Dialectal Voice : An Open-Source Voice Dataset and Automatic Speech Recognition model for Moroccan Arabic dialect

Anass Allak
Laboratoire SI2M
INSEA
Rabat-Instituts Rabat B.P. 6217 Maroc
aallak@insea.ac.ma

Abdou Mohamed Naira
Laboratoire SI2M
INSEA
Rabat-Instituts Rabat B.P. 6217 Maroc
nabdoumohamed@insea.ac.ma

Imade Benelallam
Laboratoire SI2M, AIOX Labs
INSEA
Rabat-Instituts Rabat B.P. 6217 Maroc
i.benelallam@insea.ac.ma

Kamel Gaanoun
Laboratoire SI2M
INSEA
Rabat-Instituts Rabat B.P. 6217 Maroc
kamel.gaanoun@gmail.com

Abstract

Under-represented languages such as Moroccan Arabic dialectal or Darija as it is commonly known face a lack of open systems capable of understanding them. However, a growing need for these systems by Academia, private companies and public institutions is increasingly expressed in order to better improve the human experience and ensure good productivity. We present here an automatic voice recognition system resulting from Data Centric and Transfer Learning approaches for the construction of a voice database and a Speech Recognition model for the Darija.

1 Introduction

Amazigh and Arabic are the official languages of Morocco [1]. However the majority of Moroccan speak the local variant of Arabic : Darija. Darija differs from the Modern Standard Arabic (MSA) in a multitude of facets. One of the most notable differences is the usage of a number borrowed words from Spanish, Amazigh and French which introduce a number of nonnative Arabic sound in the Darija corpus. Spoken Darija is heavily influenced by the Amazigh. The influence can be observed on the stressing of syllables in the beginning of word, starting word with "sokoun" and the grammatical use of prefixes and suffixes which absent from MSA.

Table 1: Example of differences between MSA and Darija

MSA phonetic	Darija phonetic
kitab	ktab
aqul	ngul

These differences and more make the usage of MSA dataset inadequate for Darija speech recognition. While the number of Speech Arabic dataset is growing, the number of dialectic specific dataset is limited especially for Darija.

Preprint. Under review.

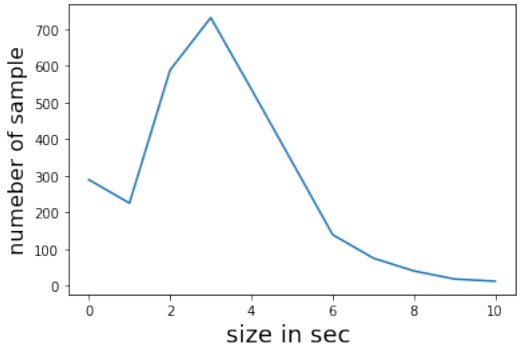

Figure 1: Distribution of audio data

2 General description

The Dvoice dataset [2] is a Darija speech dataset that was created by SI2M labs in collaboration with AIOX Lab containing audio file of Darija speech and its transcription. The distribution of the dataset is as follows :

- 2392 training files
- 600 testing files

We present in Figure 1 the distribution of the audios according to their duration which are for the most part between 2 and 6 seconds. In the next part, we will provide details on the method used to obtain this data.

3 Method

The database is built using three technics:

- Retrieving through the Dvoice platform.
- Web scrape authentic and more reliable recordings followed by their transcriptions.
- Label a set of collected recordings via the SpeechRecognition [3] library.

3.1 Reliable sources

The Dvoice Web Application [4] is a community and participatory platform that we initiated with the aim of helping research around voice technologies on African dialects, starting with Darija. Today the application allows contributors to come and give their votes by submitting recordings corresponding to texts written in Darija or to validate the recordings of other contributors. This platform has been inspired by the Common Voice initiative [5] that suffers from a lack of under represented languages.

Another method is to go and retrieve data from Darija learning sites. They contain on the one hand the text and on the other hand the corresponding recordings.

The data obtained by these methods are reliable, however it is not easy to obtain them in sufficient quantity. We then use transfer learning for the records labeling. This method is detailed in the following lines.

3.2 Transfer Learning

The idea of this step comes from the fact that building a fairly large dataset with the Dvoice initiative would risk taking a lot of time and require a great mobilization of the community. The Transfer Learning is an alternative that has successfully overcome the lack of data. It is based on videos taken from social networks and the method can be summarized as follows :

- Identification of the best potential sources of data (Youtube and Facebook videos) in which they are clearly expressed in Darija.
- Convert videos to .wav and separate channels.
- Cutting of the audios obtained by silence.
- Using the SpeechRecognition library to transcribe audio into texts. The latter is a library which uses several APIs and which allows speech recognition by offering a score for each transcription. SpeechRecognition can be used both online and offline.
- Selection of audios with transcription accuracy $\geq = 90\%$.

3.3 Data Augmentation

Data Augmentation is a technic used in Machine Learning that solves two major problems: the lack of sufficient data for training models and over fitting problems. In speech recognition, the increase in data is done by slight modifications of audio recordings, the idea is to allow the model to succeed in capturing the features in an audio even if it is presented differently. We use the SpeechBrain [6] toolkit which allows us to do the following processing on our audios : speed perturbation, time dropout, frequency dropout, clipping and augmentation lobe.

4 Experiment

As stated above, the idea of our work is to contribute to research around voice recognition in Darija. After fine-tuning XLSR-53 [7] on our finale database, we obtained a Word Error Rate (WER) of 30%. The choice of fine-tuning instead of a from-scratch approach training is justified by the fact that despite the processing provided, in particular in terms of Data Augmentation, we did not obtain enough data to be able to train a model from-scratch.

XLSR-53, being trained on several languages, allows to understand after fine-tuning and therefore on a not too large database, languages like Arabic. According to [8], the Moroccan dialect is made up of 77% of terms from Standard Arabic. We start with the hypothesis that a model capable of understanding Standard Arabic well would also be suitable for Darija.

The experimentation was carried out by following an agile methodology which allowed us to best improve the performance of our model. Here are the main lines of our experiment:

- Training the model on an Arabic subset of database retrieved from Mozilla Common Voice [5].
- Retraining on the Dvoice-v1.0 database.
- Data Dvoice-v1.0 augmentation and model retraining.

The experiment is summarized in the table below:

Table 2: Results

Version	Source	Size	Language	Speech Augmentation	WER
Beta	Mozilla Common Voice	1500	Arabic	No	70%
Beta 1	Facebook/Youtube	2400	Darija	No	90%
Version 1.0	Dvoice/Facebook /Youtube	13000	Darija	Yes	30%

Details on database preparation and model training can also be found in the official GitHub repository [9].

5 Conclusion

The Data Centric agile approach presented in this paper has improved the performance of a speech recognition model on the Darija. We have presented the different stages of construction of the dataset and the performance of the model on each. We have gone from a WER of 70% to 30% as the training

dataset evolves. Therefore, research in this area is ongoing and thus we are planning to support the continuous collection and sharing of data that can then be used of various speech recognition tasks by maintaining a release regular update to the dataset. To do this, we have two main objectives: encourage people to contribute to the Dvoice platform and continue alternative method for data collection. In order to avoid possible problems of bias, we also propose to diversify the data by adding, for example, more female voices.

References

[1] *English translation of Morocco's 2011 constitution*. Number 1. 2019.

[2] Imade Benelallam, Anass Allak, and Abdou Mohamed Naira. Dvoice : An open source dataset for Automatic Speech Recognition on Moroccan dialectal Arabic, September 2021. URL `https://doi.org/10.5281/zenodo.5482551`.

[3] Anthony Zhang, bobsayshilol, Arvind Chembarpu, Kevin Smith, haas85, DelightRun, maverick-agm, Kamus Hadenes, Sarah Braden, Bohdan Turkynewych, Steve Dougherty, and Broderick Carlin. Speech recognition (version 3.8), 2017.

[4] Dvoice web app. `https://www.dvoice.ma/`, 2021. Accessed: 2021-09-2.

[5] Common voice dataset. `https://commonvoice.mozilla.org/`. Accessed: 2010-09-30.

[6] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[7] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[8] Ridouane Tachicart, Karim Bouzoubaa, and Hamid Jaafar. Lexical differences and similarities between moroccan dialect and arabic. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, pages 331–337, 2016. doi: 10.1109/CIST.2016.7805066.

[9] AIOX Labs. Dvoice, 09 2021. URL `https://github.com/AIOXLABS/DVoice`.