Fentch – 8drtna
Project Report

Team One:
Boudaoud Amira, Abdelhak Nesrine, Mahmoudi Sarah,
Daif Oumaima, Arab Sarra

Introduction

North Africa boasts a rich tapestry of cultures, traditions, and languages. Among them, the Darija dialect stands out for its unique character. Spoken by millions across the region, Darija presents a fascinating challenge for linguists due to its inherent fluidity and resistance to strict standardization.

In this article, we present a novel approach to building a robust model capable of extracting text from Darija audio and video recordings, specifically focusing on the Algerian dialect. While acknowledging the close ties between Algerian and Moroccan Darija, we delve deeper into the specificities of Algerian speech, aiming to create a tool that captures its nuances and complexities.

Understanding the Challenges of Darija Speech Extraction

While advancements in speech extraction technology have revolutionized human-computer interaction, extracting text from spoken languages like Darija poses unique challenges. Unlike standardized Arabic, Darija lacks a formal structure and incorporates regional variations. This fluidity, while enriching the language, makes it difficult for existing speech extraction models to accurately capture the nuances of spoken Darija.

The Need for Speech Extraction Goes Beyond Human-Computer Interaction

The ability to accurately extract text from speech extends far beyond smoothing the human-computer interaction. For Darija, the extracted text can be fed to machine translation models, building a powerful bridge between the world and North Africa. It will serve as the best tool for outsiders to access many resources and literature, fostering openness towards other cultures and ideologies, thereby enriching both language and culture.

Our Motivation and the Significance of Algerian Darija

Despite all the challenges, Darija serves as the spoken language for almost all cultures in North Africa, particularly in Algeria. The Algerian Darija has a special kind of blend or tapestry and is uniquely divergent. It has adopted elements from many languages such as Arabic (especially from the center of Algeria, Oulad Nail tribe), Amazigh (the original inhabitants of North Africa and its official language), French and Spanish (due to the Spanish and French occupation during the 17th-20th centuries), and some Turkish (from when Algeria was an Ottoman protectorate).

Our Research Methodology and Forward Path

Developing a reliable speech extraction model for Algerian Darija necessitates two crucial components: a well-designed model and a high-quality dataset.

Data Acquisition: The Cornerstone of Success

Crucial to model development is the availability of high-quality training data. However, Darija presents a unique challenge in this regard due to its regional variations. This includes incorporating a wide range of accents, tones, and influences from other languages spoken in North Africa. Unfortunately, readily available Darija speech corpora are rare, necessitating a dedicated data collection strategy.

Data Collection Strategy

To address the scarcity of Darija speech corpora, we initially explored crowdsourcing platforms like Google Forms. However, the contribution rate proved to be insufficient for our needs. Consequently, we pivoted to a more controlled data collection approach.

Our Approach

We opted for a self-directed data collection method. This involved native speakers recording audio samples of spoken Darija that we could collect after scraping Instagram and YouTube and subsequently transcribing the content ourselves. While this approach was more time-consuming compared to crowdsourcing, it yielded several advantages:

- **Quality Control:** Since we were responsible for the transcription of the audios, we were capable of filtering the useful audios that involved a reasonable use of languages or accents (i.e., switching from one language to another or from one accent to another) and with minimal background noise (such as music).
- **Balanced Representation:** We were able to curate a balanced dataset by deliberately seeking recordings from diverse regions of Algeria, capturing the dialect's variety in accents and tones.
- **Unique Sample Coverage:** This method allowed us to gather a representative sample from each unique region, ensuring the dataset reflects the geographical spread of Algerian Darija.

Fine-Tuning Whisper for Algerian Darija-Kabyle

Having assured a high-quality and diverse dataset of Algerian Darija speech recordings that sum up to around 4 hours from the main regions of the country—western (Tlemcen and Oran), eastern (Annaba), southern (Adrar, Djelfa, Boussada), and northern (Algiers)—we turned our attention to the model selection process. Our chosen model is Whisper. Whisper is a transformer decoder-encoder-based model developed by OpenAI. It has recently shown great potential in speech recognition tasks and will serve as a good tool due to the following properties:

- **Multilinguality:** Unlike many speech recognition models trained solely on English, Whisper boasts native multilingual capabilities. This aligns perfectly with our goal of tackling the complexities of Darija, which incorporates influences from Arabic, French, and Berber languages.
- **Robustness to Noise and Fluency:** Whisper exhibits exceptional resilience towards background noise and variations in speech patterns. This is crucial for Darija, which is often spoken in informal settings and may exhibit regional variations in pronunciation and fluency.
- **Transfer Learning Potential:** Whisper's architecture is specifically designed to leverage transfer learning. This allows us to begin with the pre-trained English model and then fine-tune it progressively on French and Arabic datasets before finally specializing it for Algerian Darija speech recognition.

A summary on the Fine tuning process

Fine-tuning the Whisper model involves several key steps:

1. **Select a Pre-trained Model**: Choose a pre-trained Whisper model that best fits the requirements ,for our case whisper small was used.
2. **Prepare the Dataset**:Data collected,cleaned,splitted into 15 sec per each and uploaded to hugging face
3. **Tokenization and Encoding**: Kabyle data needed to add new Tokenizer,Tokenize and encode the dataset using the appropriate tokenizer provided by Hugging Face. This step converts the dataset into input features that the model can process.
4. **Define Training Parameters**: Define the training parameters, including batch size, learning rate, number of epochs, and any other relevant hyperparameters,on our machines we only set number of epochs to 2 as this was the maximum that could be run.
5. **Train the Model**: Fine-tune the pre-trained Whisper model on our dataset. During this process, the model learns from the dataset.
6. **Monitor Performance**: Monitor the model's performance during training using validation data.
7. **Evaluate on Test Data**: Evaluate the fine-tuned model was using the WER metric and running on some unseen dataset.

8. **Iterate** : Based on the evaluation results, iterate on the fine-tuning process if necessary by adjusting parameters.

Results and Findings

The model consists of two parts: one for Kabyle and one for Darija. Note that training was limited to 2 epochs on our local machines, using a maximum of 5% of the entire dataset, which was constrained by the need for powerful GPUs. During these 2 epochs, the "Word Error Rate" metric showed a high value, around 400, highlighting the need for more epochs and training.

The Result:

- Testing on the trained model provided by the **fentech** team for Darija showed correct recognition of most words, including those mixed with French and English, with a few errors in word endings or order. The exact WER (Word Error Rate) value for this final trained model was not available, but the tests yielded positive outcomes.
- For the Kabyle dataset, similar behavior was observed, with correct recognition of most text, often displaying results in French letters.

Challenges

We encountered challenges on two main levels: data and training.

- **Data Challenges** Data collection presented several challenges, including ensuring compatibility with the Whisper model's supported format and the requirement for a substantial dataset size. Additionally, adapting the data format to be compatible with Hugging Face and training required adjustments to utilize locally sourced data.
- **Model Challenges** The training process necessitated a powerful GPU, making it impractical to run locally on our machines. The model's demands exceeded our hardware capabilities.

Conclusion

In this article, we have outlined our approach and methodology for developing a speech extraction model for Algerian Darija. We discussed the challenges, the importance of high-quality data, and our data collection strategy. We also highlighted the fine tuning process and all the steps encountered

Appendix

Who did what table

Task	**Done by**
Data collection Arabic	Abdelhak Nesrine, Mahmoudi Sarah,Daif Oumaima
Data collection Kabyle	Boudaoud Amira, Arab Sarra
Data upload to hugging face	Abdelhak Nesrine, Mahmoudi Sarah
Dataset : Conversion to Parquet	Boudaoud Amira,Mahmoudi Sarah
Fine tune on Arabic only	Abdelhak Nesrine, Mahmoudi Sarah
Fine tune on French and English only	Abdelhak Nesrine
Fine tune on Darija	Abdelhak Nesrine,Arab Sarra
Fine tune on kabyle	Mahmoudi Sarah
Add tokenizer kabyle	Mahmoudi Sarah
Streamlit	Mahmoudi Sarah,Arab Sarra
Streamlit backend	Mahmoudi Sarah
Slides presentation	Boudaoud Amira,Abdelhak Nesrine
Report	Mahmoudi Sarah,Abdelhak Nesrine
Fixing bugs and debug	Mahmoudi Sarah,Abdelhak Nesrine