VoxArabica:

A Robust Dialect-Aware Arabic Speech Recognition System

Abdul Waheed[λ,⋆] **Bashar Talafha** [ξ,⋆] **Peter Sullivan**[ξ,⋆]
AbdelRahim Elmadany[ξ] **Muhammad Abdul-Mageed**[ξ,λ]
[ξ] Deep Learning & Natural Language Processing Group, The University of British Columbia
[λ]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
muhammad.mageed@ubc.ca

Abstract

Arabic is a broad language with many varieties and dialects spoken by $\sim$ 450 millions all around the world. Due to the linguistic diversity and variations, it is challenging to build a robust and generalized ASR system for Arabic. In this work, we address this gap by developing and demoing a system, dubbed VoxArabica, for dialect identification (DID) as well as automatic speech recognition (ASR) of Arabic. We train a wide range of models such as HuBERT (DID), Whisper, and XLS-R (ASR) in a supervised setting for Arabic DID and ASR tasks. Our DID models are trained to identify 17 different dialects in addition to MSA. We finetune our ASR models on MSA, Egyptian, Moroccan, and mixed data. Additionally, for the remaining dialects in ASR, we provide the option to choose various models such as Whisper and MMS in a zero-shot setting. We integrate these models into a single web interface with diverse features such as audio recording, file upload, model selection, and the option to raise flags for incorrect outputs. Overall, we believe VoxArabica will be useful for a wide range of audiences concerned with Arabic research. Our system is currently running at https://cdce-206-12-100-168.ngrok.io/.

1 Introduction

The Arabic language, with its diverse regional dialects, represents a unique linguistic spectrum with varying degrees of overlap between the different varieties at all linguistic levels (e.g., phonetic, syntactic, and semantic). In addition to Modern Standard Arabic (MSA), which is primarily used in education, pan-Arab media, and government, there are many local dialects and varieties that are sometimes categorized at regional (Zaidan and Callison-Burch, 2014; Elfardy and Diab, 2013; Elaraby and Abdul-Mageed, 2018), country (Bouamor et al., 2018; Abdul-Mageed et al., 2020a, 2021, 2022), or even province levels (Abdul-Mageed et al., 2020b). Historically, this wide and rich variation between different Arabic varieties has posed a significant challenge for automatic speech recognition (ASR) (Talafha et al., 2023; Alsayadi et al., 2022; Ali, 2020). The main focus has largely been on the recognition of MSA with very little-to-no focus on its dialects and varieties (Dhouib et al., 2022; Hussein et al., 2022; Ali et al., 2014). As such, ASR systems have conventionally been built either for MSA or individual dialects, thereby restricting their versatility and adaptability. However, the multifaceted nature of Arabic demands a robust ASR system that caters for its diverse dialects and varieties. In this work, we fill this research gap by introducing and demoing an ASR system integrated with a dialect identification model, dubbed *VoxArabica*.

VoxArabica is an end-to-end dialect-aware ASR system with dual functionality: (i) it offers a supervised dialect identification model followed by (ii) a finetuned Whisper Arabic ASR model covering multiple dialects. The dialect identification model works by assigning a country-level dialect, as well as MSA, from a set of 18 labels from input speech. This then allows the appropriate ASR model to fire. Contrary to traditional methodologies that separate dialect identification and speech recognition as two completely different tasks, our proposed pipeline integrates the two components effectively utilizing dialectal information for improved speech recognition. Such an integration not only improves the ASR output, but also establishes a framework aligned with the linguistic diversities inherent to Arabic as well. Concretely, our contributions can be summarized as follows:

- We introduce and demo our end-to-end VoxArabica system, which integrates dialect identification with state-of-the-art Arabic ASR.

[⋆]Equal contributions

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 441–449
December 7, 2023 ©2023 Association for Computational Linguistics

- Our demo is based on a user-friendly web interface characterized with rich functionalities such as audio uploading, audio recording, and user feedback options.

The rest of the paper is organized as follows: In Section 2, we overview related works. Section 3 introduces our methods. Section 4 offers a walkthrough of our demo. We conclude in Section 5.

2 Literature Review

Arabic ASR. Recent ASR research has focused on end-to-end (E2E) methods such as in Whisper (Radford et al., 2022) and the Universal Speech Model (Zhang et al., 2023). Such E2E deep learning models have significantly elevated ASR performance by allowing learning directly from the audio waveform, bypassing the need for intermediate feature extraction layers (Wang et al., 2019; Radford et al., 2022). Whisper is particularly noteworthy for its multitask training approach, incorporating ASR, voice activity detection, language identification, and speech translation. It has achieved state-of-the-art performance on multiple benchmark datasets such as Librispeech (Panayotov et al., 2015) and TEDLIUM (Rousseau et al., 2012). However, its resilience to adversarial noise has been questioned (Olivier and Raj, 2022).

For Arabic ASR specifically, the first E2E model was introduced using recurrent neural networks coupled with Connectionist Temporal Classification (CTC) (Ahmed et al., 2019). Subsequent works have built upon this foundation, including the development of transformer-based models that excel in both MSA and dialects (Belinkov et al., 2019; Hussein et al., 2022). One challenge for E2E ASR models is the substantial requirement for labeled data, particularly for languages with fewer resources such as varieties of Arabic. To address this, self-supervised and semi-supervised learning approaches are gaining traction. These models, such as Wav2vec2.0 and XLS-R, initially learn useful representations from large amounts of unlabeled or weakly labeled data and can later be finetuned for specific tasks (Baevski et al., 2020; Babu et al., 2021). W2v-BERT, another self-supervised model, employs contrastive learning and masked language modeling. It has been adapted for Arabic ASR by finetuning on the FLEURS dataset, which represents dialect-accented standard Arabic spoken by Egyptians (Chung et al., 2021; Conneau et al., 2023). Unlike Whisper, both Wav2vec2.0 and w2v-BERT necessitate a finetuning stage for effective decoding.

Arabic DID. Arabic DID has been the subject of a number of studies through recent years, enhanced by collection of spoken Arabic DID corpora such as ADI5 (Ali et al., 2017) and ADI17 (Shon et al., 2020). And advances in model architecture have mirrored changes in the larger LID research community, from i-vector (Dehak et al., 2010) based approaches (Ali et al., 2017) towards deep learning based approaches: x-vectors (Snyder et al., 2018; Shon et al., 2020), end-to-end classification using deep neural networks (Ali et al., 2019; Cai et al., 2018), and transfer learning (Sullivan et al., 2023).

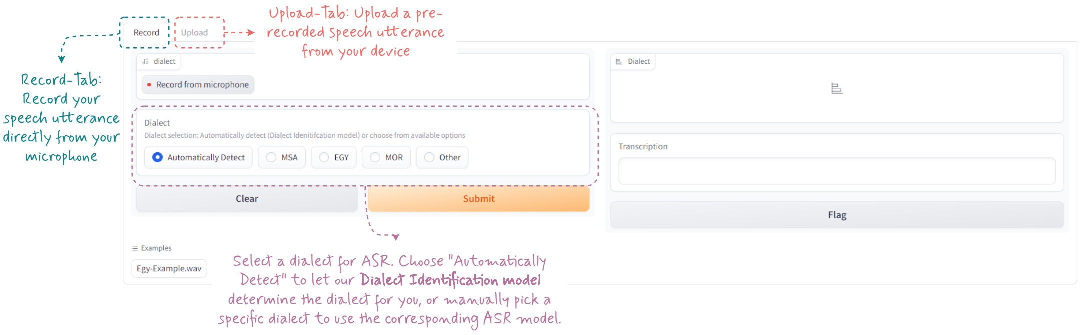
ASR and DID. Combining ASR and DID in a single pipeline remains fairly novel for Arabic. Recent works in this space has employed only limited corpora (Lounnas et al., 2020), or used ASR transcripts only to improve DID (Malmasi and Zampieri, 2017). Closest to our demonstrated system in this work is FarSpeech (Eldesouki et al., 2019), since it combines ASR and DID. However, FarSpeech is confined to coarse-grain DID and only supports MSA for ASR. In addition, compared to FarSpeech, our models are *modular* in that it allows users to run either or both ASR or DID, depending on their needs.

3 Models

3.1 DID Models

Our DID model is a transfer learning approach: finetuning HuBERT (Hsu et al., 2021) on ADI-17 (Shon et al., 2020) and the MSA portions of ADI-5 (Ali et al., 2017) and MGB-2 (Ali et al., 2016). We utilize only the MSA portions of ADI-5 due to the ambiguity of going from coarse-grain to fine-grain labels. Dialectal varieties covered in our model are *MSA, Algerian, Egyptian, Iraqi, Jordanian, Saudi, Kuwaiti, Lebanese, Libyan, Mauritanian, Moroccon, Omani, Palestinian, Qatari, Sudanese, Syrian, Emirati*, and *Yemeni*.

Training Details. Our finetuning procedure entailed performing a random search for training hyperparameters validated using the ADI-17 development set. A detailed overview of the hyperparameters searched can be found in Table 1 . We train using AdamW as optimizer, with a certain number of initial steps, *Freeze Steps*, where the original model is not updated and only the newly initialized classification layers change. After thawing, we also experiment with keeping some of the earlier layers

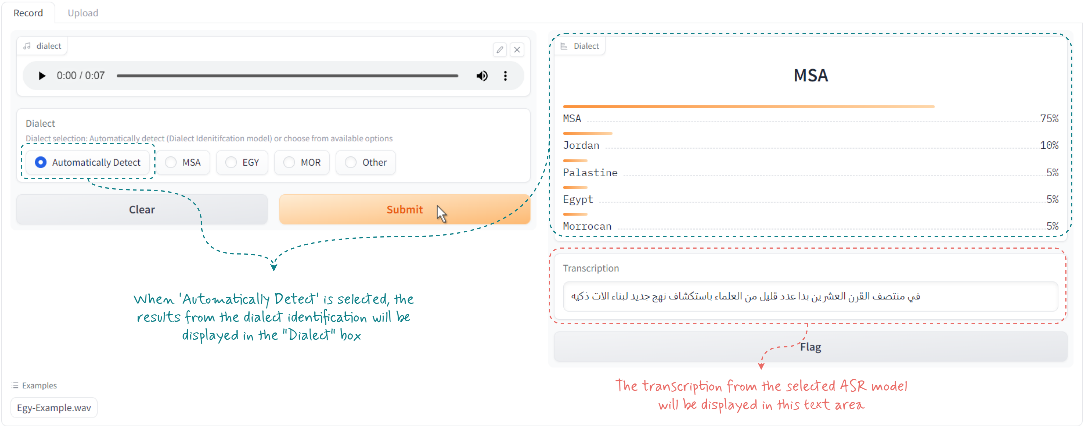

Record
Upload
dialect
Record from microphone
Dialect
Dialect selection: Automatically detect (Dialect Idenitifcation model) or choose from available options
● Automatically Detect ○ MSA ○ EGY ○ MOR ○ Other
Clear
Submit
Examples
Egy-Example.wav
Dialect
Transcription
Flag

Upload-Tab: Upload a pre-recorded speech utterance from your device

Record-Tab: Record your speech utterance directly from your microphone

Select a dialect for ASR. Choose "Automatically Detect" to let our **Dialect Identification model** determine the dialect for you, or manually pick a specific dialect to use the corresponding ASR model.

Figure 1: Users have the **option to either upload files or directly record their audio**. Additionally, the dialect can be automatically detected or manually selected for a specific ASR model.


Record
Upload
dialect
0:00 / 0:07
Dialect
Dialect selection: Automatically detect (Dialect Idenitifcation model) or choose from available options
● Automatically Detect ○ MSA ○ EGY ○ MOR ○ Other
Clear
Submit
Examples
Egy-Example.wav
Dialect
Transcription
Flag

MSA

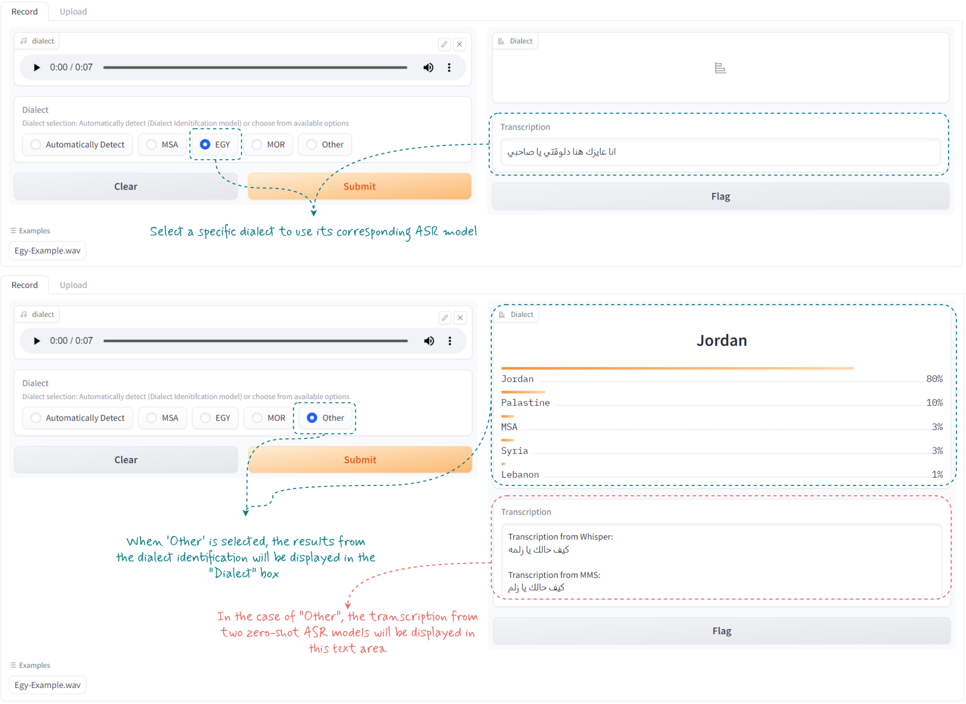
MSA	75%
Jordan	10%
Palastine	5%
Egypt	5%
Morrocan	5%

في منتصف القرن العشرين بدا عدد قليل من العلماء باستكشاف نهج جديد لبناء الات ذكيه

When 'Automatically Detect' is selected, the results from the dialect identification will be displayed in the "Dialect" box

The transcription from the selected ASR model will be displayed in this text area

Figure 2: For **automatic dialect detection**, likelihood percentages determine the ASR model choice, with transcriptions displayed in the Transcription text area.


Record
Upload
dialect
0:00 / 0:07
Dialect
Dialect selection: Automatically detect (Dialect Idenitifcation model) or choose from available options
○ Automatically Detect ○ MSA ● EGY ○ MOR ○ Other
Clear
Submit
Examples
Egy-Example.wav
Dialect
Transcription
Flag
Record
Upload
dialect
0:00 / 0:07
Dialect
Dialect selection: Automatically detect (Dialect Idenitifcation model) or choose from available options
○ Automatically Detect ○ MSA ○ EGY ○ MOR ● Other
Clear
Submit
Examples
Egy-Example.wav
Dialect
Transcription
Flag

انا عايزك هنا دلوقتي يا صاحبي

Select a specific dialect to use its corresponding ASR model

Jordan

Jordan	80%
Palastine	10%
MSA	3%
Syria	3%
Lebanon	1%

Transcription from Whisper:
كيف حالك يا زلمه

Transcription from MMS:
كيف حالك يا زلم

When 'Other' is selected, the results from the dialect identification will be displayed in the "Dialect" box

In the case of "Other", the transcription from two zero-shot ASR models will be displayed in this text area

Figure 3: When a specific dialect is manually selected, its associated **ASR model generates the transcription**. When recording in an unlisted dialect, select "Other". The dialect identification model will then detect the dialect, and both Whisper and MMS zero-shot models will produce the transcription.

Table 1: An overview of the search space of the hyperparmeter tuning for the DID model as well as optimal configuration found during the (n=30) random search. The batch size formula ensures our V100 GPUs were fully utilized during training, with a target of 75 seconds of audio regardless of the sampling duration. All values are picked from uniform distributions except for the learning rate, which was picked from a log uniform distribution.

	Range	**Conf.**
Batch Size	$4 \cdot \lfloor \frac{75}{Duration} \rfloor$	16
Freeze Steps	$[0, 1000]$	192
Learning Rate	$[1 \cdot 10^{-5}, 1 \cdot 10^{-2}]$	$6 \cdot 10^{-4}$
Max Steps	$[20k, 40k]$	29225
Duration	$[4, 18]$ seconds	4.69
Thaw Depth	$[0, 23]$	3

of the model frozen. We indicate the earliest layer that gets thawed as *Thaw Depth*. We also experiment with LayerNorm and Attention finetuning (Li et al., 2020), but our final model performed better without it.

3.2 ASR Models

We train a wide range of ASR models on a list of benchmark Arabic speech datasets. Our models include two versions of Whisper (Radford et al., 2022), *large-v2* and *small*. We also finetune XLS-R (Babu et al., 2022) for the ASR task. For MSA, we train our models on three versions of *common voice* (Ardila et al., 2019) datasets `6.1`, `9.0`, and `11.0`. We note that Talafha et al. (2023) show that Whisper *large-v2* outperforms its smaller variant as well as XLS-R trained on the same dataset. For Morrocan, Egyptian, and MSA, we fully finetune models on MGB2, MGB3, MGB5 (Ali et al., 2016, 2017, 2019). We also train ASR models on FLEURS (Conneau et al., 2023), which is accented Egyptian speech data.

Text Preprocessing. The datasets we employ exhibit various inconsistencies. For instance, within CV6.1, the utterance فَقَالَ لَهُمْ "faqaAla lahumo" is fully diacritic, whereas the utterance فإذا النجوم طمست "f<*A Alnjwm Tmst" lacks diacritic annotations, despite both originating from the Quran. Consequently, we adopt the normalization approach from (Chowdhury et al., 2021; Talafha et al., 2023), which involves: (a) discarding all punctuation marks excluding the % and @ symbols; (b) eliminating diacritics, Hamzas, and Maddas; and (c) converting eastern Arabic numerals into their western counterparts (e.g., 29 remains 29). Given that this study does not address code-switching, all Latin alphabet is excluded.

Ref (EGY)	مساء الخير اهلا ومرحبا بيكم في حلقة جديدة من برنامج بوضوح اي واحد نفسه في ثانية يطلعها قدام الكاميرا
Whisper (0-shot)	بسعي الخير أهلا ومرحبا بكم في حلقة جديدة من برنامج بوضوح أي واحد نفسه في ثانية يطلعها قدام الكاميرا
MMS (0-shot)	بساء لخير أهلن مرحباً بكم فخلى أجديدة من برنامج بوضوح أي واحد نفسفسنية يطلعها ودمك كمرة
Whisper (MSA)	بسعر الخير اهلا ومرحبا بكم في حلقة جديدة من برنامج بوضوح اي واحد نفسه سامي لا يطلعها قدم الكاميرا
Whisper(EGY)	مساء الخير اهلا ومرحبا بيكم في حلقة جديدة من برنامج بوضوح اي واحد نفسه في ثانية يطلعها قدام الكاميرا
Whisper(MOR)	مسايا الخير اهلا ومرحبا بيكم في حلقة جديدة من برنامج بوضوح اي واحد نفسو فسنية لي يطلعها قدام الكاميرا
XLS-R(MSA)	مساء الخير اهلا ومرحبا بكم في حالة جديدة من برنامج بوضوح اي واحد نفسه ثانية يطلعها قدام الكاميرا

Table 2: Example outputs produced by VoxArabica when input audio is Egyptian dialect.

Training Details. Before training, we apply preprocessing steps as mentioned above on the text. We train all of our models using AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e-5$, 500 warmup steps, and no weight decay. To prevent the model from severely overfitting, we employ early stopping with patience at 5. We use Huggingface trainer [1] with deepspeed ZeRO (Rajbhandari et al., 2019) stage-2 to parallelize our training across 8xA100 (40G) GPUs.

In our demo, we also allow users to utilize both Whisper and MMS (Pratap et al., 2023) in the zero-shot setting.

4 Walkthrough

Our demo consists of a web interface with versatile functionality. It allows users to interact with the system in multiple ways, depending on their needs.

User audio input. Users can either record their own audio through a microphone or upload a prerecorded file. In both cases, we allow different formats such as `.wav`, `.mp3`, or `.flac`, across various audio sampling rates (e.g., `16khz` or `48khz`). Figure 1 demonstrates the different options available to the user upon interacting with VoxArabica.

[1] `https://huggingface.co/docs/transformers/main_classes/trainer`

Model name	Dialect(s)	Dataset	Architecture
Whisper MSA	MSA	CV (6.1, 9.0, 11.0)	Whisper
XLS-R	MSA	CV (6.1, 9.0, 11.0)	Wav2vec 2.0
Whisper Morroco	MOR	MGB5	Whisper
Whisper Egypt	EGY	MGB3	Whisper
Whisper Zero-shot	-	-	Whisper
MMS	-	-	Wav2vec 2.0

Table 3: The utilized ASR models, their associated dialects, and respective architectures, and dataset used to train each model. Models marked with a dash are generic and not specific to a particular dialect.

Model selection. Users can choose to select an Arabic variety for transcription, or have it automatically detected using our 18-way DID system. We demonstrate this in Figure 2. Once the variety is detected, the corresponding ASR model will perform transcription and both DID transcription results will be presented on the interface (as shown in Figure 3). We offer various models: two for the EGY and MOR, respectively; two for MSA; and two generic models that can be used for any variety. We list all models in Table 3. In cases where predicted/selected variety is not covered by our ASR models, we fall back to our generic models (i.e., both Whisper zero-shot and MMS zero-shot).

User feedback. We also provide an option for users to submit *anonymous* feedback about the produced output by raising a flag. We use this information to collect high quality silver labels and discard examples where a flag is raised for incorrect outputs. It is important to note that we do not collect any external user data for any purpose, thus ensuring user privacy.

System output. Our system conveniently outputs both predicted Arabic variety and transcription across two panels as shown in Figure 3. For predicted variety, we show users all top five predictions along with model confidence for each of them. We provide outputs produced by our models in VoxArabica when the reference input is Egyptian dialect in Table 2. We also present additional examples in Appendix, Table 4.

5 Conclusion

We present a demonstration of combined DID and ASR pipeline to illustrate the potential for these systems to improve the usability of dialectal Arabic speech technologies. We report example outputs produced by our system for multiple dialects showcasing the effectiveness of integrated DID and ASR pipelines. We believe that our demo will advance the research to build a robust and generalized Arabic ASR system for a wide range of varieties and dialects and will enable a more holistic assessment of the strengths and weaknesses of these methods. For future work, we intend to add models for more dialects and varieties particularly those which are low resource.

6 Limitations

Audio classification tasks can be susceptible to out-of-domain performance degradation, which may impact real world performance. Similarly, studies on the interpretability of DID models have shown internal encoding of non-linguistic factors such as gender and channel (Chowdhury et al., 2020), which may impart bias to the models. Ensuring training corpora contain a diverse balance of speaker gender, recording conditions, as well as full coverage of the different styles of language is an ongoing challenge. We hope that by creating an online demonstration, these limitations can be further explored.

7 Ethics Statement

Intended use. We build a robust dialect identification and speech recognition system for multiple Arabic dialects as well as MSA. We showcase the capability of our system in the demo. We believe that our work will guide a new direction of research to develop a robust and generalized speech recognition system for Arabic. Through our demo, we integrate DID with ASR system which support multiple dialects.

Potential misuse and bias. Since our data is limited to a few dialects involved in finetuning DID and ASR systems, we do not expect our models to generalize all varieties and dialects of Arabic that are not supported by our models.

Acknowledgments

We gratefully 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; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[2] and UBC ARC-Sockeye.[3]

[2] https://alliancecan.ca

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

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. Nadi 2020: The first nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2010.11334*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

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.

Abbas Raza Ali. 2020. Multi-dialect arabic speech recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

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.

Ahmed M. Ali, Hamdy Mubarak, and Stephan Vogel. 2014. Advances in dialectal arabic speech recognition: a study using twitter to improve egyptian asr. In *International Workshop on Spoken Language Translation*.

Hamzah A Alsayadi, Abdelaziz A Abdelhamid, Islam Hegazy, Bandar Alotaibi, and Zaki T Fayed. 2022. Deep investigation of the recent advances in dialectal arabic speech recognition. *IEEE Access*, 10:57063–57079.

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, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

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.

Yonatan Belinkov, Ahmed Ali, and James Glass. 2019. Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition. *arXiv preprint arXiv:1907.04224*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Weicheng Cai, Zexin Cai, Wenbo Liu, Xiaoqi Wang, and Ming Li. 2018. Insights in-to-end learning scheme for language identification. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5209–5213. IEEE.

Shammur A Chowdhury, Ahmed Ali, Suwon Shon, and James R Glass. 2020. What does an end-to-end dialect identification model learn about non-dialectal information? In *INTERSPEECH*, pages 462–466.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards One Model to Rule All: Multilingual Strategy for Dialectal Code-Switching Arabic ASR. In *Proc. Interspeech 2021*, pages 2466–2470.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. Arabic automatic speech recognition: A systematic literature review. *Applied Sciences*, 12(17).

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.

Mohamed Eldesouki, Naassih Gopee, Ahmed Ali, and Kareem Darwish. 2019. Farspeech: Arabic natural language processing for live arabic speech. In *INTERSPEECH*, pages 2372–2373.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461.

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*, 29:3451–3460.

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

Xian Li, Changhan Wang, Yun Tang, Chau Tran, Yuqing Tang, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2020. Multilingual speech translation with efficient finetuning of pretrained models. *arXiv preprint arXiv:2010.12829*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Khaled Lounnas, Hassan Satori, Mohamed Hamidi, Hocine Teffahi, Mourad Abbas, and Mohamed Lichouri. 2020. Cliasr: a combined automatic speech recognition and language identification system. In *2020 1st international conference on innovative research in applied science, engineering and Technology (IRASET)*, pages 1–5. IEEE.

Shervin Malmasi and Marcos Zampieri. 2017. Arabic dialect identification using iVectors and ASR transcripts. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 178–183, Valencia, Spain. Association for Computational Linguistics.

Raphael Olivier and Bhiksha Raj. 2022. There is more than one kind of robustness: Fooling whisper with adversarial examples. *arXiv preprint arXiv:2210.17316*.

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.

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. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimization towards training A trillion parameter models. *CoRR*, abs/1910.02054.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248. IEEE.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Peter Sullivan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2023. On the Robustness of Arabic Speech Dialect Identification. In *Proc. INTERSPEECH 2023*, pages 5326–5330.

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

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages.

Appendix

Example Outputs

Ref (MSA)	يؤثر التدخين بشكل سلبي في جسم الانسان حيث ينتج عنه العديد من الاثار السلبية المؤذية للفرد وقد تؤدي بعضها الى مضاعفات تهدد الحياة
MMS	يؤثر التدخين بشكل سلبي هي جسم الإنسان حيث ينتجعنه العديد من الآثار السلبية المؤذيد الفرد وقد تؤدي بعضها إلى مضاعفات تهدد الحياة
Whisper(0-shot)	يؤثر التدخين بشكل سلبي في جسم الانسان حيث ينتج عنه العديد من الاثار السلبية المؤذية للفرد وقد تؤدي بعضها الى مضاعفات تهدد الحياة
Whisper(MSA)	يؤثر التدخين بشكل سلبي في جسم الإنسان حيث ينتج عنه العديد من الآثار السلبية المؤذية للفرد وقد تؤدي بعضها إلى مضاعفات تهدد الحياة
Whisper(MOR)	يؤثر التدخين بشكل سلبي في جسم الانسان حيت ينتج عنه العديد من الاثار السلبية المؤذية دالفرق وقد تؤدي بعضها الى مضعفات تهدد الحياة
Whisper(EGY)	يؤثر التدخين بشكل سلبي في جسم الانسان حيث ينتج عنه العديد من الاثار السلبية المؤذية للفرد وقد تؤدي بعضها الى مضاعفات تهدد الحياة
Ref (JOR - Other)	يا زلمة كيف حالك؟ شو أخبارك؟ وين هالغيبة؟ زمان عنك، ليش ما بتبين؟
MMS (0-shot)	يعزل كاف حلكشو أخبارك وانه الغاب زمان عنك لاش ما بتبين
Whisper (0-shot)	يا زلمة كيف حالك؟ شو أخبارك؟ وين هالغابة؟ زمان عنك، ليش ما بتبين؟
Whisper (MSA)	يا زلم كيف حالك شو اخبارك وانها الغيبة زمان عنك ليش ما بتبين
Whisper (MOR)	يا زلمة كيف حالك شو اخبارك وانها الغابة زمان عندك لاش مابتبين
Whisper (EGY)	يا زلمة كاف حالك شو اخبارك وانها الغابة اذا ما عنك ليش ما بتبين

Table 4: Outputs produced by VoxArabica when input is Egyptian and Jordanian. For Jordanian dialect, we do not have a finetuned model and Whisper (0-shot) performs best. Hence highlighting the lack of generalisation for various finetuned models to unseen dialects.