	قتله	قلتله	قلت له	قلت له	I told him
Yemen	ابصرت	ابصرك	ابسرت	أرأيت	did you see?

Table 8: Examples of dialect variation along with their translations in MSA and English. **Var**: variation.

Format	**Transcript**
Transliterated	أول ما يوصل! أوكي، يلا باي
Untransliterated	أول ما يوصل! okay، يلا bye
MSA	في حين وصوله! حسنًا، مع السلامه
English	As soon as he arrives! Okay, bye

Table 9: Examples of code-switching in transcription.

Table 10 shows examples of code-switching segments for each dialect, along with their transliterated versions. Code-switched terms are provided in teal color.

Dialect	Example
Algeria	كيفاه علابالكم ويلا لقمان رايح يجيب لباك أوكان ماجاتش l'affaire مخدومة كيفاه علابالكم ويلا لقمان رايح يجيب لباك أوكان ماجاتش لافار مخدومة
Egypt	أقعد، أقعد ده إنت جيت في وقتك. ال program هيبتدي بعد دقايق من دلوقتي. أقعد، أقعد ده إنت جيت في وقتك. البروجرام هيبتدي بعد دقايق من دلوقتي.
Jordan	انو هذا المطرب الهندي international يعني professional انو هذا المطرب الهندي انترناشونال يعني بروفيشنال
Mauritania	بيني و بينك Quinze بيني و بينك كوينز
Morocco	فين وصلتي ف préparation دالعرس؟ فين وصلتي فبريباراسيون دالعرس؟
Palestine	يعني maybe سبعة maximum ثمانية عبال ما الناس تتجمع يعني ميبي سبعة ماكسيموم ثمانية عبال ما الناس تتجمع
UAE	هاي الحركات مال أول، ألحين إحنا عايشين في زمن ال fast food هاي الحركات مال أول، ألحين إحنا عايشين في زمن الفاست فود
Yemen	—

Table 10: Examples of code-switching segments per dialect along with the transliterated version. Code-switched terms are provided in teal color.

A.6 Special cases

The special cases document served both as a collaborative tool for discussing and standardizing unique dialectal scenarios and as a repository for documenting dialect-specific variations and complex linguistic situations encountered during transcription. Table 11 illustrates some examples.

A.7 Preprocssing & settings

For all experiments, we utilize *transformers*[26] and *datasets*[27] libraries to load the models and datasets, respectively. We resample all audio segments to a 16kHz rate and perform the text preprocessing steps. We use a single node with A100-SXM4-40GB GPU for all evaluations. During the evaluation, we determine the WER and CER using the original reference and predicted transcriptions. Additionally, we apply text preprocessing to both the reference texts and predictions, adhering to the procedures outlined in Talafha et al. (2023). Specifically, we: (a) retain only the % and @ symbols, removing other punctuation; (b) eliminate diacritics, Hamzas, and Maddas; and (c) convert Eastern Arabic numerals to Western Arabic numerals (for instance, ٢٩ becomes 29). We keep all Latin characters as we have code-switching in *Casablanca*.

A.8 Code-switching analysis

To further understand code-switching evaluation, Tables 12 and 13 provide detailed outputs for each condition (see Section 7.3), focusing specifically on inputs involving code-switching and transliteration, respectively. We use whisper-lg-v3 for all conditions.

A.9 Error Analysis of High Error Rates

In response to the observed high error rates, particularly those exceeding 100 in our evaluations of the *Whisper-mixed* model, we perform error analysis to study the challenges contributing to these errors. This analysis is particularly focused on the Algerian dialect results, where we identify several cases (See Table 14):

- **Case 1:** Incorrect Language Base. The model frequently attempted to transcribe dialect-specific phrases by predicting phonetically similar words in MSA, despite their absence in the actual dialogue.

- **Case 2:** Inaccurate Translation Over Transcription. There were instances where the model predicted the MSA translation of phrases rather than transcribing the original dialect text.

[26] https://huggingface.co/docs/transformers/index
[27] https://huggingface.co/docs/datasets/index

Dialect	Description
Egyptian	Some speakers tend to use "ع" in the beginning of the words instead of “ه”, so we agreed on writing it as "ه". Others use the letter "ح" as in "حقولك" instead of "هقولك". We suggested writing it the way we hear. Some segments in the Egyptian dialect include urban upper Egyptian other than the Cairene one, so I wrote it as I heard. For example, a word like "أقولك" in Cairene would be "أجولك" in Upper Egyptian.
Jordanian	The word "هسا" is sometimes pronounced as "هسع", so I transcribe it based on the last letter; if "ع" is clear, I write "هسع" otherwise, I write "هسا". The word "Tomorrow" has two forms: بكرا and بكره. I decided to write بكرا to be distinguished from بكره which also means "I hate".
UAE	In many pronunciations, some Emaratis (depending on the region and tribe they belong to) put emphasis on some letters in a word. The word "علي" which means on top of me, can also be pronounced with an emphasis on the letter "ي". Another instance is where the letter "ه" is added at the end of the word "عليه". Emiratis use the word "عيل" mainly meaning "إذا ماذا ؟" or what else? However, the word has a less frequent use that means to be the cause of an issue "عيل عليه" or "عال عليه", but with a slightly different pronunciation.

Table 11: Illustrations of special cases unique to each dialect.

	Code-switching input
CS_reference	يعني maybe سبعة maximum ثمانية عبال ما الناس تتجمع بين السبعة و Maximum ثمانية عارفة بلي عندك النص .فالشركة بالحق ما عندكش الحق ف la signature. sorry سامر،أنا عمري ما نظرت لك إلا كصديق
CS - Auto	يعني ميبو 7 ماكسيموم 8 بين السبعة وماكسوم ثمان عارفة بلي عندك نصف الشركة بالحق ما عندكش حق في لا سينياتروخ ساري سامر أنا عمري ما نظرت لك إلا كصديق
CS - EN	Maybe 7, maximum 8 between 7 and maximum 8 I know you have half the company. You don’t have the right to have a seniority. Sorry, Samer. I’ve never seen you except as a friend.
CS - AR	يعني ميبو 7 ماكسيموم 8 بين السبعة وماكسوم ثمان عارفة بلي عندك نصف الشركة بالحق ما عندكش حق في لا سينياتروخ ساري سامر أنا عمري ما نظرت لك إلا كصديق

Table 12: Results of *whisper-lg-v3* on input having code-switching (Latin letters). **CS_reference**: reference transcriptions witch code-switching. **CS - Auto**: output from *whisper-lg-v3* without identifying the decoding language. **CS - EN**: output from *whisper-lg-v3* with identifying the decoding language as English. **CS - AR**: output from *whisper-lg-v3* with identifying the decoding language as Arabic.

- **Case 3:** Random Language Interference. The model sometimes generated sentences in completely unrelated languages, despite settings that specify transcription in Arabic.
- **Case 4:** Phonetic Dissimilarity in Short Utterances. Short utterances led to disproportionately high WER when the model generated MSA sentences not phonetically close to the dialect references.

	Transliterated input
Transliterated reference	يعني ميبي سبعة ماكسيموم ثمانية عبال ما الناس تتجمع بين السبعة و ماكسيموم ثمانية عارفة بلي عندك النص .فالشركة بالحق ما عندكش الحق فلا .سينياتور سوري سامر،أنا عمري ما نظرت لك إلا كصديق
Transliterated - Auto	يعني ميبو 7 ماكسيموم 8 بين السبعة وماكسوم ثمان عارفة بلي عندك نصف الشركة بالحق ما عندكش حق في لا سينياتروخ ساري سامر أنا عمري ما نظرت لك إلا كصديق
Transliterated - EN	Maybe 7, maximum 8 between 7 and maximum 8 I know you have half the company. You don’t have the right to have a seniority. Sorry, Samer. I’ve never seen you except as a friend.
Transliterated - AR	يعني ميبو 7 ماكسيموم 8 بين السبعة وماكسوم ثمان عارفة بلي عندك نصف الشركة بالحق ما عندكش حق في لا سينياتروخ ساري سامر أنا عمري ما نظرت لك إلا كصديق

Table 13: Results of *whisper-lg-v3* on input having transliterated words (Arabic letters). **Transliterated reference**: reference transcriptions with transliterated words. **Transliterated - Auto**: output from *whisper-lg-v3* without identifying the decoding language. **Transliterated - EN**: output from *whisper-lg-v3* with identifying the decoding language as English. **Transliterated - AR**: output from *whisper-lg-v3* with identifying the decoding language as Arabic.

Case #	Reference/Prediction
Case1	Reference: هام الحراير كل صباع بصنعة Prediction: أعمل حرايا بكل الصباب صن
Case2	Reference: خلاص روح للبوتيك تاعك بلوطة روح Prediction: فقط اذهب إلى بوتيكك
Case3	Reference: ما هدرتش عليك مولاي Prediction: Mă dărcea, nicmunei! Reference: نوريلك واش قادر ندير Prediction: Оңыр кел көш қадырын деп!
Case4	Reference: الله يسلمك Prediction: جيد جدا

Table 14: Samples from high error rates in the prediction of the Algerian dialect.