Moroccan Darija Speech Recognition using WHISPER

Ali ASOUAB, Mohamed AISSAM, Ahmed BENBRAHIM
Université Internationale de Rabat
{ali.asouab, mohamed.aissam, ahmed.benbrahim}@uir.ac.ma

ABSTRACT

Speech recognition technology has advanced significantly in recent years, but dialects such as Moroccan Darija remain underrepresented in automatic speech recognition (ASR) systems due to their unique linguistic characteristics and lack of resources. Moroccan Darija, a dynamic fusion of Arabic, Berber, French, and Spanish, is widely spoken but lacks comprehensive datasets for speech processing. This article introduces the Moroccan Darija Voice Corpus (MDVC), a meticulously curated dataset containing transcribed audio files from diverse sources, including Moroccan YouTube content. Leveraging this corpus, the state-of-the-art WHISPER model was fine-tuned to address the challenges of Moroccan Darija speech recognition. The resulting system demonstrates significant improvements in transcription accuracy and efficiency, bridging the gap in ASR technology for this dialect.

Keywords: Speech recognition, Moroccan Darija, WHISPER, automatic speech recognition (ASR), fine-tuning, dialectal Arabic, linguistic diversity, voice corpus, Moroccan Darija Voice Corpus (MDVC), transcription accuracy.

I. INTRODUCTION

Speech recognition technology has made significant strides, facilitating seamless human-machine interactions across various languages. However, many dialects, including Moroccan Darija, remain underrepresented in the field of automatic speech recognition (ASR). While Arabic speech recognition has been widely explored in previous works, Moroccan Darija presents unique linguistic challenges due to its fusion of Arabic, Berber, French, and Spanish influences.

Recognized as a distinct linguistic entity, Moroccan Darija plays a central role in everyday communication in Morocco. Despite its widespread use, resources for speech processing in Darija are scarce, limiting the development of robust ASR models tailored to this dialect.

Addressing this gap, we introduce an innovative and comprehensive database specifically designed for Moroccan Darija speech recognition. This dataset comprises transcribed audio files collected from diverse sources, including Moroccan YouTube content creators who communicate in Darija. The dataset is structured meticulously to enhance ASR model performance. It consists of a CSV file that maps audio waveforms (wavs) to their corresponding transcriptions, along with details such as duration and file path.

By incorporating various regional accents, this dataset provides an authentic representation of Morocco's linguistic diversity, paving the way for more accurate and inclusive speech recognition systems.

This article focuses on the creation and curation of this extensive Moroccan Darija speech corpus and its application in fine-tuning the WHISPER model for dialect-specific ASR. Through this initiative, we aim to bridge the resource gap for Moroccan Darija in language technology, fostering advancements in ASR research and practical applications.

Potential uses of this technology span multiple domains, including:

- **Marketing:** Analyzing customer opinions expressed in Moroccan Darija on social media to improve product quality and consumer engagement.
- **Healthcare:** Facilitating doctor-patient communication by enabling voice-based medical transcription and automated patient support in Moroccan Darija.
- **Education:** Assisting in literacy programs, language learning applications, and accessibility tools for students with disabilities.
- **Tourism and Hospitality:** Assisting tourists by enabling real-time translation and interaction with local services using voice technology.

By developing a comprehensive speech recognition system for Moroccan Darija, we strive to ensure linguistic inclusivity while unlocking new possibilities in natural language processing and human-computer interaction.

II. DARIJA AND ITS DIVERSE ACCENTS

The Moroccan dialect, widely known as "Darija," is a vibrant and dynamic language that embodies Morocco's cultural richness and linguistic heritage. Unlike Modern Standard Arabic (MSA), which is used in formal and official contexts, Darija is the language of daily life, spoken by the majority of Moroccans in casual conversations and informal settings. It serves as a living example of how languages evolve and adapt, blending influences from various cultures and historical eras.

At its foundation, Darija is a unique fusion of languages. While rooted in Arabic, it has been profoundly shaped by the Berber (Amazigh) languages, spoken in Morocco long before the arrival of Arabic. This influence is evident in numerous Darija words, grammatical structures, and its distinctive pronunciation. Additionally, Darija incorporates vocabulary and expressions from other languages, such as French and Spanish, reflecting Morocco's colonial history and geographical proximity to Europe. The result is a lan-

guage that is fluid, adaptive, and deeply resonant with the diverse identities of its speakers.

One of Darija's most captivating features is its regional diversity. Each region of Morocco has developed its own distinct variation of the dialect, influenced by local traditions, history, and geography. For instance:

- **Northern Regions:** Cities like Tangier and Tetouan exhibit the "Tanjawi" accent, characterized by Spanish-influenced vocabulary and melodic intonations.
- **Central Regions:** The "Fassi" accent of Fes is marked by refined pronunciation, while Casablanca ("Casawi") and Rabat ("Rabati") reflect urbanized variations of Darija.
- **Southern Regions:** In areas like Souss, the "Soussi" accent highlights Amazigh linguistic influences, while in the Sahara, the "Sahrawi" accent is distinguished by its unique rhythm and vocabulary.

These regional accents are not mere linguistic curiosities; they reflect Morocco's profound cultural diversity. The coexistence of these accents demonstrates the country's ability to celebrate its differences while maintaining a shared identity through the unifying thread of Darija.

In our Moroccan Darija speech recognition project, embracing this linguistic diversity was a fundamental priority. During the creation of the database, we ensured the collection of audio samples from across the country, capturing voices from cities, towns, and rural areas. By representing this variety of accents, our goal is to develop a speech recognition system that authentically reflects the true essence of Darija, serving all its speakers with inclusivity and accuracy. The linguistic diversity of Morocco is illustrated in the map below, providing a visual representation of the regions and their corresponding dialects:

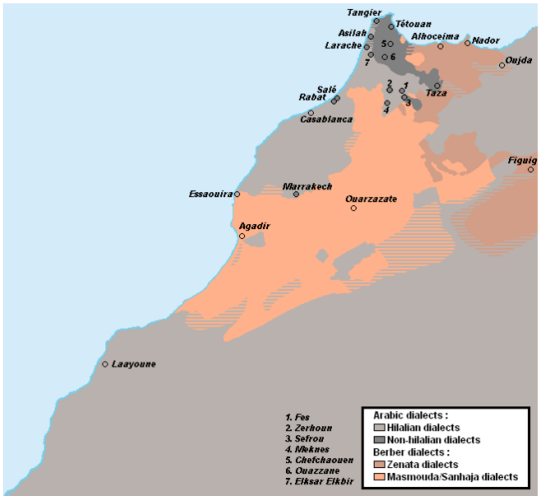

Fig. 1. Linguistic map of Morocco, highlighting the diverse regional dialects of Moroccan Darija.

III. MOROCCAN DARIJA VOICE CORPUS AND DATABASE CREATION

The Moroccan Darija Voice Corpus (MDVC) is a comprehensive database designed to capture the linguistic richness and diversity of Moroccan Darija, a dialect that blends Arabic, Berber, French, and Spanish influences. Recognized as a distinct linguistic entity, Moroccan Darija plays a central role in everyday communication in Morocco. Despite its widespread use, resources for speech processing in Darija are scarce, limiting the development of robust automatic speech recognition (ASR) models tailored to this dialect.

MDVC aims to bridge this gap by providing a large and diverse dataset that reflects the various regional accents and sub-dialects of Moroccan Darija. The dataset includes audio samples from across Morocco, capturing the full spectrum of Darija accents and sub-dialects, and accounting for the influence of Tamazight in mixed-language regions. This diversity ensures the creation of inclusive and regionally accurate ASR models that reflect the true essence of Moroccan Darija.

Moroccan Darija is divided into four main regional dialects:

- **Northern Dialects:** Spoken in cities like Tetouan and Chefchaouen, these dialects are influenced by Spanish.
- **Eastern Dialects:** Found in regions such as Oujda and Jerada, these dialects are characterized by unique local traits.
- **Western Dialects:** Predominant in urban centers like Rabat, Fes, and Casablanca, these dialects blend traditional and modern elements.
- **Southern Dialects (Hassania):** Common in areas such as Laayoune and Dakhla, these dialects have strong Tamazight and Saharan influences.

Each region also encompasses a variety of sub-dialects, adding complexity to the linguistic landscape. This diversity presents both challenges and opportunities for developing robust speech recognition systems.

A. Database Creation Process

To build a robust Moroccan Darija speech recognition system, we created a comprehensive dataset tailored specifically to this unique dialect. Given the scarcity of significant open-source speech and text resources for Moroccan Darija, we curated a dataset by collecting speech samples and their corresponding transcriptions from diverse sources. Below are the steps involved in this process:

1) Data Collection: Our primary data source was YouTube, focusing on Moroccan YouTubers who produce content in Darija. These videos cover a wide range of topics, including daily vlogs, interviews, commentary, and educational content, ensuring a diverse representation of vocabulary and speaking styles. By utilizing publicly available content, we gathered authentic and context-rich samples of spoken Darija.

2) Data Extraction: For each selected video, we extracted the audio tracks and segmented them into smaller, manageable clips. These clips were chosen to ensure that each

contained coherent and complete speech segments. The audio files were saved in standard formats, such as WAV and MP3, with metadata including duration and sampling rate.

3) Transcription: The transcriptions for the audio clips were sourced directly from the videos on YouTube. Many Moroccan YouTubers provide transcriptions as part of their content, either as subtitles or on-screen text. These transcriptions, written entirely in Arabic script, reflect the natural representation of Moroccan Darija.

4) Data Augmentation: To enhance the robustness of the dataset and improve the model's ability to generalize across various conditions, we applied several data augmentation techniques. These augmentations simulate different acoustic environments, distortions, and variations in speech delivery, which are crucial for creating a resilient speech recognition system. The following techniques were employed:

- **Augmentation Lobe:** Introduced minor variations in the audio spectrum, mimicking realistic changes in sound frequency and amplitude during recording or playback.
- **Clipping:** Introduced controlled levels of clipping in some audio segments to simulate over-modulated or distorted recordings common in real-world scenarios.
- **Frequency Dropout:** Randomly dropped certain frequency bands to mimic the effect of low-quality audio equipment or noise interference.
- **Speed Perturbation:** Varied playback speed slightly (faster or slower) to create variations in speech tempo.
- **Time Dropout:** Randomly masked small segments of audio to simulate interruptions or missing data.

These augmentation techniques were applied systematically to maintain dataset balance while introducing the variability necessary for effective model training.

5) Dataset Statistics: The finalized dataset consists of 6,827 audio clips in the training set, with a total duration of approximately 6.60 hours. The validation set includes 2,252 audio clips with a total duration of 2.19 hours, and the test set comprises 2,267 audio clips with a total duration of 2.21 hours. The diversity of speakers, topics, and acoustic conditions across these sets makes this dataset a valuable resource for training and evaluating speech recognition models tailored to Moroccan Darija.

IV. SPEECH RECOGNITION USING MDVC AND WHISPER

In the development of our speech recognition system, we leveraged the power of the Moroccan Darija Voice Corpus (MDVC) alongside the state-of-the-art WHISPER model. This combination enabled us to achieve exceptional accuracy in transcribing Moroccan Darija speech, addressing the linguistic diversity and complexity of this dialect. By fine-tuning WHISPER to the MDVC dataset, we successfully enhanced its performance, surpassing previous benchmarks in accuracy and efficiency for Moroccan Darija.

A. Presentation of WHISPER Model

1) Introduction: The WHISPER model, developed by OpenAI, is a state-of-the-art automatic speech recognition (ASR) system designed to transcribe and translate speech with remarkable accuracy. Built upon a robust transformer-based architecture, WHISPER excels in handling diverse linguistic data, making it an ideal choice for processing complex and low-resource languages like Moroccan Darija.

2) Architecture Overview: The image below (Figure 2) illustrates the encoder-decoder architecture of the WHISPER model, showcasing its key components and mechanisms:

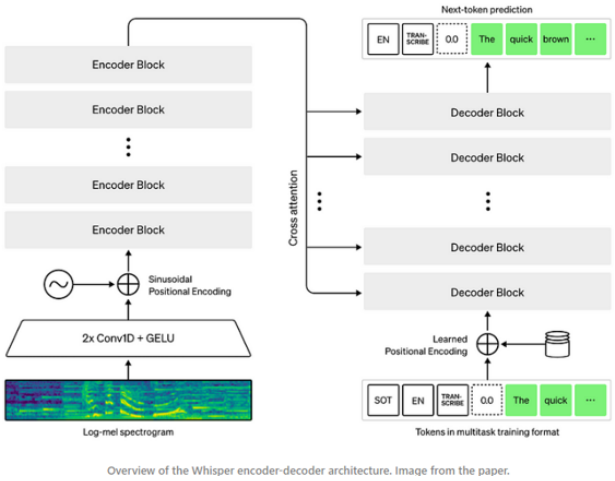

Fig. 2. Overview of the WHISPER encoder-decoder architecture.

3) Key Features of the WHISPER Model:

1) **Multilingual and Multi-task Capability:** WHISPER supports transcription and translation across multiple languages, allowing it to perform well in multilingual contexts. This capability is particularly useful for Moroccan Darija, which often integrates words and phrases from Arabic, French, Berber, and Spanish.
2) **Robustness to Noise and Accents:** Trained on a vast dataset containing various accents and acoustic conditions, WHISPER can effectively handle the diverse regional accents and challenging audio environments often encountered in Moroccan Darija speech.
3) **Open-Source Accessibility:** As an open-source model, WHISPER is freely available to researchers and developers, encouraging its application to underrepresented languages and dialects like Darija.
4) **Fine-Tuning Flexibility:** WHISPER can be fine-tuned to specialize in specific tasks or dialects. For the Moroccan Darija Voice Corpus (MDVC), fine-tuning the model ensures improved recognition performance tailored to the unique linguistic traits of Darija.
5) **Integration of Contextual Understanding:** WHISPER uses its transformer-based architecture to provide contextual understanding, which enhances its ability to recognize and transcribe naturally spoken, conversational Darija.

B. Fine-tuning WHISPER using MDVC

To enhance the accuracy of Moroccan Darija transcription, we fine-tuned the pre-trained WHISPER model using the

Moroccan Darija Voice Corpus (MDVC). Fine-tuning allows the model to adapt to the unique linguistic characteristics and acoustic variations of Moroccan Darija, ensuring better performance in recognizing and transcribing this dialect.

1) Training Setup: The fine-tuning process involved configuring several key parameters to optimize the model for Moroccan Darija:

- A small batch size was used to accommodate hardware constraints, with gradient accumulation enabling an effective batch size.
- The learning rate was set to a low value to promote generalization and prevent loss of pre-trained knowledge.
- The model was trained over three epochs, balancing accuracy and overfitting.
- Periodic evaluation was conducted during training to ensure effective convergence.

2) Data Preparation: The training dataset comprised segmented audio clips and their corresponding transcriptions. These transcriptions were prepared in Arabic script, reflecting the natural representation of Moroccan Darija. To ensure consistency, the dataset was processed to include padding for both audio features and text sequences. Padding tokens were masked appropriately during training to avoid impacting the loss calculation.

3) Training Process:

C. Fine-tuning WHISPER using MDVC

To enhance the accuracy of Moroccan Darija transcription, we fine-tuned the pre-trained WHISPER model using the Moroccan Darija Voice Corpus (MDVC). Fine-tuning allows the model to adapt to the unique linguistic characteristics and acoustic variations of Moroccan Darija, ensuring better performance in recognizing and transcribing this dialect.

1) Training Setup: The fine-tuning process involved configuring several key parameters to optimize the model for Moroccan Darija. A small batch size was chosen to accommodate hardware constraints, while gradient accumulation was employed to achieve an effective batch size that supports stable training. The evaluation was conducted periodically during training to monitor progress and ensure the model converged effectively.

To promote generalization and handle the variability in speech patterns, the learning rate was set to a low value, enabling the model to gradually adapt without losing the knowledge acquired during pre-training. The model was trained over three epochs, striking a balance between achieving high accuracy and avoiding overfitting.

2) Data Preparation: The training dataset comprised segmented audio clips and their corresponding transcriptions. These transcriptions were prepared in Arabic script, reflecting the natural representation of Moroccan Darija. To ensure consistency, the dataset was processed to include padding for both audio features and text sequences. Padding tokens were masked appropriately during training to avoid impacting the loss calculation.

3) Training Process: The fine-tuning process focused on optimizing the model's ability to recognize the linguistic diversity within Moroccan Darija. This included regional accents, variations in speech tempo, and the integration of vocabulary influenced by other languages such as French and Amazigh. The training workflow involved monitoring the model's performance on a validation set, saving checkpoints at regular intervals, and applying strategies to ensure the stability and efficiency of training.

After three epochs, the model achieved a **Training Loss** of **0.0779**, a **Validation Loss** of **0.3065**, and a **Word Error Rate (WER)** of **10.04%**. These results demonstrate significant progress in transcription accuracy and adaptability to Moroccan Darija, despite the inherent challenges posed by linguistic and acoustic variability.

The results are summarized in Table I.

Training Loss	Validation Loss	WER (%)
0.0779	0.3065	10.04

TABLE I

MODEL PERFORMANCE AFTER FINE-TUNING ON THE MDVC DATASET.

D. Challenges

Building a speech recognition system tailored to Moroccan Darija presented several unique challenges:

1) **Linguistic Diversity:** Moroccan Darija exhibits significant regional variation, influenced by geography, history, and contact with other languages such as Amazigh, French, and Spanish. Capturing this diversity in a balanced dataset was critical but complex.
2) **Lack of Standardization:** Unlike Modern Standard Arabic, Moroccan Darija lacks standardized spelling or grammar rules, making transcription and alignment of audio data more labor-intensive.
3) **Limited Resources:** Open-source datasets for Moroccan Darija were scarce, necessitating the creation of a dedicated voice corpus from scratch.
4) **Acoustic Variability:** Variability in recording conditions, such as background noise, microphone quality, and speaker intonation, posed additional challenges.

E. Evaluation

To evaluate the performance of the fine-tuned WHISPER model on Moroccan Darija, several metrics and methods were employed:

1) **Word Error Rate (WER):** WER was the primary metric used to assess the model's transcription accuracy.
2) **Regional Accent Coverage:** The model was tested on audio samples representing the four major dialect groups of Moroccan Darija.
3) **Generalization:** The model's ability to transcribe unseen data was assessed using a test set of audio clips.
4) **Qualitative Analysis:** A manual review of randomly selected transcriptions was conducted.

V. CONCLUSION

The development of a speech recognition system tailored to Moroccan Darija marks a significant milestone in advancing language technology for underrepresented dialects. Moroccan Darija, with its rich linguistic diversity and unique characteristics, presents both challenges and opportunities for ASR development. By leveraging cutting-edge tools like the WHISPER model, combined with the creation of the Moroccan Darija Voice Corpus (MDVC), this project bridges the gap in available resources for this dynamic dialect.

The fine-tuning of the WHISPER model on the MDVC dataset demonstrates the importance of incorporating regional and linguistic diversity into modern ASR systems. This approach ensures inclusivity, accuracy, and robustness across the varied accents and speech styles present in Darija. Furthermore, by embracing open-source frameworks and methodologies, this project contributes to the broader goal of democratizing AI and making speech technology accessible to all. Looking forward, the implementation of this system can serve as a foundation for numerous applications, including voice-activated technologies, automatic transcription services, and educational tools, all adapted to the Moroccan context. It also highlights the potential for extending similar approaches to other low-resource languages and dialects, fostering a more inclusive and representative global ASR landscape.

This work not only showcases the power of AI in preserving linguistic diversity but also underlines the importance of collaboration and innovation in addressing the unique challenges posed by low-resource languages like Moroccan Darija.

REFERENCES

[1] L. Lu, C. Liu, J. Li, and Y. Gong, "Exploring Transformers for Large-Scale Speech Recognition," in *Proc. Interspeech*, 2020.

[2] H. Le, J. Gu, J. Pino, D. Schwab, C. Wang, and L. Besacier, "Dual-decoder Transformer for Joint Automatic Speech Recognition and Multilingual Speech Translation," in *Proc. 28th Int. Conf. on Computational Linguistics (COLING)*, 2020, pp. 3520–3533.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020.

[4] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of Automatic Speech Recognition with Transformer Sequence-to-sequence Model," in *arXiv preprint arXiv:1910.10697*, 2019.

[5] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An Efficient Transformer for Automatic Speech Recognition," in *Proc. 36th Conf. Neural Information Processing Systems (NeurIPS)*, 2022.

[6] E. Tsunoo, Y. Kashiwagi, T. Kumakura, and S. Watanabe, "Towards Online End-to-End Transformer Automatic Speech Recognition," in *arXiv preprint arXiv:1910.11871*, 2019.

[7] A. Boumehdi and A. Yousfi, "MDVC corpus: empowering Moroccan Darija speech recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 290–301, Apr. 2024.

[8] A. Radford, J. W. Kim, T. Xu, et al., "Robust speech recognition via large-scale weak supervision," *Whisper Model Documentation*, OpenAI, 2024.